

Сафронов А.И., Иконников А.С.

Применение методов интеллектуального анализа данных к поиску рациональных вариантов автоматизированного построения графиков движения пассажирских поездов метрополитена

Аннотация: В работе выполнен анализ результатов автоматизированного построения плановых графиков движения пассажирских поездов по Кольцевой линии Московского метрополитена, полученных на базе работы Генетического Алгоритма. Описан процесс микросервисной трансформации результатов с последующим их размещением в новой базе данных *PostgreSQL*. Полученные сведения в совокупности образуют большие данные (*big data*), пригодные для обработки методами интеллектуальной аналитики.

Ключевые слова: метрополитен, интеллектуальный анализ данных, программное обеспечение, большие данные, критерий оценки, postgresql, график движения поездов

Работа посвящена вопросам развития технологии автоматизированного построения плановых графиков движения (ПГД) пассажирских поездов по Кольцевой линии Московского метрополитена. Исследования, связанные с этими вопросами, проводятся сотрудниками кафедры «Управление и защита информации» («УиЗИ») «Российского университета транспорта» (РУТ (МИИТ)).

На протяжении последних нескольких лет к работе на уровне развития интеллектуальных программных продуктов, внедренных на ГУП «Московский метрополитен», привлекаются обучающиеся. Результаты совместной деятельности обучающихся и преподавателей докладываются на конференциях, в частности, – «Интеллектуальные транспортные системы» (организовывается и проводится кафедрой «УиЗИ» [1-4]). Как итог – успешно защищаются дипломные проекты, магистерские диссертации и бакалаврские выпускные квалификационные работы [5].

Для автоматизированного построения ПГД используется система «АРМ Графиста» первого поколения. Созданный в 2013 году модуль дополнен элементами интеллектуализации. Помимо заложенных пошаговой и рекурсивной процедур автоматизированного построения ПГД внедрен механизм перебора вариантов на базе Генетического Алгоритма [6].

Результаты работы модуля формируют пласт сложно анализируемой оператором информации с большим количеством значащих признаков (метрик). Аналитическая работа в этих условиях сводится к обработке больших данных (*big data*) и, соответственно, интеллектуальному анализу данных (*data mining*). На первых этапах работы производился исключительно «майнинг» – сбор и накопление данных с характерными и значащими признаками. Получаемая «картина» данных заставила пересмотреть имеющиеся подходы к проектированию информационного обеспечения систем планирования перевозочного процесса на метрополитене [7]. Вместе с тем авторы заняты пересмотром имеющейся архитектуры программного обеспечения в направлении взаимодействующих микросервисов. В настоящее время «АРМ Графиста» представляет собой так называемое, «монолитное», прикладное программное обеспечение рабочего стола (*desktop application*).

В сфере интеллектуального анализа данных известно и изучено большое количество публикаций. Немногие посвящены прикладным аспектам разработок по предметным областям: экономики, медицины, образования и других. В основном, статьи посвящены обобщениям о существовании методов интеллектуального анализа данных и повествование в них сводится к перечислению круга задач машинного обучения. Авторы изученных публикаций независимо друг от друга сходятся во мнении, что интеллектуальный анализ данных – это способ разбора на базе компьютерных технологий информации, представленной в сложном для восприятия оператором виде. Это мнение поспособствовало выдвигению новой гипотезы: при использовании методов интеллектуального анализа данных можно найти оптимальный, с точки зрения сформулированного критерия, вариант автоматизированного построения ПГД в обозримые сроки перебора вариантов на базе Генетического Алгоритма.

Сформулированным критерием на протяжении многих лет является равномерность процессов автоматизированного построения ПГД. Внутри критерия заложено быстроедействие. На уровне формулировки и применения фитнес-функции Генетического Алгоритма под быстроедействием понимается минимальное время завершения движения поездов по одному из главных (гл.) путей Кольцевой линии.

Исходными данными для корректной работы алгоритмов автоматизированного построения ПГД являются: заданная парность движения поездов, определяемая на каждый час работы метрополитена; график оборота электроподвижного состава; точки ночной расстановки, а также последовательности их освобождения и заполнения, соответственно, ранним утром и поздним вечером. Этот набор исходных данных будем называть первичным.

При фиксированном, первичном наборе исходных данных подбираются количественные значения вводимых/снимаемых составов, определяющих, в свою очередь, количество итераций вводов/снятий в режиме однократного, пошагового запуска процедуры автоматизированного построения ПГД. Это – вторичный набор исходных данных.

Вторичный набор исходных данных необходим как для режима полного перебора вариантов рекурсивной процедуры, так и для режима перебора вариантов на базе Генетического Алгоритма. К сожалению, Генетический Алгоритм не поспособствовал устранению проблем, выявленных в схеме полного перебора, но позволил затронуть ранее недостижимые области поиска.

