МИНОБРНАУКИ РОССИИ

Федеральное государственное бюджетное образовательное учреждение

высшего образования

«САРАТОВСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ

ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ Н.Г. ЧЕРНЫШЕВСКОГО»

Кафедра Математического и компьютерного моделирования

Применение технологий многомерного анализа к исследованию экономических данных

АВТОРЕФЕРАТ МАГИСТЕРСКОЙ РАБОТЫ

студентки 2 курса 247 группы

направление 09.04.03 – Прикладная информатика

механико-математического факультета

Задворной Ирины Александровны

Научный руководитель
профессор, д. э. н., профессор Л.В. Кальянов

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

Саратов 2019

Введение

Традиционно аналитики решают задачу по извлечению полезной информации из наборов данных. Но растущий объём информации в современных исследованиях требует новых подходов к решению задач анализа. По мере увеличения размеров и сложности наборов данных неизбежно происходит переход от прямого практического анализа данных к косвенному автоматическому с использованием усовершенствованных инструментов. Современные технологии сделали сбор и организацию данных выполнимой задачей. Однако полученные данные должны быть преобразованы в информацию и знания, чтобы стать полезными. Весь процесс обработки состоит в применении различных компьютерных технологий, включая новые методы обнаружения знаний из данных.

В настоящей работе рассматриваются современные методы анализа большого объёма необработанных данных для извлечения новой информации, полезной для процесса принятия решений. Основополагающие темы данной работы — это Data Mining и OLAP. OLAP - средства выступают в качестве технологий для аналитической обработки данных. Сегодня технологии OLAP и хранилища данных становятся все более и более развитыми. Важно выполнять обработку и анализ данных предприятия, обеспечивая тем самым лучшую, более быструю и эффективную поддержку. Таким образом, технология OLAP и хранилище данных постепенно становятся преобладающими средствами для анализа. Благодаря поддержке этих двух технологий, формирование информационной системы для анализа решает все поставленные задачи и способно давать ответы на самые разные вопросы аналитиков. При помощи средств OLAP можно произвести подготовку данных большого объёма, представленных в виде многомерной структуры, для анализа с помощью Data Mining для дальнейшего использования полученных знаний в принятии управленческих решений. Исходя из этого, в ходе работы было проведено исследование возможностей применения систем анализа, основанных на технологиях OLAP и Data Mining.

Благодаря разнообразию программных продуктов и различных технологических решений, под каждую конкретную задачу можно подобрать свой набор инструментов. В качестве примера реализации рассмотрено построе-

ние многомерной структуры данных для информационной системы «Финансовое благополучие» с использованием OLAP - технологии фирмы Microsoft (Analysis Services в Microsoft SQL Server 2012).

Целью работы является изучение принципов OLAP - технологий и Data Mining, их особенности и анализ данных о финансовом благополучии.

В ходе исследования будут решены следующие задачи:

- Разбор и определение основных терминов, используемых в OLAP и Data Mining;
- Выявление областей применения и возможностей использования OLAP и Data Mining;
- Проектирование и реализация многомерной базы данных с использованием OLAP - технологий;
- Проведение анализа данных о финансовом благополучии с использованием технологий Data Mining;
- Выявление возможностей использования полученных результатов анализа для дальнейших исследований.

Итоговой задачей исследования данной темы является создание полноценной системы для анализа данных о финансовом благополучии.

Научная новизна исследования заключается в применении комбинации алгоритмов интеллектуального анализа данных для повышения достоверности полученных результатов. В ходе работы создана информационная автоматизированная система для анализа данных.

Магистерская работа состоит из введения, двух разделов, заключения, списка используемых источников и четырёх приложений. В первом разделе «Интеллектуальный анализ данных и OLAP - технологии», состоящем из семи подразделов, даётся теоретическое обоснование используемых подходов, технологий и алгоритмов. Второй раздел «Разработка информационной системы «Финансовое благополучие» и проведение многомерного анализа данных» состоит из пяти подразделов и содержит создание информационной системы, процесс подготовки и исследования данных и непосредственно анализ данных о финансовом благополучии.

В ходе исследования были сделаны публикации в соавторстве по темам, соответствующим тематике магистерской работы. Также был представлен до-

клад на VII Международной молодёжной научно-практической конференции «Математическое и компьютерное моделирование в экономике, страховании и управлении рисками».

Основное содержание работы

В первом разделе задаются определения всем используемым средствам, описываются все используемые алгоритмы интеллектуального анализа данных и указываются особенности многомерного представления данных.

