Абдрашитов Р.Т., Полищук Ю.В.
Оренбургский государственный университет

СТРУКТУРА ЗАДАЧИ ИНТЕЛЛЕКТУАЛЬНОГО АНАЛИЗА ДАННЫХ ПО БЕЗОПАСНОСТИ ДОРОЖНОГО ДВИЖЕНИЯ

В статье выявлены основные причины, способствующие распространению новых технологий, проанализированы программные средства информационного анализа данных, выделены три этапа процесса интеллектуального анализа данных, определены классы и методы интеллектуализации, рассмотрены известные операции интеллектуального анализа данных.

В Оренбургской области в течение ряда лет успешно эксплуатируется автоматизированная информационная система (АИС), обеспечивающая обработку информации по безопасности движения. Дальнейшим развитием АИС, на наш взгляд, должна являться подсистема интеллектуального анализа данных. Учитывая особенности функций ГИБДД, эта система должна комбинироваться с геоинформационной системой (ГИС) [1]. Ниже рассмотрены возможные функции (задачи) этой подсистемы. Естественно, что полное описание задач в настоящее время невозможно, т. к. по мере эксплуатации системы будут возникать новые задачи. Однако часть из них можно выделить уже сейчас. В настоящей статье приводятся результаты анализа этих задач с целью сформирования подсистемы интеллектуального анализа данных.

Основное назначение технологии информационного анализа данных (ИАД) – автоматизированный поиск ранее неизвестных закономерностей в базах данных, хранящих информацию о деятельности организации, и использование полученных знаний в процессе принятия решений. Так, например, с помощью ИАД можно предотвратить махинации с документами пли предсказать ситуацию в БДД.

Системы ИАД реализуют новую форму анализа данных, основанную на интеллектуальном подходе. Они перерабатывают большие массивы данных и автоматически выявляют скрытые правила и закономерности, которые могут быть неочевидными для пользователя. Полученные знания помогают оптимизировать процесс деятельности организации.

В качестве основных причин, способствующих распространению новой технологии, указываются следующие:

– осознание того, что в больших по объемам БД содержатся скрытые ценные знания, способствующие повышению эффективности управления;

– развитие технологии информационных хранилищ (ИХ – Data Warehousing) позволяет создать единое информационное пространство, собрав требуемые для анализа данные в центральной БД;

– снижение стоимости устройств хранения информации, до 30-40% в год, благодаря ежегодному падению цен дало возможность пользователям хранить первичные данные из транзакционных систем с высокой степенью детализации и за длительные интервалы времени;

– снижение цен на устройства хранения информации сопровождается уменьшением стоимости компьютеров (на 35% ежегодно), в том числе с SMP-архитектурой (симметричная мультипроцессорная обработка) и с MMP-архитектурой (массово-параллельная обработка), что позволяет распараллеливать выполнение DQL-запросов и существенно повысить производительность систем ИАД;

– благодаря внедрению ИХ увеличивается число сотрудников организаций, принимающих решения, их корпоративная информация становится доступной широким слоям пользователей, которые не являются профессионалами в области СУБД и программирования [2].

Существующие программные средства ИАД в общем случае можно использовать с любыми источниками данных, в том числе и с БД OLTP (Online Transaction Processing /оперативной обработки транзакций) – систем. Однако целесообразнее «приложить» этот инструментарий к информационным хранилищам, в которых неоднородные данные, полученные из разных источников, синхронизированы, очищены и приведены к единым форматам, а структура БД оптимизирована с точки зрения скорости доступа. Кроме того, при подобном подходе любые изыскания аналитиков не будут мешать работе ключевых OLTP-систем.

В общем случае информационная система на основе технологии ИХ состоит из четырех

компонентов – одного или нескольких серверов баз данных; программного обеспечения, обеспечивающего функционирование систем клиент / сервер; программы загрузки данных в ИХ из внешних источников, которая сопровождается предварительной обработкой данных; клиентских приложений, предназначенных для поддержки принятия решений.

Процесс интеллектуального анализа обычно проходит в три этапа (рис. 1).

Выбор данных. Для решения конкретной задачи нужны не все данные из ИХ. Сначала необходимо выбрать подмножество, которое будет подвергнуто анализу. При этом, возможно, потребуется объединить несколько таблиц, а полученные записи отфильтровать.

Трансформация данных. После подготовки рабочих таблиц обычно проводится предварительная обработка данных, характер которой определяется методами, применяемыми в ходе анализа. Трансформация может заключаться в удалении зашумленных данных и дублирующих записей, преобразовании типов данных, добавлении новых атрибутов и др.

