МИНОБРНАУКИ РОССИИ
Федеральное государственное бюджетное образовательное учреждение
высшего образования
**«САРАТОВСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
ИМЕНИ Н. Г. ЧЕРНЫШЕВСКОГО»**
Кафедра математической кибернетики и компьютерных наук

РЕАЛИЗАЦИЯ ПРИЛОЖЕНИЯ ПОДДЕРЖКИ ПРИНЯТИЯ РЕШЕНИЙ НА ОСНОВЕ ИНТЕЛЛЕКТУАЛЬНОГО АНАЛИЗА ДАННЫХ

АВТОРЕФЕРАТ БАКАЛАВРСКОЙ РАБОТЫ

Студентки 4 курса 411 группы
направления 02.03.02 — Фундаментальная информатика и информационные технологии
факультета КНиИТ
Муравьевой Оксаны Петровны

Научный руководитель
старший преподаватель _____ М. И. Сафрончик

Заведующий кафедрой
к. ф.-м. н., доцент _____ С. В. Миронов

Саратов 2021

СОДЕРЖАНИЕ

ВВЕДЕНИЕ

В связи с совершенствованием технологий записи и хранения данных потоки информации становятся все больше и больше с каждым годом. Деятельность любого предприятия сопровождается огромным количеством хранимых данных. Согласно оценкам экспертов, информация удваивается каждые 2-3 года.

Из-за огромного количества информации очень малая ее часть будет когда-либо увидена людьми. Понять и найти что-то полезное в этом большом количестве информации поможет широкое применение методов Data Mining.

Технология Data Mining изучает процесс нахождения новых, нетривиальных и потенциально полезных знаний в базах данных. Data Mining основан на нескольких науках, главными из которых являются искусственный интеллект, системы баз данных и статистика.

Data Mining широко применяется во многих сферах жизнедеятельности: в науке — астрономии, биологии, медицине, физике и других областях; в бизнесе — торговле, телекоммуникациях, банковском деле, промышленном производстве, а так же и в других сферах.

Целью бакалаврской работы является создание приложения поддержки принятия решений на основе интеллектуального анализа данных.

В ходе работы необходимо решить следующие задачи:

- Спроектировать и реализовать единое хранилище данных в Microsoft SQL Server.
- Реализовать ETL процесс, выполнив очистку и загрузку данных в хранилище данных из открытых источников с помощью Microsoft SQL Server Integration Services.
- Создать прогнозную модель для решения задач интеллектуального анализа данных средствами Microsoft Analysis Services.

1 Механизмы интеллектуального анализа данных

Хранилище данных (англ. Data Warehouse — DW) — это большое количество данных, используемых для помощи какой-либо организации в принятии управленческих решений. В хранилище данных поступает большой объем различных данных из различных мест, таких как маркетинг, продажи и финансы, приложения, ориентированные на клиентов и других [1].

Информация перед загрузкой в хранилище данных подвергается обработке: данные очищаются и добавляются новые атрибуты. Исходные данные из оперативных источников данных объединяются с информацией из внешних источников [2].

Извлечение, преобразование и загрузка называются ETL-процессами — это основные этапы переноса информации из одного приложения в другое. Приложения ETL извлекают информацию из исходной базы данных, преобразуют ее в формат, поддерживаемый базой данных назначения, а затем загружают в нее преобразованную информацию [3].

Data Mining — это процесс обнаружения закономерностей и знаний из больших объемов данных. Источниками данных могут быть базы данных, хранилища данных, Интернет и многое другое [4].

В интеллектуальном анализе данных применяется математический анализ для выявления закономерностей и тенденций, существующих в данных [5].

Различают следующие задачи Data Mining:

- Классификация — обнаружение различных признаков, которые характеризуют группы объектов в наборе данных, по этим признакам объект относится к тому или иному классу.
- Ассоциация — обнаружение закономерности между связанными событиями в наборе данных.
- Кластеризация — группировка элементов, которые кажутся естественными вместе. Задача состоит в том, чтобы найти эти кластеры и назначить им экземпляры, а также иметь возможность назначать новые экземпляры кластерам.
- Прогнозирование — оценка пропущенных или же будущих значений целевых численных показателей на основе особенностей исторических данных.
- Анализ связей — нахождение зависимостей в наборе данных [6] [7].

Задачи по назначению делятся на описательные и предсказательные. Описательные задачи — уделяют внимание улучшению понимания анализируемых данных. Решение предсказательных задач разбивается на два этапа. На первом этапе на основании набора данных с известными результатами строится модель. На втором этапе она используется для предсказания результатов на основании новых наборов данных.