Вначале описывается появление определения Data Mining и его особенности. Определим Data Mining как процесс обнаружения различных взаимосвязей, сводок и выводов на основе построенной модели из набора данных. «Процесс» является ключевым в этом определении. Интеллектуальный анализ данных состоит не только в том, чтобы собирать и применять различные инструменты для решения поставленной задачи, но и в автоматизации и улучшении решения.

Анализ данных, содержащихся в многомерной структуре, можно представить в виде итеративного процесса. На первом шаге анализа изучают данные и исследуют их с использованием одного из методов. Далее возникает необходимость взглянуть на данные под другим углом. Для этого модель модифицируется. Затем процесс возвращается к началу, и применяется другой инструмент анализа данных, достигающий подтверждающих или любых других результатов. Такой итеративный процесс может происходить много раз. Каждый метод используется для исследования разных аспектов данных.

Далее описываются используемые алгоритмы интеллектуального анализа данных: деревья решений, кластеризация и нейронная сеть.

Деревья решений — это один из методов интеллектуального анализа данных, применяемый для решения проблемы классификации. Классификация представляет собой процесс обучения функции, которая отображает элемент данных в один из нескольких предопределённых классов. Цель обучения — создать классификационную модель (классификатор), которая будет предсказывать по значениям своих входных атрибутов класс для некоторой сущности. Другими словами, классификация — это процесс присвоения дискретной метки (класса) немаркированной записи.

Кластерный анализ представляет собой набор методологий автоматической классификации в несколько групп с использованием меры ассоциации таким образом, что выборки внутри одной группы аналогичны, а элементы, принадлежащие разным группам, не похожи. Вход для такой системы представляет собой набор выборок и меру сходства (или несходства) между ними. Результатом анализа кластеров является количество групп, которые образуют структуру набора данных. Также производится обобщённое описание каждого кластера, что особенно важно для более глубокого анализа характеристик. Методология кластеризации особенно подходит для исследования взаимосвязей для предварительной оценки структуры выборки.

Нейронная сеть представляет собой сетевую структуру, состоящую из нескольких узлов, соединённых через направленные каналы. Каждый узел представляет собой блок обработки, а связи между узлами определяют отношения. Все узлы адаптивны (выходы зависят от изменяемых параметров). Хотя существует несколько определений и несколько подходов к концепции построения нейронной сети, остановимся на одном из них. Искусственная нейронная сеть представляет собой массивный параллельный распределённый процесс, состоящий из простых подпроцессов. Он обладает способностью учиться на основе эмпирических знаний, выраженных через параметры межсетевого соединения, и может предоставлять такие знания для использования. Высокая вычислительная мощность алгоритма обеспечивается возможностью создания параллельной распределённой структуры и обучению модели.

В общем виде каждый процесс обучения состоит из двух основных этапов. Первый этап состоит из изучения и оценки неизвестных зависимостей в системе из заданного набора. Второй этап определяется использованием расчётных зависимостей для прогнозирования нового результата для будущих входных параметров системы.

В работе также задаётся определение хранилища данных. Многомерная структура данных составляет основу для ответа на все вопросы при анализе. Хранилище данных — это основа для управления предприятием и принятия решений в различных сферах деятельности. Создание хранилища данных не является обязательным условием для проведения интеллектуального анализа

данных, однако на практике такая задача, особенно для крупных компаний, решается легче благодаря их использованию. Основная задача хранилища данных — увеличить возможности процесса извлечения знаний из данных.

Data Mining представляет собой одно из основных приложений для хранилищ данных. В отличие от других инструментов запросов и прикладных систем процесс обработки данных предоставляет конечному пользователю возможность извлекать скрытую, нетривиальную информацию. Такая информация, более сложная для извлечения, может обеспечить большие коммерческие и научные преимущества и повысить доходность инвестиций, вложенных в построение хранилища данных.

Далее описываются возможности технологий OLAP. Инструменты и методы OLAP позволяют пользователям визуализировать и анализировать данные в хранилище данных, предоставляя несколько представлений данных, поддерживаемых расширенными графическими возможностями. В этих представлениях разные объёмы данных соответствуют различным бизнес - характеристикам. Инструменты OLAP позволяют легко просматривать размерные данные под любым углом или в виде срезов. Инструменты OLAP представляют собой часть процесса анализа данных, но не являются заменой Data Mining.

Основной определяющий все инструменты OLAP фактор — это аналитический механизм, который превращает корпоративные данные в многомерную информацию для анализа. Комплексная поддержка принятия решений и индивидуальные, простые в использовании приложения с ограниченной функциональностью могут быть построены с помощью инструментов OLAP. Технология предоставляет аналитическую среду для пользователя, которая позволяет ему использовать ряд функций для изучения доступной информации, в том числе для анализа с использованием алгоритмов Data Mining.

