

## ПРИМЕНЕНИЕ БАЗ ДАННЫХ ДЛЯ ОЦЕНКИ РАБОТЫ ПАССАЖИРСКОГО ТРАНСПОРТА

О.П. Алексеев, профессор, д.т.н., С.В. Пронин, ассистент, ХНАДУ

*Аннотация.* Рассмотрен и предложен подход к созданию баз данных для оценки работы пассажирского транспорта.

*Ключевые слова:* база данных, аналитическая платформа, пассажирский транспорт, транспортная система, управление транспортом, маршрут.

### Введение

Одним из важнейших условий повышения качества удовлетворения потребности населения крупных городов в перевозке и улучшения экономических показателей работы городского общественного пассажирского транспорта есть повышение эффективности оперативного диспетчерского управления движением городского пассажирского транспорта (ГПТ). Это обеспечит наиболее полное и эффективное использование потенциальных возможностей сети общественного пассажирского транспорта города для предоставления транспортных услуг жителям города.

Повышение эффективности оперативного диспетчерского управления движением городского общественного пассажирского транспорта на основании повышения уровня автоматизации и внедрения новых информационных технологий обеспечивает разработка и внедрение в эксплуатацию автоматизированной системы управления и контроля движения городского пассажирского транспорта.

### Анализ публикаций

На сегодняшний день создание АСУ транспорта является не просто задачей автоматизации управления соответствующей подсистемой транспортного комплекса, а является процессом информатизации транспортного обслуживания жителей большого города.

Следует заметить, что сегодня приобретает большое значение применение на транспорте современных информационных технологий, [1, 2] которые можно использовать для управления подвижными объектами, повышения качества обслуживания и эффективности работы транспорта [2].

Следует заметить, что динамика развития больших городов, постоянное изменение их транспортной инфраструктуры не позволяет применить для проектирования автоматизированных систем оперативного управления готовые апробированные в других городах решения из-за индивидуальной специфики практически любого города или региона. Анализ многокритериального подхода и оценка традиционной методологии автоматизации управления на автотранспорте [1 – 4] позволяют утверждать о необходимости принципиально новых разработок, принятия новых решений относительно применения новейших прогрессивных технологий управления разными подвижными единицами общественного пассажирского транспорта больших городов. Это целиком отвечает ситуации в усовершенствовании транспортного обслуживания жителей больших городов, что сложилась на Украине, распространению средств новейших информационных, интеллектуальных технологий управления в транспортных системах [1 – 4].

Согласно вышеупомянутому в основе методологии проектирования данных систем положены принципы интеллектуализации подсистем и звеньев транспортного комплекса большого города согласно общей практике компьютеризации [4].

## Цель исследования

Из проведенного анализа можно увидеть, что в АСУ большое значение имеют базы данных [5, 6]. Применение баз данных позволит собирать информацию о потребностях населения и параметрах работы городских маршрутов. Также она дает возможность оценить уровень качества обслуживания пассажиров, надежности и эффективности работы транспорта на маршрутах, прогнозирования спроса на перевозки.

В последнее время наметилась тенденция интегрировать базы данных с так называемыми аналитическими платформами. Такой подход дает значительное преимущество, поскольку позволяет проводить экспресс-анализ, поиск и нахождение зависимостей в массивах данных в оперативном режиме. Такие методы получили название интеллектуального анализа данных и экспресс-анализа [5, 6].

## Аналитическая платформа

Работа аналитической платформы представлена на рис. 1.

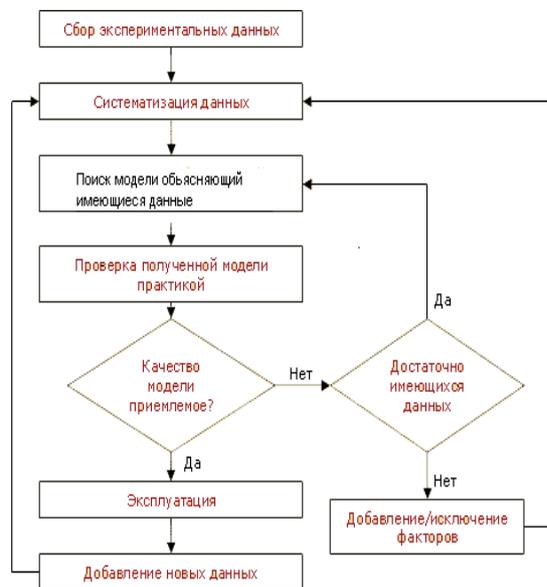


Рис. 1. Принцип работы аналитической платформы

В данных платформах применяются как правило, следующие основные методы анализа данных:

