УДК 656.13.05

А. Б. ХАЛАПСИН, магистрант Академии управления при Президенте Республики Беларусь, г. Минск; С. А. ДАНИЛЮК, инженер-программист ООО «Мотовелозавод», г. Минск

ИНФОРМАЦИОННАЯ СИСТЕМА ИНТЕЛЛЕКТУАЛЬНОГО АНАЛИЗА БОЛЬШИХ МАССИВОВ ДАННЫХ О ПРАВОНАРУШЕНИЯХ В РЕСПУБЛИКЕ БЕЛАРУСЬ

Современные цифровые технологии затронули каждого, стали неотъемлемой частью нашей жизни и оказывают влияние абсолютно на все сферы человеческой деятельности. Одним из видов социальной деятельности, отличающимся исключительной сложностью, является процесс влияния науки на обеспечение безопасности дорожного движения, связанный с получением и обработкой большого количества разнообразной информации. В последние годы объемы информации увеличились настолько, что обрабатывать их с помощью традиционных программ или аппаратных средств стало весьма затруднительно или даже невозможно. Информация о геолокации, сигналы от датчиков «интернета вещей», информация о транспортных средствах, сведения о пешеходах – всё это генерирует огромные объемы неструктурированной информации, которая может быть использована в конструктивных целях. Поэтому в качестве важнейшего технологического тренда, способного в перспективе кардинально изменить процесс поиска, анализа и использования значимой информации в обеспечении безопасности дорожного движения, можно рассматривать технологию «больших данных».

Введение. Разработка информационных систем, использующих интеллектуальный анализ для предотвращения возникновения условий совершения дорожно-транспортных правонарушений, имеет большое значение. Несмотря на наличие подобной информационной системы на республиканском уровне, было бы разумно разработать систему, позволяющую проводить такой анализ данных на местах.

Ввиду постоянного развития современных технологий и появления новых было бы разумно применить новые способы хранения и анализа данных при разработке, что существенно повысит эффективность системы.

Для более точного анализа и принятия обоснованных решений можно использовать хранилища данных (Data Warehouse). Хранилище данных представляет собой центральный репозиторий информации из разных источников и обеспечивает пользователя возможностью принимать верные решения на основе целостной информационной картины.

Постановка задачи. Существует необходимость создания информационной системы, способной работать не только на республиканском уровне, но и на местах по следующим причинам:

1) невозможность проведения интеллектуального анализа на местах;

2) ожидание результатов анализа от органов республиканского уровня и, как следствие, невозможность быстрого реагирования на ситуацию.

Такую информационную систему необходимо снабдить хранилищем данных с возможностью отправки данных в центральное хранилище данных республиканского уровня. Информационная система должна поддерживать проведение предварительной обработки данных и следующих видов интеллектуального анализа: разведочный и кластерный. Интерфейс информационной системы должен быть удобным и понятным. Система должна быть гибкой и легко масштабируемой, а технологии, используемые при создании, должны поддерживать кроссплатформенность системы.

Создание такой системы позволит выявлять причины правонарушений на дорогах Гомельской области и реагировать на них существенно быстрее, чем при использовании информационной системы республиканского уровня.

Основная часть. Основываясь на международной практике работы с большими данными в сфере дорожной безопасности, для более точного анализа и принятия обоснованных решений можно использовать хранилища данных (Data Warehouse). Согласно определению компании Oracle хранилище представляет собой разновидность системы управления данными, которая обеспечивает поддержку бизнес-аналитики. Хранилища данных предназначены только для выполнения запросов и анализа и обычно содержат большие объемы исторических данных. Хранилище объединяет данные из разных источников. С данными проводят следующие манипуляции: преобразование, унификация, обработка на наличие ошибок и приведение к определенной структуре. Такие манипуляции облегчают дальнейшую работу с данными. Главное отличие хранилища данных от транзакционной системы – сохранение историчности данных. Это имеет большое значение при анализе [1].

Для успешного функционирования информационной системы необходимо правильно подойти к подходу выбора технологий и инструментов.