По способам решения задачи разделяют на «обучение с учителем» и «обучение без учителя» [2].

1.1 Методы интеллектуального анализа данных

К базовым методам Data Mining относятся алгоритмы, основанные на переборе. При переборе с увеличением количества данных объем вычислений растет экспоненциально. При большом объеме это делает решение любой задачи почти невозможным, поэтому для уменьшения сложности в таких алгоритмах используют различного вида эвристики, приводящие к сокращению перебора.

Также, к базовым методам Data Mining можно отнести подходы, использующие элементы теории статистики. Основная идея таких методов заключается в корреляционном, регрессионном и других видах статистического анализа [2].

1.2 Основные этапы интеллектуального анализа данных

Создание аналитической модели данных является динамическим и повторяющимся процессом. Построение модели интеллектуального анализа данных можно представить как последовательность шести шагов.

1.2.1 Постановка задачи

Первый шаг процесса интеллектуального анализа данных — определение проблемы и рассмотрение способов использования данных для решения этой проблемы. Данный шаг включает анализ бизнес-требований, определение области проблемы, метрик, по которым будет выполняться оценка модели, а также определение задач для проекта интеллектуального анализа данных [5].

1.2.2 Подготовка данных

Второй шаг процесса интеллектуального анализа данных — объединение и очистка данных, определенных во время постановки задачи. Очистка

данных — это удаление недопустимых данных, интерполяция отсутствующих значений, поиск в данных скрытых зависимостей, определение источников точных данных и подбор столбцов, которые наиболее подходят для использования в анализе [5].

1.2.3 Изучение данных и построение моделей

Третий шаг процесса интеллектуального анализа данных — просмотр данных. Для принятия правильных управленческих решений при создании моделей интеллектуального анализа данных необходимо понимать данные. Методы исследования данных включают в себя расчет минимальных и максимальных значений, вычисление средневероятного и стандартного отклонения и изучение распределения данных. Стандартное отклонение и другие характеристики распределения могут сообщить полезные сведения о стабильности и точности результатов.

Четвертый шаг процесса интеллектуального анализа данных — построение моделей интеллектуального анализа данных. Знания, полученные в процессе просмотра данных, могут помочь определить и создать модели [5].

1.2.4 Исследование и проверка моделей

Пятый шаг процесса интеллектуального анализа данных — исследование построенных моделей интеллектуального анализа данных и проверка их эффективности. Перед развертыванием модели нужно проверить эффективность работы данной модели. Во время построения модели создается несколько моделей с различной конфигурацией, а затем эти модели проверяются, чтобы определить, какая из них обеспечивает лучшие результаты для поставленной задачи и имеющихся данных [5].

1.2.5 Развертывание и обновление моделей

Шестой шаг процесса интеллектуального анализа данных — развертывание моделей в рабочей среде. Развертывание важно, так как оно делает модели доступными для пользователей.

1.3 Алгоритмы

Для создания модели алгоритм анализирует полученные данные и осуществляет поиск каких-либо закономерностей и тенденций в данных. Затем

алгоритм применяет результаты этого анализа ко множеству итераций, чтобы подобрать оптимальные параметры для создания модели интеллектуального анализа данных. Эти параметры применяются ко всему набору данных, чтобы выявить полезные закономерности.

1.3.1 Алгоритм дерева принятия решений

Алгоритм дерева принятия решений используется для построения классификационных моделей. Он строит классификационные модели в виде древовидной структуры. Алгоритм дерева принятия решений используется в прогнозном моделировании дискретных и непрерывных атрибутов [8].

Алгоритм дерева принятия решений строит модель интеллектуального анализа данных путем создания разбиений в дереве, эти разбиения представляются узлами дерева. Алгоритм добавляет узел к модели каждый раз, когда входной столбец имеет значительную корреляцию с прогнозируемым столбцом [8].

1.3.2 Алгоритм временных рядов

Алгоритм временных рядов в Microsoft представляет несколько алгоритмов, оптимизированных для прогноза непрерывных значений:

- ARIMA — использует взвешенные предыдущие значения, в то время как скользящая средняя часть взвешивает ранее принятые ошибки временного ряда.
- Экспоненциальное сглаживание — состоит из базового уровня в определенный момент времени, тенденции и сезонной составляющей.
- Декомпозиция сезонного тренда — приспосабливает различные функции сезонного тренда к заданным данным и выбирает наилучшую функцию сезонного тренда в соответствии с мерой ошибки [9].

1.3.3 Алгоритм кластеризации

Алгоритм кластеризации выполняет итерацию вариантов в наборе данных, чтобы сгруппировать их в группы, содержащие подобные характеристики. Такие группировки полезны для просмотра данных, выявления в них различных аномалий и создания прогнозов. Модели кластеризации определяют связи в наборе данных, которые невозможно логически получить с помощью наблюдения [10].