Описывается, что из себя представляет многомерное представление данных. Определяющая задача для OLAP - систем — построение многомерной структуры данных («куба»). Такой «куб» данных состоит из таблиц фактов и измерений. Многомерная модель хранилища данных может быть реализована в виде схем «Звезда», «Снежинка» или «Созвездие».

К таблицам измерений относятся те сущности, в отношении которых организация хочет вести учёт. Они содержат чаще всего данные, которые мало в последствии будут подвержены изменениям. Компонентами таблицы измерений являются несколько полей с названиями (именами члена измерения) и одно целочисленное для идентификации (ключ). Также они могут содержат дополнительные поля, содержащие расширенные характеристики рассматриваемого объекта. Связь между таблицами измерений и фактов — «один ко многим».

Многомерная модель данных обычно организована вокруг одной центральной темы. Основная сущность представляет собой таблицу фактов. Факты — это числовые меры. Можно представить их в виде количества, с помощью которого необходимо проанализировать отношения между измерениями. Таблица фактов содержит имена фактов или мер, ключевые поля для каждого измерения и несколько числовых полей для вычисления агрегатных функций.

Теперь рассмотрим подходы к реализации OLAP - структуры. Можно использовать три стратегии хранения данных, которые преодолевают ограничения реляционной модели для многомерного анализа:

- Использование специализированных многомерных баз данных, которые обеспечивают оптимизированное хранение и извлечение данных для запросов OLAP реализуется в виде MOLAP - структур (Multidimensional OLAP — исходные и агрегатные данные хранятся в многомерной базе данных).
- Использование хранилища данных, построенного с использованием реляционных технологий, но оптимизированного для поддержки принятия решений, а не транзакционных операций относится к ROLAP (Relational OLAP — исходные данные остаются в реляционной базе данных, где они изначально и находились).
- Сочетание этих подходов представляет собой HOLAP (Hybrid OLAP — только результат вычисления агрегатных функций хранится в многомерной БД).

В конце описываются основные преимущества использования описанных ранее средств.

Во втором разделе описывается создание многомерной базы данных, основанной на данных о финансовом благополучии и приводятся результаты проведённого анализа данных с использованием комбинации выбранных алгоритмов Data Mining.

Информационная система в данной работе имеет определённые требования к используемым программным средствам. Они включают возможности анализа информации, хранимой в многомерной структуре, и использования реляционных баз данных для преобразования в многомерную структуру, которая обеспечивает мощные возможности хранения данных, обработки транзакций и доступа с возможностями визуализации и манипулирования многомерными данными. Самым оптимальным решением для рассматриваемой далее практической задачи является продукт Microsoft SQL Server with Analysis Services.

Основа работы — создание многомерной базы данных, построенной на основе данных исследования финансового благополучия. Данные относятся к различным временным промежуткам и показателям.

Финансовое благополучие человека зависит от его чувства финансовой безопасности и свободы выбора — как в настоящем, так и при рассмотрении будущего. Набор данных для исследования взят из открытого источника данных Бюро финансовой защиты потребителей (CFPB). В исследовании, проводимым CFPB, используется шкала финансового благополучия из 10 пунктов. Набор данных опроса включает в себя оценки респондентов по этой шкале, а также показатели индивидуальных и общих характеристик, которые, как предполагают исследования, могут влиять на финансовое благополучие взрослых, включая следующие показатели:

- Доход и занятость;
- Сбережения и возможные пути для помощи;
- Прошлый финансовый опыт;
- Финансовое поведение, навыки и знания.

В наборе данных собрано более 5000 записей. Каждая запись соответствует опросу одного респондента и исследованию условий, окружающих его, которые могут поспособствовать выяснению реальной оценки финансового благополучия данного человека. Все данные в наборе обезличены. Каждому

респонденту соответствует свой идентификационный номер. Само основное исследование проводилось с 27 октября по 5 декабря 2016 года. Начиная с этого периода, было проведено более 120 отдельных опросов, рассматривающих респондентов по различным критериям. В 2017 и 2018 годах были проведены дополнительные исследования данных по участникам.

Далее в работе приводится описание процесса построения многомерной базы данных. Разработка многомерной модели для глубокого анализа данных составляет один из главных этапов разработки системы.

