

**Федеральное государственное образовательное бюджетное учреждение
высшего профессионального образования**

**ПОВОЛЖСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
ТЕЛЕКОММУНИКАЦИЙ И ИНФОРМАТИКИ**

Кафедра электронной коммерции

**ИНФОРМАЦИОННЫЕ СИСТЕМЫ АНАЛИЗА И
ПЛАНИРОВАНИЯ ДЕЯТЕЛЬНОСТИ ОРГАНИЗАЦИИ**

**МЕТОДИЧЕСКИЕ УКАЗАНИЯ
К ПРАКТИЧЕСКИМ И ЛАБОРАТОРНЫМ РАБОТАМ
ПО ДИСЦИПЛИНЕ**

Интеллектуальные технологии в бизнесе

по подготовке магистров специальности

080500 – Бизнес-информатика

Составил: к.э.н., доцент Крюкова А.А.

Самара, 2013

Крюкова А.А. Интеллектуальные технологии в бизнесе. Методические указания к практическим и лабораторным работам. - Самара: ФГОБУ ВПО ПГУТИ, 2013. – 150

Методические указания по выполнению практических и лабораторных работ по дисциплине «Интеллектуальные технологии в бизнесе», подготовлены на кафедре «Электронная коммерция», предназначены для магистров специальности 080500 – Бизнес-информатика и являются руководством к выполнению их обучающимися. Содержат комплекс вопросов, заданий, докладов и тестов, охватывающих основные теоретические и прикладные аспекты бизнес-аналитики.

Рецензент:

Юрасов А.В. – д.э.н., проф., зав.кафедрой электронной коммерции Поволжского государственного университета телекоммуникаций и информатики
Федеральное государственное образовательное бюджетное учреждение высшего профессионального образования

Поволжский государственный университет телекоммуникаций и информатики

© Крюкова А.А., 2013

СОДЕРЖАНИЕ

Раздел 1 Материал для подготовки к факультативным занятиям.....	5
Тема 1 Понятие искусственного интеллекта. Предпосылки развития науки искусственного интеллекта.....	5
Тема 2 Современный искусственный интеллект. Связь с другими науками.....	11
Тема 3 Понятие интеллектуального анализа данных (Data Mining).....	18
Тема 4 Методы, стадии и задачи Data Mining.....	28
Тема 5 Классификация методов Data Mining.....	33
Тема 6 Сфера применения Data Mining. Применение Data Mining для решения бизнес-задач.....	40
Тема 7 Комплексный подход к внедрению Data Mining, OLAP и хранилищ данных в СППР.....	54
Тема 8 Классификация СППР.....	57
Тема 9 История создания СППР.....	65
Тема 10 Сфера применения СППР.....	68
Тема 11 Задачи, решаемые с помощью СППР.....	73

Тема 12 Задачи, которые невозможно решить при помощи СППР.....	76
Тема 13 Применение СППР в бизнесе.....	78
Тема 14 Анализ математического обеспечения существующих систем поддержки принятия решений.....	84
Раздел 2 Тематика докладов по дисциплине.....	87
Раздел 3 Практикум в системе интеллектуального анализа данных Deductor.....	91
Задание № 1.....	91
Задание № 2.....	107
Задание № 3.....	110
Задание № 4.....	118
Задания для самостоятельного выполнения.....	125
Список литературы по дисциплине.....	143

Раздел 1 МАТЕРИАЛ ДЛЯ ПОДГОТОВКИ К ПРАКТИЧЕСКИМ ЗАНЯТИЯМ

Тема 1 Понятие искусственного интеллекта. Предпосылки развития науки искусственного интеллекта

Искусственный интеллект (ИИ, англ. *Artificial intelligence, AI*) — наука и технология создания интеллектуальных машин, особенно интеллектуальных компьютерных программ. ИИ связан со сходной задачей использования компьютеров для понимания человеческого интеллекта, но не обязательно ограничивается биологически правдоподобными методами.

Поясняя своё определение, Джон Маккарти указывает: «Проблема состоит в том, что пока мы не можем в целом определить, какие вычислительные процедуры мы хотим называть интеллектуальными. Мы понимаем некоторые механизмы интеллекта и не понимаем остальные. Поэтому под интеллектом в пределах этой науки понимается только вычислительная составляющая способности достигать целей в мире».

Другие определения искусственного интеллекта:

1. Научное направление, в рамках которого ставятся и решаются задачи аппаратного или программного моделирования тех видов человеческой деятельности, которые традиционно считаются интеллектуальными.

2. Свойство интеллектуальных систем выполнять функции (творческие), которые традиционно считаются прерогативой человека. При этом интеллектуальная система — это техническая или программная система, способная решать задачи, традиционно считающиеся творческими, принадлежащие конкретной предметной области, знания о которой хранятся в памяти такой системы. Структура интеллектуальной системы включает три основных блока — базу знаний, решатель и интеллектуальный интерфейс.

3. Наука под названием «Искусственный интеллект» входит в комплекс компьютерных наук, а создаваемые на её основе технологии к информационным технологиям. Задачей этой

науки является воссоздание с помощью вычислительных систем и иных искусственных устройств разумных рассуждений и действий.

Происхождение и понимание термина «Искусственный интеллект»

Различные виды и степени интеллекта существуют у многих людей, животных и некоторых машин, интеллектуальных информационных систем и различных моделей экспертных систем с различными базами знаний. При этом, как видим, такое определение интеллекта не связано с пониманием интеллекта у человека — это разные вещи. Более того, эта наука моделирует человеческий интеллект, так как с одной стороны, можно изучить кое-что о том, как заставить машины решить проблемы, наблюдая других людей, а с другой стороны, большинство работ в ИИ касаются изучения проблем, которые требуется решать человечеству в промышленном и технологическом смысле. Поэтому ИИ-исследователи вольны использовать методы, которые не наблюдаются у людей, если это необходимо для решения конкретных проблем.

Именно в таком смысле термин ввел Джон Маккарти в 1956 году на конференции в Дармутском университете, и до сих пор несмотря на критику тех, кто считает, что интеллект — это только биологический феномен, в научной среде термин сохранил свой первоначальный смысл, несмотря на явные противоречия с точки зрения человеческого интеллекта.

Как указывает председатель Петербургского отделения российской ассоциации искусственного интеллекта Т. А. Гаврилова, в английском языке словосочетание *artificial intelligence* не имеет той слегка фантастической антропоморфной окраски, которую оно приобрело в довольно неудачном русском переводе. Слово *intelligence* означает «умение рассуждать разумно», а вовсе не «интеллект», для которого есть английский аналог *intellect*.

Одно из частных определений интеллекта, общее для человека и «машин», можно сформулировать так: «Интеллект — способность системы создавать в ходе самообучения программы

(в первую очередь эвристические) для решения задач определённого класса сложности и решать эти задачи».

Искусственный интеллект в России

Пионером искусственного интеллекта по праву можно считать коллежского советника Семёна Николаевича Корсакова (1787 - 1853), ставившего задачу усиления возможностей разума посредством разработки научных методов и устройств, переключающуюся с современной концепцией искусственного интеллекта, как усилителя естественного. В 1832 году Корсаков опубликовал описание пяти изобретенных им механических устройств, так называемых «интеллектуальные машины», для частичной механизации умственной деятельности в задачах поиска, сравнения и классификации. В конструкции своих машин Корсаков впервые в истории информатики применил перфорированные карты, игравшие у него своего рода роль баз знаний, а сами машины по существу являлись предтечами экспертных систем.

Краткий обзор истории зарождения искусственного интеллекта в СССР и России дал Президент Российской ассоциации искусственного интеллекта профессор Г.С. Осипов.

Работы в области искусственного интеллекта в России начались в 1960-х годах. В Московском университете и Академии наук был выполнен ряд пионерских исследований, возглавленных Вениамином Пушкиным и Д.А. Поспеловым.

В 1964 году была опубликована работа ленинградского логика Сергея Маслова «Обратный метод установления выводимости в классическом исчислении предикатов», в которой впервые предлагался метод автоматического поиска доказательства теорем в исчислении предикатов.

В 1966 году В.Ф. Турчиным был разработан язык рекурсивных функций Рефал.

До 1970-х годов в СССР все исследования ИИ велись в рамках кибернетики. По мнению Д.А. Поспелова, дело в том, что науки «информатика» и «кибернетика» были в это время смешаны, по причине ряда академических споров. Только в конце 1970-х в СССР начинают говорить о научном направлении «искус-

ственный интеллект» как разделе информатики. При этом родилась и сама информатика, как ни странно, подчинив себе прародительницу «кибернетику». В конце 1970-х создается толковый словарь по искусственному интеллекту, трехтомный справочник по искусственному интеллекту и энциклопедический словарь по информатике, в котором разделы «Кибернетика» и «Искусственный интеллект» входят наряду с другими разделами в состав информатики. Термин «информатика» в 80-е годы получает широкое распространение, а термин «кибернетика» постепенно исчезает из обращения, сохранившись лишь в названиях тех институтов, которые возникли в эпоху «кибернетического бума» конца 1950-х — начала 1960-х годов.

Такой взгляд на искусственный интеллект, кибернетику и информатику разделяется не всеми. Это связано с тем, что на Западе границы данных наук несколько отличаются.

Подходы и направления искусственного интеллекта

Подходы к пониманию проблемы

Единого ответа на вопрос чем занимается искусственный интеллект, не существует. Почти каждый автор, пишущий книгу об ИИ, отталкивается в ней от какого-либо определения, рассматривая в его свете достижения этой науки.

В философии не решён вопрос о природе и статусе человеческого интеллекта. Нет и точного критерия достижения компьютерами «разумности», хотя на заре искусственного интеллекта был предложен ряд гипотез, например, тест Тьюринга или гипотеза Ньюэлла – Саймона. Поэтому несмотря на наличие множества подходов как к пониманию задач ИИ, так и созданию интеллектуальных информационных систем можно выделить два основных подхода к разработке ИИ:

- нисходящий, семиотический — создание экспертных систем, баз знаний и систем логического вывода, **имитирующих** высокоуровневые психические процессы: мышление, рассуждение, речь, эмоции, творчество и т. д.;

- восходящий, биологический — изучение нейронных сетей и эволюционных вычислений, моделирующих интеллектуальное поведение на основе биологических элементов, а также

создание соответствующих вычислительных систем, таких как нейрокомпьютер или биокомпьютер.

Последний подход, строго говоря, не относится к науке о ИИ в смысле данного Джоном Маккарти — их объединяет только общая конечная цель.

Тест Тьюринга и интуитивный подход

Эмпирический тест, идея которого была предложена Аланом Тьюрингом в статье «Вычислительные машины и разум» (англ. *Computing Machinery and Intelligence*), опубликованной в 1950 году в философском журнале «*Mind*». Целью данного теста является определение возможности искусственного мышления, близкого к человеческому.

Стандартная интерпретация этого теста звучит следующим образом: «Человек взаимодействует с одним компьютером и одним человеком. На основании ответов на вопросы он должен определить, с кем он разговаривает: с человеком или компьютерной программой. Задача компьютерной программы — ввести человека в заблуждение, заставив сделать неверный выбор». Все участники теста не видят друг друга.

- Самый общий подход предполагает, что ИИ будет способен проявлять поведение, не отличающееся от человеческого, причём, в нормальных ситуациях. Эта идея является обобщением подхода теста Тьюринга, который утверждает, что машина станет разумной тогда, когда будет способна поддерживать разговор с обычным человеком, и тот не сможет понять, что говорит с машиной (разговор идёт по переписке).

- Писатели-фантасты часто предлагают ещё один подход: ИИ возникнет тогда, когда машина будет способна чувствовать и творить. Так, хозяин Э.Мартина из «Двухсотлетнего человека» начинает относиться к нему как к человеку, когда тот создаёт игрушку по собственному проекту. А Дейта из Звездного пути, будучи способным к коммуникации и научению, мечтает обрести эмоции и интуицию.

Символьный подход

Исторически символьный подход был первым в эпоху цифровых машин, так как именно после создания Лисп, первого

языка символьных вычислений, у его автора возникла уверенность в возможности практически приступить к реализации этими средствами интеллекта. Символьный подход позволяет оперировать слабоформализованными представлениями и их смыслами. От умения выделить только существенную информацию зависит эффективность и результативность решения задачи.

Но широта классов задач, эффективно решаемых человеческим разумом, требует невероятной гибкости в методах абстрагирования. А это недоступно при любом инженерном подходе, в котором исследователь выбирает методы решения, основываясь на способность быстро дать эффективное решение какой-то наиболее близкой этому исследователю задачи. То есть уже за реализованную в виде правил единственную модель абстрагирования и конструирования сущностей. Это выливается в значительные затраты ресурсов для непрофильных задач, то есть система от интеллекта возвращается к грубой силе на большинстве задач и сама суть интеллекта исчезает из проекта.

Основное применение символьной логики - это решение задач по выработке правил. Большинство исследований останавливается как раз на невозможности хотя бы обозначить новые возникшие трудности средствами выбранных на предыдущих этапах символьных системах. Тем более решить их и тем более обучить компьютер решать их или хотя бы идентифицировать и выходить из таких ситуаций.

Логический подход

Логический подход к созданию систем искусственного интеллекта направлен на создание экспертных систем с логическими моделями баз знаний с использованием языка предикатов.

Учебной моделью систем искусственного интеллекта в 1980-х годах был принят язык и система логического программирования Пролог. Базы знаний, записанные на языке Пролог, представляют наборы фактов и правил логического вывода, записанных на языке логических предикатов.

Логическая модель баз знаний позволяет записывать не только конкретные сведения и данные в форме фактов на языке Пролог, но и обобщенные сведения с помощью правил и проце-

дур логического вывода и в том числе логических правил определения понятий, выражающих определённые знания как конкретные и обобщенные сведения.

В целом исследования проблем искусственного интеллекта в рамках логического подхода к проектированию баз знаний и экспертных систем направлено на создание, развитие и эксплуатацию интеллектуальных информационных систем, включая вопросы обучения студентов и школьников, а также подготовки пользователей и разработчиков таких интеллектуальных информационных систем.

Тема 2 Современный искусственный интеллект. Связь с другими науками

В настоящий момент в создании искусственного интеллекта наблюдается вовлечение многих предметных областей, имеющих хоть какое-то отношение к ИИ. Многие подходы были опробованы, но к возникновению искусственного разума ни одна исследовательская группа так и не подошла.

Исследования ИИ влились в общий поток технологий сингулярности (видового скачка, экспоненциального развития человека), таких как информатик, экспертные системы, нанотехнология, молекулярная биоэлектроника, теоретическая биология, квантовая теория.

Некоторые из самых известных ИИ-систем:

- Deep Blue - победил чемпиона мира по шахматам. Матч Каспаров против суперЭВМ не принёс удовлетворения ни компьютерщикам, ни шахматистам, и система не была признана Каспаровым. Затем линия суперкомпьютеров IBM проявилась в проектах brute force BluGene (молекулярное моделирование) и моделирование системы пирамидальных клеток в швейцарском центре Blue Brain.

- MYCIN - одна из ранних экспертных систем, которая могла диагностировать небольшой набор заболеваний, причем часто так же точно, как и доктора.

- 20Q - проект, основанный на идеях ИИ, по мотивам классической игры «20 вопросов». Стал очень популярен после появления в Интернете на сайте 20q.net.

- Распознавание речи. Системы такие как ViaVoice способны обслуживать потребителей.

- Роботы в ежегодном турнире *RoboCup* соревнуются в упрощённой форме футбола.

Банки применяют системы искусственного интеллекта (СИИ) в страховой деятельности (актуарная математика) при игре на бирже и управлении собственностью. Методы распознавания образов (включая, как более сложные и специализированные, так и нейронные сети) широко используют при оптическом и акустическом распознавании (в том числе текста и речи), медицинской диагностике, спам-фильтрах, в системах ПВО (определение целей), а также для обеспечения ряда других задач национальной безопасности.

Разработчики компьютерных игр применяют ИИ в той или иной степени проработанности. Это образует понятие «Игровой искусственный интеллект». Стандартными задачами ИИ в играх являются нахождение пути в двумерном или трёхмерном пространстве, имитация поведения боевой единицы, расчёт верной экономической стратегии и так далее.

Перспективы

Можно выделить два направления развития ИИ:

1. решение проблем, связанных с приближением специализированных систем ИИ к возможностям человека, и их интеграции, которая реализована природой человека.

2. создание искусственного разума, представляющего интеграцию уже созданных систем ИИ в единую систему, способную решать проблемы человечества.

Связь с другими науками

Искусственный интеллект вместе с нейрофизиологией, эпистемологией и когнитивной психологией образует более общую науку, называемую когнитологией. Отдельную роль в искусственном интеллекте играет философия.

Также, с проблемами искусственного интеллекта тесно связана эпистемология - наука о знании в рамках философии. Философы, занимающиеся данной проблематикой, решают вопросы, схожие с теми, которые решаются инженерами ИИ о том, как лучше представлять и использовать знания и информацию.

Производство знаний из данных — одна из базовых проблем интеллектуального анализа данных. Существуют различные подходы к решению этой проблемы, в том числе — на основе нейросетевой технологии, использующие процедуры вербализации нейронных сетей.

Компьютерные технологии и кибернетика

В компьютерных науках проблемы искусственного интеллекта рассматриваются с позиций проектирования экспертных систем и баз знаний. Под базами знаний понимается совокупность данных и правил вывода, допускающих логический вывод и осмысленную обработку информации. В целом исследования проблем искусственного интеллекта в компьютерных науках направлено на создание, развитие и эксплуатацию интеллектуальных информационных систем, а вопросы подготовки пользователей и разработчиков таких систем решаются специалистами информационных технологий.

Психология и когнитология

Методология когнитивного моделирования предназначена для анализа и принятия решений в плохо определенных ситуациях. Была предложена Аксельродом.

Основана на моделировании субъективных представлений экспертов о ситуации и включает: методологию структуризации ситуации: модель представления знаний эксперта в виде знаково-го орграфа (когнитивной карты) (F, W) , где F — множество факторов ситуации, W — множество причинно-следственных отношений между факторами ситуации; методы анализа ситуации. В настоящее время методология когнитивного моделирования развивается в направлении совершенствования аппарата анализа и моделирования ситуации. Здесь предложены модели прогноза развития ситуации; методы решения обратных задач.

Философия

Наука «о создании искусственного разума» не могла не привлечь внимание философов. С появлением первых интеллектуальных систем были затронуты фундаментальные вопросы о человеке и знании, а отчасти о мироустройстве. С одной стороны, они неразрывно связаны с этой наукой, а с другой — приносят в неё некоторый хаос.

Философские проблемы создания искусственного интеллекта можно разделить на две группы, условно говоря, «до и после разработки ИИ». Первая группа отвечает на вопрос: «Что такое ИИ, возможно ли его создание, и, если возможно, то как это сделать?». Вторая группа (этика искусственного интеллекта) задаётся вопросом: «Каковы последствия создания ИИ для человечества?». Течение трансгуманизма считает создание ИИ одной из важнейших задач человечества.

Вопросы создания ИИ

Среди исследователей ИИ до сих пор не существует какой-либо доминирующей точки зрения на критерии интеллектуальности, систематизацию решаемых целей и задач, нет даже строгого определения науки. Существуют разные точки зрения на вопрос, что считать интеллектом. Аналитический подход предполагает анализ высшей нервной деятельности человека до низшего, неделимого уровня (функция высшей нервной деятельности, элементарная реакция на внешние раздражители (стимулы), раздражение синапсов совокупности связанных функцией нейронов) и последующее воспроизведение этих функций.

Некоторые специалисты за интеллект принимают способность рационального, мотивированного выбора, в условиях недостатка информации. То есть интеллектуальной просто считается та программа деятельности (не обязательно реализованная на современных ЭВМ), которая сможет выбрать из определённого множества альтернатив, например, куда идти в случае «налево пойдёшь...», «направо пойдёшь...», «прямо пойдёшь...».

Наиболее горячие споры в философии искусственного интеллекта вызывает вопрос возможности мышления творения человеческих рук. Вопрос «Может ли машина мыслить?», который подтолкнул исследователей к созданию науки о моделировании

человеческого разума, был поставлен Аланом Тьюрингом в 1950 году. Две основных точки зрения на этот вопрос носят названия гипотез сильного и слабого искусственного интеллекта.

Термин «сильный искусственный интеллект» ввел Джон Серль, его же словами подход и характеризуется: более того, такая программа будет не просто моделью разума; она в буквальном смысле слова сама и будет разумом, в том же смысле, в котором человеческий разум — это разум.

Напротив, сторонники слабого ИИ предпочитают рассматривать программы лишь как инструмент, позволяющий решать те или иные задачи, которые не требуют полного спектра человеческих познавательных способностей. Мысленный эксперимент «Китайская комната» Джона Сёрля — аргумент в пользу того, что прохождение теста Тьюринга не является критерием наличия у машины подлинного процесса мышления.

Мышление есть процесс обработки находящейся в памяти информации: анализ, синтез и самопрограммирование. Аналогичную позицию занимает и Роджер Пенроуз, который в своей книге «Новый ум короля» аргументирует невозможность получения процесса мышления на основе формальных систем. Существует разумный критерий отбора наиболее вероятных гипотез будущего развития (в том числе появления ИИ) — внимательное изучение развития в прошлом. В данном случае, имеет смысл обратиться к истории появления первых нервных клеток в многоклеточных организмах: первые нейроноподобные клетки появились из обычных клеток наружных слоёв первобытных многоклеточных организмов. Постепенно они мигрировали внутрь организма. ИИ (точнее электронную личность) создадут на основе человеческой личности, что будет сходно с процессом появления нервной клетки в результате трансформации обычной клетки. Пройдёт время и часть электронных личностей, будет постепенно консолидироваться в отдельные структуры, целые ансамбли из миллионов и даже миллиардов электронных единиц. Причём в специально отведённых для этого суперкомпьютерах будущего. Где-то появится и «головной мозг» нашей Цивилизации — СуперИИ. Но уже не мы станем его создателями. Он будет состоять из

совершенных электронных личностей (примерно так, как и мозг любого животного состоит из нейронов), сплотившихся под руководством единой программы, позволяющей ему ощущать себя некой сверхличностью, а всю цивилизацию — своим реальным телом.

Этика

Этот раздел содержит вопросы, касающиеся искусственного интеллекта и этики. Э.Юдковски исследует в Институте сингулярности (SIAI) в США проблемы глобального риска, которые может создать будущий сверхчеловеческий ИИ, если его не запрограммировать на дружелюбность к человеку к человеку. В 2004 году SIAI был создан сайт AsimovLaws.com, созданный для обсуждения этики ИИ в контексте проблем, затронутых в фильме «Я, робот», выпущенном лишь на два дня спустя. На этом сайте они хотели показать, что законы робототехники Азимова небезопасны, поскольку, например, могут побудить ИИ захватить власть на Земле, чтобы «защитить» людей от вреда.

Научная фантастика

В научно-фантастической литературе ИИ чаще всего изображается как сила, которая пытается свергнуть власть человека (Омниус в «Космическая одиссея 2001 года», Скайнет, «Матрица» и репликант в «Бегущий по лезвию») или обслуживающий гуманоид («Двухсотлетний человек»). Неизбежность доминирования над миром ИИ, вышедшего из под контроля, оспаривается такими его исследователями, как фантаст Айзек Азимов и кибернетик Кевин Уорвик (Kevin Warwick), известными множественными экспериментами по интеграции машин и живых существ.

Любопытное видение будущего представлено в романе «Выбор по Тьюрингу» писателя-фантаста Гарри Гаррисона и ученого Марвина Мински. Авторы рассуждают на тему утраты человечности у человека, в мозг которого была вживлена ЭВМ, и приобретения человечности машиной с ИИ, в память которой была скопирована информация из головного мозга человека.

Некоторые научные фантасты, например Вернор Виндж, также размышляли над последствиями появления ИИ, которое,

по-видимому, вызовет резкие драматические изменения в обществе. Такой период называют технологической сингулярностью.

Тема ИИ рассматривается под разными углами в творчестве Роберта Хайнлайна: гипотеза возникновения самоосознания ИИ при усложнении структуры далее определённого критического уровня и наличии взаимодействия с окружающим миром и другими носителями разума («The Moon Is a Harsh Mistress», «Time Enough For Love», персонажи Майкрофт, Дора и Ая в цикле «История будущего»), проблемы развития ИИ после гипотетического самоосознания и некоторые социально-этические вопросы («Friday»). Социально-психологические проблемы взаимодействия человека с ИИ рассматривает и роман Филипа К. Дика «Снятся ли андроидам электроовцы?», известный также по экранизации «Бегущий по лезвию».

Одни из самых глубоких исследований проблематики ИИ проявляются в творчестве фантаста и философа Станислава Лема, к примеру, в приключениях Иона Тихого неоднократно описываются взаимоотношения живых существ и машин: бунт бортового компьютера с последующими неожиданными событиями (11-ое путешествие), адаптация роботов в человеческом обществе ("Стиральная трагедия" из "Воспоминаний Ийона Тихого"), построение абсолютного порядка на планете путем переработки живых жителей (24-ое путешествие), изобретения Коркорана и Диагора (Воспоминания Ийона Тихого), психиатрическая клиника для роботов (Воспоминания Ийона Тихого). Кроме того, существует целый цикл повестей и рассказов Кибериада, где почти всеми персонажами являются роботы, которые являются далекими потомками роботов, сбежавших от людей (людей они именуют бледнотиками и считают их мифическими существами).

Среди отечественных авторов тема ИИ занимает центральное место в научно-фантастическом романе В. Л. Кузьменко «Древо жизни».

Тема 3 Понятие интеллектуального анализа данных (Data Mining)

В прошлом процесс добычи золота в горной промышленности состоял из выбора участка земли и дальнейшего ее просеивания большое количество раз. Иногда искатель находил несколько ценных самородков или мог натолкнуться на золотоносную жилу, но в большинстве случаев он вообще ничего не находил и шел дальше к другому многообещающему месту или же вовсе бросал добывать золото, считая это занятие напрасной тратой времени.

Сегодня появились новые научные методы и специализированные инструменты, сделавшие горную промышленность намного более точной и производительной. Data Mining для данных развилась почти таким же способом. Старые методы, применявшиеся математиками и статистиками, отнимали много времени, чтобы в результате получить конструктивную и полезную информацию.

Сегодня на рынке представлено множество инструментов, включающих различные методы, которые делают Data Mining прибыльным делом, все более доступным для большинства компаний.

Термин Data Mining получил свое название из двух понятий: поиска ценной информации в большой базе данных (data) и добычи горной руды (mining). Оба процесса требуют или просеивания огромного количества сырого материала, или разумного исследования и поиска искомых ценностей.

Термин Data Mining часто переводится как добыча данных, извлечение информации, раскопка данных, интеллектуальный анализ данных, средства поиска закономерностей, извлечение знаний, анализ шаблонов, "извлечение зерен знаний из гор данных", раскопка знаний в базах данных, информационная проходка данных, "промывание" данных. Понятие "обнаружение знаний в базах данных" (Knowledge Discovery in Databases, KDD) можно считать синонимом Data Mining.

Понятие Data Mining, появившееся в 1978 году, приобрело высокую популярность в современной трактовке примерно с первой половины 1990-х годов. До этого времени обработка и анализ данных осуществлялся в рамках прикладной статистики, при этом в основном решались задачи обработки небольших баз данных.