Интегрированная среда разработки (IDE) – это инструмент, используемый для разработки приложений простым, быстрым и надежным способом.

В качестве IDE был выбран PyCharm от компании JetBrains, который имеет полный комплект средств, необходимых для эффективного программирования на Python.

PyCharm предлагает большой набор инструментов из коробки: встроенный отладчик и инструмент запуска тестов, профилировщик Python, полнофункциональный встроенный терминал, инструменты для работы с базами данных. IDE интегрирована с популярными системами контроля версий, содержит встроенный SSH-терминал, поддерживает возможности удаленной разработки и удаленные интерпретаторы, а также интеграцию с Docker и Vagrant.

Поддерживает работу со всеми популярными базами данных: Oracle, SQL Server, PostgreSQL, MySQL. PyCharm помогает редактировать SQL-код, выполнять запросы, просматривать данные и изменять схемы.

Данная IDE также предоставляет широкие возможности для отладки, тестирования и профилирования.

Основным языком программирования был выбран Python. Python – универсальный современный язык программирования высокого уровня, к преимуществам которого относят высокую производительность программных решений и структурированный, хорошо читаемый код. Имеет широкий перечень встроенных библиотек, позволяет применять внушительный набор полезных функций и возможностей [2].

Основными областями применения Python являются:

1) веб-разработка;

2) машинное обучение;

3) автоматизация процессов.

Также выбранный язык программирования предоставляет широкие возможности для работы с базами данных. В рабочей среде языка находится программный интерфейс, который позволяет пользоваться базами прямо из сценария с помощью запросов SQL. Также код, написанный на Python, может с минимальными доработками использоваться для баз данных PostgreSQL, MySQL и Oracle.

Помимо этого, Python широко используется для работы со сложными вычислительными процессами и позволяет удобно визуализировать полученные данные.

В качестве СУБД была выбрана PostgreSQL. Данная СУБД является передовая база данных с открытым исходным кодом, и может служить простой реляционной базой данных, базой данных временных рядов и выступать в роли эффективного недорого хранилища данных. Поддерживает интеграцию с различными аналитическими инструментами [3].

Для построения хранилища данных были написаны скриптовые файлы `create_schema.py` и `sql_queries.py` на языке программирования Python.

Главная задача файла `create_schema.py`– создание базы данных и таблиц, в которые в дальнейшем будут загружены обработанные данные. Для установления соединения с PostgreSQL был использован драйвер psycopg, который в качестве модуля psycopg2 импортируется в файл Python. `sql_queries.py` содержит SQL-запросы создания необходимых таблиц, «обернутые» в интерфейс языка Python.

Приведем пример нескольких функций скриптового файла `create_schema.py`:

Функция `create_db` при помощи модуля psycopg2 настраивает подключение к серверу PostgreSQL, а также принимает в качестве аргументов информацию, необходимую для подключения: хост и порт, имя пользователя, предполагаемое имя базы данных. Далее создается база данных с заданным пользователем именем.

```
def create_db(db_credential_info):
    db_host, db_user, db_password, db_name = db_credential_info
    print('Creating new database.')
    conn = psycopg2.connect(host=db_host, database='postgres', user=db_user, password=db_password)

conn.set_isolation_level(ISOLATION_LEVEL_AUTOCOMMIT)
    cur = conn.cursor()
    try:
        cur.execute("DROP DATABASE IF EXISTS %s ;" % db_name)
        cur.execute("CREATE DATABASE %s ;" % db_name)
    except psycopg2.errors.lookup("55006"):
        locked = True
    cur.close()
```

Следующая функция при помощи модуля psycopg2 настраивает подключение к серверу PostgreSQL, принимает в качестве аргументов информацию, необходимую для подключения: хост и порт, имя пользователя, предполагаемое имя базы данных. Далее запускаются написанные в файле `sql_queries.py` SQL-запросы создания таблиц, обернутые в интерфейс Python.

```
def create_tables(db_credential_info):
    db_host, db_user, db_password, db_name = db_credential_info
    conn = None
    try:
        for command in create_table_queries:
            print('Building tables.')
            conn = psycopg2.connect(host=db_host, database=db_name, user=db_user, password=db_password)
            cur = conn.cursor()
            cur.execute(command)
            conn.commit()
            cur.close()
    except (Exception, psycopg2.DatabaseError) as error:
        print(error)
        cur.close()
    finally:
        if conn:
            conn.close()
```