Анализ. Трансформированные данные последовательно обрабатываются по одной или нескольким методикам с целью извлечения требуемой информации или знаний.

Классы операций и методы интеллектуального анализа данных. В ходе ИАД могут выполняться различные операции, которые, в свою очередь, реализуются при помощи разнообразных алгоритмов. В основе большинства из них лежит аппарат математической статистики. Методы ИАД условно делят на два класса: операции проверки гипотез и операции поиска зависимостей, направленные на автоматическое выявление закономерностей пли правил, которым подчиняются данные ИХ.

К недостаткам процедур первого класса можно отнести ограниченность анализа жесткими рамками заранее принятой гипотезы. Пользователь предполагает, например, что факты двух видов нарушения правил БДД связаны. В процессе анализа будут проверены исторические данные и сделан вывод, верна гипотеза или нет. Проблема заключается в том, что другие возможные корреляции попросту выпадут из рассмотрения, если для аналитика они априори не очевидны.

Для второго класса системы ИАД самостоятельно обрабатывают информацию с целью обнаружения внутренних закономерностей. Полученные результаты часто оказываются весьма неожиданными и ведут к нетривиальным выводам.

Комбинируя операции двух классов, описанных выше, возможно реализовать самые различные стратегии анализа.

Рассмотрим известные операции ИАД.

Проверка гипотез. Операции этого типа выполняют генераторы отчетов, системы обработки SQL-запросов, приложения многомерных БД и модули статистического анализа.

Генерация отчетов и обработка запросов. Является наиболее распространенной и про-

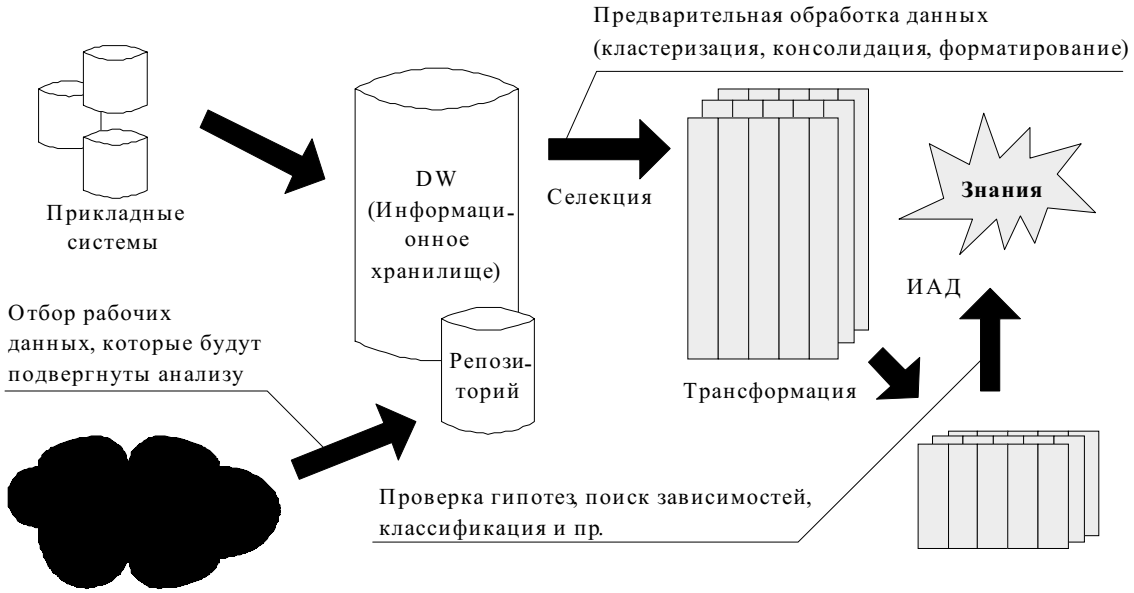

Рисунок 1. Архитектура системы ИАД.

стой формой анализа. Ее основное назначение – подтвердить правильность гипотез, сформулированных пользователем, который последовательно генерирует несколько разных уточняющих SQL-запросов, призванных помочь в подтверждении правильности исходного предположения. Результаты обработки запросов обычно оформляются в виде таблиц или графиков. Последовательность выполненных запросов и правила построения таблиц или графиков образуют так называемый сценарий анализа, который впоследствии можно распространять среди других пользователей.