О популярности Data Mining говорит и тот факт, что результат поиска термина "Data Mining" в поисковой системе Google (на сентябрь 2005 года) - более 18 миллионов страниц.

Data Mining - мультидисциплинарная область, возникшая и развивающаяся на базе таких наук как прикладная статистика, распознавание образов, искусственный интеллект, теория баз данных и др., см. рисунок 1.

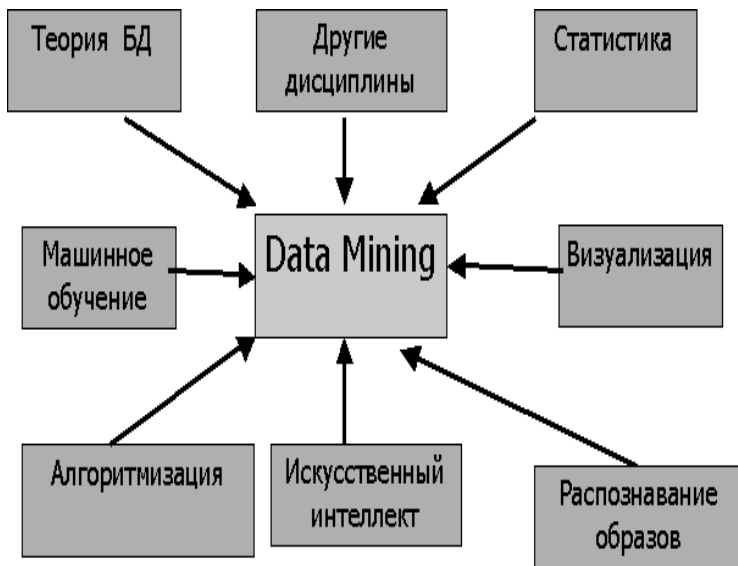


Рисунок 1 - Data Mining как мультидисциплинарная область

Data Mining как часть рынка информационных технологий

Классификация аналитических систем

Агентство Gartner Group, занимающееся анализом рынков информационных технологий, в 1980-х годах ввело термин "Business Intelligence" (BI), деловой интеллект или бизнес-интеллект. Этот термин предложен для описания различных концепций и методов, которые улучшают бизнес решения путем использования систем поддержки принятия решений.

В 1996 году агентство уточнило определение данного термина. *Business Intelligence* – программные средства, функционирующие в рамках предприятия и обеспечивающие функции доступа и анализа информации, которая находится в хранилище данных, а также обеспечивающие принятие правильных и обоснованных управленческих решений.

Понятие BI объединяет в себе различные средства и технологии анализа и обработки данных масштаба предприятия.

На основе этих средств создаются BI-системы, цель которых - повысить качество информации для принятия управленческих решений.

BI-системы также известны под названием *Систем Поддержки Принятия Решений* (СППР, DSS, Decision Support System). Эти системы превращают данные в информацию, на основе которой можно принимать решения, т.е. поддерживающую принятие решений.

Gartner Group определяет состав рынка систем Business Intelligence как набор программных продуктов следующих классов:

- средства построения хранилищ данных (data warehousing, ХД);
- системы оперативной аналитической обработки (OLAP);
- информационно-аналитические системы (Enterprise Information Systems, EIS);
- средства интеллектуального анализа данных (data mining);

- инструменты для выполнения запросов и построения отчетов (query and reporting tools).

Классификация Gartner базируется на методе функциональных задач, где программные продукты каждого класса выполняют определенный набор функций или операций с использованием специальных технологий.

Рынок инструментов Data Mining

На рынке программного обеспечения Data Mining существует огромное разнообразие продуктов, относящихся к этой категории. И не растеряться в нем достаточно сложно. Для выбора продукта следует тщательно изучить задачи, поставленные перед Вами, и обозначить те результаты, которые необходимо получить.

Приведем цитату из Руководства по приобретению продуктов Data Mining (Enterprise Data Mining Buying Guide) компании Aberdeen Group: "Data Mining - технология добычи полезной информации из баз данных. Однако в связи с существенными различиями между инструментами, опытом и финансовым состоянием поставщиков продуктов, предприятиям необходимо тщательно оценивать предполагаемых разработчиков Data Mining и партнеров".

Существуют различные варианты решений по внедрению инструментов Data Mining, например:

- покупка готового программного обеспечения Data Mining;
- покупка программного обеспечения Data Mining, адаптированного под конкретный бизнес;
- разработка Data Mining-продукта на заказ сторонней компанией;
- разработка Data Mining-продукта своими силами;
- различные комбинации вариантов, описанных выше, в том числе использование различных библиотек, компонентов и инструментальных наборов для разработчиков создания встроенных приложений Data Mining.

В этой лекции мы рассмотрим, что предлагает рынок готового программного обеспечения, в частности, оценим рынок в разрезе задач Data Mining.

Поставщики Data Mining

В начале 90-х годов прошлого столетия рынок Data Mining насчитывал около десяти поставщиков. В середине 90-х число поставщиков, представленных компаниями малого, среднего и большого размера, насчитывало более 50 фирм.

Сейчас к аналитическим технологиям, в том числе к Data Mining, проявляется огромный интерес. На этом рынке работает множество фирм, ориентированных на создание инструментов Data Mining, а также комплексного внедрения Data Mining, OLAP и хранилищ данных. Инструменты Data Mining во многих случаях рассматриваются как составная часть BI-платформ, в состав которых также входят средства построения хранилищ и витрин данных, средства обработки неожиданных запросов (ad-hoc query), средства отчетности (reporting), а также инструменты OLAP.

Разработкой в секторе Data Mining всемирного рынка программного обеспечения заняты как всемирно известные лидеры, так и новые развивающиеся компании. Инструменты Data Mining могут быть представлены либо как самостоятельное приложение, либо как дополнения к основному продукту.

Последний вариант реализуется многими лидерами рынка программного обеспечения. Так, уже стало традицией, что разработчики универсальных статистических пакетов, в дополнение к традиционным методам статистического анализа, включают в пакет определенный набор методов Data Mining. Это такие пакеты как SPSS (SPSS, Clementine), Statistica (StatSoft), SAS Institute (SAS Enterprise Miner). Некоторые разработчики OLAP-решений также предлагают набор методов Data Mining, например, семейство продуктов Cognos. Есть поставщики, включающие Data Mining решения в функциональность СУБД: это Microsoft (Microsoft SQL Server), Oracle, IBM (IBM Intelligent Miner for Data).

Рынок поставщиков Data Mining активно развивается. Постоянно появляются новые фирмы-разработчики и новые инструменты.

Относительно цен на инструменты, редактор отмечает, что они имеют тенденцию изменяться, а также отличаются по стоимости для бизнес-пользователей и научных работников, так как последние иногда могут получить бесплатную лицензию для исследований.

Представленные выше продукты, согласно предполагаемой цене для бизнес-пользователей на май 2010 года, сгруппированы следующим образом:

- *Уровень предприятия: (US \$10000 и больше)*

Fair Isaac, IBM, Insightful, KXEN, Oracle, SAS, SPSS.

- *Уровень отдела: (от \$1000 до \$9999)*

Angoss, CART/MARS/TreeNet/Random Forests, Equibits, GhostMiner, Gornik, Mineset, MATLAB, Megaputer, Microsoft SQL Server, Statsoft Statistica, ThinkAnalytics.

- *Личный уровень: (от \$1 до \$999): Excel, See5.*

• *Свободно распространяемое программное обеспечение: C4.5, R, Weka, Xelopes.*

Инструменты Data Mining можно оценивать по различным критериям. Оценка программных средств Data Mining с точки зрения конечного пользователя определяется путем оценки набора его характеристик. Их можно поделить на две группы: бизнес-характеристики и технические характеристики. Это деление является достаточно условным, и некоторые характеристики могут попадать одновременно в обе категории.

Характеристика № 1. Интуитивный интерфейс.

Интерфейс - среда передачи информации между программной средой и пользователем, диалоговая система, которая позволяет передать человеку все необходимые данные, полученные на этапе формализации и вычисления.

Интерфейс подразумевает расположение различных элементов, в т.ч. блоков меню, информационных полей, графических блоков, блоков форм, на экранных формах.

Для удобства работы пользователя необходимо, чтобы интерфейс был интуитивным.

Интуитивный интерфейс позволяет пользователю легко и быстро воспринимать элементы интерфейса, благодаря чему диалог "программная среда-пользователь" становится проще и доступней.

Понятие интуитивного интерфейса включает также понятие знакомой окружающей среды и наличие внятной нетехнической терминологии (например, для сообщения пользователю о совершенной ошибке).

Характеристика № 2. Удобство экспорта/импорта данных.

При работе с инструментом Data Mining-пользователь часто применяет разнообразные наборы данных, работает с различными источниками данных. Это могут быть текстовые файлы, файлы электронных таблиц, файлы баз данных. Инструмент Data Mining должен иметь удобный способ загрузки (импорта) данных. По окончании работы пользователь также должен иметь удобный способ выгрузки (экспорта) данных в удобную для него среду. Программа должна поддерживать наиболее распространенные форматы данных: txt, dbf, xls, csv и другие.

Дополнительное удобство для пользователя создается при возможности загрузки и выгрузки определенной части (по выбору пользователя) импортируемых или экспортируемых полей.

Характеристика № 3. Наглядность и разнообразие получаемой отчетности

Эта характеристика подразумевает получение отчетности в терминах предметной области, а также в качественно спроектированных выходных формах в том количестве, которое может предоставить пользователю всю необходимую результативную информацию.

Характеристика № 4. Легкость обучения работы с инструментарием

Характеристика № 5. Прозрачные и понятные шаги Data Mining-процесса

Характеристика № 6. Руководство пользователя. Существенно упрощает работу пользователя наличие руководства пользователя, с пошаговым описанием шагов генерации моделей Data Mining.

Характеристика № 7. Удобство и простота использования. Существенно облегчает работу начинающего пользователя возможность использовать Мастер или Визард (Wizard).

Характеристика № 8. Для пользователей, не владеющих английским языком, важной характеристикой является **наличие русифицированной версии инструмента**, а также документации на русском языке.

Характеристика № 9. Наличие демонстрационной версии с решением конкретного примера.

Характеристика № 10. Возможности визуализации. Наличие графического представления информации существенно облегчает интерпретируемость полученных результатов.

Характеристика № 11. Наличие значений параметров, заданных по умолчанию. Для начинающих пользователей - это достаточно существенная характеристика, так как при выполнении многих алгоритмов от пользователя требуется задание или выбор большого числа параметров. Особенно много их в инструментах, реализующих метод нейронных сетей. В нейросимуляторах чаще всего заранее заданы значения основных параметров, иной раз неопытным пользователям даже не рекомендуется изменять эти значения. Если же такие значения отсутствуют, пользователю приходится перепробовать множество вариантов, прежде чем получить приемлемый результат.

Характеристика № 12. Количество реализуемых методов и алгоритмов. Во многих инструментах Data Mining реализовано сразу несколько методов, позволяющих решать одну или несколько задач. Если для решения одной задачи (классификации) предусмотрена возможность использования нескольких методов (деревьев решений и нейронных сетей), пользователь получает возможность сравнивать характеристики моделей, построенных при помощи этих методов.

Характеристика № 13. Скорость вычислений и скорость представления результатов.

Характеристика № 14. Наличие квалифицированного ассистента (консультации по выбору методов и алгоритмов), консультационная поддержка.

Характеристика № 15. Возможности поиска, сортировки, фильтрации.

Такая возможность полезна как для входных данных, так и для выходной информации. Применяется сортировка по различным критериям (полям), с возможностью накладывания условий.

При условии фильтрации входных данных появляется возможность построения модели Data Mining на одной из выборок набора данных. Необходимость и польза от проведения такого анализа была описана в одной из лекций, посвященных процессу Data Mining. Фильтрация выходной информации полезна с точки зрения интерпретации результатов. Так, например, иногда при построении деревьев решений результаты получаются слишком громоздкими, и здесь могут оказаться полезными функция как фильтрации, так и поиска и сортировки. Дополнительное удобство для пользователя - цветовая подсветка некоторых категорий записей.

Характеристика № 16. Защита, пароль. Очень часто при помощи Data Mining анализируется конфиденциальная информация, поэтому наличие пароля доступа в систему является желательной характеристикой для инструмента.

Характеристика № 17. Платформы, на которых поддерживается работа инструмента, в частности: PC Standalone (95/98/2000/NT), Unix Server, Unix Standalone, PC Client, NT Server.

Описанные характеристики являются критериями функциональности, удобства, безопасности инструмента Data Mining. При выборе инструмента следует руководствоваться потребностями, а также задачами, которые необходимо решить.

Так, например, если точно известно, что фирме необходимо решать исключительно задачи классификации, то возмож-

ность решения инструментом других задач совсем не является критичной. Однако, следует учитывать, что внедрение Data Mining при серьезном подходе требует серьезных финансовых вложений, поэтому необходимо учитывать все возможные задачи, которые могут возникнуть в перспективе.

Тема 4 Методы, стадии и задачи Data Mining

Основная особенность Data Mining - это сочетание широкого математического инструментария (от классического статистического анализа до новых кибернетических методов) и последних достижений в сфере информационных технологий. В технологии Data Mining гармонично объединились строго формализованные методы и методы неформального анализа, т.е. количественный и качественный анализ данных.

К методам и алгоритмам Data Mining относятся следующие: искусственные нейронные сети, деревья решений, символьные правила, методы ближайшего соседа и k-ближайшего соседа, метод опорных векторов, байесовские сети, линейная регрессия, корреляционно-регрессионный анализ; иерархические методы кластерного анализа, неиерархические методы кластерного анализа, в том числе алгоритмы k-средних и k-медианы; методы поиска ассоциативных правил, в том числе алгоритм Apriori; метод ограниченного перебора, эволюционное программирование и генетические алгоритмы, разнообразные методы визуализации данных и множество других методов.

Большинство аналитических методов, используемые в технологии Data Mining - это известные математические алгоритмы и методы. Новым в их применении является возможность их использования при решении тех или иных конкретных проблем, обусловленная появившимися возможностями технических и программных средств. Следует отметить, что большинство методов Data Mining были разработаны в рамках теории искусственного интеллекта.

Метод (method) представляет собой норму или правило, определенный путь, способ, прием решений задачи теоретического, практического, познавательного, управленческого характера.

Понятие алгоритма появилось задолго до создания электронных вычислительных машин. Сейчас алгоритмы являются основой для решения многих прикладных и теоретических задач в различных сферах человеческой деятельности, в большинстве -

это задачи, решение которых предусмотрено с использованием компьютера.

Алгоритм (algorithm) - точное предписание относительно последовательности действий (шагов), преобразующих исходные данные в искомый результат.

Классификация стадий Data Mining

Data Mining может состоять из двух или трех стадий:

Стадия 1. Выявление закономерностей (свободный поиск).

Стадия 2. Использование выявленных закономерностей для предсказания неизвестных значений (прогностическое моделирование).

В дополнение к этим стадиям иногда вводят стадию валидации, следующую за стадией свободного поиска. Цель валидации - проверка достоверности найденных закономерностей. Однако, мы будем считать валидацию частью первой стадии, поскольку в реализации многих методов, в частности, нейронных сетей и деревьев решений, предусмотрено деление общего множества данных на обучающее и проверочное, и последнее позволяет проверять достоверность полученных результатов.

Стадия 3. Анализ исключений - стадия предназначена для выявления и объяснения аномалий, найденных в закономерностях.

Задачи Data Mining

Напомним, что в основу технологии Data Mining положена концепция шаблонов, представляющих собой закономерности. В результате обнаружения этих, скрытых от невооруженного глаза закономерностей решаются задачи Data Mining. Различным типам закономерностей, которые могут быть выражены в форме, понятной человеку, соответствуют определенные задачи Data Mining.

Задачи (tasks) Data Mining иногда называют закономерностями (regularity) или техниками (techniques).

Единого мнения относительно того, какие задачи следует относить к Data Mining, нет. Большинство авторитетных источников перечисляют следующие: классификация, кластеризация,

прогнозирование, ассоциация, визуализация, анализ и обнаружение отклонений, оценивание, анализ связей, подведение итогов.

Цель описания, которое следует ниже, - дать общее представление о задачах Data Mining, сравнить некоторые из них, а также представить некоторые методы, с помощью которых эти задачи решаются. Наиболее распространенные задачи Data Mining - классификация, кластеризация, ассоциация, прогнозирование и визуализация - будут подробно рассмотрены в последующих лекциях. Таким образом, задачи подразделяются по типам производимой информации, это наиболее общая классификация задач Data Mining. Дальнейшее детальное знакомство с методами решения задач Data Mining будет представлено в следующем разделе курса.

Задачи Data Mining

Классификация (Classification)

Краткое описание. Наиболее простая и распространенная задача Data Mining. В результате решения задачи классификации обнаруживаются признаки, которые характеризуют группы объектов исследуемого набора данных - классы; по этим признакам новый объект можно отнести к тому или иному классу.

Методы решения. Для решения задачи классификации могут использоваться методы: ближайшего соседа (Nearest Neighbor); k-ближайшего соседа (k-Nearest Neighbor); байесовские сети (Bayesian Networks); индукция деревьев решений; нейронные сети (neural networks).

Кластеризация (Clustering)

Краткое описание. Кластеризация является логическим продолжением идеи классификации. Это задача более сложная, особенность кластеризации заключается в том, что классы объектов изначально не predetermined. Результатом кластеризации является разбиение объектов на группы.

Пример метода решения задачи кластеризации: обучение "без учителя" особого вида нейронных сетей - самоорганизующихся карт Кохонена.

Ассоциация (Associations)

Краткое описание. В ходе решения задачи поиска ассоциативных правил отыскиваются закономерности между связанными событиями в наборе данных.

Отличие ассоциации от двух предыдущих задач Data Mining: поиск закономерностей осуществляется не на основе свойств анализируемого объекта, а между несколькими событиями, которые происходят одновременно.

Наиболее известный алгоритм решения задачи поиска ассоциативных правил - алгоритм Apriori.

Последовательность (Sequence), или последовательная ассоциация (sequential association)

Краткое описание. Последовательность позволяет найти временные закономерности между транзакциями. Задача последовательности подобна ассоциации, но ее целью является установление закономерностей не между одновременно наступающими событиями, а между событиями, связанными во времени (т.е. происходящими с некоторым определенным интервалом во времени). Другими словами, последовательность определяется высокой вероятностью цепочки связанных во времени событий. Фактически, ассоциация является частным случаем последовательности с временным лагом, равным нулю. Эту задачу Data Mining также называют задачей нахождения последовательных шаблонов (sequential pattern).

Правило последовательности: после события X через определенное время произойдет событие Y.

Пример. После покупки квартиры жильцы в 60% случаев в течение двух недель приобретают холодильник, а в течение двух месяцев в 50% случаев приобретается телевизор. Решение данной задачи широко применяется в маркетинге и менеджменте, например, при управлении циклом работы с клиентом (Customer Lifecycle Management).

Прогнозирование (Forecasting)

Краткое описание. В результате решения задачи прогнозирования на основе особенностей исторических данных оцениваются пропущенные или же будущие значения целевых численных показателей.

Для решения таких задач широко применяются методы математической статистики, нейронные сети и др.

Тема 5 Классификация методов Data Mining

Классификация методов Data Mining. Далее мы рассмотрим несколько известных классификаций методов Data Mining по различным признакам.

Классификация технологических методов Data Mining

Все методы Data Mining подразделяются на две большие группы по принципу работы с исходными обучающими данными. В этой классификации верхний уровень определяется на основании того, сохраняются ли данные после Data Mining либо они дистиллируются для последующего использования.

1. Непосредственное использование данных, или сохранение данных.

В этом случае исходные данные хранятся в явном детализированном виде и непосредственно используются на стадиях прогностического моделирования и/или анализа исключений. Проблема этой группы методов - при их использовании могут возникнуть сложности анализа сверхбольших баз данных.

Методы этой группы: кластерный анализ, метод ближайшего соседа, метод k-ближайшего соседа, рассуждение по аналогии.

2. Выявление и использование формализованных закономерностей, или дистилляция шаблонов.

При технологии дистилляции шаблонов один образец (шаблон) информации извлекается из исходных данных и преобразуется в некие формальные конструкции, вид которых зависит от используемого метода Data Mining. Этот процесс выполняется на стадии свободного поиска, у первой же группы методов данная стадия в принципе отсутствует. На стадиях прогностического моделирования и анализа исключений используются результаты стадии свободного поиска, они значительно компактнее самих баз данных. Напомним, что конструкции этих моделей могут быть трактуемыми аналитиком либо нетрактуемыми ("черными ящиками").

Методы этой группы: логические методы; методы визуализации; методы кросс-табуляции; методы, основанные на урав-

нениях. Логические методы, или методы логической индукции, включают: нечеткие запросы и анализы; символьные правила; деревья решений; генетические алгоритмы.

Методы этой группы являются, пожалуй, наиболее интерпретируемыми - они оформляют найденные закономерности, в большинстве случаев, в достаточно прозрачном виде с точки зрения пользователя. Полученные правила могут включать непрерывные и дискретные переменные. Следует заметить, что деревья решений могут быть легко преобразованы в наборы символьных правил путем генерации одного правила по пути от корня дерева до его терминальной вершины. Деревья решений и правила фактически являются разными способами решения одной задачи и отличаются лишь по своим возможностям. Кроме того, реализация правил осуществляется более медленными алгоритмами, чем индукция деревьев решений.

Методы кросс-табуляции: агенты, баесовские (доверительные) сети, кросс-табличная визуализация. Последний метод не совсем отвечает одному из свойств Data Mining - самостоятельному поиску закономерностей аналитической системой. Однако, предоставление информации в виде кросс-таблиц обеспечивает реализацию основной задачи Data Mining - поиск шаблонов, поэтому этот метод можно также считать одним из методов Data Mining.

Методы на основе уравнений.

Методы этой группы выражают выявленные закономерности в виде математических выражений - уравнений. Следовательно, они могут работать лишь с численными переменными, и переменные других типов должны быть закодированы соответствующим образом. Это несколько ограничивает применение методов данной группы, тем не менее они широко используются при решении различных задач, особенно задач прогнозирования.

Основные методы данной группы: статистические методы и нейронные сети. Статистические методы наиболее часто применяются для решения задач прогнозирования. Существует множество методов статистического анализа данных, среди них, например, корреляционно-регрессионный анализ, корреляция

рядов динамики, выявление тенденций динамических рядов, гармонический анализ. Другая классификация разделяет все многообразие методов Data Mining на две группы: статистические и кибернетические методы. Эта схема разделения основана на различных подходах к обучению математических моделей.

Следует отметить, что существует два подхода отнесения статистических методов к Data Mining. Первый из них противопоставляет статистические методы и Data Mining, его сторонники считают классические статистические методы отдельным направлением анализа данных. Согласно второму подходу, статистические методы анализа являются частью математического инструментария Data Mining. Большинство авторитетных источников придерживается второго подхода.

В этой классификации различают две группы методов:

- статистические методы, основанные на использовании усредненного накопленного опыта, который отражен в ретроспективных данных;
- кибернетические методы, включающие множество разнородных математических подходов.

Недостаток такой классификации: и статистические, и кибернетические алгоритмы тем или иным образом опираются на сопоставление статистического опыта с результатами мониторинга текущей ситуации.

Преимуществом такой классификации является ее удобство для интерпретации - она используется при описании математических средств современного подхода к извлечению знаний из массивов исходных наблюдений (оперативных и ретроспективных), т.е. в задачах Data Mining.

Рассмотрим подробнее представленные выше группы.

Статистические методы Data Mining

Эти методы представляют собой четыре взаимосвязанных раздела:

- предварительный анализ природы статистических данных (проверка гипотез стационарности, нормальности, независимости, однородности, оценка вида функции распределения, ее параметров и т.п.);

- выявление связей и закономерностей (линейный и нелинейный регрессионный анализ, корреляционный анализ и др.);
- многомерный статистический анализ (линейный и нелинейный дискриминантный анализ, кластерный анализ, компонентный анализ, факторный анализ и др.);
- динамические модели и прогноз на основе временных рядов.

Арсенал статистических методов Data Mining классифицирован на четыре группы методов:

1. Дескриптивный анализ и описание исходных данных.
2. Анализ связей (корреляционный и регрессионный анализ, факторный анализ, дисперсионный анализ).
3. Многомерный статистический анализ (компонентный анализ, дискриминантный анализ, многомерный регрессионный анализ, канонические корреляции и др.).
4. Анализ временных рядов (динамические модели и прогнозирование).

Кибернетические методы Data Mining

Второе направление Data Mining - это множество подходов, объединенных идеей компьютерной математики и использования теории искусственного интеллекта.

К этой группе относятся такие методы:

- искусственные нейронные сети (распознавание, кластеризация, прогноз);
- эволюционное программирование (в т.ч. алгоритмы метода группового учета аргументов);
- генетические алгоритмы (оптимизация);
- ассоциативная память (поиск аналогов, прототипов);
- нечеткая логика;
- деревья решений;
- системы обработки экспертных знаний.

Методы Data Mining также можно классифицировать по задачам Data Mining.

В соответствии с такой классификацией выделяем две группы. Первая из них - это подразделение методов Data Mining

на решающие задачи сегментации (т.е. задачи классификации и кластеризации) и задачи прогнозирования.

В соответствии со второй классификацией по задачам методы Data Mining могут быть направлены на получение описательных и прогнозирующих результатов.

Описательные методы служат для нахождения шаблонов или образцов, описывающих данные, которые поддаются интерпретации с точки зрения аналитика.

К методам, направленным на получение описательных результатов, относятся итеративные методы кластерного анализа, в том числе: алгоритм k-средних, k-медианы, иерархические методы кластерного анализа, самоорганизующиеся карты Кохонена, методы кросс-табличной визуализации, различные методы визуализации и другие.

Прогнозирующие методы используют значения одних переменных для предсказания/прогнозирования неизвестных (пропущенных) или будущих значений других (целевых) переменных.

К методам, направленным на получение прогнозирующих результатов, относятся такие методы: нейронные сети, деревья решений, линейная регрессия, метод ближайшего соседа, метод опорных векторов и др.

Свойства методов Data Mining

Свойства методов Data Mining

Различные методы Data Mining характеризуются определенными свойствами, которые могут быть определяющими при выборе метода анализа данных. Методы можно сравнивать между собой, оценивая характеристики их свойств.

Среди основных свойств и характеристик методов Data Mining рассмотрим следующие: точность, масштабируемость, интерпретируемость, проверяемость, трудоемкость, гибкость, быстрота и популярность.

Масштабируемость - свойство вычислительной системы, которое обеспечивает предсказуемый рост системных характеристик, например, быстроты реакции, общей производительности и пр., при добавлении к ней вычислительных ресурсов. В табл. 1. приведена сравнительная характеристика некоторых распростра-

ненных методов. Оценка каждой из характеристик проведена следующими категориями, в порядке возрастания: чрезвычайно низкая, очень низкая, низкая/нейтральная, нейтральная/низкая, нейтральная, нейтральная/высокая, высокая, очень высокая.

Как видно из рассмотренной таблицы, каждый из методов имеет свои сильные и слабые стороны. Но ни один метод, какой бы не была его оценка с точки зрения присущих ему характеристик, не может обеспечить решение всего спектра задач Data Mining.