2 Программные средства для реализации приложения интеллектуального анализа данных

Microsoft SQL Server — это система управления реляционными базами данных, разработанная корпорацией Microsoft. Система позволяет решать важные задачи внутри предприятия [11].

Microsoft SQL Server Management Studio — это интегрированная среда для управления какой-либо инфраструктурой SQL. SQL Server Management Studio предоставляет единую программу, которая содержит в себе множество графических инструментов для доступа к службе Microsoft SQL Server для разработчиков и администраторов [12].

Microsoft SQL Server Integration Services — это службы для построения решений по интеграции и преобразованию данных на уровне предприятия. Службы Integration Services можно использовать для решения различных задач путем копирования и загрузки файлов, загрузки хранилищ данных, очистки и интеллектуального анализа данных, а также управления объектами и данными [13].

Microsoft Analysis Services — это средство аналитических данных, используемое в службе поддержки принятия решений и бизнес-аналитики, предоставляющее возможности моделирования данных корпоративного уровня для Bussines Intelligence, анализа данных и создания отчетов [14].

Microsoft Visual Studio — это платформа для написания, отладки и сборки кода. Это интегрированная среда разработки, которая представляет собой многофункциональную программу [15].

DMX — это язык, который можно использовать для создания моделей интеллектуального анализа данных и работы с ними в Microsoft SQL Server. Расширения интеллектуального анализа данных могут использоваться для создания структуры новых моделей интеллектуального анализа данных, обучения этих моделей, а также для осуществления обзора, управления и прогнозирования по этим моделям [16].

3 Реализация приложения интеллектуального анализа данных

3.1 Постановка задач

В работе создается приложение интеллектуального анализа данных, позволяющее на основе больших объемов данных выявлять общие закономерности и делать прогнозы. В качестве источника данных для приложения были взяты данные по победителям школьных олимпиад на портале открытых данных правительства Москвы.

В рамках реализации данного приложения необходимо спроектировать и реализовать единое хранилище данных, выполнить очистку и загрузку данных в него из открытых источников с помощью ETL процесса и создать прогнозную модель интеллектуального анализа данных, позволяющую решать ряд следующих задач:

- Выявить предметную направленность различных общеобразовательных учреждений, основываясь на результатах прошедших олимпиад.
- Проанализировав результаты олимпиад, составить рейтинг общеобразовательных учреждений.
- Проанализировав направленность различных общеобразовательных учреждений по различным предметам по результатам прошедших олимпиад дать прогноз на будущий результат.
- Проанализировав количество проведенных олимпиад в различные месяца выявить сезонные проявления и дать прогноз на ближайший год.
- Выявить общие сходства участия в олимпиадах различных общеобразовательных учреждений и сформировать группы.

3.2 Подготовка данных

В SQL Server была создана база данных «OlympiadData». После чего данные по победителям олимпиад были загружены в базу. Создано хранилище данных «OlympiadDW». Был создан проект служб Integration Services «OWSSIS». Созданы соединения с базой «OlympiadData» и «OlympiadDW».

С помощью различных инструментов: источников и назначений данных, уточняющих запросов, команд к DB, сортировок, статистических запросов, созданы схемы загрузки данных в хранилище.

После этого проект был запущен и данные загружены в хранилище данных.

3.3 Анализ данных

В среде Microsoft Visual Studio 2017 был создан проект многомерных данных и интеллектуального анализа служб Analysis Services. Добавлен источник данных — хранилище данных «OlympiadDW». Добавлено представление источника данных — таблицы хранилища данных «OlympiadDW». Создана структура интеллектуального анализа данных на основе хранилища данных «OlympiadDW». В качестве метода выбран алгоритм дерева принятия решений. В качестве таблицы вариантов выбрана таблица «Statistic». В качестве ключа выбран столбец «GlobalID», в качестве входных данных выбраны столбцы «Number Of Winners In The School By Subject», «School» и «Status». В качестве прогнозируемого столбца выбран столбец «Subject». Развернуто решение на сервере и произошла обработка структуры и модели интеллектуального анализа данных.

На первых уровнях алгоритм производит разбиения на несколько групп по количеству победителей по различным предметам в различных школах. А затем происходят разбиения по школам. Таким образом, например, школа с номером 159843 имеет большую направленность в изучении литературы по результатам прошедших олимпиад. С помощью конструктора прогнозирования написан DMX-запрос с использованием прогнозирующей функции «Predict», выдающий для различных школ предметы, по котором школа выигрывала олимпиады большее количество раз по сравнению с другими предметами. Таким образом выявляется предметная направленность школы.

