Министерство образования и науки Российской Федерации

ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ
ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ
ВЫСШЕГО ОБРАЗОВАНИЯ
«САРАТОВСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
ИМЕНИ Н.Г. ЧЕРНЫШЕВСКОГО»

Кафедра Математического и компьютерного моделирования

Проектирование и реализация многомерной базы данных с использованием технологий Data Mining

АВТОРЕФЕРАТ БАКАЛАВРСКОЙ РАБОТЫ

студентки 4 курса 441 группы

направление 09.03.03 — Прикладная информатика

механико-математического факультета

Задворной Ирины Александровны

Научный руководитель
доцент, к. ф. – м. н. О.М. Ромакина

Зав. кафедрой
зав. кафедрой, д. ф. – м. н. Ю.А. Блинков

Саратов 2017

ВВЕДЕНИЕ

Данная работа раскрывает тему OLAP - технологий, а также продуктов ее реализации. На данный момент существует множество разработок по данной проблеме. OLAP — это универсальный инструмент, который может быть использован в любой прикладной области. В качестве объекта исследований для анализа представленных OLAP - продуктов были выбраны одни из лидирующих компаний на рынке: Microsoft, Oracle, Cognos и других. OLAP - технологии выступают в качестве технологий оперативной аналитической обработки данных, например, всех отчетов какой - либо компании. Практическое применение задачи сейчас распространено во многих областях, связанных с большими объемами анализируемой информации. С помощью OLAP - средств производится сбор, хранение и анализ многомерных данных в целях поддержки процессов принятия аналитических решений. Аспекты всех возможностей данной технологии и будут рассмотрены.

В последние годы очевидна тенденция к централизованной обработке информации и управлению данными. Это связано с ростом производительности вычислительных систем в целом и быстродействием средств коммуникации. Уже сегодня уровень развития и многообразие возможностей предлагаемых аналитических систем позволяют с выгодой использовать накопленные данные и определять оптимальную стратегию развития бизнеса.

В бакалаврской работе раскрываются вопросы по темам многомерного хранения данных и OLAP - технологий. Определяется, что представляют собой технологии, лежащие в основе информационных систем, реализующих принципы данных средств. Рассматриваются концепции хранилищ данных и OLAP, требования к хранилищам данных и OLAP - средствам, логическую организацию OLAP - данных, а также основные термины и понятия, применяемые в многомерном анализе.

Благодаря разнообразию различных OLAP - продуктов и различных технологических решений, каждый разработчик может подобрать свой набор подходящих инструментов для внедрения в своей компании. В качестве примера реализации будет рассмотрен пример построения хранилища данных для информационной системы с использованием OLAP - технологии фирмы Microsoft (Analysis Services в Microsoft SQL Server 2012).

Целью работы является изучение принципов OLAP - технологий и методов Data Mining, их особенности и практической реализации в продуктах компаний.

В ходе исследования будут решены следующие задачи:

- Разбор и определение основных терминов, используемых в OLAP;
- Рассмотрение различных архитектурных особенностей реализации OLAP - технологий;
- Выявление областей применения и возможностей использования данных средств;
- Анализ представленных на рынке средств для реализации OLAP;
- Проектирование и реализация многомерной базы данных для практического примера;
- Проведение анализа данных с использованием технологий Data Mining.

Итоговой задачей исследования данной темы является получение теоретических знаний в этой области и возможности их применения для решения частных и общих задач на практике.

Работа состоит из введения, трёх глав, заключения, списка использованных источников и приложения. В первой главе «OLAP - технологии и анализ данных» даются общие сведения об OLAP - средствах и технологиях Data Mining, описывается постановка основных понятий многомерного представления информации, рассматривается применение и решаемые задачи, а также преимущества использования. Вторая глава «Средства реализации OLAP - технологий» содержит сравнительный анализ основных программных продуктов и составляется перечень критериев выбора подходящего средства для применения на практике. В третьей главе, которая называется Информационная система «Маркетинговая кампания банка», посвященной практическому примеру использования представленных технологий, проводится построение многомерной базы данных и осуществляется анализ данных с помощью технологий Data Mining.

КРАТКОЕ СОДЕРЖАНИЕ РАБОТЫ

В первой главе задаются определения основных понятий.

Вначале рассматриваются хранилища данных. В основе их концепции лежит идея разделения данных, используемых для оперативной обработки и анализа данных. Это позволяет оптимизировать структуры данных, используемые для хранения, для выполнения операций ввода, модификации, удаления, поиска и структуризации данных и для анализа. Их называют соответственно оперативными источниками данных и хранилищем данных.