После запуска данных скриптов в PostgreSQL будет построено хранилище данных по схеме «звезда», включающее в себя следующие таблицы:

- dim_member – таблица измерений, содержащая информацию об участнике ДТП;
- dim_viol – таблица измерений, содержащая информацию, характеризующую конкретное ДТП;
- dim_place – таблица измерений, характеризующая место ДТП;
- dim_videos – таблица измерений, содержащая видеоматериал;
- dim_img – таблица измерений, содержащая изображения;
- fact_protocol – таблица фактов, содержащая ключевые поля таблиц измерений.

Посмотреть, правильно ли создались необходимые таблицы, можно в PgAdmin4 – платформе с открытым исходным кодом для администрирования и разработки для PostgreSQL – и связанных с ней систем управления базами данных. Платформа написана на Python и jQuery и поддерживает все функции PostgreSQL.

С помощью PgAdmin4 была создана ERD-схема хранилища данных.

Построенное хранилище позволит хранить всю необходимую информацию для разведочного и кластерного анализа. В хранилище реализована возможность хранить видеофайлы, что позволит быстро и удобно просмотреть необходимые моменты совершения правонарушения, например, видеозаписи с камер наблюдения или с видеорегистраторов, установленных в автомоби-

лях. Также реализована возможность хранить необходимые изображения, которые также помогут в расследовании правонарушений. В будущем предполагается расширить функционал информационной системы с помощью добавления функций анализа видеофайлов и изображений, что позволит быстрее получать необходимую и более точную информацию при расследовании правонарушений.

Следующим этапом реализации информационной системы является построение ETL- или ELT-процесса.

ETL является сокращением от Extract (Извлечение), Transform (Преобразование) и Load (Загрузка). В таком процессе данные извлекаются из разных исходных систем, файлов, а затем происходит преобразование данных, например, операции вычисления, конкатенации, валидации.

В качестве источника данных используются файлы формата xsls, поэтому для обработки данных была выбрана библиотека Pandas.

Pandas является высокоуровневой библиотекой Python для анализа и обработки данных. Построена поверх более низкоуровневой библиотеки NumPy, которая написана на языке программирования C, что существенно повышает производительность. Pandas является продвинутой и быстроразвивающейся библиотекой для обработки и анализа данных в экосистеме Python [4].

Основными структурами данных в Pandas являются DataFrame и Series. Структура Series представляет собой объект, похожий на одномерный массив, но отличительной его чертой является наличие ассоциированных меток – индексов, вдоль каждого элемента из списка. Такая особенность превращает его в ассоциативный массив или словарь в Python.

Объект DataFrame можно представить в виде обычной таблицы, так как DataFrame является табличной структурой данных. Столбцами в объекте DataFrame выступают объекты Series, строки которых являются их непосредственными элементами.

Pandas поддерживает все самые популярные форматы хранения данных: csv, excel, sql, буфер обмена, html.

Для построения ETL-процесса был написан скриптовый файл `etl.py`, где при помощи библиотеки Pandas реализован процесс извлечения данных из исходных файлов формата «.xsls», а также трансформация, валидация данных. Далее на этапе загрузки при помощи ранее упомянутого модуля psycopg2 было установлено соединение с построенным хранилищем данных. Ниже приведен пример функции, которая при помощи SQL-запроса, возможностей Pandas и установленного соединения с PostgreSQL загружает обработанные данные в хранилище:

```
    def execute_values(conn, df, table):
        tuples = [tuple(x) for x in df.to_numpy()]
        cols = ','.join(list(df.columns))
        # SQL query to execute
        query = "INSERT INTO %s(%s) VALUES %%s" %
(table, cols)
        cursor = conn.cursor()
        try:
            extras.execute_values(cursor, query, tuples)
            conn.commit()
        except (Exception, psycopg2.DatabaseError) as error:
            print("Error: %s" % error)
            conn.rollback()
            cursor.close()
            return 1
        print("the dataframe is inserted")
        cursor.close()
```

Для визуализации данных и их удобного восприятия было решено построить дашборд.