- 1 Data Mining;
- 2 Knowledge Discovery in Databases (KDD) [5, 6]

Подробно остановимся на каждом из них. Развитие методов записи и хранения данных привело к бурному росту объемов собираемой и анализируемой информации. Объемы данных настолько внушительны, что человеку просто не по силам проанализировать их самостоятельно, хотя необходимость проведения такого анализа вполне очевидна, ведь в этих «сырых» данных заключены знания, которые могут быть использованы при принятии решений. Для того чтобы провести автоматический анализ данных, используется Data Mining.

Data Mining – это процесс обнаружения в «сырых» данных ранее неизвестных нетривиальных практически полезных и доступных интерпретации знаний, необходимых для принятия решений в различных сферах человеческой деятельности. Data Mining является одним из шагов Knowledge Discovery in Databases.

Информация, найденная в процессе применения методов Data Mining, должна быть нетривиальной и ранее неизвестной, например, средние продажи не являются таковыми. Знания должны описывать новые связи между свойствами, предсказывать значения одних признаков на основе других и т.д. Найденные знания должны быть применимы и на новых данных с некоторой степенью достоверности. Полезность заключается в том, что эти знания могут приносить определенную выгоду при их применении. Знания должны быть изложены в доступном виде и для пользователя-не математика. Например, проще всего воспринимаются человеком логические конструкции «если ... то ...». Более того, такие правила могут быть использованы в различных СУБД в качестве SQL-запросов. В случае, когда извлеченные знания непрозрачны для пользователя, должны существовать методы постобработки, позволяющие привести их к интерпретируемому виду.

Алгоритмы, используемые в Data Mining, требуют большого количества вычислений. Раньше это являлось сдерживающим фактором широкого практического применения Data Mining, однако сегодняшний рост производительности современных процессоров снял остроту этой проблемы. Теперь за приемлемое время можно провести качественный анализ сотен тысяч и миллионов записей.

Задачи, решаемые методами Data Mining:  
Классификация – это отнесение объектов (наблюдений, событий) к одному из заранее известных классов.

Регрессия, в том числе задачи прогнозирования. Установление зависимости непрерывных выходных от входных переменных.

Кластеризация – это группировка объектов (наблюдений, событий) на основе данных (свойств), описывающих сущность этих объектов. Объекты внутри кластера должны быть «похожими» друг на друга и отличаться от объектов, вошедших в другие кластеры. Чем больше похожи объекты внутри кластера и чем больше отличий между кластерами, тем точнее кластеризация.

Ассоциация – выявление закономерностей между связанными событиями. Примером такой закономерности служит правило, указывающее, что из события  $X$  следует событие  $Y$ . Такие правила называются ассоциативными. Впервые эта задача была предложена для нахождения типичных шаблонов покупок, совершаемых в супермаркетах, поэтому иногда ее еще называют анализом рыночной корзины (market basket analysis).  
Последовательные шаблоны – установление закономерностей между связанными во времени событиями, т.е. обнаружение зависимости, что если произойдет событие  $X$ , то спустя заданное время произойдет событие  $Y$ .

Анализ отклонений – выявление наиболее нехарактерных шаблонов.

Для решения вышеописанных задач используются различные методы и алгоритмы Data Mining. Ввиду того, что Data Mining развивалась и развивается на стыке таких дисциплин, как статистика, теория информации, машинное обучение, теория баз данных, вполне закономерно, что большинство алгоритмов и методов Data Mining были разработаны на основе различных методов из этих дисциплин. Например, процедура кластеризации k-means была просто заимствована из статистики. Большую популярность получили следующие методы Data Mining: нейронные сети, деревья решений, алгоритмы кластеризации, в том числе и масштабируемые, алгоритмы обнаружения ассоциативных связей между событиями и т.д.