Многомерная база данных создаётся на основе предварительно созданной в SQL Server Managment Studio реляционной базе данных. В выстраиваемой системе будет функционировать многомерная БД, состоящая из одного многомерного куба. Так как в программной среде нет ограничений на количество многомерных объектов, таких как атрибуты, иерархии, измерения, кубы и группы мер, для реализации будем опираться на описание предметной области и построим схему, состоящую из 24 таблиц измерений и 1 таблицы фактов, в соответствии с рисунком 1.

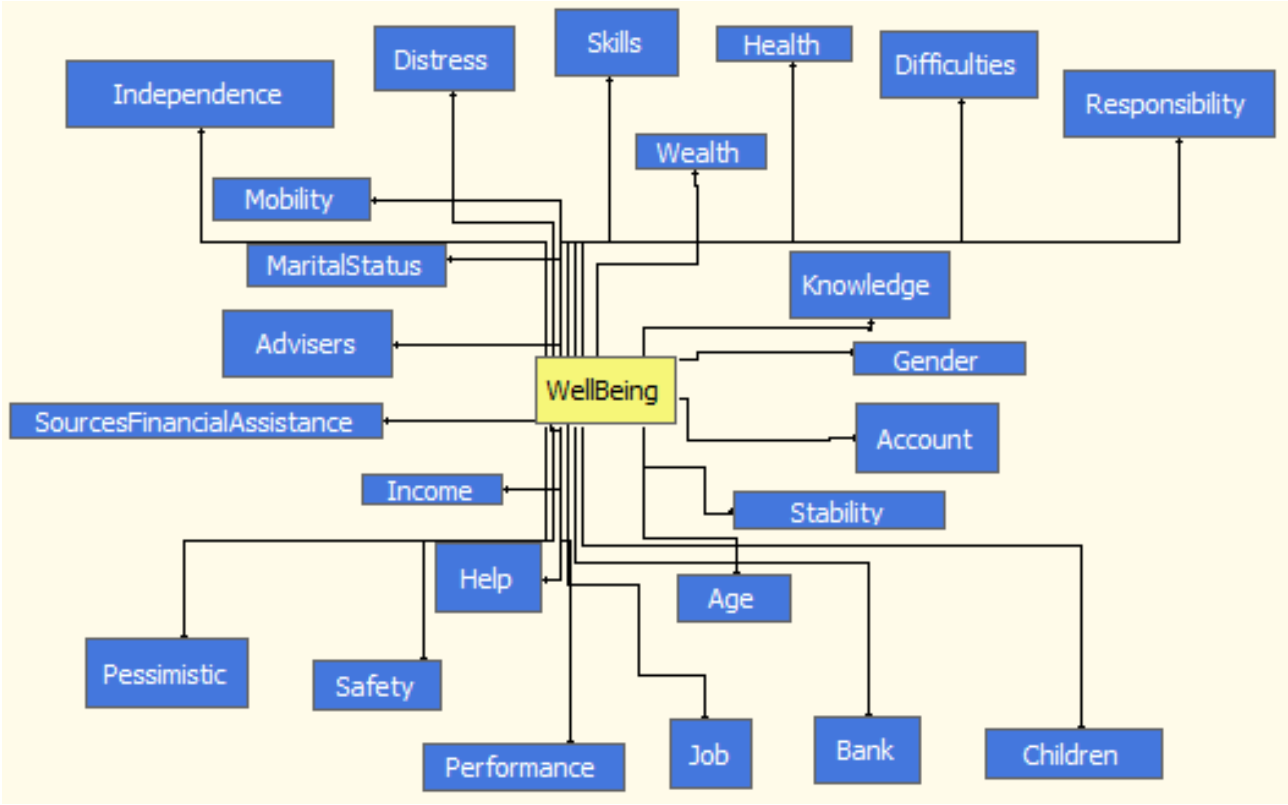

Рисунок 1 — Схема хранилища данных

Первоначально в работе проводится подготовка и исследование данных. Все данные в наборе обезличены. Отсутствуют прямые идентификаторы респондентов, которые были использованы в рамках исследования. Все косвенные идентификаторы были либо удалены, либо замаскированы путём объединения небольших категорий в более общие (например, возраст по возрастным группам). Файл данных в работе состоит из одного набора с разделенными запятыми значениями (файл формата CSV). В первоначальном файле с набором данных содержится 6394 строки с 217 параметрами.

Первоначально в наборе данных были удалены параметры, дающие дополнительные характеристики для исследователей, но не несущие смысловой нагрузки. К таким параметрам относятся коррелирующие с другими данными столбцы, финальные и созданные уже на основе опросов дополнительные нумераторы и вспомогательная информация о проведении самого тестирования. Среди параметров первоначального набора были удалены 45, после чего осталось 172 информативных параметра.