Данный термин обозначает инструмент, выполняющий идентичную функцию: презентует данные о состоянии каких-либо процессов. Дашборд – наиболее эффективный способ представления данных для анализа и управления системой [6].

Целью создания дашборда является представление аналитических данных в компактном виде. Благодаря дашбордам можно определять взаимозависимость между разными показателями, выявлять тенденции и предотвращать потенциальные проблемы.

Для построения дашборда была использована библиотека Streamlit. Streamlit является платформой веб-приложений, которая позволяет создавать и разрабатывать веб-приложения на основе Python, которые используются для обмена результатами анализа, создания многофункциональных интерактивных интерфейсов и иллюстрации новых моделей машинного обучения.

Был написан удобный интерфейс, позволяющий переключаться между различными режимами анализа, опциями, выбирать временные рамки (рисунок 1).

При помощи библиотеки Pandas были реализованы следующие виды анализов: разведочный, кластерный.

Разведочный анализ данных (EDA), расшифровывающийся как Exploratory Data Analysis – один из первых и определяющих шагов проекта науки о данных, который приводит в движение весь проект. Он придает проекту конкретное направление и формирует план его реализации [7].

Разведочный анализ данных означает изучение данных для получения из них практической информации. Включает в себя анализ и обобщение массивных наборов данных в форме диаграмм и графиков. Это один из самых важных этапов в науке о данных. В построенном дашборде можно выбрать режим EDA и перейти к разведочному анализу данных в реальном времени.

Разработанный интерфейс позволяет увеличивать выбранные отрезки диаграммы, скачивать диаграмму в формате «png», разворачивать диаграмму на полный экран. По построенным диаграммам можно сделать выводы о количестве смертельных исходов, травм в различные периоды времени и, соответственно, принять меры, способные минимизировать число жертв ДТП.

Следующий вид анализа, реализованный в информационной системе, – кластерный.

Под кластерным анализом понимается совокупность многомерных статистических методов классификации объектов по характеризующим их признакам, разделение совокупности объектов на однородные группы, близкие по определяющим критериям, выделение объектов определенной группы.

Совокупность многомерных статистических методов кластерного анализа можно разделить на иерархические методы и неиерархические (метод k-средних, двухэтапный кластерный анализ, метод ближайших соседей). Однако общепринятой классификации методов кластерного анализа не существует, и к ним относят множество алгоритмов машинного обучения, решающих задачу разделения совокупности на однородные группы.

а)

б)

Рисунок 1 – Интерфейс информационной системы:
а – режим EDA; *б* – кластерный режим

В качестве метода кластерного анализа в работе был использован метод k-средних (рисунок 2).

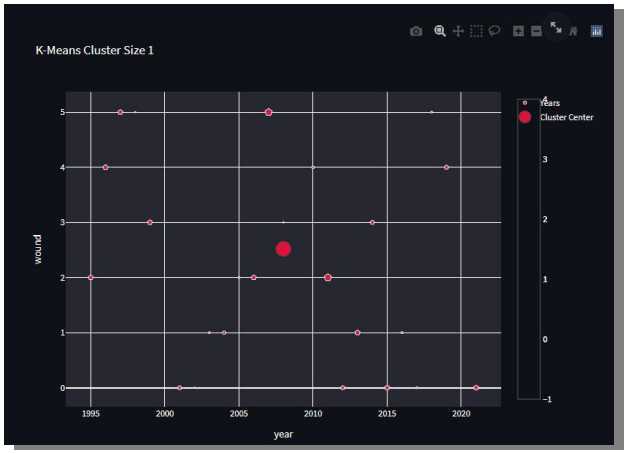

Рисунок 2 – Пример реализации k-средних

Метод k-средних используется для кластеризации данных на основе алгоритма разбиения векторного пространства на заранее определенное число кластеров k [2].