Главной задачей хранилища является представление изначального материала для анализа в одном месте и в простой, понятной для дальнейшего разбора структуре. Основные требования: поддержка высокой скорости доступа к данным, обеспечение достоверности и полноты данных, поддержка непротиворечивости данных, возможность получения и сравнения различных данных, обеспечение просмотра данных, поддержка процесса пополнения данных.

Далее рассматривается сама технология OLAP. Технология комплексного многомерного анализа данных получила название OLAP (On - Line Analytical Processing). Это ключевой компонент организации хранилищ данных. Можно определить OLAP как совокупность средств многомерного анализа данных, накопленных в хранилище. При этом OLAP - функциональность может быть представлена различными способами.

Определяются следующие требования к приложениям для многомерного анализа: предоставление пользователю результатов анализа за приемлемое время, пусть даже ценой менее детального анализа, возможность осуществления любого логического и статистического анализа, характерного для данного приложения, и его сохранения в доступном для конечного пользователя виде, многопользовательский доступ к данным с поддержкой соответствующих механизмов блокировок и средств авторизованного доступа, многомерное концептуальное представление данных, включая полную поддержку для иерархий и множественных иерархий, возможность обращаться к любой нужной информации независимо от ее объема и места хранения.

После в работе определяется, что представляют собой кубы OLAP с точки зрения структуры. OLAP - структура, созданная из рабочих данных, на-

зывается OLAP - кубом. Основными составляющими структуры хранилищ данных являются таблица фактов и таблицы измерений. Таблица фактов содержит сведения о тех событиях, которые будут использоваться при анализе. Таблицы измерений содержат неизменяемые либо редко изменяемые данные. Каждая таблица измерений должна находиться в отношении «один ко многим» с таблицей фактов. Если каждое измерение содержится в одной таблице, такая схема хранилища данных носит название «Звезда». Если же хотя бы одно измерение содержится в нескольких связанных таблицах, такая схема хранилища данных носит название «Снежинка». Куб создаётся из соединения таблиц с применением одной из схем.

На основании того, что как исходные, так и агрегатные данные могут храниться либо в реляционных, либо в многомерных структурах, различают три способа хранения данных:

- MOLAP — исходные и агрегатные данные хранятся в многомерной базе данных.
- ROLAP — исходные данные остаются в реляционной базе данных, где они изначально и находились. Агрегатные же данные помещают в специально созданные для их хранения служебные таблицы в той же базе данных.
- HOLAP — исходные данные остаются в той же реляционной базе данных, где они изначально находились, а агрегатные данные хранятся в многомерной базе данных.

После этого в работе проводится изучение технологий Data Mining. Они представляют собой мощный аппарат для проведения современного бизнес - анализа информации, накапливаемой в процессе деятельности компаниями. Data Mining («добыча данных») — процесс обнаружения в первоначальных данных ранее неизвестных, нетривиальных, практически полезных и доступных интерпретации знаний, необходимых для принятия решений в различных сферах человеческой деятельности. Это не конкретная технология, а скорее сам процесс поиска различных взаимосвязей и закономерностей посредством самых разных математических и статистических алгоритмов. Цель их поиска состоит в представлении информации в форме, которая напрямую отражает протекающие бизнес - процессы, а также построить модель, при использова-

нии которой можно составить прогноз результативности процессов, важных для компании.

Выделяют пять основных типов закономерностей, которые определяются методами Data Mining: ассоциация, последовательность, классификация, кластеризация, временные закономерности. Технологии Data Mining реализуют довольно большое количество разнообразных методов исследования данных. Выделяют основные группы инструментов: регрессионный и корреляционный анализ; методы, основанные на эмпирических моделях; нейросетевые алгоритмы; деревья решений; кластерные модели; эволюционное программирование.

После этого рассматриваются применение, решаемые задачи, преимущества и недостатки.

Основные аспекты, решение которых становится более эффективным при использовании хранилищ данных и OLAP - технологий: аналитические, визуальные, имитационные, управленческие, оптимизационные, зависимые. Определяется список основных отраслей применения OLAP - технологий: статистика, телекоммуникации, промышленность, геологоразведка, банковское дело, бизнес, маркетинг, финансы, образование, медицина, фармакология, экономика, социология, страхование. Технологии применимы везде, где может быть определена задача анализа многофакторных данных.