Большинство инструментов Data Mining, предлагаемых сейчас на рынке программного обеспечения, реализуют сразу несколько методов, например, деревья решений, индукцию правил и визуализацию, или же нейронные сети, самоорганизующиеся карты Кохонена и визуализацию.

В универсальных прикладных статистических пакетах (например, SPSS, SAS, STATGRAPHICS, Statistica, др.) реализуется широкий спектр разнообразнейших методов (как статистических, так и кибернетических). Следует учитывать, что для возможности их использования, а также для интерпретации результатов работы статистических методов (корреляционного, регрессионного, факторного, дисперсионного анализа и др.) требуются специальные знания в области статистики.

Универсальность того или иного инструмента часто накладывает определенные ограничения на его возможности. Преимуществом использования таких универсальных пакетов является возможность относительно легко сравнивать результаты построенных моделей, полученные различными методами. Такая возможность реализована, например, в пакете Statistica, где сравнение основано на так называемой "конкурентной оценке моделей". Эта оценка состоит в применении различных моделей к одному и тому же набору данных и последующем сравнении их характеристик для выбора наилучшей из них.

Таблица 1

Сравнительная характеристика методов Data Mining

Алгоритм	Точность	Масштабируемость	Интерпретируемость	Пригодность к использованию	Трудоемкость	Быстрота	Популярность, широта использования
<i>классические методы (линейная регрессия)</i>	Нейтральная	высокая	высокая / нейтральная	высокая	нейтральная	высокая	низкая
<i>нейронные сети</i>	высокая	низкая	низкая	низкая	нейтральная	очень низкая	низкая
<i>методы визуализации</i>	высокая	очень низкая	высокая	высокая	очень высокая	чрезвычайно низкая	высокая / нейтральная
<i>деревья решений</i>	низкая	высокая	высокая	высокая / нейтральная	высокая	высокая / нейтральная	высокая / нейтральная
<i>полиномиальные нейронные сети</i>	высокая	нейтральная	низкая	высокая / нейтральная	нейтральная / низкая	низкая / нейтральная	нейтральная
<i>k-ближайшего соседа</i>	низкая	очень низкая	высокая / нейтральная	нейтральная	нейтральная / низкая	высокая	низкая

Тема 6 Сфера применения Data Mining. Применение Data Mining для решения бизнес-задач

Следует отметить, что на сегодняшний день наибольшее распространение технология Data Mining получила при решении бизнес-задач. Возможно, причина в том, что именно в этом направлении отдача от использования инструментов Data Mining может составлять, по некоторым источникам, до 1000% и затраты на ее внедрение могут достаточно быстро окупиться.

Сейчас технология Data Mining используется практически во всех сферах деятельности человека, где накоплены ретроспективные данные.

Мы будем рассматривать четыре основные сферы применения технологии Data Mining подробно: наука, бизнес, исследования для правительства и Web-направление.

- Применение Data Mining для решения бизнес-задач. Основные направления: банковское дело, финансы, страхование, CRM, производство, телекоммуникации, электронная коммерция, маркетинг, фондовый рынок и другие.

- Применение Data Mining для решения задач государственного уровня. Основные направления: поиск лиц, уклоняющихся от налогов; средства в борьбе с терроризмом.

- Применение Data Mining для научных исследований. Основные направления: медицина, биология, молекулярная генетика и геномная инженерия, биоинформатика, астрономия, прикладная химия, исследования, касающиеся наркотической зависимости, и другие.

- Применение Data Mining для решения Web-задач. Основные направления: поисковые машины (search engines), счетчики и другие.

Банковское дело

Технология Data Mining используется в банковской сфере для решения ряда типичных задач.

Задача "Выдавать ли кредит клиенту?"

Классический пример применения Data Mining в банковском деле – решение задачи определения возможной некредитоспособности клиента банка. Эту задачу также называют анализом кредитоспособности клиента или "Выдавать ли кредит клиенту?".

Без применения технологии Data Mining задача решается сотрудниками банковского учреждения на основе их опыта, интуиции и субъективных представлений о том, какой клиент является благонадежным. По похожей схеме работают системы поддержки принятия решений и на основе методов Data Mining. Такие системы на основе исторической (ретроспективной) информации и при помощи методов классификации выявляют клиентов, которые в прошлом не вернули кредит.

Задача "Выдавать ли кредит клиенту?" при помощи методов Data Mining решается следующим образом. Совокупность клиентов банка разбивается на два класса (вернувшие и не вернувшие кредит); на основе группы клиентов, не вернувших кредит, определяются основные "черты" потенциального неплательщика; при поступлении информации о новом клиенте определяется его класс ("вернет кредит", "не вернет кредит").

Задача привлечения новых клиентов банка

С помощью инструментов Data Mining возможно провести классификацию на "более выгодных" и "менее выгодных" клиентов. После определения наиболее выгодного сегмента клиентов банку есть смысл проводить более активную маркетинговую политику по привлечению клиентов именно среди найденной группы.

Другие задачи сегментации клиентов

Разбивая клиентов при помощи инструментов Data Mining на различные группы, банк имеет возможность сделать свою маркетинговую политику более целенаправленной, а потому - эффективной, предлагая различным группам клиентов именно те виды услуг, в которых они нуждаются.

Задача управления ликвидностью банка. Прогнозирование остатка на счетах клиентов

Проводя прогнозирования временного ряда с информацией об остатках на счетах клиентов за предыдущие периоды, применяя методы Data Mining, можно получить прогноз остатка на счетах в

определенный момент в будущем. Полученные результаты могут быть использованы для оценки и управления ликвидностью банка.

Задача выявления случаев мошенничества с кредитными карточками

Для выявления подозрительных операций с кредитными карточками применяются так называемые "подозрительные стереотипы поведения", определяемые в результате анализа банковских транзакций, которые впоследствии оказались мошенническими. Для определения подозрительных случаев используется совокупность последовательных операций на определенном временном интервале. Если система Data Mining считает очередную операцию подозрительной, банковский работник может, ориентируясь на эту информацию, заблокировать операции с определенной карточкой.

Страхование

Страховой бизнес связан с определенным риском. Здесь задачи, решаемые при помощи Data Mining, сходны с задачами в банковском деле.

Информация, полученная в результате сегментации клиентов на группы, используется для определения групп клиентов. В результате страховая компания может с наибольшей выгодой и наименьшим риском предлагать определенные группы услуг конкретным группам клиентов.

Задача выявления мошенничества решается путем нахождения некоего общего стереотипа поведения клиентов-мошенников.

Телекоммуникации

В сфере телекоммуникаций достижения Data Mining могут использоваться для решения задачи, типичной для любой компании, которая работает с целью привлечения постоянных клиентов, - определения лояльности этих клиентов. Необходимость решения таких задач обусловлена жесткой конкуренцией на рынке телекоммуникаций и постоянной миграцией клиентов от одной компании в другую. Как известно, удержание клиента намного дешевле его возврата. Поэтому возникает необходимость выявления определенных групп клиентов и разработка наборов услуг, наиболее привлекательных именно для них. В этой сфере, так же как и во мно-

гих других, важной задачей является выявление фактов мошенничества.

Помимо таких задач, являющихся типичными для многих областей деятельности, существует группа задач, определяемых спецификой сферы телекоммуникаций.

Электронная коммерция

В сфере электронной коммерции Data Mining применяется для формирования рекомендательных систем и решения задач классификации посетителей Web-сайтов. Такая классификация позволяет компаниям выявлять определенные группы клиентов и проводить маркетинговую политику в соответствии с обнаруженными интересами и потребностями клиентов. Технология Data Mining для электронной коммерции тесно связана с технологией Web Mining.

Промышленное производство

Особенности промышленного производства и технологических процессов создают хорошие предпосылки для возможности использования технологии Data Mining в ходе решения различных производственных задач. Технический процесс по своей природе должен быть контролируемым, а все его отклонения находятся в заранее известных пределах;

т.е. здесь мы можем говорить об определенной стабильности, которая обычно не присуща большинству задач, встающих перед технологией Data Mining.

Основные задачи Data Mining в промышленном производстве:

- комплексный системный анализ производственных ситуаций;
- краткосрочный и долгосрочный прогноз развития производственных ситуаций;
- выработка вариантов оптимизационных решений;
- прогнозирование качества изделия в зависимости от некоторых параметров технологического процесса;
- обнаружение скрытых тенденций и закономерностей развития производственных процессов;

- прогнозирование закономерностей развития производственных процессов;
- обнаружение скрытых факторов влияния;
- обнаружение и идентификация ранее неизвестных взаимосвязей между производственными параметрами и факторами влияния;
- анализ среды взаимодействия производственных процессов и прогнозирование изменения ее характеристик;
- выработку оптимизационных рекомендаций по управлению производственными процессами;
- визуализацию результатов анализа, подготовку предварительных отчетов и проектов допустимых решений с оценками достоверности и эффективности возможных реализаций.

Маркетинг

В сфере маркетинга Data Mining находит очень широкое применение.

Основные вопросы маркетинга "Что продается?", "Как продается?", "Кто является потребителем?"

В лекции, посвященной задачам классификации и кластеризации, подробно описано использование кластерного анализа для решения задач маркетинга, как, например, сегментация потребителей.

Другой распространенный набор методов для решения задач маркетинга - методы и алгоритмы поиска ассоциативных правил.

Также успешно здесь используется поиск временных закономерностей.

Розничная торговля

В сфере розничной торговли, как и в маркетинге, применяются:

- алгоритмы поиска ассоциативных правил (для определения часто встречающихся наборов товаров, которые покупатели покупают одновременно). Выявление таких правил помогает размещать товары на прилавках торговых залов, вырабатывать стратегии закупки товаров и их размещения на складах и т.д.

- использование временных последовательностей, например, для определения необходимых объемов запасов товаров на складе.

- методы классификации и кластеризации для определения групп или категорий клиентов, знание которых способствует успешному продвижению товаров.

Фондовый рынок

Вот список задач фондового рынка, которые можно решать при помощи технологии Data Mining:

- прогнозирование будущих значений финансовых инструментов и индикаторов по их прошлым значениям;

- прогноз тренда (будущего направления движения - рост, падение, флэт) финансового инструмента и его силы (сильный, умеренно сильный и т.д.);

- выделение кластерной структуры рынка, отрасли, сектора по некоторому набору характеристик;

- динамическое управление портфелем;

- прогноз волатильности;

- оценка рисков;

- предсказание наступления кризиса и прогноз его развития;

- выбор активов и др.

Кроме описанных выше сфер деятельности, технология Data Mining может применяться в самых разнообразных областях бизнеса, где есть необходимость в анализе данных и накоплен некоторый объем ретроспективной информации.

Применение Data Mining в CRM

Одно из наиболее перспективных направлений применения Data Mining - использование данной технологии в аналитическом CRM.

CRM (Customer Relationship Management) - управление отношениями с клиентами. При совместном использовании этих технологий добыча знаний совмещается с "добычей денег" из данных о клиентах.

Важным аспектом в работе отделов маркетинга и отдела продаж является составление целостного представления о клиентах, информация об их особенностях, характеристиках, структуре клиентской базы. В CRM используется так называемое профилирование клиентов, дающее полное представление всей необходимой информации о клиентах. Профилирование клиентов включает следующие компоненты: сегментация клиентов, прибыльность клиентов, удержание клиентов, анализ реакции клиентов. Каждый из этих компонентов может исследоваться при помощи Data Mining, а анализ их в совокупности, как компонентов профилирования, в результате может дать те знания, которые из каждой отдельной характеристики получить невозможно.

В результате использования Data Mining решается задача сегментации клиентов на основе их прибыльности. Анализ выделяет те сегменты покупателей, которые приносят наибольшую прибыль. Сегментация также может осуществляться на основе лояльности клиентов. В результате сегментации вся клиентская база будет поделена на определенные сегменты, с общими характеристиками. В соответствии с этими характеристиками компания может индивидуально подбирать маркетинговую политику для каждой группы клиентов.

Также можно использовать технологию Data Mining для прогнозирования реакции определенного сегмента клиентов на определенный вид рекламы или рекламных акций - на основе ретроспективных данных, накопленных в предыдущие периоды.

Таким образом, определяя закономерности поведения клиентов при помощи технологии Data Mining, можно существенно повысить эффективность работы отделов маркетинга, продаж и сбыта. При объединении технологий CRM и Data Mining и грамотном их внедрении в бизнес компания получает значительные преимущества перед конкурентами.

Data Mining в исследованиях для правительства

В планах правительства США стоит создание системы, которая позволит отслеживать всех иностранцев, приезжающих в страну. Задача этого комплекса: начиная с пограничного терминала, на основе технологии биометрической идентификации лично-

сти и различных других баз данных контролировать, насколько реальные планы иностранцев соответствуют заявленным ранее (включая перемещения по стране, сроки отъезда и др.). Предварительная стоимость системы составляет более 10 млрд. долларов, разработчик комплекса - компания Accenture.

По данным аналитического отчета Главного контрольного управления американского Конгресса, правительственные ведомства США участвуют приблизительно в двухстах проектах на основе анализа данных (Data Mining), собирающих разнообразную информацию о населении. Более ста из этих проектов направлены на сбор персональной информации (имена, фамилии, адреса e-mail, номера соцстрахования и удостоверений водительских прав), и на основе этой информации осуществляют предсказания возможного поведения людей. Поскольку в упомянутом отчете не приведена информация о секретных отчетах, надо полагать, что общее число таких систем значительно больше.

Несмотря на пользу, которую приносят системы отслеживания, эксперты упомянутого управления, так же как и независимые эксперты, предупреждают о значительном риске, с которым связаны подобные проекты. Причина опасений - проблемы, которые могут возникнуть при управлении и надзоре за такими базами.

Data Mining для научных исследований

Биоинформатика

Одна из научных областей применения технологии Data Mining - биоинформатика, направление, целью которого является разработка алгоритмов для анализа и систематизации генетической информации. Полученные алгоритмы используются для определения структур макромолекул, а также их функций, с целью объяснения различных биологических явлений.

Медицина

Несмотря на консервативность медицины во многих ее аспектах, технология Data Mining в последние годы активно применяется для различных исследований и в этой сфере человеческой деятельности. Традиционно для постановки медицинских диагнозов используются экспертные системы, которые построены на ос-

нове символьных правил, сочетающих, например, симптомы пациента и его заболевание. С использованием Data Mining при помощи шаблонов можно разработать базу знаний для экспертной системы.

Фармацевтика

В области фармацевтики методы Data Mining также имеют достаточно широкое применение. Это задачи исследования эффективности клинического применения определенных препаратов, определение групп препаратов, которые будут эффективны для конкретных групп пациентов. Актуальными здесь также являются задачи продвижения лекарственных препаратов на рынок.

Молекулярная генетика и генная инженерия

В молекулярной генетике и генной инженерии выделяют отдельное направление Data Mining, которое имеет название анализ данных в микро-массивах (Microarray Data Analysis, MDA). Некоторые применения этого направления:

- ранняя и более точная диагностика;
- новые молекулярные цели для терапии;
- улучшенные и индивидуально подобранные виды лечения;
- фундаментальные биологические открытия.

Примеры использования Data Mining - молекулярный диагноз некоторых серьезных заболеваний; открытие того, что генетический код действительно может предсказывать вероятность заболевания; открытие некоторых новых лекарств и препаратов.

Основные понятия, которыми оперирует Data Mining в областях "Молекулярная генетика и генная инженерия" - маркеры, т.е. генетические коды, которые контролируют различные признаки живого организма.

На финансирование проектов с использованием Data Mining в рассматриваемых сферах выделяют значительные финансовые средства.

Химия

Технология Data Mining активно используется в исследованиях органической и неорганической химии. Одно из возможных применений Data Mining в этой сфере - выявление каких-либо спе-

цифических особенностей строения соединений, которые могут включать тысячи элементов.

Далее мы рассмотрим технологии, в основу которых также положено понятие Mining или "добыча".

Web Mining

Web Mining можно перевести как "добыча данных в Web". Web Intelligence или Web Интеллект готов "открыть новую главу" в стремительном развитии электронного бизнеса. Способность определять интересы и предпочтения каждого посетителя, наблюдая за его поведением, является серьезным и критичным преимуществом конкурентной борьбы на рынке электронной коммерции.

Системы Web Mining могут ответить на многие вопросы, например, кто из посетителей является потенциальным клиентом Web-магазина, какая группа клиентов Web-магазина приносит наибольший доход, каковы интересы определенного посетителя или группы посетителей.

Технология Web Mining охватывает методы, которые способны на основе данных сайта обнаружить новые, ранее неизвестные знания и которые в дальнейшем можно будет использовать на практике. Другими словами, технология Web Mining применяет технологию Data Mining для анализа неструктурированной, неоднородной, распределенной и значительной по объему информации, содержащейся на Web-узлах.

Согласно таксономии Web Mining, здесь можно выделить два основных направления: Web Content Mining и Web Usage Mining.

Web Content Mining подразумевает автоматический поиск и извлечение качественной информации из разнообразных источников Интернета, перегруженных "информационным шумом". Здесь также идет речь о различных средствах кластеризации и аннотировании документов.

В этом направлении, в свою очередь, выделяют два подхода: подход, основанный на агентах, и подход, основанный на базах данных.

Подход, основанный на агентах (Agent Based Approach), включает такие системы:

- интеллектуальные поисковые агенты (Intelligent Search Agents);
- фильтрация информации / классификация;
- персонифицированные агенты сети.

Примеры систем интеллектуальных агентов поиска:

- Harvest (Brown и др., 1994),
- FAQ-Finder (Hammond и др., 1995),
- Information Manifold (Kirk и др., 1995),
- OCCAM (Kwok and Weld, 1996), and ParaSite (Spertus, 1997),
- ILA (Information Learning Agent) (Perkowitz and Etzioni, 1995),
- ShopBot (Doorenbos и др., 1996).

Подход, основанный на базах данных (Database Approach), включает системы:

- многоуровневые базы данных;
- системы web-запросов (Web Query Systems);

Примеры систем web-запросов:

- W3QL (Konopnicki и Shmueli, 1995),
- WebLog (Lakshmanan и др., 1996),
- Lorel (Quass и др., 1995),
- UnQL (Buneman и др., 1995 and 1996),
- TSIMMIS (Chawathe и др., 1994).

Второе направление Web Usage Mining подразумевает обнаружение закономерностей в действиях пользователя Web-узла или их группы.

Анализируется следующая информация:

- какие страницы просматривал пользователь;
- какова последовательность просмотра страниц.

Анализируется также, какие группы пользователей можно выделить среди общего их числа на основе истории просмотра Web-узла.

Web Usage Mining включает следующие составляющие:

- предварительная обработка;
- операционная идентификация;

- инструменты обнаружения шаблонов;
- инструменты анализа шаблонов.

При использовании Web Mining перед разработчиками возникает два типа задач. Первая касается сбора данных, вторая - использования методов персонификации. В результате сбора некоторого объема персонифицированных ретроспективных данных о конкретном клиенте, система накапливает определенные знания о нем и может рекомендовать ему, например, определенные наборы товаров или услуг. На основе информации о всех посетителях сайта. Web-система может выявить определенные группы посетителей и также рекомендовать им товары или же предлагать товары в рассылках.

Задачи Web Mining согласно можно подразделить на такие категории:

- Предварительная обработка данных для Web Mining.
- Обнаружение шаблонов и открытие знаний с использованием ассоциативных правил, временных последовательностей, классификации и кластеризации;
- Анализ полученного знания.

Text Mining

Text Mining охватывает новые методы для выполнения семантического анализа текстов, информационного поиска и управления. Синонимом понятия Text Mining является KDT (Knowledge Discovering in Text - поиск или обнаружение знаний в тексте).

В отличие от технологии Data Mining, которая предусматривает анализ упорядоченной в некие структуры информации, технология Text Mining анализирует большие и сверхбольшие массивы неструктурированной информации.

Программы, реализующие эту задачу, должны некоторым образом оперировать естественным человеческим языком и при этом понимать семантику анализируемого текста. Один из методов, на котором основаны некоторые Text Mining системы, - поиск так называемой подстроки в строке.

Call Mining

По словам Энн Беднарц, "добыча звонков" может стать популярным инструментом корпоративных информационных систем. Технология Call Mining объединяет в себя распознавание речи, ее анализ и Data Mining. Ее цель - упрощение поиска в аудио-архивах, содержащих записи переговоров между операторами и клиентами. При помощи этой технологии операторы могут обнаруживать недостатки в системе обслуживания клиентов, находить возможности увеличения продаж, а также выявлять тенденции в обращениях клиентов.

Среди разработчиков новой технологии Call Mining ("добыча" и анализ звонков) - компании CallMiner, Nexidia, ScanSoft, Witness Systems. В технологии Call Mining разработано два подхода - на основе преобразования речи в текст и на базе фонетического анализа.

Примером реализации первого подхода, основанного на преобразовании речи, является система CallMiner. В процессе Call Mining сначала используется система преобразования речи, затем следует ее анализ, в ходе которого в зависимости от содержания разговоров формируется статистика телефонных вызовов. Полученная информация хранится в базе данных, в которой возможен поиск, извлечение и обработка. Пример реализации второго подхода - фонетического анализа - продукция компании Nexidia. При этом подходе речь разбивается на фонемы, являющиеся звуками или их сочетаниями. Такие элементы образуют распознаваемые фрагменты. При поиске определенных слов и их сочетаний система идентифицирует их с фонемами. Аналитики отмечают, что за последние годы интерес к системам на основе Call Mining значительно возрос. Это объясняется тем фактом, что менеджеры высшего звена компаний, работающих в различных сферах, в т.ч. в области финансов, мобильной связи, авиабизнеса, не хотят тратить много времени на прослушивание звонков с целью обобщения информации или же выявления каких-либо фактов нарушений. По словам Дэниэла Хонг, аналитика компании Datamonitor: "Использование этих технологий повышает оперативность и снижает стоимость обработки информации".

Типичная инсталляция продукции от разработчика Nexidia обходится в сумму от 100 до 300 тыс. долл. Стоимость внедрения системы CallMiner по преобразованию речи и набора аналитических приложений составляет около 450 тыс. долл. По мнению Шоллера, приложения Audio Mining и Video Mining найдут со временем гораздо более широкое применение, например, при индексации учебных видеофильмов и презентаций в медиабibliothеках компаний. Однако технологии Audio Mining и Video Mining находятся сейчас на уровне становления, а практическое их применение - на самой начальной стадии.

Тема 7 Комплексный подход к внедрению Data Mining, OLAP и хранилищ данных в СППР

Процесс Data Mining, который как раз и заключается в движении вверх по этой информационной пирамиде, неразрывно связан с процессом принятия решений, его можно рассматривать как неотъемлемую часть систем поддержки принятия решений (СППР).

Таким образом, Data Mining можно рассматривать как процесс поддержки принятия решений, при этом накопленные сведения автоматически обобщаются до информации, которая может быть охарактеризована как знания.

С пониманием решений и принятием решений мы уже кратко познакомились в одной из первых лекций курса.

СППР возникли в результате развития управленческих информационных систем и систем управления базами данных в начале 70-х годов прошлого века.

На данный момент существует огромное количество СППР, разработанных и внедренных в различных областях человеческой деятельности. Темпы их разработок постоянно возрастают.

Однако на сегодняшний день, несмотря на распространенность данных систем, общепризнанное определение данного термина пока не найдено. Следует отметить, что хотя СППР широко применяется во всем мире, на просторах СНГ системам этого типа пока еще не уделяется должное внимание.

Рассмотрим, что же представляет собой система поддержки принятия решений. Как уже было отмечено, данный вопрос является дискуссионным, так же как и вопрос отнесения различных типов систем к классу СППР; мнения по этому поводу часто даже противоречат друг другу. Приведем несколько определений СППР.

Основу СППР составляет комплекс взаимосвязанных моделей с соответствующей информационной поддержкой исследования, экспертные и интеллектуальные системы, включающие опыт решения задач управления и обеспечивающие участие коллектива экспертов в процессе выработки рациональных решений.

Система поддержки принятия решений - это диалоговая автоматизированная система, использующая правила принятия реше-

ний и соответствующие модели с базами данных, а также интерактивный компьютерный процесс моделирования.

СППР - это средство для "вычисления решений", которое основано "на использовании ряда процедур по обработке данных и суждений, помогающих лицу, принимающему решение (далее - ЛПР), в принятии решения".

СППР - "интерактивные автоматизированные системы, которые помогают ЛПР использовать данные и модели, чтобы решать неструктурированные проблемы".

СППР - "компьютерная информационная система, используемая для поддержки различных видов деятельности при принятии решения в ситуациях, где невозможно или нежелательно иметь автоматические системы, которые полностью выполняют весь процесс принятия решения". СППР не заменяет ЛПР, автоматизируя процесс принятия решения, а оказывает ему помощь в ходе решения поставленной задачи.

Следует заметить, что, начиная с первых определений СППР, круг задач, решаемых при их помощи, ограничился слабоструктурированными и неструктурированными.

Определим СППР таким образом: СППР - интерактивная компьютерная система, предназначенная для поддержки принятия решений в слабоструктурированных и неструктурированных проблемах различных видов человеческой деятельности.

Существенными концепциями этого определения являются:

- компьютерная интерактивная (т.е. не обуславливающая обязательного непосредственного использования ЛПР системы поддержки принятия решений);
- поддержка принятия решений (решение принимает человек);
- слабоструктурированных и неструктурированных проблем (именно такими проблемами занимаются руководители).

Рассмотрим, что же представляет собой классификация проблем на слабоструктурированные, неструктурированные и структурированные.

Неструктурированные задачи имеют только качественное описание, основанное на суждениях ЛПР, количественные зависимости между основными характеристиками задачи не известны.

Структурированные задачи характеризуются существенными зависимостями, которые могут быть выражены количественно.

Слабоструктурированные задачи занимают промежуточное положение и являются "сочетающими количественные и качественные зависимости, причем малоизвестные и неопределенные стороны задачи имеют тенденцию доминировать".

Можно выделить три компонента, составляющие основу классической структуры СППР, которыми она отличается от других типов информационных систем: подсистему интерфейса пользователя, подсистему управления базой данных и подсистему управления базой моделей.

Если посмотреть на СППР с функциональной стороны, можно выделить следующие ее компоненты:

- сервер хранилища данных;
- инструментарий OLAP;
- инструментарий Data Mining.

Эти компоненты СППР рассматривают такие основные вопросы: вопрос накопления данных и их моделирования на концептуальном уровне, вопрос эффективной загрузки данных из нескольких независимых источников и вопрос анализа данных.

Можно сказать, что использование оперативной аналитической обработки (систем OLAP) на сегодня ограничивается обеспечением доступа к многомерным данным.

Технология Data Mining представляет в СППР наибольший интерес, поскольку с ее помощью можно провести наиболее глубокий и всесторонний анализ данных и, следовательно, принимать наиболее взвешенные и обоснованные решения.

Тема 8 Классификация СППР

Современные системы поддержки принятия решения (СППР) представляют собой системы, максимально приспособленные к решению задач повседневной управленческой деятельности, являются инструментом, призванным оказать помощь лицам, принимающим решения (ЛПР). С помощью СППР может производиться выбор решений некоторых неструктурированных и слабоструктурированных задач, в том числе и многокритериальных.