Под кластеризацией понимается разделение множества входных векторов на группы (кластеры) по степени «схожести».

Кластер – это группы объектов, выделенные в результате кластерного анализа на основе заданной меры сходства или различий между объектами [5].

Основная идея заключается в том, что на каждой итерации заново вычисляется центр масс для каждого кластера, полученного на предыдущем шаге, затем векторы разбиваются на кластеры вновь в соответствии с тем, какой из новых центров оказался ближе по выбранной метрике. Алгоритм завершается, когда на какой-то итерации не происходит изменения кластеров.

Заключение и выводы. В ходе работы был представлен и проанализирован материал, касающийся возможности применения методов интеллектуального анализа больших данных при работе с информацией о правонарушениях в мировой практике и в Республике Беларусь. Создано хранилище данных, позволяющее хранить все необходимые данные для разведочного и кластерного анализа. В хранилище также реализована возможность хранить видеофайлы, что позволит быстро

и удобно просмотреть необходимые моменты совершения правонарушения. Также реализована возможность хранить необходимые изображения. Кроме этого, был разработан удобный веб-интерфейс, интуитивно понятный и позволяющий переключаться между различными режимами анализа, опциями, выбирать временные рамки. Интерфейс информационной системы позволит сравнивать разные блоки информации в реальном времени и быстро получать результат в компактном, настроенном под себя виде.

Преимущества от внедрения заключаются в том, что разработанная информационная система позволяет хранить, обрабатывать и анализировать данные на высокой скорости и с большой точностью, а также принимать обоснованные решения, способные минимизировать число погибших и пострадавших в результате дорожно-транспортных происшествий, повысить безопасность дорожного движения.

Список литературы

1 **Quinlan, J. R.** Simplifying ETL-processes / J. R. Quinlan // Int. J. Man-Mach. Stud. – 1987. – P. 321–334.

2 Объем данных всего мира [Электронный ресурс] / Информационный портал About data. – Режим доступа : https://aboutdata.ru/2017/04/27/volume-of-data-by-2025/. – Дата доступа : 20.04.2022.

3 **Гусятников, В. Н.** Стандартизация и разработка программных систем / В. Н. Гусятников, А. И. Безруков. – М. : Финансы и статистика, 2010. – 288 с.

4 Интеллектуальный анализ данных. – Data Mining [Электронный ресурс] / ITstan. – 2011. – Режим доступа : https://www. itstan.ru/. – Дата доступа : 16.04.2022.

5 **Петров, А. И.** Оценка корреляционно-регрессионных связей между уровнем автомобилизации и тяжестью ДТП в европейских странах / А. И. Петров // Фундаментальные исследования. – 2016. – № 6, ч. 2. – С. 439–443.

6 Об органах внутренних дел Республики Беларусь [Электронный ресурс] : Закон Респ. Беларусь от 17 июля 2007 г. № 263-З : в ред. Закона Респ. Беларусь от 17.05.2021 г. // КонсультантПлюс. Беларусь / ООО «ЮрСпектр», Нац. центр правовой информ. Респ. Беларусь. – Минск, 2021.

7 **Михайличенко, О. В.** Создание информационных хранилищ данных / О. В. Михайличенко. – М. : Информационные технологии, 2018. – 689 с.

Получено 01.12.2022

A. B. Khalapsin, S. A. Daniliuk. Information system for intelligent analysis of large data on offenses in the Republic of Belarus.

Modern digital technologies have affected every person, an exceptional part of life and the impact on absolutely all types of activities. One of the social activities of exceptional diversity is the process of studying the science of road safety, associated with the receipt and processing of a very large variety of diverse information. In recent years, the amount of information has increased dramatically, making it very difficult or impossible to use it with software or hardware. Geolocation information, IoT sensor signals, vehicle information, pedestrian information all generate vast amounts of unstructured information that can be used for constructive purposes. Therefore, “big data” technology can be considered as the most important technological trend that can radically change the process of searching, analyzing and using significant information in ensuring road safety in the future.