Для исследуемого первичного набора исходных данных вторичным набором является ситуация, при которой:

- на 1-й, 2-й и 3-й итерациях ввода составов на утренний час «пик» из депо на I гл. путь выводится по 2 состава (на каждой итерации), на II гл. путь – по 3 состава (на каждой итерации);
- на 1-й итерации снятия составов после утреннего часа «пик» в депо с I и II гл. путей убирается по 2 состава (с каждого пути);
- на 2-й итерации снятия составов после утреннего часа «пик» в депо с I и II гл. путей убирается по 1 составу (с каждого пути);
- на 1-й итерации ввода составов на вечерний час «пик» из депо на I и II гл. пути выводится по 2 состава (с каждого пути);

- на 2-й итерации ввода составов на утренний час «пик» из депо на I и II гл. пути выводится по 1 составу (с каждого пути);
- на 1-й и 2-й итерациях снятия составов после вечернего часа «пик» в депо с I гл. пути убирается по 2 состава (на каждой итерации), со II гл. пути – по 3 состава (на каждой итерации);
- на 3-й итерации снятия составов после вечернего часа «пик» в депо с I и II гл. путей убирается по 1 составу (с каждого пути);
- на 4-й итерации снятия составов после вечернего часа «пик» в депо с I гл. пути убирается 1 состав, со II гл. пути – 2 состава.

Для фиксированных первичного и вторичного набора исходных данных предусмотрен прямой переход к интересующему варианту автоматизированного построения ПГД по его уникальному идентификатору, именуемому «вектором-кодом» варианта, формируемому по принципу, проиллюстрированному на рисунке 1.

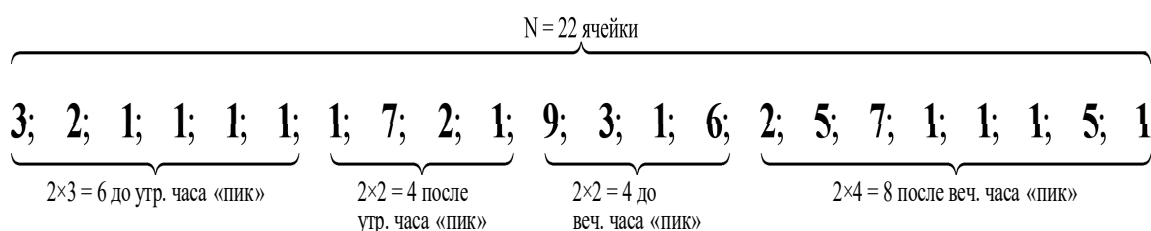


Рисунок 1 – Кортёж, формирующий «вектор-код» варианта автоматизированного построения ПГД

В 22-х ячейках представлен последовательный набор значений, описывающих сдвиг на i «ниток» относительно момента начала/окончания итерации снятия/ввода:

- 6 значений до утреннего часа «пик» (по 3 на гл. путь);
- 4 значения после утреннего часа «пик» (по 2 на гл. путь);
- 4 значения до вечернего часа «пик» (по 2 на гл. путь);
- 8 значений после вечернего часа «пик» (по 4 на гл. путь).

На каждой позиции записана реализация из диапазона $[1 \dots Na_i]$. Na_i варьируется в зависимости от количества «ниток», оставшихся на ПГД к рассматриваемому моменту времени автоматизированного составления ПГД.

В качестве «вектора-кода» исходно рассматривалась строка текста [2]. Позже, в [6] «вектор-код» определен как объект, качественные характеристики которого необходимо хранить для нужд последующего анализа:

$$VarVect = (ID, Seq, Criteria), \quad (1)$$

где ID – уникальный идентификатор исходного набора данных, необходимых для автоматизированного составления ПГД;

Seq – кортеж целых чисел, составленный по принципу, изображенному на рисунке 1;

$Criteria$ – значение фитнес-функции оценки качества варианта автоматизированного составления ПГД, измеряемое в «метросекундах» (на метрополитене расчеты показателей времени ведутся с дискретизацией в 5 секунд).

Информация о первом поколении (первичная популяция фиксированного размера R) хранится в файле / ином внешнем источнике информации / генерируется псевдослучайным образом в «АРМ Графиста».

Информация о популяции любого поколения может быть представлена в виде двумерных массивов, в которых количество строк равно R – размер популяции, количество столбцов равно Nl или $Nl + 1$ в зависимости от формы представления исходных данных. Элемент матрицы при размерности $[R \times Nl]$:

$$Z_{i,j} \in [1..Na_i], \quad i \in 1..R, \quad j \in 1..Nl, \quad (2)$$

где Na – количество вариантов смещения последовательности ввода/снятия составов относительно первично рассчитанной последовательности – мощность множества аллелей;

Nl – количество итераций равномерных вводов/снятий составов, реализуемых в ходе расчета варианта автоматизированного составления ПГД – мощность множества локусов.