Произведена проверка результатов. Проверка представляет собой процесс оценки соответствия моделей интеллектуального анализа данных фактическим данным. Ранее, в мастере создания структуры интеллектуального анализа данных, 30 процентов данных выделялись под обучение.

Для решения следующей задачи — определения рейтинга школ по результатам прошедших олимпиад в различные годы используется так же алгоритм дерева принятия решений. Модель интеллектуального анализа данных была скорректирована. В качестве ключа используется столбец «GlobalID», в качестве входных данных используется столбец «Number Of Winners In The School By Subject In Year», а в качестве прогнозируемого «School». Развернуто решение на сервере и произошла обработка структуры и модели интеллектуального анализа данных.

Для составления рейтинга школ с помощью конструктора прогнозов, используя прогнозирующую функцию «PredictHistogram» был написан запрос, для каждого года выдающий рейтинг школ на основе анализа количества победителей в различных школах в различных годах.

Для решения следующей задачи — анализа направленности различных школ по различным предметам по результатам прошедших олимпиад и прогноза на будущее используется алгоритм дерева принятия решений. Модель интеллектуального анализа данных была скорректирована. В качестве ключа используется столбец «GlobalID», в качестве входных данных используется столбец «Year», «Subject» и «School», а в качестве прогнозируемого «Number Of Winners In The School By Subject». Развернуто решение на сервере и произошла обработка структуры и модели интеллектуального анализа данных.

Для анализа направленности различных школ по различным предметам по результатам прошедших олимпиад и прогноза на будущее с помощью конструктора прогнозов был написан запрос, выдающий вероятность различных школ занимать победные места по различным предметам в будущем.

Для решения следующей задачи — анализа количества проведенных олимпиад в различные месяца, выявления сезонных проявлений и прогноза на ближайшие месяцы, используется алгоритм временных рядов. Создана структура интеллектуального анализа данных на основе хранилища данных «OlympiadDW». В качестве метода выбран алгоритм временных рядов. В качестве таблицы вариантов выбрана таблица «Statistic». В качестве ключа выбран столбец «Time». В качестве прогнозируемого столбца выбран столбец «Number Of Olympiad In Month». Развернуто решение на сервере и произошла обработка структуры и модели интеллектуального анализа данных.

Диаграмма отражает месяца различных годов по горизонтали и количество проведенных олимпиад в этот месяц по вертикали. Количество олимпиад резко возрастает в определенные месяца, а затем снова снижается. Таким образом, олимпиады проводятся в течение года неравномерно, что может значительно увеличивать нагрузку в определенные месяца на учащихся. Большинство олимпиад в учебном году проводятся в ноябре каждый год.

Для анализа количества проведенных олимпиад в различные месяца, выявления сезонных проявлений и прогноза на будущее с помощью конструктора прогнозов был написан запрос, выдающий прогноз количества олимпиад

на ближайшие месяцы. Так как последний месяц за который есть данные — это декабрь 2019, следовательно прогноз начинается с января 2020. Так как в прошлые годы в начале года проводилось около 1-3 олимпиад, то прогноз на следующий год отражает примерно такие же показатели.

Для решения следующей задачи — выявления общих сходств участия в олимпиадах различных общеобразовательных учреждений и формирования групп используется алгоритм кластеризации. Создана структура интеллектуального анализа данных на основе хранилища данных «OlympiadDW». В качестве метода выбран алгоритм кластеризации. В качестве таблицы вариантов выбрана таблица «Statistic». В качестве ключа выбран столбец «GlobalID». В качестве выходных данных выбраны столбцы «Class», «Class», «Number Of Winners Per Year At School», «Status», «School» и «Subject». Развернуто решение на сервере и произошла обработка структуры и модели интеллектуального анализа данных.

Диаграмма разбила школы на 10 групп, каждая из которых имеет свои характерные признаки. С помощью профилей кластеров можно определить эти признаки. В большинстве групп в олимпиадах занимают призовые места учащиеся девятых и одиннадцатых классов, а в кластерах 9 и 10 — учащиеся пятых и шестых классов. По диаграмме так же можно выявить наиболее «сильные» школы. «Кластер 4» содержит в себе школы с высоким количеством победителей различных олимпиад.

ЗАКЛЮЧЕНИЕ

Интеллектуальный анализ данных является одним из наиболее актуальных и востребованных сегодня направлений. Технологии Data Mining — это сильнейший аппарат современного бизнес-анализа и исследования данных для обнаружения скрытых закономерностей и построения предсказательных моделей.