СППР, как правило, являются результатом мультидисциплинарного исследования, включающего теории баз данных, искусственного интеллекта, интерактивных компьютерных систем, методов методов имитационного моделирования.

С момента появления первых разработок по созданию СППР, не было дано четкого определения СППР.

Ранние определения СППР (в начале 70-х годов прошлого века) отражали следующие три момента: (1) возможность оперировать с неструктурированными или слабоструктурированными задачами, в отличие, от задач, с которыми имеет дело исследование операций; (2) интерактивные автоматизированные (то есть реализованные на базе компьютера) системы; (3) разделение данных и моделей. Приведем определения СППР: СППР — совокупность процедур по обработке данных и суждений, помогающих руководителю в принятии решений, основанная на использовании моделей.

СППР — это интерактивные автоматизированные системы, помогающие ЛПР, использовать использовать данные и модели для решения слабоструктурированных проблем.

СППР — это система, которая обеспечивает пользователям доступ к данным и/или моделям, так что они могут принимать лучшие решения.

Последнее определение не отражает участия компьютера в создании СППР, вопросы возможности включения нормативных моделей в состав СППР и др.

В настоящее время нет общепринятого определения СППР, поскольку конструкция СППР существенно зависит от вида задач, для решения которых она разрабатывается, от доступных данных, информации и знаний, а также от пользователей системы. Можно привести, тем не менее, некоторые элементы и характеристики, общепризнанные, как части СППР:

СППР – в большинстве случаев – это интерактивная автоматизированная система, которая помогает пользователю (ЛПР) использовать данные и модели для идентификации и решения задач и принятия решений. Система должна обладать возможностью работать с интерактивными запросами с достаточно простым для изучения языком запросов.

СППР обладает следующими четырьмя основными характеристиками:

1. СППР использует и данные, и модели;
2. СППР предназначены для помощи менеджерам в принятии решений для слабоструктурированных и неструктурированных задач;

3. Они поддерживают, а не заменяют, выработку решений менеджерами;

4. Цель СППР — улучшение эффективности решений.

Идеальная СППР:

1. оперирует со слабоструктурированными решениями;
2. предназначена для ЛПР различного уровня;
3. может быть адаптирована для группового и индивидуального использования;

4. поддерживает как взаимозависимые, так и последовательные решения;

5. поддерживает 3 фазы процесса решения: интеллектуальную часть, проектирование и выбор;

6. поддерживает разнообразные стили и методы решения, что может быть полезно при решении задачи группой ЛПР;

7. является гибкой и адаптируется к изменениям как организации, так и ее окружения;

8. проста в использовании и модификации;

9. улучшает эффективность процесса принятия решений;

10.позволяет человеку управлять процессом принятия решений с помощью компьютера, а не наоборот;

11.поддерживает эволюционное использование и легко адаптируется к изменяющимся требованиям;

12.может быть легко построена, если может быть сформулирована логика конструкции СППР;

13.поддерживает моделирование;

14.позволяет использовать знания.

Вопрос классификаций СППР на сегодняшний день является актуальным, продолжают разрабатки новых таксономий. Рассмотрим две из них.

Ниже приведена классификация СППР по сходству некоторых признаков (D.J. Power, 2000):

- СППР, ориентированные на данные (Data-driven DSS, Data-oriented DSS);
- СППР, ориентированные на модели (Model-driven DSS);
- СППР, ориентированные на знания (Knowledge-driven DSS);
- СППР, ориентированные на документы (Document-driven DSS);
- СППР, ориентированные на коммуникации и групповые СППР;
- Интер-организованные и Интра-организованные СППР (Inter-Organizational или Intra-Organizational DSS);
- Специфически функциональные СППР или СППР общего назначения (Function-Specific или General Purpose DSS);
- СППР на базе Web (Web-Based DSS).

В зависимости от данных, с которыми работают СППР, выделяют два основных их типа СППР: EIS и DSS.

EIS (Execution Information System) - информационная система Руководства, ИСР.

СППР этого типа являются оперативными, предназначенными для немедленного реагирования на текущую ситуацию. В большинстве они ориентированы на неподготовленного пользователя, потому имеют упрощенный интерфейс, базовый набор предлагаемых возможностей, фиксированные формы представления информации и перечень решаемых задач. Такие системы основаны на типичных запросах, число которых относительно невелико; от-

четы, полученные в результате таких запросов, представляются в максимально удобном виде.

DSS (Decision Support System). К системам этого типа относят многофункциональные системы анализа и исследования данных. Они предполагают глубокую проработку данных, которую можно использовать в процессе принятия решений.

Системы этого типа, в отличие от EIS, рассчитаны на пользователей, имеющих как знания в предметной области, так и возможности использования современных компьютерных технологий. Этим системам присущи черты искусственного интеллекта, за счет возможности проработки исходных данных в конкретные выводы по поставленной задаче. Такие системы имеет смысл создавать, если есть основания для обобщения и анализа данных и процессов их обработки.

В последнее время к СППР относят только второй тип, т.е. DSS.

Системы этого типа иногда называют динамическими, т.е. они должны быть ориентированы на обработку неожиданных (ad hoc) запросов. Поддержка принятия решений на основе накопленных данных может выполняться в трех базовых сферах:

1. Область детализированных данных (OLTP-системы).

Целью большинства таких систем является поиск информации, это так называемые информационно-поисковые системы. Они могут использоваться в качестве надстроек над системами обработки данных или как хранилища данных.

2. Сфера агрегированных показателей (OLAP-системы).

Задачами OLAP систем является обобщение, агрегация, гиперкубическое представление информации и многомерный анализ. Это могут быть многомерные СУБД или же реляционные базы с предварительной агрегацией данных.

3. Сфера закономерностей (Data Mining).

Такое деление систем на EIS и DSS не обязательно означает реализацию СППР одного из типов. Они могут существовать параллельно, когда каждая из систем предоставляет свои функции определенной категории пользователей.

Общая схема поддержки принятия решений включает:

- помощь ЛПП при оценке состояния управляемой системы и воздействий на нее; выявление предпочтений ЛПП;
- генерацию возможных решений;
- оценку возможных альтернатив, исходя из предпочтений ЛПП;
- анализ последствий принимаемых решений и выбор лучшего с точки зрения ЛПП.

OLAP-системы

В основе концепции OLAP, или оперативной аналитической обработки данных (On-Line Analytical Processing), лежит многомерное концептуальное представление данных (Multidimensional conceptual view).

Термин OLAP введен Коддом (E. F. Codd) в 1993 году. Главная идея данной системы заключается в построении многомерных таблиц, которые могут быть доступны для запросов пользователей. Эти многомерные таблицы или так называемые многомерные кубы строятся на основе исходных и агрегированных данных. И исходные, и агрегированные данные для многомерных таблиц могут храниться как в реляционных, так и в многомерных базах данных. Взаимодействуя с OLAP-системой, пользователь может осуществлять гибкий просмотр информации, получать различные срезы данных, выполнять аналитические операции детализации, свертки, сквозного распределения, сравнения во времени. Вся работа с OLAP-системой происходит в терминах предметной области.

OLAP-продукты

Сейчас на рынке представлено огромное многообразие OLAP-систем. Разработано несколько классификаций продуктов этого типа: например, классификация по способу хранения данных, по месту нахождения OLAP-машины, по степени готовности к применению. Рассмотрим первую из приведенных классификаций.

Существует три способа хранения данных в OLAP-системах или три архитектуры OLAP-серверов:

- MOLAP (Multidimensional OLAP);
- ROLAP (Relational OLAP);
- HOLAP (Hybrid OLAP).

Таким образом, согласно этой классификации OLAP-продукты могут быть представлены тремя классами систем.

- В случае MOLAP, исходные и многомерные данные хранятся в многомерной БД или в многомерном локальном кубе. Такой способ хранения обеспечивает высокую скорость выполнения OLAP-операций. Но многомерная база в этом случае чаще всего будет избыточной. Куб, построенный на ее основе, будет сильно зависеть от числа измерений. При увеличении количества измерений объем куба будет экспоненциально расти. Иногда это может привести к "взрывному росту" объема данных, парализующему в результате запросы пользователей.

- В ROLAP-продуктах исходные данные хранятся в реляционных БД или в плоских локальных таблицах на файл-сервере. Агрегатные данные могут помещаться в служебные таблицы в той же БД. Преобразование данных из реляционной БД в многомерные кубы происходит по запросу OLAP-средства. При этом скорость построения куба будет сильно зависеть от типа источника данных, и поэтому время отклика системы порой становится неприемлемо большим.

- В случае использования гибридной архитектуры, т.е. в HOLAP-продуктах, исходные данные остаются в реляционной базе, а агрегаты размещаются в многомерной. Построение OLAP-куба выполняется по запросу OLAP-средства на основе реляционных и многомерных данных. Такой подход позволяет избежать взрывного роста данных. При этом можно достичь оптимального времени исполнения клиентских запросов.

Следующая классификация - по месту размещения OLAP-машины. По этому признаку OLAP-продукты делятся на OLAP-серверы и OLAP-клиенты.

В серверных OLAP-средствах вычисления и хранение агрегатных данных выполняются отдельным процессом – сервером. Клиентское приложение получает только результаты запросов к многомерным кубам, которые хранятся на сервере. Некоторые OLAP-серверы поддерживают хранение данных только в реляционных базах, другие - только в многомерных. Многие современные OLAP-серверы поддерживают все три способа хранения данных:

MOLAP, ROLAP и HOLAP. Одним из самых распространенным в настоящее время серверным решением является OLAP-сервер корпорации Microsoft. OLAP-клиент устроен по-другому. Построение многомерного куба и OLAP-вычисления выполняются в памяти клиентского компьютера.

С помощью OLAP-сервера может быть организовано физическое хранение обработанной многомерной информации, что позволяет быстро выдавать ответы на запросы пользователя. Кроме того, предусматривается преобразование данных из реляционных и других баз в многомерные структуры в режиме реального времени. Каким образом реляционные и многомерные средства работают совместно? OLAP продукты вливаются в существующую корпоративную инфраструктуру путем интегрирования с реляционными системами. Администраторы баз данных либо загружают реляционные данные в многомерный кэш, либо настраивают кэш для доступа к SQL-данным.

В табл.2 приведены сравнительные характеристики различных моделей управления данными:

Таблица 2

Сравнительные характеристики различных моделей управления
данными

Характери- стики	Реляци- онные СУБД OLTP	Реляционные СУБД СППР/Хранил ища данных	Многомер- ные СУБД OLAP
Типовая опе- рация	Обновле- ние	Отчет	Анализ
Уровень ана- литических требований	Низкий	Средний	Высокий
Экраны	Неизме- няемые	Определяемые пользователем	Определяе- мые пользо- вателем
Объем данных на транзакцию	Неболь- шой	От малого до большого	Большой
Уровень дан- ных	Деталь- ные	Детальные и суммарные	В основном суммарные
Сроки хране- ния данных	Только текущие	Исторические и текущие	Историче- ские, теку- щие и про- гнозируемые
Структурные элементы	Записи	Записи	Массивы

Тема 9 История создания СППР

До середины 60-х годов прошлого века создание больших информационных систем (ИС) было чрезвычайно дорогостоящим, поэтому первые ИС менеджмента (так называемые Management Information Systems — MIS) были созданы в эти годы лишь в достаточно больших компаниях. MIS предназначались для подготовки периодических структурированных отчетов для менеджеров.

В конце **60-х годов** появляется новый тип ИС — модель-ориентированные СППР (Model-oriented Decision Support Systems — DSS) или системы управленческих решений (Management Decision Systems — MDS). По мнению первооткрывателей СППР Keen P. G. W., Scott Morton M. S.[16] (1978), концепция поддержки решений была развита на основе «теоретических исследований в области принятия решений... и технических работ по созданию интерактивных компьютерных систем».

В **1971 г.** — опубликована книга Scott Morton'а, в которой впервые были описаны результаты внедрения СППР, основанной на использовании математических моделей.

1974 г. — в одной из работ было дано определение ИС менеджмента — MIS (Management Information System): «MIS — это интегрированная человеко-машинная система обеспечения информацией, поддерживающая функции операций, менеджмента и принятия решений в организации. Системы используют компьютерную технику и программное обеспечение, модели управления и принятия решений, а также базу данных».

1975 г. — J.D.C.Little в своей работе предложил критерии проектирования СППР в менеджменте.

1978 г. — опубликован учебник по СППР, в котором исчерпывающе описаны аспекты создания СППР: анализ, проектирование, внедрение, оценка и разработка.

1980 г. — опубликована диссертация S. Alter, в которой он дал основы классификации СППР.

1981 г. — Bonczek, Holsapple и Whinston в своей книге создали теоретические основы проектирования СППР. Они выделили 4 необходимых компонента, присущих всем СППР: 1) Языковая

система (Language System — LS) — СППР может принимать все сообщения; 2) Система презентаций (Presentation System (PS)) (СППР может выдавать свои сообщения); 3) Система знаний (Knowledge System — KS) — все знания СППР сохраняет; 4) Система обработки задач (Problem-Processing System (PPS)) — программный «механизм», который пытается распознать и решить задачу во время работы СППР.

1981 г. — В своей книге R.Sprague и E.Carlson описали, каким образом на практике можно построить СППР. Тогда же была разработана информационная система руководителя (Executive Information System (EIS)) — компьютерная система, предназначенная для обеспечения текущей адекватной информации для поддержки принятия управленческих решений менеджером.

Начиная с **1990-х**, разрабатываются так называемые Data Warehouses — хранилища данных. В **1993 г.** Е. Коддом для СППР специального вида был предложен термин OLAP (Online Analytical Processing)- оперативный анализ данных, онлайн-аналитическая обработка данных для поддержки принятия важных решений. Исходные данные для анализа представлены в виде многомерного куба, по которому можно получать нужные разрезы — отчёты. Выполнение операций над данными осуществляется OLAP-машиной. По способу хранения данных различают MOLAP, ROLAP и HOLAP. По месту размещения OLAP-машины различаются OLAP-клиенты и OLAP-серверы. OLAP-клиент производит построение многомерного куба и вычисления на клиентском ПК, а OLAP-сервер получает запрос, вычисляет и хранит агрегатные данные на сервере, выдавая только результаты.

В начале нового тысячелетия была создана СППР на основе Web. **27 октября 2005** года в Москве на Международной конференции «Информационные и телемедицинские технологии в охране здоровья» (ITHC 2005), А. Пастухов (Россия) представил СППР нового класса — PSTM (Personal Information Systems of Top Managers). Основным отличием PSTM от существующих СППР является построение системы для конкретного лица, принимающее решение, с предварительной логико-аналитической обработкой информации в автоматическом режиме и выводом информации на один экран.

Интересно отметить создание предтечи СППР коллежским советником С. Н. Корсаковым, опубликовавшим еще в 1832 году описание механических устройств, так называемых «интеллектуальных машин», которые "могли быть использованы при решении различных задач в повседневной жизни, для того, чтобы сделать какой бы то ни было вывод", например помочь принять решение о наиболее подходящих лекарствах по наблюдаемым у пациента симптомам заболевания.

Тема 10 Сфера применения СППР

СППР призваны помогать людям, принимающим решение в сложных для полного и объективного анализа областях деятельности. Системы поддержки принятия решений применяются, в основном, на верхнем уровне управления, имеющего стратегическое долгосрочное значение. К таким задачам относят формирование стратегии предприятия, разработка плана привлечения экономических ресурсов и т.д.

СППР необходимы всем крупным организациям, которые хотят получать своевременную, полную и достоверную информацию о своей деятельности из единого источника, а также иметь возможность комплексного анализа этой информации и извлечения из большого объема данных показателей, требующихся для принятия управленческих решений. Это хорошо понятно на примере сетей розничной торговли: владелец маленького магазинчика все помнит сам, и ему не нужна СППР для понимания, какой ассортимент нужно покупать, а крупная розничная сеть без автоматизации просто не справится.

СППР, как следует из самого названия, в первую очередь нужны лицам, принимающим решения. Это топ-менеджмент, руководители отделов, да и все сотрудники, принимающие важные решения для стабильной ежедневной работы предприятия.

Учитывая, что любое предприятие является цельным организмом, не совсем удачно «отдавать на откуп» СППР какую-либо часть деятельности компании. Ведь для решения даже одной задачи требуется целевой поиск разнообразной информации, вполне возможно эту информацию использует и другой отдел. Максимальную пользу принесет внедрение СППР не в одном конкретном отделе, а на всем предприятии: это ведет к сокращению затрат — как временных, так и финансовых (не говоря уже об упущенных выгодах компании, не консолидирующей многообразную информацию о своей деятельности в доступном для обработки формате).

Большие просторы открывают СППР для маркетинговой деятельности предприятия: прогноз продаж, сегментация клиентов и др. В этом случае основными потребителями СППР будут марке-

тологи и аналитики. Продуктивно использование СППР для принятия важных решений в области прогнозирования поломок оборудования и, соответственно, плановых ремонтов. Оценка поставщиков в отделах закупок — актуальная задача, решаемая также с помощью СППР.

СППР необходимы там, где существует необходимость выбора из множества альтернатив. Но это только общие описания, которые в действительности не раскрывают сущность вопроса. В настоящее время системы поддержки принятия решений востребованы почти во всех областях и на всех стадиях экономики и управления, как в бизнесе, так и в государственных организациях. А в работе с СППР задействованы не только лица, наделенные правом принятия решений, но и менеджеры среднего звена.

Во всех сферах, где необходимы системы, существуют общие характеристики:

- стратегическое планирование и управление;
- большие массивы обрабатываемой и хранящейся информации;
- слабо структурированные или непрограммируемые процессы;
- быстрота реагирования на изменяющийся рынок или бизнес-процессы;
- большое количество аналитических процессов.

Изначально СППР были активно востребованы банками и финансовыми структурами, в работе которых стратегическое планирование преобладает над операционным. В бизнес-процессах банков изначально заложено большое количество данных, требующих постоянного анализа и последующего проектирования, в зависимости от множества критериев поведения как потребителей, конкурентов, так и окружающей среды вообще.

Но в последнее время системы поддержки принятия решений начинают активно использовать даже государственные ВУЗы. Это связано с тем, что учебные заведения активно стремятся стать бизнес-единицами и перейти на новый уровень управления. В качестве резюме можно констатировать следующее: СППР необходима людям и организациям, управляющим своим капиталом с по-

мощью стратегического планирования и стремящимся максимально исключить риски.

Методики, заложенные в СППР

Для анализа и выработки предложений в СППР используются разные методы. Среди них: информационный поиск, интеллектуальный анализ данных, поиск знаний в базах данных, рассуждение на основе прецедентов, имитационное моделирование, генетические алгоритмы, нейронные сети и др. Некоторые из них были разработаны в рамках искусственного интеллекта. Если в основе работы системы лежит один или несколько таких методов, то говорят об интеллектуальной СППР (ИСППР).

Функционально в систему поддержки принятия решений должны входить как минимум две составляющие:

- база знаний как основа для анализа;
- аналитический аппарат, который формирует советы, их обоснования, отчеты, проводит обработку информации и расчеты.

Такая теория является достаточно общей и абстрактной, не позволяющей вывести ни стандартов для систем поддержки принятия решений, ни общей оболочки. Если же делалась попытка формализовать системы этого класса, то в результате появлялась либо надстройка над системой, либо система отчетности, либо более-менее структурированная база знаний.

В 80-е годы и в начале 90-х, когда интерес к системам этого класса был наиболее высок, «поддержка принятия решений» воспринималась как попытка присвоить компьютеру статус высшего разума. К сожалению, а может быть и к счастью, этого не произошло. Очевидно, не было наработано достаточно опыта, чтобы переосмыслить создание принципиально нового думающего устройства из программируемого железного ящика. В теории это было так. А на практике мы не заметили, как нас окружили вполне интеллектуальные системы, которые, безусловно, помогают принимать решения. Однако мы не квалифицируем эти системы как класс «систем поддержки принятия решений».

В настоящее время разработано большое количество методов и методик для СППР. Под методологией принятия решения понимается логическая организация деятельности по разработке

управленческого решения, включающая в себя формулирование цели принятия решения, методы разработки решений, критерии оценки качества решения и выбор альтернативных вариантов.

Управленческие решения целесообразно классифицировать с методологической точки зрения, среди которых можно выделить:

1. Решения по составлению плановых заданий. Программы поддержки решений такого рода представляют собой имитационные модели системы с детерминированными переменными. Целью поддержки принятия решения является обеспечение возможности быстрого «проигрывания» вариантов альтернативных сценариев планов и выбор наилучшего из них.

2. Решения инвестиционного характера по крупным изменениям в существующей технологии. Их отличие от решений, описанных выше, заключается в том, что, как правило, не существует обязательного регламента их принятия и оценки эффекта от их применения. Поэтому решения такого рода зачастую могут вообще не приниматься либо приниматься на интуитивной основе, что может привести к тому, что их претворение в жизнь или не приведет к улучшению работы предприятия, или не будет осуществлено из-за нехватки ресурсов и финансовых средств. Назначение систем поддержки принятия решений состоит в том, чтобы руководитель быстро нашел такое, в правильности которого он уверен. Наличие в системе управления предприятием его имитационной модели позволяет с достаточной степенью точности и незначительными затратами времени «проиграть» различные варианты и выбрать из них наилучший, а для формализуемых задач — оптимальный.

3. Решения по отслеживанию соблюдения балансовых и затратных норм на всех уровнях производства и управления. Программы поддержки решений такого рода должны включать в себя помимо имитационных моделей модули автоматизации процесса получения выводов в части соответствия расчетных заданий ходу их выполнения. Такие модули получения выводов по накопленным знаниям должны обращать внимание на необходимость принятия решения в той или иной ситуации, требующей обязательного реагирования.

4. Решения по совершенствованию уровня технологических и управленческих процессов. Программы поддержки принятия таких решений должны включать в себя базы накопленных знаний по передовым разработкам в соответствующей предметной области с возможностью автоматизированного «поднятия вопроса» о возможности использования более прогрессивных, чем применяемые на предприятии.

Работа с СППР — это, в значительной степени, интерактивный и итерационный процесс. Цель любой модели или методики, закладываемой в СППР, — описание в математических терминах тех или иных событий или объектов. Компьютерная система, используя различные данные, исходную информацию, правила и алгоритмы, предлагает набор решений. Пользователь системы оценивает полученные результаты исходя из своих представлений и, в случае необходимости, уточняет запросы, задействует альтернативные сценарии и т.д.

Ценность информации, получаемой от СППР, во многом определяется заложенными в основу ее работы моделями и сценариями, а также полнотой и достоверностью исходных данных. И если показатели деятельности компании сравнительно легко можно получить из «внутренних» бизнес-приложений, то подключение, постоянное обновление и обработка данных из внешних источников часто является достаточно сложной задачей — как при постановке, так и в процессе реализации.

Перечень используемых моделей работы с данными (Data Mining) достаточно широк: это и статистические методы, и нейронные сети, и так называемое «дерево решений» и многие другие. В целом можно сказать, что их настройка и применение требуют от пользователя высокого уровня подготовки. Поэтому, учитывая, что часто пользователь системы не является специалистом в ИТ и математике, большое значение имеют грамотная организация интерфейса и предварительная проработка сценариев использования системы.

Тема 11 Задачи, решаемые с помощью СППР

Наиболее часто решаемые задачи: автоматический сбор информации, обогащение данных, отчетность, прогнозирование, анализ рисков, анализ отклонений. Но не всегда возможно именно такое сочетание — «проще и эффективнее». Придется выбирать, так как чем эффективнее, тем сложнее.

Есть простые и понятные механизмы анализа, например визуализация (такая, как OLAP). Это удобно, легко запускается, дает быстрый эффект и помогает принимать решения. Но такими простыми способами решаются относительно несложные задачи: анализ динамики, выявление самых ходовых товаров и т.п. Это доступно всем, поэтому с их помощью трудно обеспечить серьезное конкурентное преимущество компании.

Есть задачи, способные принести огромный эффект, например оптимизация запасов, но они сложны. Для принятия решений в области управления запасами нужно прогнозировать спрос, учитывать особенности потребления, влияние внешних факторов и много всего остального. Когда удастся решить данную задачу, то отдачу организация получает большую, но эту проблему простой не назовешь.

Основная задача системы поддержки принятия решения — предоставить аналитикам инструмент для выполнения углубленного анализа данных. По степени интеллектуальности обработки данных при анализе выделяют три класса задач анализа:

1. **Информационно-поисковый.** Система осуществляет поиск необходимых данных в соответствии с заранее определенными запросами. Задачи этого класса решаются построением систем информационно-поискового анализа на базе реляционных СУБД и статических запросов с использованием языка SQL.

2. **Оперативно-аналитический.** Система производит группировку и обобщение данных в любом виде, необходимом аналитику. Причем в этом случае заранее невозможно предсказать необходимые аналитику запросы. Для этого класса задач необходимо построение систем оперативного анализа с применением тех-

нологии оперативной аналитической обработки данных OLAP, использующей концепцию многомерного анализа данных.

3. **Интеллектуальный.** Система осуществляет поиск функциональных и логических закономерностей в накопленных данных, построение моделей и правил, которые объясняют найденные закономерности и/или с определенной вероятностью прогнозируют развитие некоторых процессов.

Такого рода задачи решаются построением систем интеллектуального анализа, реализующего методы и алгоритмы Data Mining.

Системы поддержки принятия решений могут применяться в различных сферах деятельности. Проще и эффективнее СППР справляются с задачами, требующими проработки больших массивов информации. А учитывая повальную информатизацию бизнеса, когда накопление информации растет лавинообразно, это и многие аналитические задачи.

Проще, наверное, указать направления, где эффективно использование СППР:

- торговля, как оптовая, так и розничная (прогноз продаж, сегментация клиентов, ассортиментная политика, анализ потребительской корзины, программы лояльности, оценка эффективности маркетинговых действий, анализ аномалий, пресечение мошеннических действий персонала и многое другое);
- интернет-бизнес (построение рекомендательных систем для персонализации пользователей веб-сайтов, с целью повышения лояльности покупателей, и, как следствие, повышение продаж, выявление случаев мошенничества и т.д.);
- телекоммуникационный бизнес (к примеру, анализ доходности и риска клиентов);
- промышленное производство (прогнозирование качества производимого изделия в зависимости от измеряемых параметров технологического процесса, планирование ремонтов);
- медицина (диагностика заболеваний, оценка диагностических тестов, выявление побочных эффектов);
- банковская деятельность (наиболее распространенная задача — кредитный скоринг, прогноз остатков на счетах и др.);

- энергетика (главная задача — прогноз потребления электроэнергии);
- страховой бизнес;
- государственные учреждения;
- маркетинговые, исследовательские агентства - и многие другие компании.

СППР может применяться практически везде, где требуется решить задачу на основании анализа данных.

Наиболее эффективным образом СППР справляются с задачами, где можно применить четкие математические модели. Это могут быть задачи на оптимизацию различных функций (например, логистических операций, планирования товарных запасов, оптимизации торгового ассортимента и его размещения в торговом зале и пр.), а также поиск тенденций, закономерностей, причинноследственный анализ, кластеризация и типизация объектов (в частности, клиентов).

СППР хорошо зарекомендовали себя и в качестве инструмента наглядной визуализации сложных многофакторных процессов с использованием слабо структурированных данных. В простейшем случае, СППР — это просто удобный инструмент руководителя для операционного контроля и принятия тактических решений.