Все недостатки технологий прямо вытекают из выбранного подхода к реализации OLAP - технологий. Технология OLAP призвана повысить эффективность информационно-аналитической и управленческой деятельности руководящего персонала. Из этой идеологии и вытекают основные преимущества. Во - первых, предметная ориентированность подобных систем. Во - вторых, многопользовательский режим работы. В - третьих, удобные средства доступа, просмотра, визуализации и анализа информации. В - четвертых, неизменность данных позволяют формировать и в дальнейшем использовать для анализа массивы заранее обработанных данных. В - пятых, быстрая детализация итоговых данных. В - шестых, высокая скорость формирования отчетов.

Во второй главе проводится сравнительный анализ основных программных продуктов, реализующих OLAP - технологии, и составляется перечень критериев выбора подходящего средства для применения на практике.

Многомерный анализ данных может быть произведен с помощью различных средств, которые делятся на две категории: клиентские OLAP - средства и серверные OLAP - средства. Главным преимуществом применения серверных OLAP - средств по сравнению с клиентскими OLAP - средствами является то, что в случае применения серверных средств вычисление и хранение агрегатных данных происходят на сервере, а клиентское приложение получает лишь результаты запросов к ним. Средства анализа и обработки данных масштаба предприятия, как правило, базируются именно на серверных OLAP - средствах. Выбор их достаточно велик. Рассматриваются подходы каждой из компаний (Microsoft Analysis Services, Oracle OLAP Services, Oracle Hyperion Essbase, SAS OLAP Server, Cognos PowerPlay, Microstrategy OLAP Services, IBM Cognos TM1, Mondrian Pentaho Analysis Services, Palo) к реализации, процесс их реального функционирования, архитектура используемого решения, конкурентоспособные преимущества, недостатки с точки зрения эксплуатации системы и способность продукта выполнять ту или иную задачу.

В ходе рассмотрения самых распространенных OLAP - средств определяются основные параметры, которые следует учитывать при выборе программного средства. От функциональных возможностей продукта целиком и полностью будут зависеть и перспективы развития системы. Различия продуктов связаны с самыми разными возможностями, начиная от функциональных и технических и заканчивая архитектурными. Были выделены важные аспекты, которые стоит определять для наиболее оптимального выбора OLAP - продукта. Во - первых, следует оценить потребности предприятия, связанные с масштабом предполагаемой системы. Во - вторых, необходимо определить возможности компании. В - третьих, следует определить, откуда будут поступать данные и какого они будут объема. В - четвертых, нужно выявить задачи, которые будут стоять перед OLAP - средством. В - пятых, сложность изучения внедряемого средства должна соответствовать персоналу.

В третьей главе в качестве примера реализации приводится создание многомерной базы данных, построенной на основе разрозненных данных по проведению телефонных маркетинговых кампаний банка. Данные относятся к прямым маркетинговым кампаниям (проводимым с помощью телефонных звонков) банка. Данная система собирает статистические данные о проведенных маркетинговых мероприятиях, качественных параметров успешности их проведения и различные экономические показатели, влияющие на результативность. На основе данной системы можно проводить статистический анализ данных с целью выявления тенденций к результативности предоставления услуг банка. Построенную систему можно определить как систему статистического сбора данных для дальнейшей оценки и анализа, синтезирующую знания из различных источников.

В составе системы реализованы следующие сущности для построения многомерной базы данных: Возрастная группа, Образование, Работа, Семья, Кредитная история, Банковская история, Время, Экономическая ситуация.

Сама система подразумевает хранение следующих показателей: длительность связи с клиентом в анализируемый период; количество произведенных контактов в период проведения данной акции на основе сформированной истории; количество дней, прошедших с последней связи с клиентом; число сотрудников, занятых в период проведения маркетинговой кампании; результативность связи с конкретным клиентом.

Далее производится выбор программного средства. Основной потребностью системы является хранение и удобное представление данных. Важным условием является наличие бесплатной версии продукта. Потребности в объемах разрабатываемой системы — один многомерный куб данных. Основные задачи — хранение, представление и анализ данных. Самым оптимальным решением оказался продукт Microsoft SQL Server with Analysis Services.

В выстраиваемой системе функционирует схема «Звезда», состоящая из 8 таблиц измерений и 1 таблицы фактов в соответствии с рисунком 1. Для дальнейшего анализа информации с помощью технологии Data Mining нам создается представление источника данных, которое будет содержать в себе информацию об основных показателях.