Статистический анализ. Простейшая статистическая обработка возможна при работе с данными на уровне SQL-запросов, однако для выполнения более содержательного статистического анализа, например при проверке сложных гипотез, требуются специализированные средства, которые не только поддерживают соответствующие методы анализа (регрессионный, факторный, дисперсионный, кластерный и др.), но и имеют наглядные средства визуализации результатов. Многие пакеты статистической обработки позволяют проверять гипотезы и автоматически их генерировать.

Поиск зависимостей. К этому типу операций относятся прогнозное моделирование, анализ связей, сегментация данных и идентификация отклонений.

Прогнозное моделирование. Благодаря развитию различных методов автоматического построения моделей (методы индукции, нейронные сети) прогнозное моделирование стало самым распространенным типом операций ИАД. Основой для всевозможных систем прогнозирования служит историческая информация, хранящаяся в БД в виде временных рядов, которые отражают динамику исследуемой системы в прошлом. Если удается построить математическую модель, адекватно описывающую эту динамику, есть вероятность, что с ее помощью можно предсказать и поведение системы в будущем. Ценность некоторых прогнозных моделей, особенно базирующихся на индукционных методах, заключается в простоте представления результатов. Если поведение сложной системы описывается в терминах простых логических выражений типа if-then-else, соответствующие модели легко воспринимаются пользователями, их удобно изучать и модифицировать.

Анализ связей. Если цель прогнозного моделирования заключается в построении обобщенной модели данных, то в задачах анализа связей требуется найти специфические связи между различными записями в БД. Классические алгоритмы выявления связей базируются на статистических методах корреляционного и регрессионного анализов.

Сегментация баз данных. Разбиение записей БД на несколько групп (сегментов или коллекций) часто проводится в качестве предварительного этапа с целью сузить поиск и сократить период дальнейшей обработки. Например, выбираются все данные, относящиеся к определенному промежутку времени или региону. К подготовленным данным применяются другие методы – прогнозного моделирования или анализа связей.

Идентификация отклонений. Цель этой операции – во-первых, выявить данные, которые не входят ни в один из имеющихся сегментов, а во-вторых, установить, являются ли они «шумом» или отражают пока неизвестные закономерности. Часто идентификация отклонений выполняется в сочетании с сегментацией базы данных, когда указанный сегмент считается нормой, а все остальные – отклонением от нее. Применяемые здесь алгоритмы аналогичны критериям проверки статистических гипотез, а также основываются на методах дисперсионного анализа [3].

Существующие операции ИАД поискового типа поддерживаются большим числом всевозможных методик: одна и та же операция может быть реализована различными способами. Рассмотрим их.

Индукция – это процесс автоматической генерации классификационной модели на основе специально подготовленных тестовых (обучающих) данных, содержащихся в БД или ИХ. Обучающие выборки, как правило, представляют собой небольшие наборы данных, соответствующие описаниям уже известных классов. Индуцированная модель выглядит как совокупность образцов (реализаций) данных, по которым идентифицируется этот класс. Однажды созданную модель можно применять в дальнейшем для обнаружения классов среди еще не классифицированных записей.

Существует два типа индукции: нейронная и символическая. В основе нейронных методик лежат сетевые архитектуры узлов (нейронные сети), соединенных между собой связями, имеющими различные веса. Методы символической индукции сводятся к построению де-

ревьев решений, ветви которых задаются при помощи логических формул.

Индукционные методы позволяют построить качественные модели даже в том случае, когда обучающие данные неполны или сильно зашумлены. Форма представления моделей легко читаема, и пользователь почти всегда может проследить путь, по которому двигалась система при построении окончательной классификации. Кроме того, индукционные системы способны накапливать знания, что существенно улучшает качество вновь генерируемых моделей.

Поиск ассоциаций выглядит следующим образом. Пусть имеется коллекция элементов и набор записей. Каждая из записей содержит какие-то элементы коллекции. Выявленные зависимости выражаются посредством правил, например: «89% всех записей, в которые входят элементы А, Б и В, включают также элементы Г и Д». Вероятность выполнения (истинности) правила называется доверительной (другое ее название – уровень значимости). Приведенное правило иногда формулируют иначе: «Элементы А, Б и С находятся на противоположной стороне ассоциации по отношению к элементам Г и Д». В общем случае в процессе поиска может быть задействовано любое число элементов по обе стороны ассоциации. Поиск ассоциаций предполагает, например, выявление корреляций в исходных данных и последующее интервальное оценивание для проверки соответствующей гипотезы.