Тема 12 Задачи, которые невозможно решить при помощи СППР

Полностью вообще ни одну задачу невозможно решить при помощи СППР-системы. Даже при получении, казалось бы, простых отчетов возникает множество нюансов. Ведь неспроста в названии используется слово «поддержки». Самые же сложные вещи касаются задач прогнозирования, моделирования, то есть тех, которые требуют предвидения чего-либо. Можно говорить, что чем больше технология работы компании похожа на конвейер, тем больше пользы от СППР-систем. Например, хорошо это работает в розничных сетях, у операторов телекоммуникационных услуг, в банках. В этих случаях удастся формализовать сам процесс принятия решений. Какие для этого используются технологии, не столь важно, главное, что формализованные решения можно применить на практике.

Чем более сложные продукты предлагает компания, чем больше уделяется внимания нюансам, тем менее эффективны и менее применимы СППР-решения. В этом случае, пожалуй, работают только различные инструменты визуализации и механизмы обогащения данных. Другими словами, автоматически принять решение не получается, но можно предоставить потребителю как можно больше информации в удобном виде, а дальше он сам должен на основе предоставленных данных сделать вывод.

Чаще всего речь идет не о том, что ту или иную задачу невозможно решить с помощью СППР, а о том, что невозможно или сложно ее решить каким-то конкретным программным продуктом. К счастью, сегодня выбор программных продуктов, на базе которых и строится система поддержки принятия решений, предоставляет возможность удовлетворить свои запросы пусть не одним, но, во всяком случае, комплексом программных средств.

СППР имеют ряд как технических, так и принципиальных ограничений. К «техническим» можно отнести значительную ресурсоемкость некоторых алгоритмов работы с данными, особенно на больших массивах и при включении в анализ большого количества факторов. Это может существенно затруднить применение

таких алгоритмов на стандартных вычислительных платформах. К принципиальным ограничениям относится невозможность получения доступа к исчерпывающему объему данных. Самое главное, что должен понимать пользователь системы СППР, — это невозможность полностью устранить риски при принятии решений.

Тема 13 Применение СППР в бизнесе

Крупноформатная торговля

Крупноформатная торговля и компании электронной коммерции (B2C, B2B) явились первыми институциональными заказчиками на СППР. Основными задачами, решаемыми в данном секторе, являются:

- анализ ассортимента (селективный маргинальный доход, оборачиваемость запасов, статистическое управление запасами, фондоотдача);
- распределение площадей, раскладка;
- анализ эффективности деятельности менеджеров и мотивация персонала;
- планирование и анализ эффективности рекламы, акций, распродажи и т.п.;
- управление ценообразованием.

В части управления раскладкой можно привести известный пример с корреляцией покупок пива и памперсов. Или так называемая «ловушка на кассе» — это мелкие товары, которые выкладываются непосредственно в кассовой зоне. Площадь этой зоны ограничена. Что туда положить? Опять «нет ничего практичнее хорошей теории» — нужен анализ потребительских предпочтений, который, в частности, дает многомерный статистический анализ чеков.

В мелкооптовой торговле ситуация попроще, т.к. там потребитель идентифицирован и учтен в базе данных торговой компании, что позволяет непосредственно анализировать клиентское поведение. В розничной торговле покупатель анонимный, хотя многие компании изначально это исключают, например, METRO Cash & Carry.

Вообще основная тенденция развития прикладных информационных систем в последние пять лет — это ассимиляция систем управления взаимоотношениями с клиентами, возникших в качестве самостоятельных, в контур ERP, причем обе при этом только выигрывают.

Банки и финансовые компании

Рынок СППР в финансовых институтах сейчас самый емкий. Сфера применения DSS-систем в банках касается прежде всего:

- банковского ритейла (платежные пластиковые карты и чеки);
- анализа рисков;
- предотвращения мошенничества (прежде всего с пластиковыми картами);
- анализа потребительского поведения и проектирования новых финансовых услуг.

Последнее, прежде всего, основано на анализе и формировании потребительских групп, которые характеризуются сходным поведением. Результатом этой работы являются проекты, например, молодежных жилищных кредитов, условия овердрафтов, VIP-программы клиентского обслуживания. Предотвращение мошенничества — это перспективная зона использования методов искусственного интеллекта, которая никогда не будет исчерпана, как никогда не будет исчерпано воображение у мошенников.

В страховых компаниях DSS-системы еще не имеют такого широкого распространения, но это только подчеркивает потенциальную перспективность данного рынка.

Телекоммуникации

В телекоммуникационных компаниях, прежде всего мобильной связи, роль DSS-систем связана с проектированием новых услуг, которое основано на выявлении устойчивых клиентских групп и преимущественного клиентского поведения. Этот рынок по времени жизни можно считать неисчерпаемым.

Промышленность

В промышленности к сферам применения DSS-систем можно отнести:

- управление взаимоотношениями с клиентами;
- статистическое управление запасами;
- финансовое и бюджетное планирование и управление;
- анализ и управление рисками.

Какие изменения в парадигме управления промышленностью произошли за последние 50 лет? До 60-х годов промышленно-

ное производство развивалось главным образом за счет развития технологии, что выражалось тезисом: «производить и продавать». В тот период, безусловно, предложение явно формировало спрос. При этом основные производственные фонды были преимущественно материальными: здания, сооружения, оборудование, за которым стояли патентованные технологии.

К концу 20-го века признанным тезисом, выражающим рациональное рыночное поведение, стала парадигма «воспринимать и реагировать». Темп появления новых революционных технологий замедлился, технологии в основном находятся на этапе эволюции. А фронт конкурентной борьбы переместился в область проектирования новых продуктов и услуг. При этом преобладающим стали намерения и пожелания клиентов: явно или неявно выраженные. В качестве примеров можно привести практически полный переход на заказное конфигурирование автомобильной промышленности, постоянно возрастающий спектр предложений услуг в сфере телекоммуникаций при том же самом оборудовании и т.д.

Все большее и большее значение приобретает информация и методы работы с ней. Это тем более актуально в развитых странах мира на фоне сохраняющейся тенденции переноса непосредственно материального производства в развивающиеся страны с низкой стоимостью рабочей силы, энергетических и сырьевых ресурсов. Концепция DSS-систем прямо соответствует задаче информационного обеспечения данной парадигмы.

Каковы сегодня основные промышленные тенденции? Это:

- глобализация;
- укрупнение;
- специализация (для средних компаний);
- интеграция в поставочные сети;
- фокусировка на разработке новых продуктов и услуг;
- необходимость одновременно конкурировать как по качеству, так и по цене.

Промышленность сегодня фокусируется на:

- разработке новых продуктов;
- коммерциализации;

- использовании преимуществ консолидации и интеграции в поставочные сети;
- управлении людскими ресурсами.

Анализируя причины отставания США в промышленном развитии, Комиссия Министерства внешней торговли США считает, что для подъема конкурентоспособности, в частности, необходимо (автор приводит только те пункты рекомендаций, которые имеют отношение к предмету рассмотрения, сам исходный перечень немного шире):

- уделять больше внимания стратегическому планированию и больше инвестировать в исследования и разработки;
- изучать стратегию иностранных конкурентов и совершенствовать собственную;
- уделять больше внимания производственной функции и больше инвестировать в оборудование и кадры;
- устранить коммуникативные барьеры в пределах организации;
- признать ценность развития информационных связей с поставщиками и потребителями.

Информационная поддержка реализации вышеперечисленных рекомендаций со стороны DSS-систем может выглядеть следующим образом:

- «уделять ... внимание стратегическому планированию...» — анализировать исторические данные по структуре себестоимости, динамике цен;
- «изучать стратегию иностранных конкурентов» — анализировать динамику рынков;
- «уделять больше внимания производственной функции» — анализировать затраты по управлению активами, динамику тарифов, эффективность использования оборудования и фондоотдачу;
- «устранить коммуникативные барьеры» — анализировать исторические данные по параметрам реализации внутренних бизнес-процессов и эффективность результатов;

- «признать ценность развития информационных связей» — анализировать исторические данные взаимоотношений с клиентами и поставщиками.

Эффективное решение данных задач требует углубленного анализа как рыночного окружения, так и динамики использования всех внутренних ресурсов.

Особое значение в конкурентной борьбе при практически равной ситуации по возможности доступа к технологиям приобретает персонал и подходы к управлению. В развитых странах мира персонал, по крайней мере, ведущий в стратегическом планировании, переместился из категории «Затраты» (Cost) в категорию «Фонды» — первые надо неуклонно сокращать, а вторые надо развивать и инвестировать.

Также следует отметить, что в настоящее время в мире действует общая глобальная тенденция преимущественного развития рынка услуг по сравнению со сферой непосредственно производства. Экономика все более и более становится информационной, а не материальной.

Рассматривая корпоративный рынок, очень показательным является анализ того, что могут и чего не могут наследуемые системы, прежде всего типов ERP и Project Management.

Оборона

В оборонной области аналитические системы класса DSS развиваются в решении задач:

- планирования и управления операциями;
- планирования и управления эксплуатацией.

Так, по результатам первой войны в Ираке экономический эффект от использования систем искусственного интеллекта был оценен в сумму порядка 100 млн. долларов. Это привело приблизительно к трехкратному увеличению ассигнований на развитие данных информационных технологий в интересах Министерства обороны США. Сегодня в данной области ассигнования уже оцениваются суммами в миллиарды долларов.

Государство

В области государственного строительства роль DSS-систем пока невелика. Потенциально их область использования

связана с оценкой эффективности государственных и муниципальных программ. Это связано, прежде всего, с тем, что государственные и муниципальные программы не сводятся к экономическому эффекту как таковому. Развитие информационных систем в данной сфере в большой мере зависят от философского осмысления роли и места государства в будущем мире, т.е. основополагающую роль в данном процессе имеет выработка критериев и подходов к их оценке.

Тема 14 Анализ математического обеспечения существующих систем поддержки принятия решений

Рассмотрим более подробно средства интеллектуального анализа данных (ИАД, Data Mining), применяемые в системах поддержки принятия решений.

В качестве первого направления развития средств ИАД следует выделить методы статистической обработки данных, которые можно разделить на четыре взаимосвязанных раздела:

- предварительный анализ природы статистических данных (проверка гипотез стационарности, нормальности, независимости, однородности, оценка вида функции распределения и ее параметров);
- выявление связей и закономерностей (линейный и нелинейный регрессионный анализ, корреляционный анализ);
- многомерный статистический анализ (линейный и нелинейный дискриминантный анализ, кластер-анализ, компонентный анализ, факторный анализ);
- динамические модели и прогноз на основе временных рядов.

Среди наиболее известных и популярных средств статистического анализа следует назвать пакеты Statistica, SPSS, Systat, Statgraphics, SAS, BMDP, TimeLab, Data-Desk, S-Plus, Scenario (BI), «Мезозавр».

Особое направление в спектре аналитических средств ИАД составляют методы, основанные на нечетких множествах. Их применение позволяет ранжировать данные по степени близости к желаемым результатам, осуществлять так называемый нечеткий поиск в базах данных. Однако платой за повышенную универсальность является снижение уровня достоверности и точности получаемых результатов. Поэтому число специализированных приложений данного метода по-прежнему невелико, несмотря на то, что на протяжении последних 35 лет математики-прикладники проявляли к нему повышенный интерес.

Второе крупное направление развития составляют кибернетические методы оптимизации, основанные на принципах самораз-

вивающихся систем — методы нейронных сетей, эволюционного и генетического программирования.

Однако новые достоинства порождают и новые проблемы. В частности, решения, полученные кибернетическими методами, часто не допускают наглядных интерпретаций, что в определенной степени усложняет жизнь предметным экспертам.

К программным продуктам, использующим кибернетические методы ИАД, относятся системы PolyAnalyst, NeuroShell, GeneHunter, BrainMaker, OWL, 4Thought (BI).

Непосредственно к кибернетическим методам ИАД примыкают синергетические методы. Их применение позволяет реально оценить горизонт долгосрочного прогноза. Особенный интерес вызывают исследования, связанные с попытками построения эффективных систем управления в неустойчивых режимах функционирования.

К третьему крупному разделу ИАД следует отнести совокупность традиционных методов решения оптимизационных задач — вариационные методы, методы исследования операций, включающие в себя различные виды математического программирования (линейное, нелинейное, дискретное, целочисленное), динамическое программирование, принцип максимума Понтрягина, методы теории систем массового обслуживания. Программные реализации большинства этих методов входят в стандартные пакеты прикладных программ, например Math CAD и MatLab.

В четвертый раздел средств ИАД входят средства, которые назовем условно экспертными, т. е. связанными с непосредственным использованием опыта эксперта. К их числу относят метод «ближайшего соседа», который лег в основу таких программных продуктов, как Pattern Recognition Workbench или KATE tools.

Другой подход к выбору решения связан с построением последовательного логического вывода — дерева решений, в каждом узле которого эксперт осуществляет простейший логический выбор («да» — «нет»). В зависимости от принятого выбора, поиск решения продвигается по правой или левой ветви дерева и в конце концов приходит к терминальной ветви, отвечающей конкретному окончательному решению. Здесь процесс статистического обуче-

ния выведен за пределы программы и сконцентрирован в виде некоторого априорного опыта, заключенного в наборе ветвей-решений.

Одной из разновидностей метода деревьев решений является алгоритм деревьев классификации и регрессии, предлагающий набор правил для дихотомической классификации совокупности исходных данных. Данный метод обычно применяется для предсказания того, какие последовательности событий будут иметь заданный исход. На основе деревьев решений разработаны такие программные продукты, как IDIS, C5.0 и SIPINA.

К экспертным методам следует отнести и предметно-ориентированные системы анализа ситуаций и прогноза, основанные на фиксированных математических моделях, отвечающих той или иной теоретической концепции. Роль эксперта состоит в выборе наиболее адекватной системы и интерпретации полученного алгоритма. Достоинства и недостатки таких систем очевидны — предельная простота и доступность применения и расплата достоверностью и точностью за эту простоту. Примерами программных продуктов, отвечающих предметно-ориентированным системам в области финансов, являются Wall Street Money, MetaStock, SuperCharts, Candlestick Forecaster.

В завершение обзора экспертных методов ИАД следует упомянуть методы визуализации данных и результатов их анализа, позволяющие наглядно отображать полученные выводы для создания у предметных экспертов и/или руководителей проектов единой картины ситуации. К программным продуктам, позволяющим формировать предварительные отчеты и визуализировать результаты, следует отнести системы Mineset и Impromptu (BI). В частности, система Mineset содержит в себе такие инструменты, как ландшафтный визуализатор, визуализаторы дисперсии, деревьев, правил и свидетельств.

Формировать сложные нелинейные отображения средствами цветной графики позволяет новое направление визуализации результатов, основанное на идеях фрактальной математики.

Раздел 2 ТЕМАТИКА ДОКЛАДОВ ПО ДИСЦИПЛИНЕ

1. *Искусственный интеллект*

- Современный искусственный интеллект. Текущее положение дел и перспективы развития;
- Интеллектуальная робототехника;
- Искусственный интеллект в художественной литературе;
- Искусственный интеллект в России;
- Биологическое моделирование искусственного интеллекта;
- Применение искусственного интеллекта в медицине;
- Айзек Азимов – автор законов робототехники;
- Искусственный интеллект в киноиндустрии;
- Андроиды - роботы нового поколения;
- Бионические информационные системы;
- Искусственные нейронные сети;
- Философские проблемы искусственного интеллекта и искусственной жизни;
- Сознание и искусственный интеллект;
- Тенденции развития систем автоматического распознавания речи;
- Искусственный интеллект и теоретические вопросы психологии;
- История развития науки об искусственном интеллекте;
- Современные компьютерные технологии и интеллект;
- Современная наука и искусственный интеллект;
- Обзор современных роботов;
- Искусственный интеллект в военной технике;
- Применение робототехники для космических исследований;
- Робототехника в быту;
- Промышленная робототехника;
- Роботы-гиноиды;

- Aiko - женский робот-андроид с искусственным интеллектом.

2. *Интеллектуальный анализ данных (Data Mining)*

- Сегментация клиентской базы с использованием технологии Data Mining;

- Интеллектуальный анализ данных в CRM;

- Оценка кредитоспособности клиента с помощью технологии Data Mining;

- Деревья решений;

- Яркие представители рынка интеллектуального анализа данных: PolyAnalyst

- Задачи Data Mining;

- Интеллектуальный анализ данных в медицине

- Интеллектуальный анализ данных в химии

- Применение нейронных сетей для анализа операций на рынке Forex;

- Технология OLAP.

- Роль метода «деревьев решений» в СППР.

3. *Методология систем поддержки принятия решений*

- СППР в электронном правительстве

- СППР в медицине

- СППР в маркетинге

- Методики, применяемые в системах поддержки принятия решений

- Создание СППР с помощью Excel

- СППР «Выбор»: описание, возможности

- SPSS Statistica

- Интеграция ГИС и СППР

- История возникновения СППР в России

- Интеллектуальный анализ данных в СППР

- Возникновение СППР на Западе

- СППР в концепции CRM

- Программные продукты класса СППР: рынок, основные разработчики

- Когнитивное моделирование как один из методов, лежащих в основе СППР
- Рассуждение на основе прецедентов как один из методов, лежащих в основе СППР
- Экспертные системы как ИС близкие СППР: основные возможности, сферы применения и решаемые задачи
- Имитационное моделирование в СППР
- С. Н. Корсаков – основоположник СППР в России

Требования к докладу:

1. Доклад должен быть представлен в виде презентации (MS PowerPoint), обязательным условием является устное выступление, наличие в докладе таблиц, диаграмм и схем, присутствие не более 6-7 строк текста на одном слайде. Выступление докладчика должно содержать минимум лишней, малозначимой информации, максимум интересной и неочевидной. Тема доклада должна быть глубоко раскрыта, выступление не должно быть описательным, а должно содержать точку зрения автора и собственный вывод. Общее время выступления не должно превышать 5-7 минут.

Правила оформления презентаций:

Общий порядок слайдов презентации должен быть следующим: 1) титульный слайд; 2) план презентации (5-6 пунктов максимум); 3) основная часть; 4) заключение (выводы).

Дизайн презентации должен быть простым и лаконичным; должны присутствовать два типа слайдов: для титульных, планов и т. п. и для основного текста. Каждый слайд презентации должен иметь заголовок и свой номер. Точку в конце заголовка ставить не рекомендуется, кроме этого сами заголовки не должны повторяться и не быть длинными.

Текст на слайдах должен носить тезисный характер, поскольку его основная цель - сопровождать подробное изложение мыслей докладчика.

Если в презентации имеет место диаграмма, то у нее должно быть название или таким названием может служить заголовок слайда; диаграмма должна занимать все место на слайде, все ее линии и подписи должны быть хорошо видны.

Что касается таблиц презентации, то обязательным условием их оформления является наличие названия и отличие шапки от основных данных.

Использовать встроенные эффекты анимации в презентации можно только тогда, когда без этого не обойтись (например, последовательное появление элементов диаграммы), в противном случае – от них лучше отказаться.

Раздел 3 ПРАКТИКУМ В СИСТЕМЕ ИНТЕЛЛЕКТУАЛЬНОГО АНАЛИЗА ДАННЫХ DEDUCTOR

Deductor является аналитической платформой - основой для создания законченных прикладных решений в области анализа данных. Реализованные в Deductor технологии позволяют на базе единой архитектуры пройти все этапы построения аналитической системы от создания хранилища данных до автоматического подбора моделей и визуализации полученных результатов.

ЗАДАНИЕ № 1

Знакомство с аналитической платформой Deductor

Цель работы - ознакомиться с архитектурой, основными частями и пользовательским интерфейсом Deductor, получить навыки создания сценариев обработки и визуализации данных, создания и наполнения хранилища данных. Deductor состоит из 3-х частей – многомерного хранилища данных Deductor Warehouse, аналитического приложения Deductor Studio и рабочего места конечного пользователя Deductor Viewer.

Deductor Warehouse – многомерное хранилище данных, аккумулирующее всю необходимую для анализа предметной области информацию. Использование единого хранилища позволяет обеспечить непротиворечивость данных, их централизованное хранение и автоматически обеспечивает всю необходимую поддержку процесса анализа данных. Deductor Warehouse оптимизирован для решения именно аналитических задач, что положительно сказывается на скорости доступа к данным.

Deductor Studio – программа, реализующая функции импорта, обработки, визуализации и экспорта данных. Deductor Studio может функционировать и без хранилища данных, получая информацию из любых других источников, но наиболее оптимальным является их совместное использование. В Deductor Studio включен полный набор механизмов, позволяющий получить информацию из произвольного источника данных, провести весь цикл обработки (очистку, трансформацию данных, построение моделей), отобразить полученные результаты наиболее удобным образом (OLAP,

диаграммы, деревья...) и экспортировать результаты на сторону. Это полностью соответствует концепции извлечения знаний из баз данных (KDD).

Deductor Viewer – рабочее место конечного пользователя. Позволяет отделить процесс построения моделей от использования уже готовых моделей. Все сложные операции по подготовке моделей выполняются аналитиками-экспертами при помощи *Deductor Studio*, а *Deductor Viewer* обеспечивает пользователям простой способ работы с готовыми результатами, скрывает от них все сложности построения моделей и не предъявляет высоких требований к квалификации сотрудников.

Архитектура системы построена таким образом, что вся работа по анализу данных в *Deductor Studio* базируется на выполнении следующих действий:

- импорт данных;
- обработка данных;
- визуализация;
- экспорт данных.

Процесс построения моделей в *Deductor* основывается на следующих трех принципах:

1. Использование обработчиков;
2. Использование визуализаторов;
3. Создание сценариев.

Обработка и визуализация – две атомарные операции с данными в *Deductor*. Под обработкой понимаются любые манипуляции над набором данных: от самых простых (например, сортировка) до сложных (построение нейронной сети). бработчик можно представить в виде «черного ящика», на вход которого подается набор данных, а на выходе формируется преобразованный набор данных (рис. 2).

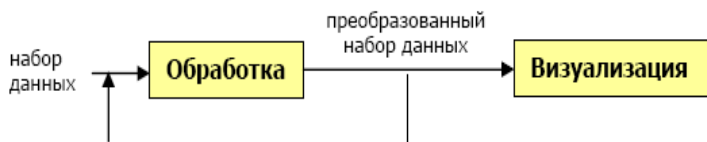


Рисунок 2 – Обработка и визуализация

Реализованные в Deductor обработчики покрывают основную потребность в анализе данных и создания законченных аналитических решений на базе Data Mining. Их классификация приведена на рис. 3.

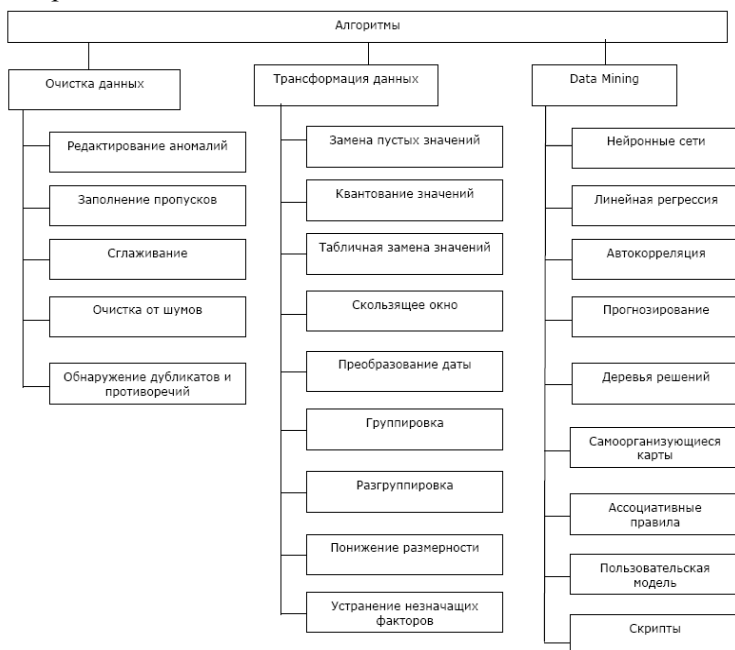


Рисунок 3 - Классификация алгоритмов (обработчиков) в Deductor

Любой набор данных можно визуализировать каким-либо доступным способом или несколькими способами, поскольку визуализация помогает интерпретировать построенные модели.

В Deductor предусмотрены следующие способы визуализации данных.

- **OLAP.** Многомерное представление данных. Любые данные, используемые в программе, можно посмотреть в виде кросс-таблицы и кросс-диаграммы.

- **Таблица.** Стандартное табличное представление с возможностью фильтрации данных.

- **Диаграмма.** График изменения любого показателя.

- **Гистограмма.** График разброса показателей.

- **Статистика.** Статистические показатели набора данных.

- **Диаграмма рассеяния.** График отклонения прогнозируемых при помощи модели значений от реальных. Может быть построена только для непрерывных величин и только после использования механизмов построения модели, например, нейросети или линейной регрессии. Используется для визуальной оценки качества построенной модели.

- **Таблица сопряженности.** Предназначена для оценки результатов классификации вне

- зависимости от используемой модели. Таблица сопряженности отображает результаты сравнения категориальных значений исходного выходного столбца и категориальных значений рассчитанного выходного столбца. Используется для оценки качества классификации.

- **«Что-если».** Таблица и диаграмма. Позволяют «прогнозировать» через построенную модель любые интересующие пользователя данные и оценить влияние того или иного фактора на результат.

- **Обучающая выборка.** Набор данных, используемый для построения модели.

- **Диаграмма прогноза.** Применяется после использования метода обработки.

- **Прогнозирование.** Прогнозные значения выделяются цветом.

- **Граф нейросети.** Визуальное отображение обученной нейросети. Отображается структура нейронной сети и значения весов;

- **Дерево решений.** Отображение дерева решений, полученного при помощи соответствующего алгоритма.

- **Дерево правил.** Отображение в иерархическом виде (в виде дерева) ассоциативных правил.

- **Правила.** Отображает в текстовом виде правила, полученные при помощи алгоритма построения деревьев решений или поиска ассоциаций.

- **Карта Кохонена.** Отображение карт, построенных при помощи соответствующего алгоритма.

- **Описание.** Текстовое описание параметров импорта/обработки/экспорта в дереве сценариев обработки.

Сценарий представляет собой иерархическую последовательность обработки и визуализации наборов данных. После импорта может следовать произвольное число обработчиков любой степени глубины и вложенности. Каждой операции обработки соответствует отдельный узел дерева, или объект сценария. Любой объект можно визуализировать тем или иным доступным средством. Набор данных служит механизмом, соединяющим все объекты сценария. Можно сказать, что сценарий – наиболее естественный с точки зрения аналитика способ представления этапов построения модели. Это позволяет быстро создавать модели, обладающие большой гибкостью и расширяемостью, сравнивать несколько моделей. На рис. 4 изображен пример сценария.