Набор данных был также очищен от строк, содержащих пустые данные, включая просто отсутствующие ответы респондентов, ошибки в ответах и противоречивые данные. После данной очистки из 6394 строк данных в наборе осталось 5025. Далее рассматриваются все полученные параметры системы и даётся их подробная характеристика.

Microsoft SQL Server позволяет создавать несколько моделей в одной структуре Data Mining, поэтому в рамках одного решения можно использовать различные алгоритмы, чтобы получить разные представления о данных. В рамках поставленной практической задачи рассматривается комбинирование трёх основных алгоритмов Data Mining (дерево решений, кластеризация, нейронная сеть).

Деревья решений используются в системе для анализа каждой отдельной группы показателей. Все модели деревьев решений представляют собой лучший результат работы алгоритма для каждого класса факторов, влияющих на финансовое благополучие человека. Во всех построенных для рассматриваемого набора данных деревьях решений в качестве выходного поля выбран основной показатель финансового благополучия. Всего построено 18 моделей деревьев решений.

Кластеризация используется для всестороннего анализа финансового положения респондентов. С помощью данного алгоритма проверяются все основные группы факторов на основе данных моделей деревьев решений. Проведение кластеризации на всём объёме выборки по всем параметрам позволяет увидеть основные особенности анализируемых данных. Алгоритм подразумевает разбиение на однородные группы, которые формируются на основе значений всех рассматриваемых параметров. Для проведения кластеризации были использованы все параметры моделей деревьев решений. Использовались весовые коэффициенты, используемые в предыдущих моделях. На основе 56 параметров, влияющих на результат в предыдущих моделях, сформировано 10 кластеров.

Заключительным этапом работы является создание нейронной сети для всей системы с использованием тех параметров ранее описанных моделей, которые оказывают существенное влияние на результат. Используются весовые коэффициенты моделей деревьев решений и кластеризации. Построенная нейронная сеть состоит из 273 входных нейронов, соответствующих всем возможным значениям 56 параметров модели. В качестве входных используются все параметры, получившие наибольшие весовые коэффициенты в предыдущих алгоритмах. В модели также используется один скрытый слой, состоящий из 56 нейронов. На выходном слое находится один выходной узел (показатель финансового благополучия).

Благодаря высокой скорости получения результатов работы моделей, терпимости к сложности данных алгоритмов интеллектуального анализа данных, проработанной многомерной модели данных, а также подходу к использованию комбинации нескольких моделей для улучшения результатов анализа удалось разработать систему, основанную на изучаемых данных, которая способна на основе тестирования респондента предсказать степень его финансового благополучия.

Также были выявлены основные категории, склонные к наиболее высоким и низким показателям финансового благополучия. И был составлен перечень наиболее перспективных параметров для проведения подобных исследований в других регионах, которые наиболее полно проиллюстрируют респондентов с минимальными затратами времени на проведения тестирования. Результа-

ты работы могут быть использованы как для изучения вопросов финансового благополучия, так и для составления анкетирования для проведения независимого исследования.

Заключение

Современное положение развития OLAP - систем подразумевает их перспективное развитие в дальнейшем. Многомерная обработка информации — важная часть систем, работающих с большим объёмом данных. Построение систем для анализа данных, основанных на технологиях OLAP и хранилищах данных, используемых в этом исследовании, получил расширенное применение и достиг хороших результатов.

В ходе работы были рассмотрены возможности и способы реализации OLAP - технологий и анализа данных с использованием Data Mining. Были рассмотрены и разобраны основные термины и понятия. Также для понимания всех возможностей структурных решений для OLAP - систем были описаны различные архитектурные особенности их реализации. Особенности технологий Data Mining были разобраны на основе комбинации трёх алгоритмов: дерево решений, кластеризация и нейронная сеть. В итоге были выявлены основные преимущества использования данных средств.

Была приведена конкретная реализация технологий на практическом примере. В соответствии с этим был проведено построение для информационной системы «Финансовое благополучие» 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 использованием выбранного подхода к комбинированному использованию алгоритмов Data Mining.

Результаты анализа, проведённого в рамках данной работы, могут стать основой для изучения вопросов финансового благополучия, для проведения дополнительных исследований с использованием выведенных параметров и для улучшения финансового положения населения с помощью развития определённых сфер жизни и решения основополагающих вопросов, выявленных в рамках исследования.

В итоге, в ходе рассмотрения были изучены все аспекты темы и выполнены все поставленные задачи.