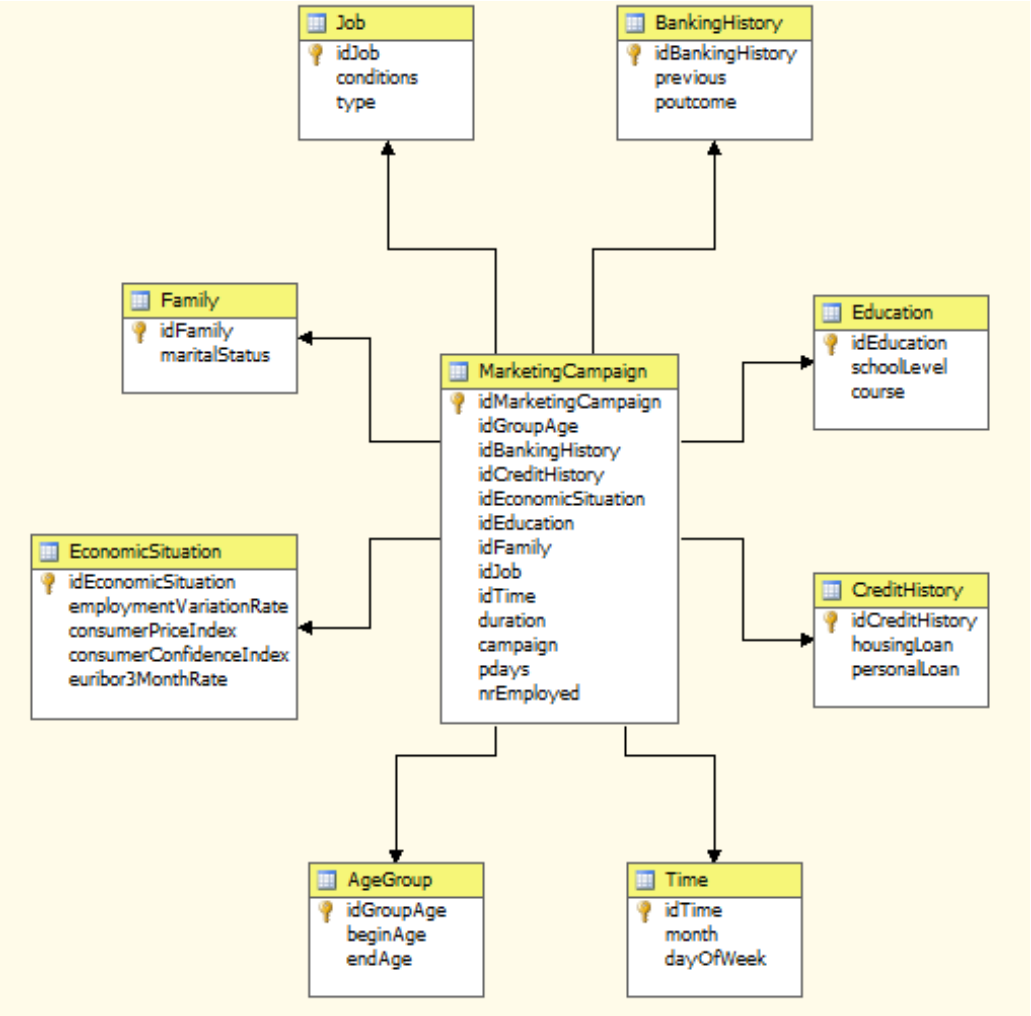

Рисунок 1 — Схема хранилища данных

После этого проводится анализ информации построенной системы с помощью технологии Data Mining. Этапы анализа: постановка задачи, сбор и подготовка данных, построение модели и оценка точности, применение модели.

В рамках поставленной практической задачи рассматриваются самые распространенные возможности Data Mining, представленные в рамках программного продукта Microsoft SQL Server 2012.

Алгоритм дерева решений Microsoft является алгоритмом классификации для использования в интеллектуальном моделировании как дискретных, так и непрерывных атрибутов. Алгоритм «Деревья решений Microsoft» создает модель интеллектуального анализа данных, проводя серию разбиений в дереве. Эти разбиения представлены как узлы. Алгоритм добавляет узел к мо-

дели каждый раз, когда найденный столбец ввода значительно коррелирует с прогнозируемым столбцом.

Итак, в рамках изучения имеющихся данных в созданной системе построено два дерева решений для определения факторов, влияющих на результативность проведения маркетинговых кампаний. Для первого дерева решений используются следующие данные: информация о возрасте клиента, его образовании, наличии кредитов, и типе работы. Для второго же дерева используются данные, характеризующие предыдущие и текущее взаимодействия клиента с банком, такие, как: количество произведенных звонков в период текущей маркетинговой кампании, время в днях с момента последнего обращения клиента в банк, результативность предыдущей кампании для конкретного клиента, количество звонков за весь период времени.