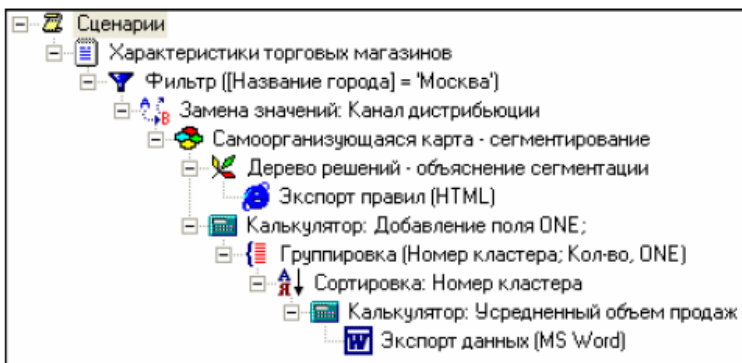


Рисунок 4 – Пример сценария в Deductor

Интерфейс Deductor Studio состоит из главного окна, внутри которого располагаются панели сценариев, отчетов, источников данных и результаты моделирования (таблицы, графики, кросс-диаграммы, правила и т.д.). Все сценарии создаются на основе запуска мастеров. В распоряжение аналитика имеется 4 мастера: импорт, экспорт, обработка, отображение. *Мастер импорта* предназначен для автоматизации получения данных из любого источника, предусмотренного в системе. На первом шаге мастера импорта открывается список всех предусмотренных в системе типов источников данных. Число шагов мастера импорта, а также набор настраиваемых параметров отличается для разных типов источников.

Мастер обработки предназначен для настройки всех параметров выбранного алгоритма. *Мастер отображений* позволяет в пошаговом режиме выбрать и настроить наиболее удобный способ представления данных. В зависимости от обработчика, в результате которого была получена ветвь сценария, список доступных для него видов отображений будет различным. Например, после построения деревьев решений их можно отобразить с помощью визуализаторов «Деревья решений» и «Правила». Эти способы отображения не доступны для других обработчиков.

Мастер экспорта позволяет в пошаговом режиме выполнить экспорт данных в файлы наиболее распространенных форматов.

Создание хранилища данных

Хранение данных в многомерном виде в специальной структуре – хранилище данных – облегчает последующий доступ к данным, их анализ и обработку. Хранилище данных Deductor Warehouse основано на реляционной базе данных (Firebird), которая содержит таблицы для хранения информации и таблицы связей, обеспечивающие целостное хранение сведений. Поверх реляционной базы данных реализован специальный слой, который преобразует реляционное представление к многомерному.

Для создания нового хранилища данных в Deductor необходимо выполнить следующую последовательность действий.

1. Откройте панель «Источники данных» (одноименный пункт в меню «Вид»).

2. На дереве источников данных вызовите контекстное меню и выберите действие «Создать локальное хранилище данных». На экране появится диалоговое окно, в котором нужно задать имя, метку (например, Lab1) и расположение хранилища данных, или согласиться с предлагаемыми по умолчанию (рис. 5).

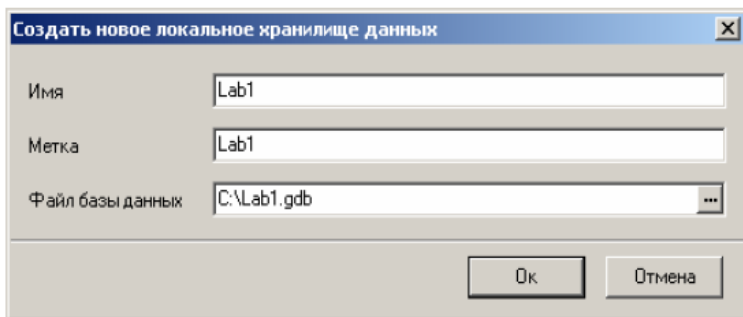


Рисунок 5 – Окно создания хранилищ данных

После нажатия на кнопку «Ок» по указанному пути будет создано пустое хранилище данных. С помощью кнопки проверьте подключение к хранилищу.

Наполнение хранилища данных

Рассмотрим на примере нескольких таблиц последовательность действий, необходимую для наполнения хранилища данных информацией. В папке \Lab1 располагаются две таблицы с именами «Товары» (goods.dbf) и «Группы» (groups.dbf), содержащие информацию о номенклатуре фармацевтической продукции некоторого торгового предприятия.

Таблица 3
Товары

КодТовара	Количество_в_партии	Производитель	КодГруппы
20	6	САГМЕЛ	4
21	15	БЕРЛИН-ХЕМИ АГ (ГРУППА МЕНАРИНИ)	5
22	270	БЕРЛИН-ХЕМИ АГ (ГРУППА МЕНАРИНИ)	5
26	1	АРМАВИРСКАЯ БИОФАБРИКА	6
27	1	ГЕКСАЛ АГ (САЛЮТАС ФАРМА)	4
...

Таблица 4
Группы

КодГруппы	Название
1	Витамины
2	Сердечно-сосудистые средства
3	Противоревматические и жаропонижающие
4	Респираторные заболевания
5	Гормоны
6	Противоопухолевые средства
...	...

Для того чтобы строить многомерные отчеты без создания хранилища данных, нужно объединить таблицы, поскольку, во-первых, они находятся в нормализованном виде, во-вторых, с стр. 10 из 31 кодами групп товаров работать неудобно. Для этого пер-

вым шагом в новом сценарии создается узел, импортирующий таблицу «Товары» (запускается «Мастер импорта»). После этого к узлу импорта таблицы «Товары» применяется обработчик «Слияние». В полях «Метка столбца» предусмотрена возможность задания альтернативных названий столбцов. При слиянии на шаге 6 мастера задаются общие поля двух таблиц, которые называются измерениями. После слияния к исходной таблице добавятся поля-факты. В нашем случае измерением служит поле «КодГруппы», фактом - поле «Наименование» (рис. 6).

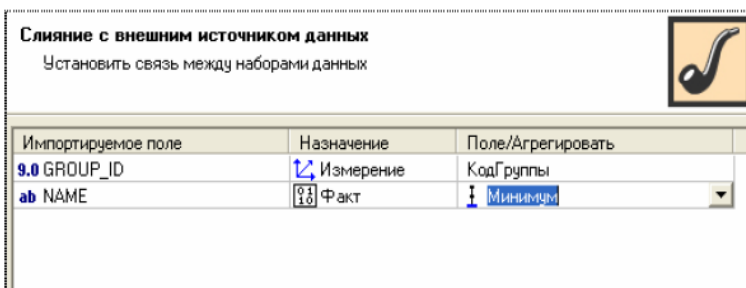


Рисунок 6 – Установка параметров слияния

В результате будет создана новая таблица с пятью полями («Код товара», «Количество», «Производитель», «Код группы», «Имя группы»), готовая для дальнейшей обработки. Сценарий будет иметь два узла (рис. 7).

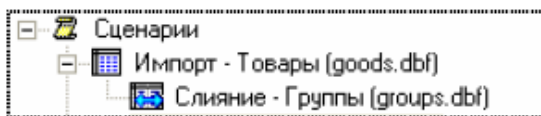


Рисунок 7 – Сценарий обработки

При первоначальном наполнении пустого хранилища данных информацией о товарах рекомендуется придерживаться определенной последовательности действий (рис. 8). Перед этим следует определиться, какие поля являются измерениями, а какие – фактами, какие таблицы представляют собой процессы.



Рисунок 8 - Последовательность загрузки данных в хранилище

Под процессом понимается определенное действие, например, продажи товара, отгрузки, поступления денежных средств и прочее. Можно сказать, что с каждым процессом связан определенный бизнес-процесс. Измерение может иметь свойства. Соответственно, загружать измерения отдельно вне процесса имеет смысл, если оно имеет свойства. При загрузке процесса измерение со свойствами загружается по его идентификатору, а при загрузке измерения загружаются также и его свойства. В таблице «Группы» поле «КодГруппы» является измерением, поле «Название» - его свойством. Как правило, свойства присутствуют в таблицах-справочниках, а измерение указывает на ключевое (уникальное) поле. Для загрузки измерений и их свойств запускается «Мастер экспорта», источником выбирается «Deductor(измерение)» и в списке доступных хранилищ данных указывается то, в которое необходимо выполнить загрузку. Первоначально никаких измерений нет, поэтому создадим измерение, нажав на кнопку . Появится окно редактора измерений, в котором сделаем поле GROUP_ID измерением, NAME - свойством (рис. 9). В колонке «Название» введем метки полей.



Назначение	Тип поля	Название	Поля в источник...
 Измерение	9.0 Вещественный	КодГруппы	GROUP_ID
 Свойство	ab Строковый	Группа	NAME

Рисунок 9 - Создание измерения

Данные из таблицы «Товары» загружаются в хранилище в виде нового процесса (например, процесс «Распределение товаров»). Снова запускается «Мастер экспорта», но источником выбирается «Deductor(процесс)». Фактом будет поле «Количество», все остальные – измерениями. Измерение «КодГруппы» уже есть в хранилище, т.к. оно имело свойство и загружалось отдельно. Поэтому эта строка закрашена в серый цвет (рис. 10).





Назначение	Тип поля	Название	Поля в источник...
 Измерение	9.0 Вещественный	КодТовара	TOVAR_ID
 Измерение	ab Строковый	Производитель	MAKER
 Измерение	9.0 Вещественный	КодГруппы	GROUP_ID
 Факт	9.0 Вещественный	Кол-во	UPAC

Рисунок 10 - Создание процесса «Товары»

Таким образом, с помощью сценария (рис. 11) будет создана структура хранилища и загружены данные о распределении номенклатуры товаров по производителям и группам.

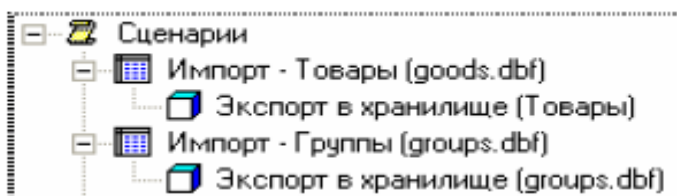


Рисунок 11 - Сценарий наполнения хранилища

Извлечение информации из хранилища данных

После создания и наполнения хранилища информацию из него можно извлекать Мастером импорта, выбрав в качестве источника «Deductor Warehouse» и указав необходимые измерения и факты процесса (рис. 12).

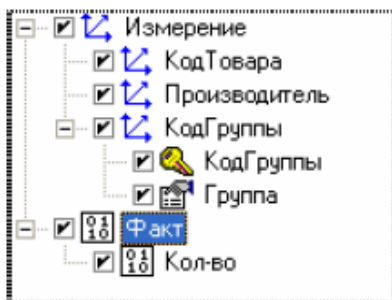


Рисунок 12 - Выбор импортируемых данных

Создание многомерных отчетов

По окончании импорта мастер предложит настроить способы отображения данных. Для создания многомерного отчета следует выбрать «Куб», снова указать измерения и факты, задать размещение измерений, способ агрегации фактов и заголовок для ветки сценария. Например, чтобы получить многомерный отчет распределения продаж товаров в разрезе производителя и групп товаров, то на 3 и 4 шаге «Мастера настройки отображения» следует выбрать параметры, как это показано на рис. 13.

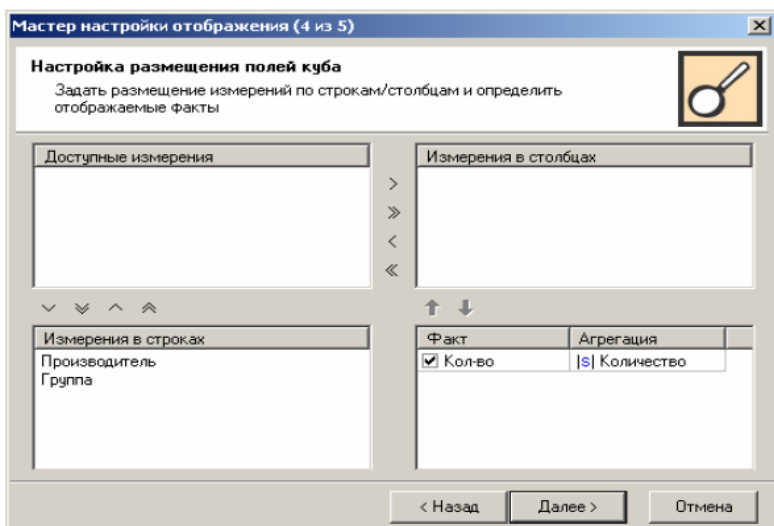


Рисунок 13 – Мастер настройки отображения

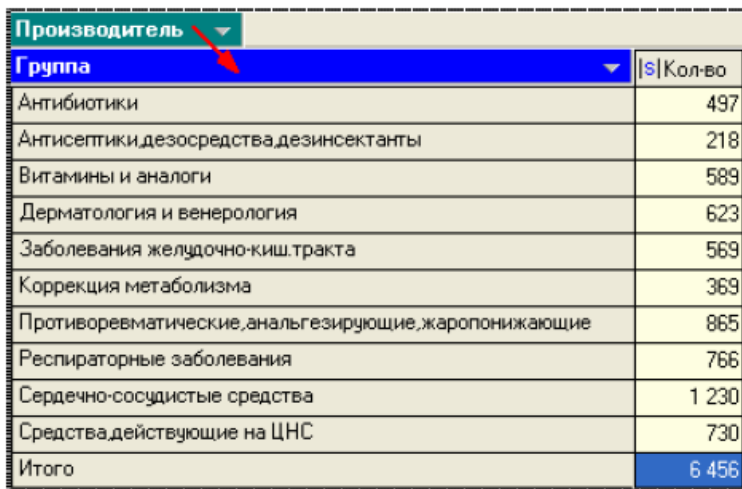
Производитель	Группа	[S] Кол-во
ОРИОН КОРПОРЕЙШНЛ	Противоревматические,анальгезирующие,жаропонижающие	3
	Сердечно-сосудистые средства	8
	Итого	11
ОСТФАРМ АРНЕЙМИТЕЛЬ ГМБХ	Сердечно-сосудистые средства	1
	Итого	1
ПАБ (ХОРВАТИЯ)	Респираторные заболевания	1
	Итого	1
ПАНАЦЕЯ БИОТЕК	Антибиотики	1
	Витамины и аналоги	1
	Заболевания желудочно-киш.тракта	1
	Противоревматические,анальгезирующие,жаропонижающие	6
	Итого	9

Рисунок 14 – Пример многомерного отчета

Измерения в кросс-таблице изображаются специальными полями. Синие поля показывают измерения, участвующие в построении таблицы. Зелеными полями отображаются скрытые измерения, не участвующие в построении таблицы. Имеется возмож-

ность перестраивать таблицу с помощью мыши «на лету». Сделать это можно, если перетаскивать поля с заголовками измерений.

Например, на рис. 15 скрытое измерение «Производитель» с помощью операции перетаскивания становится участвующим в построении, как это изображено на рис. 14.



Группа	Кол-во
Антибиотики	497
Антисептики,дезодоранты,дезинсектанты	218
Витамины и аналоги	589
Дерматология и венерология	623
Заболевания желудочно-киш.тракта	569
Коррекция метаболизма	369
Противоревматические,анальгетирующие,жаропонижающие	865
Респираторные заболевания	766
Сердечно-сосудистые средства	1 230
Средства,действующие на ЦНС	730
Итого	6 456

Рисунок 15 - Скрытое измерение «Производитель»

Изменять расположение измерений можно, используя операцию транспонирования таблицы.

В результате транспонирования данные, ранее отображавшиеся в строках, отображаются в столбцах, а данные в столбцах преобразуются в строки. Транспонирование во многих случаях позволяет оперативно сделать таблицу более удобной для восприятия. Пример транспонирования таблицы представлен на рис. 16.

Производитель	Группа	Количество
АБОЛМЕД	Антибиотики	7
	Итого	7

	Группа			
Производитель	Антибиотики	Антисептики	Витамины	Дерматология
"ЛАБОРАТОРИЯ ФАРМАЧЕУТИКО"				
АБОЛМЕД	7			
АВД ФАРМ И КО/ВИАТРИС ГМБХ				
АВЕНТИС ФАРМА				1
АВЕНТИС/СОТЕКС				
АДЖИО				5
АЙ СИ ЭН АЛКАЛОИДА				
АЙ СИ ЭН ГАЛЕНИКА	1		5	2
АЙ СИ ЭН ЛЕКСРЕДСТВА		1	3	
АЙ СИ ЭН МАРБИОФАРМ			20	2
АЙ СИ ЭН ОКТЯБРЬ	1		28	
АЙ СИ ЭН ПОЛИФАРМ			2	3

Рисунок 16 - Транспонирование измерения «Группа»

Предусмотрено пять способов объединения (агрегирования) фактов в кросс-таблице:

сумма – вычисляется сумма объединяемых фактов;

минимум – среди всех объединяемых фактов в таблице отображается только минимальный;

максимум - среди всех объединяемых фактов в таблице отображается только максимальный;

среднее – вычисляется среднее значение объединяемых фактов;

количество – в кросс-таблице будет отображаться количество объединенных фактов.

Для изменения способа агрегации фактов нужно вызвать окно «Настройка размещения», нажав кнопку «Изменение размещения» на панели инструментов.

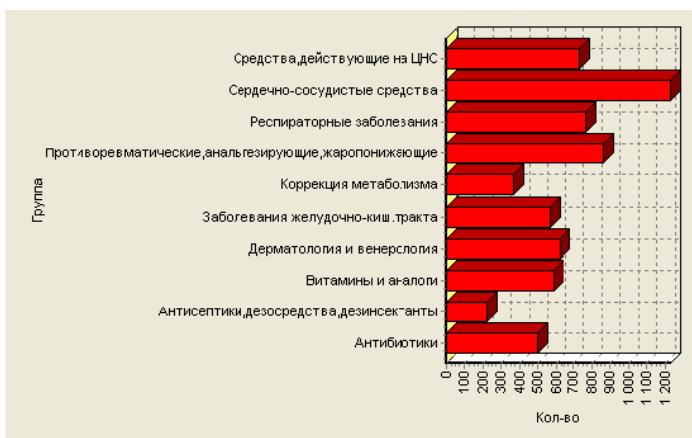


Рисунок 17 - Кросс-диаграмма

Кросс-диаграмма представляет собой диаграмму заданного типа, построенную на основе кросс-таблицы. Основное отличие кросс-диаграммы от обычной диаграммы в том, что она однозначно соответствует текущему состоянию кросс-таблицы и при любых ее изменениях изменяется соответственно. Например, для многомерного отчета на рис. 15 соответствующая кросс-диаграмма выглядит как это показано на рис. 17.

Автоматическая загрузка данных в хранилище

При повторной загрузке процессов в ХД каждому измерению процесса сопоставляется измерение источника данных, каждому факту процесса – факты источника. Также можно указать измерение, по которому будет происходить удаление фактов при повторной загрузке. Например, указав в качестве такого измерения дату отгрузки, из хранилища будут удалены все факты отгрузки, в случае, если эта же дата есть в источнике. Таким образом, достигается устранение дублирующих значений по какому-либо измерению.

Задание. Повторите все описанные действия с прилагаемыми таблицами, что в конечном итоге приведет к созданию и

наполнению хранилища данных, расположенного на сетевом или локальном диске.

ЗАДАНИЕ № 2

Многомерные отчеты и простая аналитика

Цель работы – освоить и закрепить навыки создания хранилища данных и извлечения из него информации, построения многомерных отчетов и кросс-диаграмм и их анализа.

В папке \Lab2\Part1 (варианты 6-10 – \Lab2\Part2) находятся 4 таблицы (файлы с расширением *.dbf). Это таблицы со следующей информацией:

«Товары» (produces.dbf) – номенклатура товаров (фармацептика);

«Товарные группы» (groups.dbf) – группы товаров, например, витамины, желчегонные средства, иммуномодуляторы, анестетики, адаптогены и т.д.

«Торговые отделы» (stores.dbf) – информация о торговых отделах

«Продажи» (sales.dbf) – история продаж товаров за 1 год. Каждая транзакция содержит дату и время продажи, товар, количество, сумму и торговый отдел.

Структура таблиц с описанием каждого поля приведена в приложении.

Задание. Построить сценарии и на их основе, а также при помощи операций транспонирования измерений и агрегирования фактов, сформировать отчеты и ответить на вопросы в заданиях.

Для выполнения заданий понадобятся сведения из лабораторной работы № 1 и дополнительные обработчики:

- преобразование даты/времени;
- фильтрация;

Обработчик «Преобразование даты/времени»

Разбиение даты служит для анализа всевозможных показателей за определенный период (день, неделя, месяц, квартал, год). Суть разбиения заключается в том, что на основе столбца с информацией о дате формируется другой столбец, в котором указывается, к какому заданному интервалу времени принадлежит строка дан-

ных. Тип интервала задается аналитиком, исходя из того, что он хочет получить – данные за год, квартал, месяц, неделю, день или сразу по всем интервалам.

Значения нового столбца, полученного после применения преобразования даты, могут быть одного из трех типов: строка, число или дата. Например, нужно преобразовать дату «10.04.2011» и заменить ее месяцем. Тогда в столбце строкового типа будет содержаться «2011-M04» и его уже нельзя использовать как дату, например, к нему нельзя снова применить преобразование даты. А в столбце типа «дата» будет значение «01.04.2011». К нему снова можно применить преобразование и заменить, например, номером квартала. Новый столбец будет содержать значение 2 числового типа.

Пример использования преобразования даты приведен в табл. 5. Первый столбец «Дата» – это исходный столбец. Остальные получены после обработки.

Обработчик «Фильтрация»

С помощью операции фильтрации можно оставить в таблице только те записи, которые удовлетворяют заданным условиям, а остальные скрыть.

Список вопросов

1. Построить куб по трем измерениям (торговая точка, месяц года, товарная группа), в ячейках которого отображается сумма и объем (количество проданных единиц продукции) продаж за все периоды, имеющиеся в базе данных. Какая торговая точка приносит наибольшую сумму продаж? Какая товарная группа имеет максимальную сумму продаж? Постройте кросс-диаграмму сумм продаж: общие продажи, продажи по торговым точкам, продажи по товарным группам.

Таблица 5

Пример использования преобразования таблицы

Дата	Год + Квартал	Год + Месяц	Год + Неделя	Кварта л	Месяц	Неделя	День года	День недели (число)	День недели (строка)
01.01.2004	01.01.2004	01.01.2004	01.01.2004	1	1	1	1	4	4 Четверг
09.01.2004	01.01.2004	01.01.2004	05.01.2004	1	1	2	9	5	5 Пятница
17.01.2004	01.01.2004	01.01.2004	12.01.2004	1	1	3	17	6	6 Суббота
25.01.2004	01.01.2004	01.01.2004	19.01.2004	1	1	4	25	7	7 Воскресенье
02.02.2004	01.01.2004	01.02.2004	02.02.2004	1	2	6	33	1	1 Понедельник
10.02.2004	01.01.2004	01.02.2004	09.02.2004	1	2	7	41	2	2 Вторник
18.02.2004	01.01.2004	01.02.2004	16.02.2004	1	2	8	49	3	3 Среда
26.02.2004	01.01.2004	01.02.2004	23.02.2004	1	2	9	57	4	4 Четверг
05.03.2004	01.01.2004	01.03.2004	01.03.2004	1	3	10	65	5	5 Пятница
13.03.2004	01.01.2004	01.03.2004	08.03.2004	1	3	11	73	6	6 Суббота

2. То же, что в п.1, но за последние три месяца от имеющихся данных.

3. То же, что в п.1, но за последние три недели от имеющихся данных.

4. Найти сумму максимальной и средней стоимости покупки за последний месяц от имеющихся данных.

5. Сформировать многомерный отчет и график загруженности торговых точек по времени суток и торговым точкам. На какие часы приходятся пики продаж?

6. То же, что в п. 5, но за три месяца от имеющихся данных.

7. Сформировать многомерный отчет и график загруженности торговых точек по дням недели.

8. То же, что в п. 7, но за последний месяц от имеющихся данных.

9. Сформировать многомерный отчет и график загруженности торговых точек по дням месяца. Постройте линию тренда.

10. То же, что в п. 9, но за последние три месяца от имеющихся данных.

11. 20 самых продаваемых товаров.

12. То же, что в п. 11, но за последние три недели от имеющихся данных.

13. 10 самых продаваемых товаров по воскресеньям.

14. 5 самых популярных товаров в каждой товарной группе.

15. То же, что и п. 14, но за последнюю неделю.
 16. Товары, дающие 50% объема продаж.
 17. То же, что и п. 16, но за последние 3 месяца от имеющихся данных.
 18. То же, что и п. 16, но за последнюю неделю.
 Распределение вопросов по вариантам(Табл. 6).

Таблица 6
Вопросы по вариантам

№ вар № задания	1(6)	2(7)	3(8)	4(9)	5(10)
1	+			+	
2		+			+
3			+		
4	+	+	+	+	+
5	+			+	
6	+				+
7		+			+
8		+		+	
9			+	+	
10			+		+
11	+				
12		+			
13			+		
14	+	+	+		+
15		+		+	+
16	+		+		+
17		+	+	+	
18	+			+	

ЗАДАНИЕ № 3

Задачи сегментации и классификации

Цель работы – научиться применять методы Data Mining для решения задач сегментирования и классификации на примере задачи банковского кредитования (скоринга). В папке \Lab3 расположено 2 файла:

WhCredit.gdb – хранилище данных, содержащее информацию о выдаче и возврате кредитов физическим лицам (кредитная история);

Credit.ded – файл сценария Deductor 4.

Задание.

1. Ознакомьтесь с приведенным ниже необходимым теоретическим материалом, который содержит актуальность решения

задачи банковского кредитования методами Data Mining и ее методике, описания обработчиков и визуализаторов Deductor для выполнения индивидуального задания.

2. Сценарий в файле Credit.ded (настроен на хранилище данных с именем Credit) производит сегментацию заемщиков на 6 кластеров с помощью самоорганизующихся карт. Сегментирование производилось по следующим входным параметрам:

- ☐ цель кредитования;
- ☐ сумма кредита;
- ☐ срок кредита;
- ☐ возраст;
- ☐ среднемесячный доход;
- ☐ среднемесячный расход;
- ☐ количество иждивенцев.

Запустите сценарий сегментации. Проинтерпретируйте результаты сегментации, проведя визуальный анализ карт признаков. Дайте каждому сегменту заемщиков название. Оцените численность каждого сегмента и постройте соответствующую диаграмму. Постройте дерево решений для объяснения результатов сегментации. Для обучения используйте 60% от всех данных, остальные – для тестирования. Дальнейшее задание выполните согласно варианту.

Вариант № 1

Постройте многомерный отчет и кросс-диаграмму распределения по целям кредитования.

Постройте модель дерева решений для оценки кредитоспособности заемщика для сегмента 0.

Вариант № 2

Постройте многомерный отчет и кросс-диаграмму распределения заемщиков по возрастным группам. Постройте модель дерева решений для оценки кредитоспособности заемщика для сегмента 1.

Вариант № 3

Постройте многомерный отчет и кросс-диаграмму возрастных групп, на которые приходится 50% выдаваемых кредитов.

Постройте модель дерева решений для оценки кредитоспособности заемщика для сегмента 2.

Вариант № 4

Постройте многомерный отчет и кросс-диаграмму распределения заемщиков по целям кредитования и полу заемщика.

Постройте модель дерева решений для оценки кредитоспособности заемщика для сегмента 3.

Вариант № 5

Постройте многомерный отчет и кросс-диаграмму распределения заемщиков по целям кредитования и должностям.

Постройте модель дерева решений для оценки кредитоспособности заемщика для сегмента 4.

Для каждой модели проведите оценку качества и точности. Результатом проделанной работы должен стать сценарий Deductor.