Кластеризация – это процесс разбиения БД на ряд сегментов, или кластеров, объединяющих данные, имеющие общие характеристики. Результаты кластеризации применяются для подведения суммарных итогов по кластерам и в качестве входных данных для какого-либо другого метода анализа. Поскольку размеры кластеров меньше размеров БД и в них содержатся относительно «чистые» данные, время обработки заметно сокращается, а качество моделей данных повышается. Конечная цель кластеризации – разбиение исходного множества элементов (записей БД) на классы, в этом случае нередко именуемые кластерами. Основное отличие кластеризации от традиционных методов классификации заключается в отсутствии обучающей выборки и вообще каких-либо априорных сведений о структуре и статических свойствах классифицируемых данных. В последние годы, помимо многочисленных традиционных методов, для проведения кластерного анализа все шире применяются нейронные сети.

Архитектура нейронной сети включает взаимосвязанные вычислительные элементы (нейроны), каждый из которых генерирует выходной сигнал в ответ на несколько входных. Выход элемента является входом для других. Каждый вход получает вес (в виде коэффициента в соответствующем уравнении), который корректируется в процессе обучения сети. Обучение сводится к подбору таких весов, при котором нейронная сеть безошибочно распознает эталонную выборку. Как правило, для реализации алгоритмов нейронных сетей требуются мощные вычислительные ресурсы, поскольку во время обучения тестовые данные приходится обрабатывать сотни тысяч раз. Иногда после первого этапа обучения нейронной сети предъявляются более «тонкие» тестовые данные, чтобы улучшить параметры ее настройки. Наивысшие результаты распознавания (и классификации) достигаются при дополнительном структурировании сети путем разбиения всего множества нейронов на два или большее число слоев.

Поскольку каждый элемент нейронной сети частично изолирован от своих соседей, у нейронных алгоритмов имеется хороший потенциал для распараллеливания вычислений. С помощью алгоритмов нейронных сетей можно решать многие задачи ИАД: прогнозировать поведение объекта в будущем, основываясь на данных о его динамике в прошлом, производить факторный анализ, выявлять аномалии и сходства.

Генетические алгоритмы были предложены в начале 70-х годов Джоном Холландом (John Holland) с целью имитации эволюционных процессов в живой природе. Холланд предпринял попытку формализовать законы эволюции и использовать их для решения задач оптимизации. Генетические алгоритмы оперируют такими понятиями, как ген, хромосома, популяция, мутация и пр. С их помощью решены многие прикладные задачи, в частности известная «задача коммивояжера». В системах ИАД генетические алгоритмы используются для поиска зависимостей [4].

Возможность ИАД увеличивается при сочетании с информационными (ГИС) технологиями. Это обусловлено тем, что деятельность ГИБДД пространственно распределена. Отра-

жение ситуаций, возникающих в процессе ее функционирования, может дать новую информацию и возможности управления. Рассматривая задачи ИАД безопасности движения, можно выделить следующие классы задач.

Прогнозирование – одна из самых распространенных задач ИАД. В частности, при планировании и составлении бюджета необходимо прогнозировать уровень безопасности движения и другие параметры с учетом многочисленных взаимосвязанных факторов – сезонных, региональных, общеэкономических и т. д. Можно также выявлять корреляции между различными данными.

Анализ работы персонала: эффективность труда служащих зависит от уровня подготовки, оплаты труда, опыта работы, взаимоотношений с руководством и т. д. Проанализировав влияние этих факторов, можно выработать методику повышения эффективности деятельности каждого сотрудника, а также предложить оптимальную стратегию подбора кадров в будущем.

Профилирование участков ДД: с помощью нейросетевых моделей анализируя множество участков ДД, можно получить представление о типичном участке ДД по аналогии с портретом «типичного клиента компании». Кроме того, можно выяснить, почему движение по некоторым участкам ДД стало неэффективным (уменьшилась пропускная способность, увеличилась аварийность и т. д.). Результатом анализа является выбор стратегии определения требуемых параметров участков ДД. Кроме того, можно выяснить причины аварийности на конкретных участках ДД.

Оценка потенциальных участков ДД: планируя строительство новых участков ДД, имеет смысл прогнозировать параметры их эксплуатации (пропускную способность, аварийность и т. д.). Проводимый анализ позволяет вырабатывать оптимальные проектные решения ДД.

Анализ работы региональных отделений ГИБДД: с помощью нейросетевых моделей можно сравнивать результаты деятельности региональных отделений и определять, от чего зависит эффективность их работы (географическое положение, численность персонала и т. д.). Результаты используются для оптимизации работы «отстающих» отделений.