В ходе бакалаврской работы были рассмотрены возможности интеллектуального анализа данных и создано приложение поддержки принятия решений на основе интеллектуального анализа данных.

Подробно были рассмотрены BI-системы и их компоненты, основные этапы анализа данных, а так же алгоритмы интеллектуального анализа данных.

В результате бакалаврской работы было спроектировано и реализовано хранилище данных, с помощью ETL процессов очищены и загружены данные из открытых источников в хранилище данных, проведен анализ данных с помощью различных алгоритмов, а так же проведено тестирование моделей.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1 What is a Data Warehouse and Why Does It Matter To Your Business? [Электронный ресурс]. — URL: https://www.talend.com/resources/what-is-data-warehouse (Дата обращения 01.05.2021). Загл. с экр. Яз. рус.

2 *Барсегян, А. А.* Анализ данных и процессов / А. А. Барсегян, М. С. Куприянов, И. И. Холод, М. Д. Тесс, С. И. Елизаров. — Санкт-Петербург: БХВ-Петербург, 2009.

3 Порядок разработки ETL-процессов [Электронный ресурс]. — URL: http://www.olap.ru/basic/etl.asp (Дата обращения 08.05.2021). Загл. с экр. Яз. рус.

4 *Han, Jiawei*. Data Mining: Concepts and Techniques / Jiawei Han, Micheline Kamber, Jian Pei. — USA: Elsevier, 2012.

5 Основные понятия интеллектуального анализа данных [Электронный ресурс]. — URL: https://docs.microsoft.com/ru-ru/analysis-services/data-mining/data-mining-concepts?view=asallproducts-allversions (Дата обращения 06.05.2021). Загл. с экр. Яз. рус.

6 Что такое Data Mining? [Электронный ресурс]. — URL: https://intuit.ru/studies/courses/6/6/lecture/158 (Дата обращения 01.05.2021). Загл. с экр. Яз. рус.

7 *Witten, Ian H.* Data Mining. Practical Machine Learning. Tools and Techniques / Ian H. Witten, Eibe Frank, Mark A. Hall. — USA: Elsevier, 2011.

8 Microsoft Naive Bayes Algorithm [Электронный ресурс]. — URL: https://docs.microsoft.com/ru-ru/analysis-services/data-mining/microsoft-naive-bayes-algorithm?view=asallproducts-allversions (Дата обращения 17.05.2021). Загл. с экр. Яз. рус.

9 Алгоритм временных рядов (Майкрософт) [Электронный ресурс]. — URL: https://docs.microsoft.com/ru-ru/analysis-services/data-mining/microsoft-time-series-algorithm?view=asallproducts-allversions (Дата обращения 17.05.2021). Загл. с экр. Яз. рус.

10 Алгоритм кластеризации (Майкрософт) [Электронный ресурс]. — URL: https://docs.microsoft.com/ru-ru/analysis-services/data-mining/microsoft-clustering-algorithm?view=asallproducts-allversions (Дата обращения 17.05.2021). Загл. с экр. Яз. рус.

11 Microsoft SQL Server [Электронный ресурс]. — URL: https://navicongroup.ru/platforms/4025/ (Дата обращения 26.05.2021). Загл. с экр. Яз. рус.

12 Что такое SQL Server Management Studio (SSMS)? [Электронный ресурс]. — URL: https://docs.microsoft.com/ru-ru/sql/ssms/sql-server-management-studio-ssms?view=sql-server-ver15 (Дата обращения 26.05.2021). Загл. с экр. Яз. рус.

13 Основные сведения об Analysis Services [Электронный ресурс]. — URL: https://docs.microsoft.com/ru-ru/analysis-services/analysis-services-overview?view=asallproducts-allversions (Дата обращения 26.05.2021). Загл. с экр. Яз. рус.

14 Основные сведения об Analysis Services [Электронный ресурс]. — URL: https://docs.microsoft.com/ru-ru/analysis-services/analysis-services-overview?view=asallproducts-allversions (Дата обращения 26.05.2021). Загл. с экр. Яз. рус.

15 Добро пожаловать в интегрированную среду разработки Visual Studio [Электронный ресурс]. — URL: https://docs.microsoft.com/ru-ru/visualstudio/get-started/visual-studio-ide?view=vs-2019 (Дата обращения 26.05.2021). Загл. с экр. Яз. рус.

16 Справочник по расширениям интеллектуального анализа данных [Электронный ресурс]. — URL: https://docs.microsoft.com/ru-ru/sql/dmx/data-mining-extensions-dmx-reference?view=sql-server-ver15 (Дата обращения 26.05.2021). Загл. с экр. Яз. рус.