Data Mining в банковском кредитовании

Одной из важнейших задач в банковском кредитовании является анализ потенциальных заемщиков. В настоящее время большинство российских банков решают вопрос снижения своих кредитных рисков путем простого переноса их на поручителей заемщика. В современных российских условиях стремительного спроса на услуги банковского кредитования банк, который умеет оценить кредитный риск как можно точнее, получит преимущество над конкурентами, дополнительную прибыль, возможность управлять уровнем риска. Одним из доступных инструментов для оценки кредитного риска, особенно в условиях отсутствия экспертов по оценке риска, являются методы Data Mining. Эксперты в области банковского кредитования выделяют несколько факторов, которые влияют на кредитоспособность человека (табл. 7).

Таблица 7

Факторы, влияющие на кредитоспособность

Категория	Некоторые факторы категории
Базовая персональная информация	Пол, возраст, образование ...
Информация о семейном положении	Состояние в браке, количество детей ...
Регистрационная информация	Регистрация, срок проживания по данному адресу ...
Информация о занятости	Специальность, сфера деятельности предприятия ...
Информация о финансовом положении	Зарплата, другие начисления и удержания
Информация по обеспеченности	Имущество, ценные бумаги...
Информация о кредитной истории	Количество прошлых кредитов, текущие обязательства ...

Тем самым должно достигаться и отнесение потенциально-го заемщика к способным вернуть кредит или не способным. При наличии статистических данных (кредитной истории) модель классификации строится с использованием дерева решений.

Для выполнения заданий понадобятся дополнительные обработчики:


- ☐ дерево решений;
- ☐ группировка;
- ☐ сортировка и визуализаторы:
- ☐ карта Кохонена;
- ☐ дерево решений;
- ☐ правила;
- ☐ таблица сопряженности;
- ☐ «Что-Если»;
- ☐ диаграмма.

Обработчик и визуализатор «Дерево решений»

Построение дерева решений производится в процессе обучения. Настройки параметров обучения можно изменить в окне мастера (рис. 18)

Настройка параметров обучения дерева решений

Укажите значения параметров обучения дерева решений



Параметры ранней остановки

Минимальное количество примеров в узле, при котором будет создан новый

2

☒ Строить дерево с более достоверными правилами в ущерб компактности дерева
Очередной узел будет разбиваться на подузлы, если количество нераспознаваемых примеров в узле больше значения параметра "минимальное количество примеров в узле".

Параметры отсекаания

☒ Отсекать узлы дерева
Уровень доверия используемый при отсекании узлов дерева, %

20.0

Чем меньше уровень доверия, тем больше узлов будет отсечено и тем компактнее будет дерево.

Рисунок 18 - Параметры обучения обработчика «Дерево решений»

Параметры обучения дерева решений следующие:

☐ Минимальное количество примеров, при котором будет создан новый узел. Задается минимальное количество примеров, которое возможно в узле. Если примеров, которые попадают в данный узел, будет меньше заданного - узел считается листом (т.е. дальнейшее ветвление прекращается). Чем больше этот параметр, тем менее ветвистым получается дерево.

☐ Строить дерево с более достоверными правилами в ущерб сложности. Включает специальный алгоритм, который, усложняя структуру дерева, увеличивает достоверность результатов классификации. При этом дерево получается, как правило, более ветвистым.

☐ Уровень доверия, используемый при отсекании узлов дерева. Значение этого параметра задается в процентах и должно лежать в пределах от 0 до 100. Чем больше уровень доверия, тем более ветвистым получается дерево, и, соответственно, чем меньше уровень доверия, тем больше узлов будет отсечено при его построении. Для просмотра дерева решений предназначен одноименный визуализатор (рис. 19).

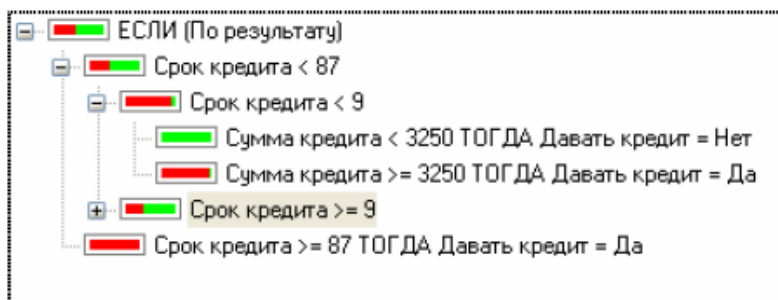


Рисунок 19 - Пример дерева решений

Обработчик «Группировка»

Группировка позволяет объединять записи по полям - измерениям, агрегируя данные в полях-фактах для дальнейшего анализа. Для настройки группировки требуется указать, какие поля являются измерениями, а какие – фактами. Для каждого факта требуется указать функцию агрегации. Это может быть сумма, среднее, максимум, минимум, количество. При выполнении группировки в таблице данных ищутся записи с одинаковыми полями-измерениями. К полям- фактам таких записей применяются функции агрегации. Группировка осуществляется и при построении OLAP-куба. Однако, в отличие от куба, при использовании обработчика «Группировка» формируется таблица со сгруппированными значениями, которую можно в дальнейшем использовать для обработки другими алгоритмами (обработчиками) Deductor.

Обработчик «Сортировка»

С помощью сортировки можно изменять порядок следования записей в исходной выборке данных в соответствии с заданным пользователем алгоритмом сортировки. Результатом выполнения сортировки будет новый набор данных, записи в которой будут следовать в соответствии с заданными параметрами сортировки.

В окне настройки параметров сортировки представлен список условий сортировки, в котором содержатся две графы:

☐ Имя поля - содержит имя полей, по которым следует выполнить сортировку.

□ Порядок сортировки - содержит порядок сортировки данных в соответствующем поле – по возрастанию или по убыванию.

Визуализатор «Карта Кохонена»

Данный визуализатор обеспечивает просмотр построенной в результате обучения самоорганизующейся карты, которую можно представить в виде слоеного пирога, каждый слой которого представляет собой раскраску, порожденную одной из компонент исходных данных.

Полученный набор раскрасок может использоваться для анализа закономерностей, имеющих между компонентами набора данных (рис. 20).

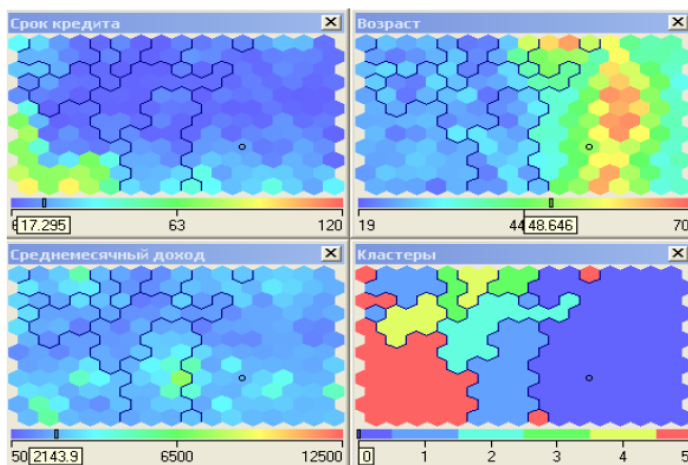


Рисунок 20 - Самоорганизующиеся карты

Эксперт, последовательно просматривая карты, выдвигает гипотезы, объясняющие объединение прецедентов в отдельные группы (кластеры). Например, карты на рис. 20 подтверждают гипотезу, что кредиты на длительный срок востребованы исключительно у заемщиков молодого возраста.

Визуализатор «Правила»

Данный визуализатор является альтернативой дереву решений – правила отображаются не в иерархичном, а обычном продукционном виде «Если-то».

Визуализатор «Таблица сопряженности»

Для того чтобы оценить качество классификации данных, обычно используют таблицу сопряженности. Для решения задачи классификации используется таблица, в которой уже есть выходной столбец, содержащий класс объекта. После применения алгоритма добавляется еще один столбец с выходным полем, но его значения уже вычисляются, используя построенную модель. При этом значения в столбцах могут отличаться. Чем больше таких отличий, тем хуже построенная модель классификации. Ниже изображен пример таблицы сопряженности.

Таблица 8

Пример таблицы сопряженности

	Классифицировано			
Фактически	Класс 1	Класс 2	Класс 3	Итого
Класс 1	239			239
Класс 2	7	10		17
Класс 3	4	1	17	22
Итого	250	11	17	278

В данном примере три класса, поэтому таблица сопряженности имеет размер 3 на 3 ячейки. На главной диагонали показано количество правильно классифицированных примеров (зеленый цвет). Красным цветом выделены неправильно распознанные примеры.

Таблицу сопряженности удобно применять для оценки качества модели, построенной с помощью обработчика «Дерево решений». Если количество неправильно классифицированных примеров довольно велико, это говорит о плохо построенной модели и нужно либо изменить параметры построения модели, либо увеличить обучающую выборку, либо изменить набор входных полей. Если же количество неправильно классифицированных примеров

мало, это может быть почвой для дальнейшего анализа и говорит о том, что пример является аномалией. В этом случае можно посмотреть, чем же характеризуются такие примеры и возможно добавить новый класс для отнесения этих примеров.

Визуализатор «Что-если»

Анализ по методу «Что-если» позволяет исследовать как будет вести себя построенная система обработки при подаче на ее вход тех или иных данных. Проще говоря, проводится эксперимент, в котором, изменяя значения входных полей обучающей или рабочей выборки нейронной сети или дерева решений, пользователь наблюдает за изменением значений на выходе (рис. 21).

Поле	Тип	Значение	Минимум	Максимум	Кол-во
Входные					
Сумма кредита	12	7000	2000	69500	
Возраст	12	37	19	70	
Образование	ab	специальное			3
Площадь квартиры	12	37	12	70	
Автомобиль	ab	отечественная			3
Срок проживания	12	22	2	43	
Выходные					
Давать кредит	ab	Да			

Рисунок 21 - Таблица «Что-если»

С использованием диаграммы «Что-если» можно решать обратную задачу – то есть визуально наблюдать, при каких значениях входных переменных будет достигнуто желаемое выходное значение.

ЗАДАНИЕ № 4

Задачи регрессии. Прогнозирование объема продаж

Цель работы – научиться применять методы Data Mining для решения задач прогнозирования временных рядов на примере построения модели прогноза объема продаж. В папке \Lab4 распо-

ложен файл sales.dbf – данные, содержащие историю продаж за некоторый период, имеющий следующую структуру:

Дата Группа товара Товар Количество Сумма

10.01.2011 Группа1 Товар1 768 2418.00

12.01.2011 Группа1 Товар1 64 211.11

13.01.2011 Группа1 Товар2 346 1042.00

13.01.2011 Группа2 Товар3 6 21.7

Задание

Требуется на основе исторических данных построить прогноз количества и сумм продаж на будущие два периода (период - месяц) по каждой товарной позиции. Оцените точность прогноза, сравнив результаты с реальными данными будущих продаж, которые находятся в файле fact.txt.

Для выполнения задания понадобятся следующие обработчики:

- ☐ фильтрация;
 - ☐ преобразование даты;
 - ☐ группировка;
 - ☐ разгруппировка;
 - ☐ автокорреляция;
 - ☐ удаление аномалий и сглаживание (парциальная предобработка);
 - ☐ скользящее окно;
 - ☐ нейросеть;
 - ☐ прогноз;
- и визуализаторы:
- ☐ таблица и OLAP-куб;
 - ☐ диаграмма;
 - ☐ диаграмма прогноза.

Порядок выполнения лабораторной работы

Последовательность действий для построения прогноза продаж по каждому товару включает в себя следующие шаги.

1. Выдвижение гипотезы – какие факторы оказывают влияние на будущий объем продаж.

2. Сбор данных и приведение их к пригодному для дальнейшей обработки виду.

3. Выделение сезонности в истории продаж.

4. Удаление аномалий и шумов.

5. Выбор метода для построения модели прогноза продаж.

6. Оценка качества построенной модели.

В данном случае уже выдвинута гипотеза о том, что на объеме продаж будущего периода в основном влияют продажи за прошлые месяцы. Для остальных этапов необходимо воспользоваться соответствующими обработчиками Deductor.

Обработчик «Автокорреляция»

Целью автокорреляционного анализа является выяснение степени статистической зависимости между различными значениями (отсчетами) случайной последовательности, которую образует поле выборки данных. В процессе автокорреляционного анализа рассчитываются коэффициенты корреляции (мера взаимной зависимости) для двух значений выборки, отстоящих друг от друга на определенное количество отсчетов, называемые также лагом.

Применительно к анализу временных рядов автокорреляция позволяет выделить месячную и годовую сезонность в данных (рис. 4.1). Видно, что пик зависимости на данных приходится на 12 месяц, что свидетельствует о годовой сезонности. Поэтому величину продаж годовой давности необходимо обязательно учитывать при построении модели (если используется __нейронная сеть – то подавать на вход).

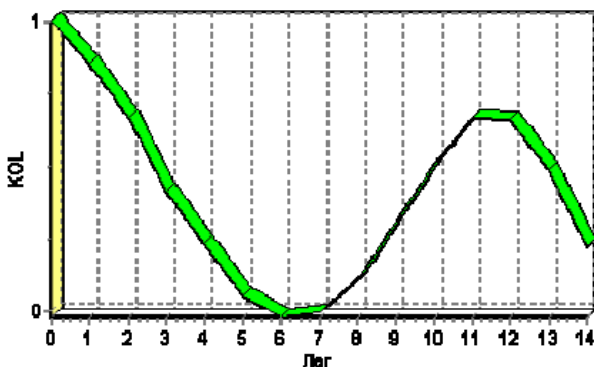


Рисунок 22 - Месячная и годовая сезонность в данных

Обработчик «Парциальная предобработка»

Позволяет удалить аномалии и шумы в исходных данных. Шумы в данных (в данном случае временной ряд истории продаж) не только скрывают общую тенденцию, но и проявляют себя при построении модели прогноза. Из-за них модель может получиться плохого качества. Аналогичная ситуация с аномалиями, т.е. резкими отклонениями величины от ее ожидаемого значения.

Однако следует понимать, что решение о сглаживании, удалении шумов и аномалий целиком зависит от специфики решаемой задачи, бизнес-правил предметной области и т.д. Иными словами, при сглаживании временного ряда выдвигается гипотеза о зашумленности исходных данных.

Аномалии, или случайные всплески в отгрузке товаров могут быть вызваны такими факторами, как задержка транспорта в пути и другими форс-мажорами. Кроме того, при агрегации продаж по месяцам уже происходит частичное сглаживание данных, и, может быть, дополнительное удаление шумов не требуется. В любом случае, необходимо строить несколько моделей.

В обработчике «Парциальная обработка» предусмотрен выбор степени подавления аномалий и вычитания шумов: малая, средняя, большая. Кроме того, в обработчике предусмотрена возможность заполнения пропусков, а также сглаживание данных, в

том числе вейвлет-преобразованием. Выбор того или иного алгоритма определяется спецификой конкретной задачи.

Обработчик «Скользящее окно»

При прогнозировании временных рядов при помощи обучающихся алгоритмов (в том числе искусственной нейронной сети), требуется подавать на вход анализатора значения нескольких, смежных, отсчетов из исходного набора данных. Например, в случае когда необходимо подавать на вход значения сумм продаж за последние 3 месяца и сумму продаж год назад. При этом эффективность реализации заметно повышается, если не выбирать данные каждый раз из нескольких последовательных записей, а последовательно расположить данные, относящиеся к конкретной позиции окна, в одной записи.

Значения в одном из полей записи будут относиться к текущему отсчету, а в других – смещены от текущего отсчета «в будущее» или «в прошлое». Следовательно, преобразование скользящего окна имеет два параметра: «глубина погружения» - количество «прошлых» отсчетов (включая текущий отсчет), попадающих в окно, и «горизонт прогнозирования» – количество «будущих» отсчетов. Такой метод отбора данных называется скользящим окном (поскольку это окно «перемещается» по всему набору).

Обработчик «Нейросеть»

Обработчик предназначен для решения задач регрессии и прогнозирования. В данном случае нейросеть строится для прогнозирования будущих значений временного ряда. Для проверки обобщающей способности нейросети рекомендуется разбить имеющееся множество данных на две части: обучающее и тестовое. Как правило, при прогнозировании временных рядов, доля тестового множества составляет не более 10-20%.

С помощью визуализатора «Диаграмма» оценивается способность построенной нейросетевой модели к обобщению. Для этого в одном окне выводятся графики исходного и спрогнозированного временных рядов. На рис. 23 изображен пример таких графиков. Видно, что построенная модель обеспечивает приемлемую точность.

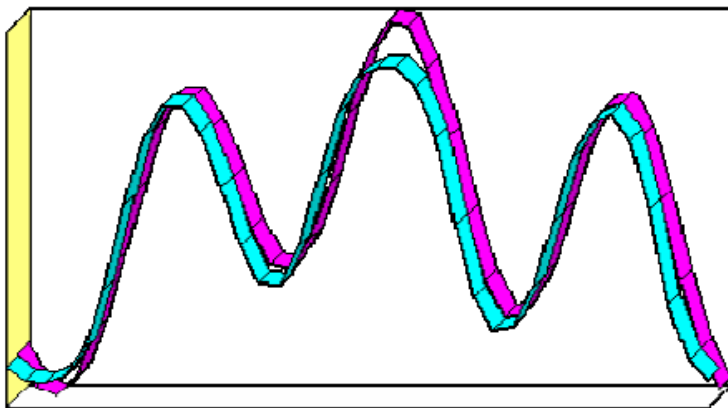


Рисунок 23 - Оценка качества построенной модели

Обработчик «Прогнозирование» и визуализатор «Диаграмма прогноза»

Прогнозирование позволяет получать предсказание значений временного ряда на число отсчетов, соответствующее заданному горизонту прогнозирования. Алгоритм прогнозирования работает следующим образом. Пусть в результате преобразования методом скользящего окна была получена последовательность временных отсчетов:

$$X(-n), \dots, X(-2), X(-1), X_0, X(+1)$$

где $X(+1)$ прогнозируемое значение, полученное с помощью предыдущего этапа обработки (например нейронной сети) на основе n предыдущих значений. Тогда, чтобы построить прогноз для значения $X(+2)$ нужно сдвинуть всю последовательность на один отсчет влево, чтобы ранее сделанный прогноз $X(+1)$ тоже вошел в число исходных значений. Затем снова будет запущен алгоритм расчета прогнозируемого значения - $X(+2)$ будет рассчитан с учетом $X(+1)$ и так далее в соответствии с заданным горизонтом прогноза.

Диаграмма прогноза становится доступной в списке способов представления только для тех ветвей сценария, которые содержат прогноз временного ряда. Основное отличие диаграммы про-

гноза от обычной диаграммы в том, что на ней, кроме исходных данных отображаются результаты прогноза, при этом исходные данные и прогноз отличаются по цвету.

Обработчик «Разгруппировка»

Применительно к задаче прогнозирования объема продаж разгруппировка позволяет распределить прогнозные значения, рассчитанные моделью (например, нейросетью) для определенной группы товара, по каждой товарной позиции в данной группе. Основой для такого попозиционного распределения служит гипотеза о том, что каждый товар в группе Просмотр результатов попозиционного прогноза удобно реализовать в многомерном виде, особенно если прогноз осуществляется более чем на один период (рис. 24)

Шаг прогноза ▼			
ТОВАР ▼	1	2	Итого
1818	20 026	32 515	52 540
2473	20 590	33 431	54 021
2844	8 141	13 218	21 358
3061	8 776	14 249	23 025
3068	1 021	1 657	2 678
4397	6 176	10 027	16 203
4530	2 033	3 302	5 335
4531	1 844	2 994	4 838
4718	5 308	8 618	13 926
Итого	73 915	120 011	193 926

Рисунок 24 - Прогноз по каждой товарной позиции после разгруппировки

Таким образом, типичный сценарий построения модели объема продаж в Deductor выглядит следующим образом (рис. 25).

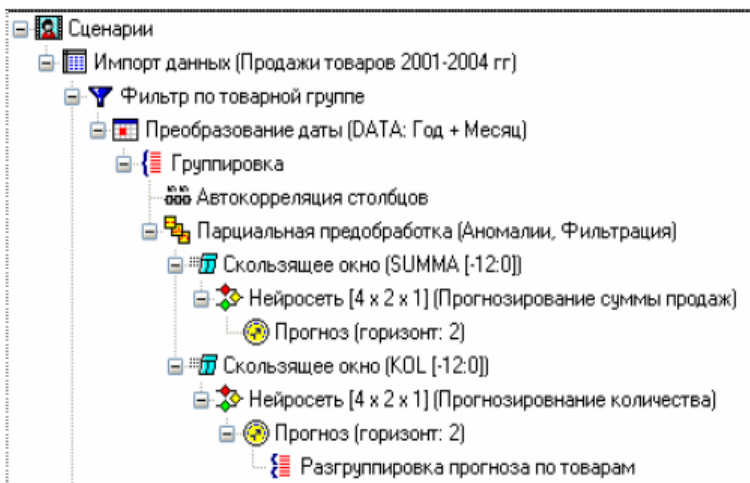


Рисунок 25 - Типичный сценарий построения прогноза продаж

ЗАДАНИЯ ДЛЯ САМОСТОЯТЕЛЬНОГО ВЫПОЛНЕНИЯ

1. Анализ данных CRM-систем

Цель работы – научиться применять методы Data Mining для решения задач управления клиентской базой: анализ воздействия рекламы, сегментация клиентской базы, поиск признаков прибыльных клиентов, анализ предпочтений товаров, прогнозирование успеха сделки и т.д.

Задание

На основе произвольно сформированной клиентской БД с помощью возможностей платформы Deductor решить **не менее 3-х задач** по управлению клиентской базой.

Методический материал

Любая организация в процессе своей деятельности стремится максимизировать прибыль. Для этого ей необходимо правильно организовать работу, ориентируясь на те вещи, которые приносят наибольшую прибыль с наименьшими затратами. Неко-

торое время назад считалось, что товар или предоставляемые организацией услуги приносят ей прибыль. Сейчас же все большее распространение приобретает клиентно-ориентированная стратегия, где во главу угла ставится клиент и механизмы взаимодействия с ним. Данная стратегия получила название CRM – Customer Relationship Management. На рынке существует большое разнообразие CRM систем [1, 2], однако почти все они предназначены для автоматизации работ по сбору и систематизации данных о клиентах и практически не обладают развитыми средствами анализа. Только в наиболее дорогих имеются средства OLAP. Из-за игнорирования вопросов анализа CRM данных, часто организации даже не подозревают о каких-то закономерностях в поведении клиентов, в то время как знание подобных закономерностей и их учет в своих действиях могут принести значительную практическую пользу.

Ниже рассмотрен кейс – аналитическое решение для анализа информации о клиентах на базе платформы Deductor. Кейс включает в себя хранилище данных о клиентах, готовые механизмы аналитической отчетности, а также сценарии решения более сложных задач: анализ воздействия рекламы, сегментация клиентов, установление признаков клиентов, приносящих наибольшую прибыль, определение предпочтения товаров, прогнозирования успеха сделки и объемов продаж.

Хранилище данных по клиентам

Важным этапом на пути построения аналитической системы является создание единого централизованного хранилища данных по клиентам. В нем должна содержаться наиболее полная информация о клиентах, об истории взаимоотношений, открытых или завершенных сделках с клиентом.

Существует много направлений деятельности организаций, работающих с клиентами. И для разных направлений, вероятно, требуется хранить различную информацию о потребителях. Например, если организация занимается торговлей, то наиболее интересной информацией о клиенте будет сфера его деятельности, категория клиента, географический регион его расположения, а сделки будут отражать продажи некоторого товара. Если же это

банки, работающие с населением, то информацией о клиенте будут его возраст, пол, семейное положение, другие личные характеристики, социальный статус, заработок и другое, а сделки будут отражать предоставление каких-либо услуг, например, предоставление кредита. Описываемое решение ориентировано на использование в торговых организациях.

В хранилище данных информация хранится в измерениях и процессах. Измерение – это объект анализа, который может характеризоваться свойствами, присущими только ему и имеет уникальный идентификатор. Процесс представляет собой звезду, в центре которой хранятся факты, а лучи являются измерениями. Хранилище данных будет содержать следующую информацию о клиентах: название, вид (юридическое или физическое лицо), сфера деятельности, географический регион расположения клиента, тип клиента по классификации ABC, тип клиента по классификации XYZ, потенциальный клиент или клиент.

Процесс отображает сделку с клиентом, то есть продажи клиенту некоторого товара. Фактом процесса будет сумма сделки и количество закупаемого товара. Измерения – клиент, менеджер организации, курирующий сделку, дата совершения сделки, состояние сделки (открыта, отказ, успех), причина отказа в случае неуспешной сделки, источник информации о приобретаемом товаре, товар.

Например, процесс "Сделка" содержит следующее измерения:

- номер клиента;
- состояние сделки;
- причина отказа в случае неуспешной сделки;
- менеджер, курирующий сделку;
- дата сделки;
- источник информации, способствующий сделке;
- номер покупаемого товара;
- номер региона;
- сфера деятельности клиента.

и факты:

- количество купленного по сделке товара;

- сумма сделки.

Зачем требовалось хранилище данных? Во-первых для централизованного хранения данных о клиентах и сделках. Данные в ХД могут собираться из любых источников, например, из учетной CRM системы, бухгалтерской или складской программы. Во-вторых данные будут представлены в удобном для анализа виде. В-третьих, обеспечивается полнота и непротиворечивость данных с точки зрения анализа.

Теперь, когда есть хранилище, можно приступить к анализу данных по клиентам.

Создание аналитической отчетности

Благодаря аналитической отчетности данные из хранилища представляются в виде, удобном для дальнейшего анализа. Наиболее удобным инструментом для получения аналитической отчетности являются OLAP-кубы (On Line Analytical Processing – оперативная аналитическая обработка данных). OLAP дает возможность в реальном времени генерировать описательные и сравнительные сводки данных и получать ответы на различные другие аналитические запросы. OLAP-кубы представляют собой проекцию исходного куба данных на куб данных меньшей размерности. При этом значения ячеек объединяются. Такие проекции или срезы исходного куба представляются на плоскости в виде кросс-таблицы.

На рисунке 26 представлена кросс таблица с двумя измерениями: менеджер и статус сделки. По такой таблице легко определить наиболее успешных менеджеров.