Сравнительный анализ с ГИБДД других регионов: необходимо сравнить деятельность рассматриваемых подразделений и выяснить, какие факторы определяют эффективность их деятельности.

В данном направлении можно выделить следующие виды задач: вспомогательные, традиционные и новые (нетрадиционные).

Вспомогательные задачи должны включать в себя обмен с хранилищем данных, анализ и проверка исходных данных на полноту и достоверность, формирование массивов данных, подлежащих обработке, хранение результатов обработки (в том числе и промежуточных).

При этом важную роль должна сыграть концепция активных данных. Суть этой концепции заключается в том, что в настоящее время функционирует концепция активных алгоритмов, когда алгоритмы для проведения вычисления инициируют поиск данных. В концепции активных данных анализ их на полноту и противоречивость инициирует поиск новых данных для анализа и новых алгоритмов для их обработки. Таким образом, видно, что так называемые «вспомогательные задачи» имеют важное, а иногда и самостоятельное значение.

Традиционные задачи. К ним относятся задачи, которые выполняются уже в настоящее время, закреплены в нормативных документах. Как правило, задачи статистической обработки данных.

К новым задачам относятся, на наш взгляд, те, которые ранее не использовались в практике.

При создании информационной системы необходимо разработать удобный пользовательский интерфейс, так как именно удобный интерфейс увеличивает скорость принятия решений. При построении информационных систем широкое распространение получила технология взаимодействия пользователя с системой через Web-интерфейс в INTRANET технодогии.

Привлекательность технологии INTRANET для построения корпоративных информационных систем обуславливается рядом важных факторов.

Во-первых, несмотря на обилие компьютеров и работающих на них программ, информационные технологии остаются по-прежнему в основе своей «бумажными». Эта ситуация вряд ли кардинально изменится в ближайшем будущем.

Во-вторых, в организации чаще всего нет и не было сколько-нибудь осмысленного подхода к управлению информацией. Все, что с ней

происходит: ее создание, передача, потребление, принятие решений на ее основе – есть результат несистематизированных и слабо согласованных действий сотрудников и руководителей, выполняемых без учета дисциплины работы с информацией.

Реальная информационная технология должна быть «бесшовно» вплетена в сложную ткань жизнедеятельности организации. Для этого она должна обладать особыми уникальными свойствами, которые ей предоставляет технология INTRANET.

Внедрение технологии INTRANET связано в первую очередь с резким улучшением качества потребления информации, напрямую влияющим на производительность труда сотрудников организации. Для системы INTRANET ключевыми становятся новые понятия – **публикация информации**, **потребители информации**, **предоставление информации**. Результат применения INTRANET – резкое сокращение бумажных архивов, легкость и простота публикации информации, универсальный и естественный доступ к информации с помощью навигаторов, существенное сокращение затрат на администрирование приложений на рабочих местах пользователей, немедленная актуализация любых изменений в ИХ организации, смещение акцентов от создания информации к ее эффективному потреблению.

Ключевыми качествами INTRANET, напрямую связанными с экономическими аспектами деятельности современной организации, являются: простота и естественность технологии; низкий риск и быстрая отдача инвестиций; интеграционный и «каталитический» характер технологии; эффективное управление и коммуникации в организации.

При создании систем INTRANET отмечено еще одно качество новой технологии, которое заключается в том, что усложнение системы, расширение сервиса, детализация функций не требуют от пользователя специальных знаний. Он учится работе с информацией один раз, а далее, пользуясь в своей повседневной работе средствами навигации по информационному пространству, он обнаруживает новые возможности выполнения задач. Психологически это исключительно важно. Человек начинает по-другому относится к работе с ИХ. Происходит изменение культуры работы с информацией. Близость информации к потребителю обеспечивает успех технологии INTRANET.

Существенным свойством внедрения INTRANET является быстрая отдача. Это резко упрощает внедрение технологии, поскольку пользователь сразу видит отдачу и поэтому начинает охотно взаимодействовать с разработчиками и помогать внедрению системы.

Будучи применимой практически в любых условиях, обладая уникальным интеграционным качеством, INTRANET-технология в предельно сжатые сроки позволяет получить конкретный, видимый, эффективный для каждодневной работы результат.

Рассмотрим основные принципы управления информацией внутри организации.