Knowledge Discovery in Databases (KDD) – это процесс поиска полезных знаний в «сырых» данных. KDD включает в себя вопросы: подготовки данных, выбора информативных признаков, очистки данных, применения методов Data Mining (DM), постобработки данных и интерпретации полученных результатов. Безусловно, «сердцем» всего этого процесса являются методы DM, позволяющие обнаруживать знания.

Этими знаниями могут быть правила, описывающие связи между свойствами данных (деревья решений), часто встречающиеся шаблоны (ассоциативные правила), а также результаты классификации (нейронные сети) и кластеризации данных (карты Кохонена) и т.д.

Процесс Knowledge Discovery in Databases состоит из следующих шагов:

Подготовка исходного набора данных. Этот этап заключается в создании набора данных, в том числе из различных источников, выбора обучающей выборки и т.д. Для этого должны существовать развитые инструменты доступа к различным источникам данных. Желательно иметь поддержку работы с хранилищами данных и наличие семантического слоя, позволяющего использовать для подготовки исходных данных не технические термины, а бизнес-понятия.

Предобработка данных. Для того чтобы эффективно применять методы Data Mining, следует обратить внимание на вопросы предобработки данных. Данные могут содержать пропуски, шумы, аномальные значения и т.д. Кроме того, данные могут быть избыточны, недостаточны и т.д. В некоторых задачах требуется дополнить данные некоторой априорной информацией. Наивно предполагать, что если подать данные на вход системы в существующем виде, то на выходе получим полезные знания. Данные должны быть качественными и корректными с точки зрения используемого метода DM. Поэтому первый этап KDD заключается в предобработке данных. Более того, иногда размерность исходного пространства может быть очень большой, и тогда желательно применять специальные алгоритмы понижения размерности. Это как отбор значимых

признаков, так и отображение данных в пространство меньшей размерности.

Трансформация, нормализация данных. Этот шаг необходим для приведения информации к пригодному для последующего анализа виду. Для чего нужно проделать такие операции, как приведение типов, квантование, приведение к «скользящему окну» и прочее. Кроме того, некоторые методы анализа требуют, чтобы исходные данные были в каком-то определенном виде. Нейронные сети, скажем, работают только с числовыми данными, причем они должны быть нормализованы.

Data Mining. На этом шаге применяются различные алгоритмы для нахождения знаний. Это нейронные сети, деревья решений, алгоритмы кластеризации, установления ассоциаций и т.д.

Постобработка данных. Интерпретация результатов и применение полученных знаний в практических целях оптимизации работы.

### **Выводы**

В статье был предложен подход к построению баз данных работы городского общественного транспорта. Данные базы могут быть использованы в качестве алгоритмического обеспечения автоматизированных систем управления и контроля работы городского пассажирского транспорта. В качестве математического аппарата в данных платформах предполагается использовать аппа-

рат интеллектуальных технологий, который хорошо зарекомендовал себя именно при применении его в системах поддержки принятия решений в различных областях экономики, промышленности, медицины и т.д.

### **Литература**

1. Алексієв О.П., Алексієв В.О., Серіков С. А. Підвищення ефективності управління громадським пасажирським автотранспортом / Весник ХНАДУ. – Харків: ХНАДУ. – Вип. 22. – 2003. – С. 56 – 61.
2. Алексеев В.О. Интеллектуальная технология организации движения транспортных средств / Автомобильный транспорт. – Харьков: ХНАДУ.– 2002. – Вып. 10. – С. 305 – 311.
3. Макаров И.П., Ямпольский В.З. Автоматизация управления городским транспортом. – М.: Транспорт, 1981. – 256 с.
4. Рева В.М., Лигум Ю.С. и др. Оперативное управление городским пассажирским транспортом. – К.: Техника, 1982. – 175 с.
5. Барсегян А.А., Куприянов М.С. Методы и модели анализа данных: OLAP и Data Mining – СПб.: «БХВ Петербург», 2004.
6. Чубукова И.А. Data Mining: Учебное пособие. – М.: Интернет-университет информационных технологий: БИНОМ: Лаборатория знаний, 2006. – 382 с.

Рецензент: А.В. Бажинов, профессор, д.т.н., ХНАДУ.

Статья поступила в редакцию 5 июня 2009 г.