	DEAL_STATE ▾			
MANAGER ▾	Отказ	Открыта	Успех	Итого
Баранов	8 105 280,00	7 724 140,00	46 459 960,00	62 289 380,00
Дмитриева	22 065 220,00	19 748 960,00	15 393 080,00	57 207 260,00
Иванова	20 820 540,00	23 102 400,00	14 714 560,00	58 637 500,00
Костина	20 866 180,00	23 381 780,00	14 251 360,00	58 499 320,00
Медведев	7 356 480,00	8 457 580,00	45 713 460,00	61 527 520,00
Чернова	7 990 900,00	8 014 340,00	46 028 440,00	62 033 680,00
Итого	87 204 600,00	90 429 200,00	182 560 860,00	360 194 660,00

Рисунок 26 - Кросс-таблица

Более наглядное представление кросс-таблицы дает кросс-диаграмма:

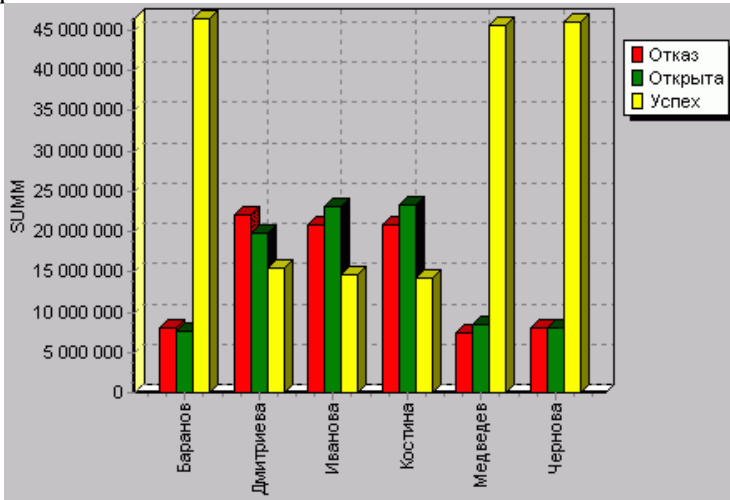


Рисунок 27- Кросс-диаграмма

В Deductor в OLAP-кубе с помощью инструмента "селектор" можно агрегировать факты по какому-либо измерению, оставляя только те объекты, которые соответствуют указанному условию. Например, можно объединить сумму сделок по клиентам, оставив только тех, которые в сумме приносят 50% прибыли. На рисунках 28 и 29 представлены селектор и клиенты, приносящие 50% прибыли.

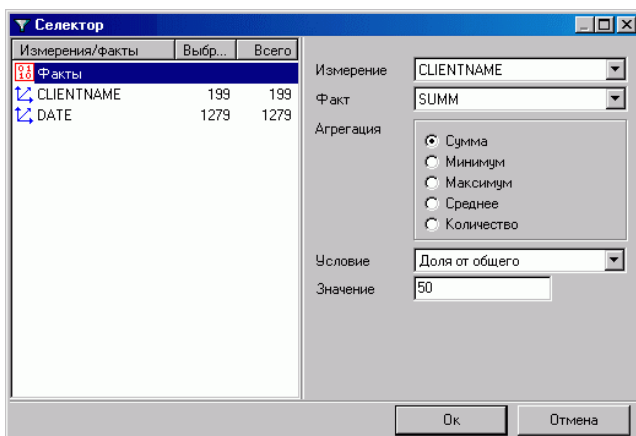


Рисунок 28 – Селектор

CLIENTNAME	Σ SUMM
Покупатель 1	3 044 620,00
Покупатель 10	2 735 880,00
Покупатель 12	3 144 920,00
Покупатель 2	15 923 320,00
Покупатель 3	5 767 360,00
Покупатель 35	2 811 700,00
Покупатель 4	10 285 840,00
Покупатель 5	5 326 060,00
Покупатель 6	14 545 360,00
Покупатель 7	8 896 760,00
Покупатель 8	19 803 680,00
Итого	92 285 500,00

Рисунок 29 - Клиенты, приносящие 50% прибыли

Это небольшой пример применения аналитической отчетности. На практике можно таким образом анализировать любую информацию о клиентах, имеющуюся в хранилище в произвольных разрезах. Иногда может потребоваться некоторая предобработка данных, например, очистка данных , для чего используются механизмы анализа, заложенные в Deductor.

Анализ воздействия рекламы

Для привлечения новых клиентов торговые организации активно используют рекламу в средствах массовой информации. Однако не во всех источниках реклама создает одинаковое воздействие на людей. Кроме того, стоимость рекламы может очень сильно отличаться. Поэтому важно знать, какую пользу какая реклама приносит и на какие категории клиентов воздействует.

В данных о сделках содержится информация об источнике рекламы, который способствовал совершению сделки. С помощью OLAP-куба можно получать отчеты с объемами сделок в разрезе различных источников. Не менее интересна информация в разрезе источника и каких-либо характеристик клиента. На рисунке 5 приведен пример отчета, из которого видно, что сильное воздействие оказывают семинары, а также видно, как это воздействие распределяется по регионам.

	SOURCE								
REGION	Выставка	Газета	Друзья	Интернет	Не указ.	Радио	Семинар	ТВ	Итого
Брянск	27.04%	3.31%	0.88%	1.35%	0.39%	3.43%	34.88%	28.73%	100%
Владимир	41.44%	1.6%	3.13%	1.36%	1.64%	2.86%	39.78%	8.19%	100%
Волгоград	11.2%	10.81%	14.76%	7.49%	3.45%	14.4%	11.7%	26.19%	100%
Воронеж	14.29%		13.37%	20.83%	5.13%			46.38%	100%
Калуга	26.29%	6.15%	4.99%	1.66%	6.15%	4.81%	33.58%	16.37%	100%
Липецк	18.52%	8.97%	6.3%	5.72%	4.16%	11.25%	21.58%	23.5%	100%
Москва	12.14%	9.36%	11.31%	8.06%	8.73%	10.71%	16.72%	22.97%	100%
Московская обл.	20.71%	10.85%	7.03%	7.16%	5.74%	7.9%	25.98%	14.62%	100%
Нижний Новгород								100%	100%
Рязань	40.31%	4.6%	3.61%	0.6%	0.96%	2.19%	40.69%	7.04%	100%
Тамбов	48.7%	2.77%	2.25%	1.74%	2.41%	1.8%	37.04%	3.29%	100%
Тула	28.6%	4.32%	4.38%	2.88%	2.02%	4.45%	27.98%	25.36%	100%
Итого	29.7%	6.04%	5.41%	3.86%	3.77%	5.53%	30.18%	15.51%	100%

Рисунок 30 - Воздействие рекламы в регионах

Для того чтобы понять, какими же характеристиками обладают клиенты, совершающие сделки по тому или иному источнику рекламы, удобно воспользоваться самоорганизующимися картами Кохонена. Для их обучения необходимо использовать в качестве входных поля со всевозможными характеристиками клиентов и источник рекламы. После обучения нейросети будут построены карты по одной на каждое поле. Выделив на карте с источником интересующую рекламу, можно будет посмотреть на других картах, какие же клиенты попали в эту область.

Сегментация клиентов

Наиболее активные фирмы, занимающиеся торговлей, используют как пассивную рекламу для привлечения новых клиентов, то есть, например, рекламу на телевидении, на радио, в прессе, так и рассылку с прямыми коммерческими предложениями. Для повышения эффективности подобных мероприятий необходимо учитывать интересы клиентов, объектов воздействия, т.е. предлагать клиентам именно тот товар, который они предпочитают. Но нельзя учесть предпочтения каждого клиента. Необходимо выделять некоторые группы – сегменты – клиентов и уже этим группам предлагать конкретную категорию товаров.

Выделять сегменты клиентов можно по нескольким группам признаков. Это могут быть сегменты по сфере деятельности, по географическому расположению. Для чего это нужно? После сегментирования можно узнать, какие именно сегменты являются наиболее активными, какие приносят наибольшую прибыль и где находятся наиболее лояльные клиенты, выделить характерные для них признаки. Для решения этой задачи воспользуемся мощным механизмом кластеризации – самоорганизующимися картами Кохонена.

На рисунке 6 представлены карты Кохонена, полученные после кластеризации клиентов по характеристикам: вид (KIND), удаленность региона (DISTANCE), сфера деятельности (BRANCH), среднегодовые суммы сделок (SUMM) и частота сделок – количество сделок в неделю (DEAL_STATE). Поле BRANCH является строковым и использовалось как выходное, поэтому в процессе обучения нейросети оно не участвовало. Но после группировки по остальным полям значения поля BRANCH также могли быть сгруппированы на карте.

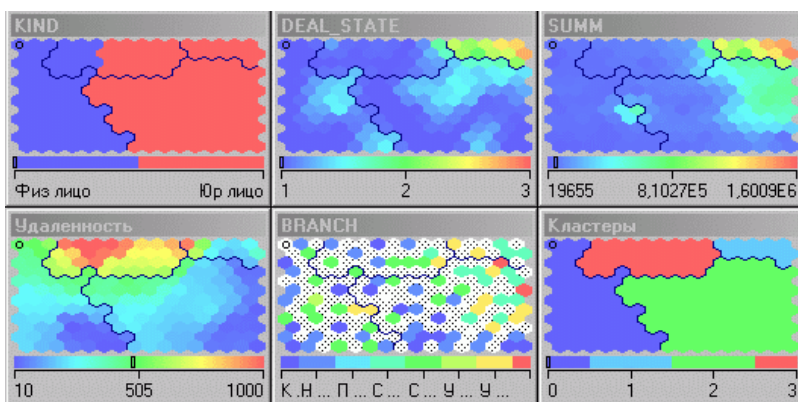


Рисунок 31 - Сегментация клиентов

Видно достаточно четкое разделение на 4 сегмента. Воспользуемся теперь деревом решений для получения правил отнесения клиентов к этим сегментам. Входными для дерева решений будут те же характеристики, а выходным – номер кластера. Результат представлен на рисунке 7.

Полученные сегменты можно интерпретировать следующим образом. Сегмент 0 – физические лица, расположенные ближе 600 км – клиенты, совершающие сделки на небольшую сумму и не очень часто. Сегмент 1 – клиенты, приносящие большую прибыль и часто совершающие сделки – это юридические лица в сфере связи, расположенные от 175 до 600 км. Клиенты сегмента 3 редко совершают сделки и приносят маленькие прибыли. Эти клиенты расположены дальше 600 км. Сегмент 2 обладает средними показателями

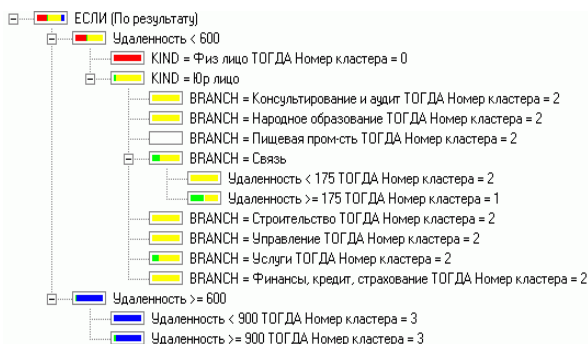


Рисунок 32 - Правила отнесения к сегменту

Поиск признаков прибыльных клиентов

Какие же из выделенных выше сегментов являются наиболее прибыльными? Список клиентов, приносящих большую прибыль, удобно посмотреть с помощью OLAP-куба. На рисунке 8 показана кросс-таблица, на которой по горизонтальной оси отложены номера сегментов клиентов, полученных при помощи карт Кохонена, а по вертикальной – среднегодовые суммы сделок, приходящиеся на каждого клиента. Видно, что особое внимание нужно уделять клиентам первого сегмента.

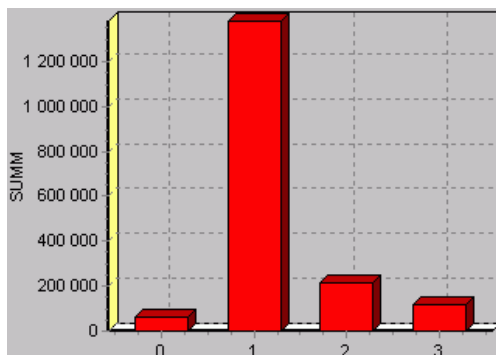


Рисунок 33 - Продажи по сегментам

Анализ предпочтений товаров

Как уже говорилось выше, анализ предпочтений товаров нужен для успешного проведения маркетинговых кампаний по рекламной рассылке клиентам. Для этого выделяются сегменты клиентов, предпочитающих определенную группу товаров. При этом один и тот же клиент может попасть в разные сегменты.

Наиболее простой и наглядный способ определения, какие клиенты предпочитают ту или иную группу товара, – это применение карт Кохонена. Для анализа используется информация обо всех сделках с клиентами на приобретение товара, т.е. есть таблица с клиентами, их характеристиками и наименованием группы товара, который они приобретали. Эта таблица используется при обучении карт Кохонена. На вход нейросети подаются характеристики клиента, а на выход – группа товара. Надо отметить, что выходные поля в обучении нейросети не участвуют. После группировки клиентов в некоторые области группы товаров также должны сгруппироваться в области, если, конечно, существует зависимость между характеристиками клиентов и покупаемым товаром. Если после обучения карт Кохонена группы товаров не разбросаны по карте, а сгруппированы в области, то зависимость есть, и далее можно сделать группировку по сегменту клиентов, клиентам и группе товаров. Для такой группировки удобно воспользоваться OLAP-кубом, подсчитав количество клиентов в каждом сегменте по каждой группе товара. Результат такой группировки представлен на рисунке 34.

На этом рисунке по горизонтали отложены номера сегментов клиентов, а по вертикали – количество клиентов в каждом сегменте по каждой группе товара. Например, в нулевом сегменте 90 клиентов предпочитают третью группу товара, а 65 – вторую. Причем, один и тот же клиент мог внести вклад и в ту и в другую группу.

Видно, что в первом сегменте нет какого-то определенного предпочтения товаров, во втором сегменте явно пользуется спросом группа 1, в нулевом предпочитают группы 2 и 3, в третьем – 1 и 2. Теперь ясно, каким клиентам и какой товар предлагать.

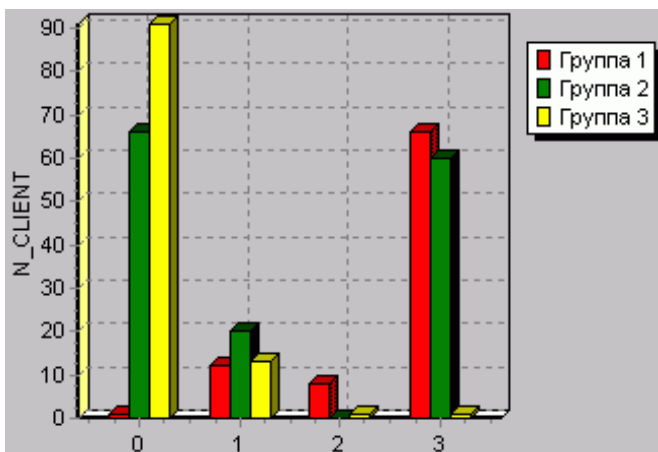


Рисунок 34 - Количество клиентов в каждом сегменте по каждой группе товара

В торговых организациях постоянно обновляется ассортимент товаров. Если ассортимент вовремя не обновится, организация вскоре может потерять часть клиентов. С другой стороны, чрезмерное пополнение ассортимента новинками может не соответствовать текущим требованиям рынка. Алгоритм поиска компромисса может быть следующим. В каждой сегментной группе клиентов ищутся те, которые в большей степени предпочитают данную группу товара, и именно им предлагается новинка из той товарной группы, которая наиболее предпочтительна для данного сегмента. В случае успешного предложения этот товар можно будет предложить остальным представителям данного сегмента.

В OLAP-кубе из предыдущего примера возьмем только третий сегмент и с помощью селектора оставим только первые 10 наиболее активных клиентов. Ими оказались клиенты, представленные на рисунке 35.

		GROUP ▾			
Segm ▾	CLIENTNAME ▾	Группа 1	Группа 2	Группа 3	Итого
3	Покупатель 133	12,84%	4,87%		5,79%
	Покупатель 135	17,19%	0,99%		5,44%
	Покупатель 136	14,36%	5,03%		6,3%
	Покупатель 140	10,01%	6,02%		5,44%
	Покупатель 143	11,1%	5,18%		5,41%
	Покупатель 15		50,5%		21,1%
	Покупатель 172	17,41%	3,35%		6,49%
	Покупатель 30			100%	28,96%
	Покупатель 41		17,9%		7,48%
	Покупатель 62	17,08%	6,17%		7,57%
	Итого	100%	100%	100%	100%
Итого		100%	100%	100%	100%

Рисунок 35 - Клиенты третьего сегмента, в большей степени предпочитающие товар групп 1 и 2

В случае появления нового товара можно сначала предложить его этим клиентам, а в случае положительных откликов распространить предложение на остальных клиентов этой группы.

Прогнозирование успеха сделки

Заранее спрогнозировать успех сделки важно потому, что в случае отказа клиента организация теряет не только затраченное на его привлечение и работу с ним время, но и затраченные на него средства, а также самого клиента, который, видимо, уйдет к конкуренту.

Когда у организации накопилось достаточно много информации о завершенных сделках, успешных и неуспешных, можно использовать эту информацию для выяснения факторов, которые в большей степени влияют на конечное состояние сделки.

Для решения этой задачи воспользуемся деревом решений. В качестве входных полей для его обучения используем следующую информацию о сделках: KIND (вид клиента), ABC (тип по

ABC классификации), MANAGER (менеджер), REGION (регион), DISTANCE (удаленность), BRANCH (сфера деятельности), E-MAIL, FAX, FONE (количество писем, факсов, телефонных разговоров перед сделкой). В качестве выходного используем поле DEAL_STATE (состояние сделки).

N	Условие	Решение (DEAL_STATE)	Поддержка		Достоверность	
			%	Кол-во	%	Кол-во
1	FONE < 29,5	Отказ	33,36	1215	100,00	1215
2	FONE >= 29,5 И E-MAIL < 2,5	Отказ	0,55	20	100,00	20
3	FONE >= 29,5 И E-MAIL >= 2,5 И FONE < 30,5 И FAX < 3,5	Отказ	0,22	8	50,00	4
4	FONE >= 29,5 И E-MAIL >= 2,5 И FONE < 30,5 И FAX >= 3,5	Успех	0,88	32	100,00	32
5	FONE >= 29,5 И E-MAIL >= 2,5 И FONE >= 30,5	Успех	64,99	2367	100,00	2367

Рисунок 36 - Правила прогнозирования успеха сделки.

После обучения получаем дерево решений, которое представляет в иерархическом виде правила, следуя им, можно спрогнозировать успех сделки. На верхних уровнях дерева располагаются характеристики сделки, в наибольшей степени влияющие на ее состояние. Правила из дерева решений можно представить в виде таблицы, представленной на рисунке 36.

Прогнозирование объемов продаж

Торговые организации стремятся свести к минимуму время, которое товар лежит на складе, а также место, которое он там занимает. С другой стороны, необходимо, чтобы на складе всегда лежал требуемый в настоящее время товар. Прогнозирование объемов продаж является важным шагом на пути принятия решения по оптимизации работы предприятия.

Важными данными для построения прогноза объемов продаж является статистика продаж за предыдущие периоды. Такая информация есть в нашем хранилище. Не всегда имеет смысл делать общий прогноз продаж. Дело в том, что существуют случайные клиенты, которые нарушают общую тенденцию, покупая товар в "неподходящий" сезон или случайный товар. Поэтому перед про-

гнозированием нужно выделить сделки по соответствующему сегменту клиентов и определенной группе товара, например, предпочитаемого этим сегментом.

Для примера сделаем фильтрацию всех сделок по третьему сегменту клиентов группе товара номер 1 (предпочитаемый третьим сегментом), оставив только успешные сделки. Прогнозировать будем количество продаваемого товара. Рассматривать сделки по дням не имеет смысла, так как каждый день может очень сильно отличаться от другого по объему продаж. Но продажи в среднем по неделям или по месяцам отличаются не сильно. Сгруппируем все сделки по неделям. После такой группировки тенденция наблюдается лучше, но все еще существуют различные выбросы и шумы на кривой продаж. Для их устранения в Deductor применяется инструмент «парциальная обработка». Воспользуемся им, указав удаление аномальных значений и вейвлет преобразование для сглаживания кривой продаж. Результат всех этих операций показан на рисунке 12.

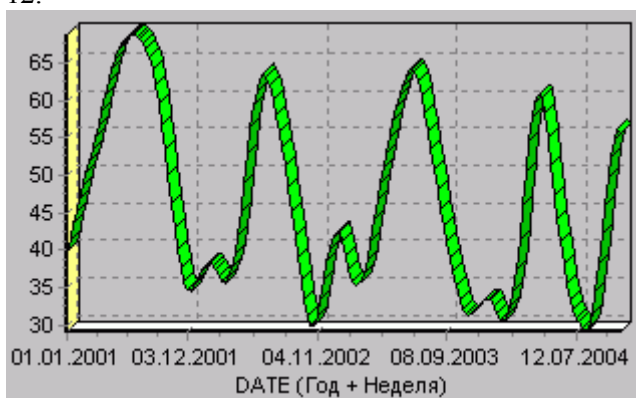


Рисунок 37 - Сглаженная кривая продаж

Прогноз строится на предположении сезонности продаж и общего развития рынка. Для оценки сезонности используют автокорреляцию. Она показывает степень зависимости некоторого значения от предыдущих значений этой величины. В нашем примере

максимум линии автокорреляции приходится на 43 недели, что приблизительно соответствует одному году (рисунок 38).

После оценки сезонности необходимо построить модель. В нашем примере, когда объемы продаж зависят от предыдущих продаж, можно использовать линейную регрессию. Однако, в более сложных ситуациях лучшие результаты даст использование нейронных сетей для построения более адекватных моделей. Обучать модель на данных за полгода назад не имеет смысла, так как зависимость между ними и текущими продажами очень слабая (рисунок 38). При обучении линейной регрессии будем подавать на ее вход объем продаж за две предыдущие недели текущего года, чтобы учесть общее развитие рынка, и две недели, соответствующие им год назад, чтобы учесть сезонность. Таким образом учитываются спуски и подъемы кривой продаж. Качество построенной модели можно наглядно посмотреть на диаграмме рассеяния (рисунок 39). Истинные значения располагаются вдоль прямой линии, а полученные по модели – выше или ниже ее.

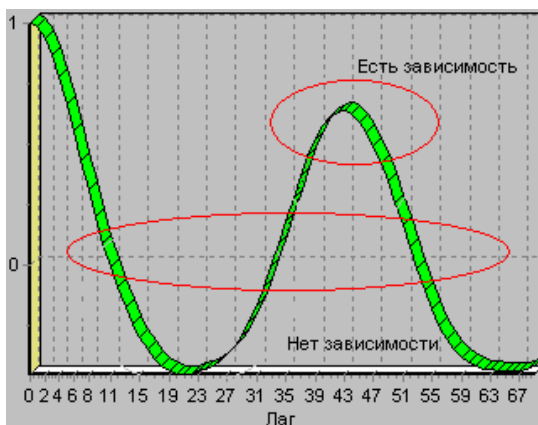


Рисунок 38 - Автокорреляция объемов продаж

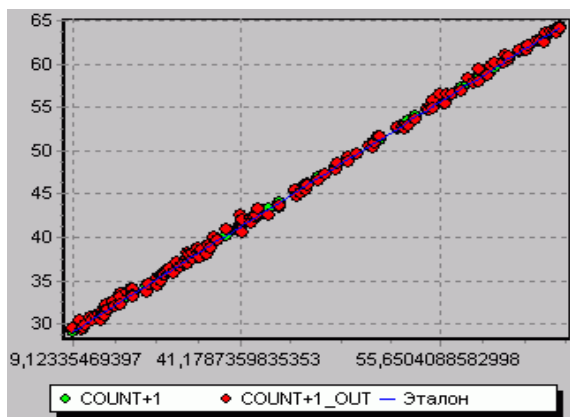


Рисунок 39 - Диаграмма рассеяния.

После того, как построена модель, можно построить прогноз, прогнав через нее историю продаж. Сделаем прогноз на 4 недели вперед. Прогноз на более длительный период будет в гораздо меньшей степени соответствовать действительности. Результат можно посмотреть на диаграмме прогноза (рисунок 40), либо в таблице (рисунок 41).

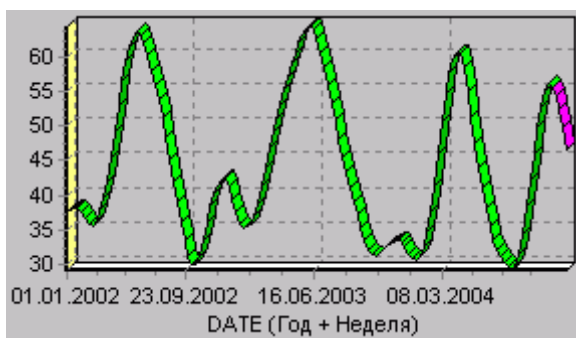


Рисунок 40 - Прогноз объемов продаж на 4 недели (диаграмма прогноза)

COUNT+1	Шаг прогноза
55,6504088582998	
54,3683652006523	1
52,3521976068863	2
49,7009969563775	3
46,5215618447132	4

Рисунок 41 - Прогноз объемов продаж на 4 недели (таблица)

После того, как сделан прогноз по группам товаров, можно получить прогноз по каждому товару группы. Делается это путем разгруппировки пропорционально вкладу каждого товара группы в общий объем за предыдущие периоды времени. Результат такой разгруппировки показан на рисунке 42.

	Шаг прогноза				
GOODS	1	2	3	4	Итого
Товар10	12,60	12,13	11,52	10,78	47,02
Товар14	5,75	5,53	5,25	4,92	21,45
Товар15	3,76	3,62	3,43	3,21	14,02
Товар16	5,53	5,32	5,05	4,73	20,62
Товар2	3,76	3,62	3,43	3,21	14,02
Товар20	5,97	5,75	5,45	5,11	22,27
Товар3	7,51	7,24	6,87	6,43	28,05
Товар5	5,75	5,53	5,25	4,92	21,45
Товар9	3,76	3,62	3,43	3,21	14,02
Итого	54,37	52,35	49,70	46,52	202,94

Рисунок 42 - Прогноз объемов продаж на 4 недели по каждому товару

Сделав таким способом прогноз по всем сегментам клиентов и всем группам товаров, можно получить общий прогноз объемов продаж. Такой прогноз учитывает сезонность продаж каждой группы товаров и активность каждого сегмента клиентов в разные периоды времени.

Список литературы по дисциплине

Основная литература

1. Интеллектуальные модели анализа экономической информации: электронный курс лекций. – BaseGroup Labs, 2005.
2. Аналитическая платформа Deductor 4. Руководство пользователя. – BaseGroup Labs, 2005. – 101 с.
3. Deductor 4. Описание демопримеров. – BaseGroup Labs, 2004.
4. Чубукова И.А. Data Mining. – М.: Лаборатория знаний Интуит, 2008. – 382 с.

Дополнительная литература

1. Барсегян А. А. Технология анализа данных: Data Mining, Visual Mining, Text Mining, OLAP/ А. А. Барсегян [и др.]. - 2-е изд. – СПб. : БХВ-Петербург, 2007. – 375 с.
2. Дюк В., Самойленко А. Data Mining. – СПб.: Питер, 2001.
3. Люггер Джордж Ф. Искусственный интеллект. Стратегии и методы решения сложных проблем. – М.: Вильямс, 2005. – 864 с.
4. Сигел Э. Ф. Практическая бизнес-аналитика. – М.: Вильямс, 2002.

Федеральное государственное образовательное бюджетное
учреждение высшего профессионального образования
“Поволжский государственный университет
телекоммуникаций и информатики”
443010, г. Самара, ул. Льва Толстого 23

Подписано в печать 19.03.12 г. Формат 60 x 84/16
Бумага офсетная №1. Гарнитура Таймс.
Заказ 1193. Печать оперативная. Усл. печ. л. 8,77. Тираж 100 экз.

Отпечатано в издательстве учебной и научной литературы
Поволжского государственного университета
телекоммуникаций и информатики
443090, г. Самара, Московское шоссе 77, т. (846) 228-00-44