INTRANET предполагает такую организацию информационной системы, которая опирается на принцип предоставления информации всем нуждающимся в ней, доставку информации по инициативе ее потребителя, а не поставщика. Причем требование, чтобы информация всегда была актуальной, предъявляется только к поставщику, гарантируя ее немедленную доставку. При таком подходе, с одной стороны, для организации необходимы источники информации, а с другой стороны, для ее сотрудников необходим единый универсальный инструмент для работы с информацией. В качестве первого естественно рассматривать один или несколько Web-серверов, а в качестве второго – навигатор.

Отметим, что доставка информации по инициативе потребителя не исключает возможности принудительного информирования последнего в случае необходимости. Отметим также, что любые информационные системы, в основу которых положено переложение на компьютерную технику технологии «бумажного» документооборота, этими свойствами не обладают. В них изначально закладывается структура, обладающая огромной инерцией, которая будет оказывать пассивное сопротивление любым нововведениям.

В целом Web-технология предлагает определенную концепцию предоставления информационных услуг потребителям, которая отличается следующими особенностями:

- информация предоставляется потребителю в виде публикаций;
- публикация может объединять информационные источники различной природы и географического расположения;
- изменения в информационных источниках мгновенно отражаются в публикациях;

• в публикациях могут содержаться ссылки на другие публикации без ограничения на местоположение и источники последних (гипертекстовые ссылки);

• потребительские качества публикаций соответствуют современным стандартам мультимедиа (доступны текст, графика, звук, видео, анимация).

Применение Web-технологии как средства публикации информации имеет следующие отличительные черты:

• публикатор не заботится о процессе доставки информации к потребителю;

• затраты публикатора не зависят от «тиража» публикации;

• количество потенциальных потребителей информации практически не ограничено;

• презентационные качества публикаций соответствуют современным запросам потребителей;

• публикации отражают текущую информацию, время запаздывания определяется исключительно скоростью подготовки электронного документа;

• информация, представленная в публикации, легко доступна благодаря гипертекстовым ссылкам и средствам контекстного поиска;

• информация легко усваивается потребителем благодаря широкому спектру изобразительных возможностей, предоставляемых Web-технологией;

• Web-технология не предъявляет особых требований к типам и источникам информации;

• Web-технология масштабируема: увеличение числа одновременно обслуживаемых потребителей не требует радикальной перестройки системы [5].

В современных навигаторах, опирающихся на идеи гипертекста, создан унифицированный интерфейс для доступа к самым разнообразным источникам информации. Для того чтобы обратиться к файлу, к таблице базы данных или к результатам работы какого-либо прикладного пакета, используется одна и та же программа с одними и теми же средствами управления. Работа с навигатором быстро осваивается. После этого сотрудник имеет доступ и работает с любой информацией, получаемой через Web.

Web-технологии позволяют интегрировать различные источники информации и различные ее типы (файлы, БД, электронные таблицы, электронную почту и т. д.). Технология позволяет получать результаты в виде данных принципиально различной природы – текст, таблицы, графика (рисунки, чертежи, схемы) и т. д.

Поэтому для доступа к данным различных типов нужны специализированные приложения, которые обрабатывают данные определенного типа. В связи с этим возникает проблема выбора приложений, необходимых пользователю для решения поставленных задач.

Проведенный нами анализ дает основание сделать вывод о том, что ИАД с использованием Web-интерфейса является на настоящий момент одним из самых перспективных средств построения интеллектуальных аналитических систем в обеспечении безопасности ДД, что связано с ужесточением условий дорожного движения, вызванных ростом количества участников дорожного движения, повышением скоростных характеристик транспортных средств, сложностью и неоднозначностью дорожных ситуаций, повышением криминализации дорожного движения.

В дальнейшем нами предполагается проектирование систем ИАД для решения задач этого класса, адаптированных к требованиям ГИБДД и аналогичных организаций.

Список использованной литературы:

1. Хавилов В.А., Пуковский А.И. Пути совершенствования деятельности ГИБДД УВД Оренбургской области на основе интеллектуальных технологий // Бюллетень главного управления ГИБДД РФ, №24, январь 2004 г.
2. Дюк В., Самойленко А. Data mining: учебный курс. – СПб.: Питер, 2001. – 386 с.
3. Филиппов В.А. Интеллектуальный анализ данных: методы и средства. – М.: Эдиториал УРСС, 2001. – 52 с.
4. David Hand, Heikki Mannila, Padhraic Smyth. Principles of Data Mining. The MIT Press, 2001. 546 pages.
5. Филиппов В.А., Щукин Б.А., Лукин Н.В. Интеллектуальный анализ данных в многомерных СУБД с использованием WEB-технологий. – М.: АПП, 2000.