Для наглядной оценки достоверности построенной модели используем графики точности прогнозов. Они имеют следующий вид: ось X представляет собой процентную долю из набора данных, выбранных для прогноза, а на оси Y указывается процент точных прогнозов для этих данных. По отклонению мы можем давать оценку достоверности построенных моделей. Максимальное отклонение построенных деревьев решений не превосходит 15%. Модели не являются идеальными, но достаточно точно проводят прогноз основного результирующего показателя.

Алгоритм кластеризации Microsoft представляет собой алгоритм сегментации, который выполняет итерацию по случаям в наборе данных для группировки их в кластеры, которые содержат сходные характеристики. Алгоритм кластеризации Microsoft сначала идентифицирует отношения в наборе данных и генерирует серию кластеров на основе этих отношений. Группа кластеров указывает на график и иллюстрирует отношения, которые идентифицирует алгоритм. После первого определения кластеров алгоритм вычисляет, насколько хорошо кластеры представляют группировки, а затем пытается переопределить группировки для создания кластеров, которые лучше представляют данные. Алгоритм выполняет итерации этого процесса, пока он не сможет улучшить результаты за счет переопределения кластеров.

В рамках изучения имеющихся данных в построенной системе построены две различные кластеризации. Первая кластеризация проводена на основе

всех основных данных клиента: информация о возрасте клиента, его образовании, типе работы, количество произведенных звонков в период текущей маркетинговой кампании, день недели для текущего звонка, длительность разговора с клиентом, время в днях с момента последнего обращения клиента в банк, результативность предыдущей кампании для конкретного клиента, количество звонков за весь период времени и количество клиентов банка, занятых в рамках текущей маркетинговой кампании. Вторая кластеризация проведена на основе личной информации о клиентах: информация о возрасте клиента, его образовании, типе работы и его семейном положении. Подобная кластеризация позволяет оценить типичные группы клиентов банка. Максимальное отклонение по построенной модели около 13%.

Алгоритм нейронной сети Microsoft представляет собой реализацию популярной и адаптируемой архитектуры нейронной сети для машинного обучения. Алгоритм работает, проверяя каждое возможное состояние входного атрибута на каждое возможное состояние прогнозируемого атрибута и вычисляя вероятности для каждой комбинации на основании данных обучения. Алгоритм Microsoft Neural Network создает сеть, состоящую из минимум трех уровней узлов (иногда называемых нейронами). Этими слоями являются входной слой, скрытый слой и выходной слой.

Для наглядной оценки достоверности построенной модели воспользуемся графиком точности прогнозов. Модель также достаточно достоверна.

В работе приводятся выводы по всем методам анализа информации для поставленных задач.

ЗАКЛЮЧЕНИЕ

Несмотря на то, что стоимость аналитических систем даже сегодня остается достаточно высокой, а методологии и технологии реализации таких систем находятся ещё в стадии их становления, уже сегодня, экономический эффект, обеспечиваемый ими, существенно превышает эффект от традиционных оперативных систем.

Анализ современного состояния OLAP - технологий позволяет говорить о серьезных перспективах их развития. Многомерная обработка информации становится необходимым компонентом любого хранилища данных. В ходе работы были рассмотрены возможности и способы реализации OLAP - технологий.

В теоретическом блоке были рассмотрены и разобраны основные термины и понятия, используемых в OLAP - технологиях. Также для понимания всех возможностей структурных решений для OLAP - систем были рассмотрены различные архитектурные особенности их реализации. В итоге были выявлены области применения и возможности использования данных средств. Также были рассмотрены особенности технологий Data Mining.

В ходе работы был проведен анализ представленных на рынке средств для реализации OLAP. В то же время, широкое разнообразие подходов к реализации таких систем заставляет тщательно подходить к выбору и внедрению OLAP - систем. Внедрение таких систем на предприятие должны привести не только к обеспечению структурного OLAP - проектирования его информационной системы, но и дать возможность оптимизации, а также автоматизированного проектирования систем многомерной обработки данных, обеспечивающего надежную и оперативную работу хранилища данных.

Итогом исследования стало определения оптимального набора критериев для выбора средства реализации. К тому же представленные способы реализации и решения в рамках данной темы являются оптимальными (каждый для своей области), так как являются наиболее яркими представителями всех возможных способов реализации данных технологий.

В процессе изучения была приведена конкретная реализация технологий на практическом примере. В соответствии с этим был приведен пример

построения для информационной системы «Маркетинговая кампания банка» многомерной базы данных с использованием OLAP - технологий фирмы Microsoft (Analysis Services в Microsoft SQL Server 2012). После этого был проведён комплексный анализ представленной информации с использование основных методов Data Mining.

В итоге, в ходе рассмотрения были изучены все аспекты темы.