При размерности $[N \times Nl + 1]$ в матрицу добавляется столбец, содержащий числовое значение выбранного критерия (фитнес-функции) для каждой строки матрицы:

$$Criteria_i = \begin{cases} t_{np}^{\max}, & \text{в случае успеха построения} \\ F, & \text{при неудачном построении} \end{cases}, \quad (3)$$

где t_{np}^{\max} – время завершения движения в соответствии с успешно реализованным вариантом автоматизированного построения ПГД;
 F – расчетная величина «штрафа» (детализирована в [6]).

В результате работы Генетического Алгоритма система «АРМ Графиста» формирует csv-файл (*CSV – comma-separated values*, значения, разделенные запятыми), представляющий собой журнал событий (*event log*), содержащий все изменения, происходящие с популяцией. Как правило, спустя десятки итераций работы Генетического Алгоритма – это уже глубокие потомки по сравнению с первичной популяцией указанного размера R . Под особями и потомками все еще понимаются «векторы-коды» вариантов автоматизированного построения ПГД, рассматриваемые как объекты согласно формуле (1). Генетический Алгоритм в настоящей работе запускался многократно с различными наборами особей в первичной популяции, но одного и того же размера R . Csv-файлы в работе рассматриваются как локальные текстовые базы данных. Далее с полученными csv-файлами начинают работать новые микросервисы.

1. Микросервис «*CSV_Cutter*» отвечает за выкройку из журнала событий последнего полного перечня потомков с их характеристиками. На выходе укороченный csv-файл.

2. Микросервис «*CSV_Glue*» реализует объединение потомков из укороченных csv-файлов в один, новый, общий csv-файл.

3. Микросервис «*CSV_Merge_Duplicates*» модифицирует общий csv-файл, исключая из него повторные вхождения «векторов-кодов», добавляя к их структуре столбец, содержащий количество упоминаний каждого «вектора-кода», предоставляя возможность для оценки процента повторов вариантов, возникающих в работе независимых сессий запуска Генетического Алгоритма.

4. Микросервис «*CSV_To_PostgreSQL*» реализует перенос данных из csv-файла в базу данных *PostgreSQL*.

Создаваемая в *PostgreSQL* таблица (отношение базы данных) состоит из 26 столбцов: «*id*» – численный идентификатор варианта построения ПГД (номер по порядку), «*v1*»–«*v22*» – значение N_{a_i} из формулы (2), «*count*» – количество повторных упоминаний «вектора-кода», «*criterion*» – значение фитнес-функции

(сформулированного критерия), «vct» – исходная, текстовая запись «вектора-кода».

Нормализация базы данных намеренно не проводилась. Именно такой вид отношения базы данных обеспечивает удобство работы механизмов интеллектуального анализа, заложенных в микросервис «*PGD_Analysis*». Следует отметить, что для альтернативного вторичного набора исходных данных количество итераций вводов / снятий составов может меняться. Это потребует использования и накопления другой базы данных близкой, но все равно иной структуры. На решении частной, рассматриваемой в статье, задачи с четко обозначенными первичным и вторичным наборами исходных данных отсутствие нормализации базы данных никак не сказывается.

В настоящее время авторы статьи продолжают работу над развитием микросервиса «*PGD_Analysis*», предназначенного для интеллектуального анализа результатов автоматизированного построения ПГД согласно формируемым им гистограммам групп данных, зависящих от управляемого набора значимых признаков (метрик). Ожидается, что такое управление позволит выявить область рационального поиска вариантов автоматизированного построения ПГД, внутри которой за обозримое время перебора найдется оптимальный, с точки зрения сформулированного критерия, вариант автоматизированного построения ПГД пассажирских поездов по Кольцевой линии Московского метрополитена.

Литература:

1. Баранов Л.А., Сидоренко В.Г., Сафронов А.И. Особенности автоматизированного планирования движения поездов в интеллектуальных системах управления внеуличного городского транспорта в условиях интенсивного движения / Интеллектуальные транспортные системы: материалы Международной научно-практической конференции. Москва, 26 мая 2022 г. – М.: РУТ, 2022. – С. 59-70.

2. Сафронов А.И., Дудкин А.В. Анализ результатов автоматизированного построения плановых графиков движения поездов по кольцевой линии Московского метрополитена / Интеллектуальные транспортные системы: материалы

Международной научно-практической конференции. Москва, 26 мая 2022 г. – М.: РУТ, 2022. – С. 290-297.

3. *Сафронов А.И., Иконников А.С.* Множество оцениваемых параметров результата автоматизированного построения плановых графиков движения поездов и графоаналитические способы их сопоставления / Интеллектуальные транспортные системы: материалы Международной научно-практической конференции. Москва, 26 мая 2022 г. – М.: РУТ, 2022. – С. 306-315.

4. *Сафронов А.И., Старовойтова У.А.* Антология задач организации графического пользовательского интерфейса в интеллектуальной системе «АРМ графиста» / Интеллектуальные транспортные системы: материалы Международной научно-практической конференции. Москва, 26 мая 2022 г. – М.: РУТ, 2022. – С. 326-337.

5. *Логинова Л.Н., Новиков В.Г., Сафронов А.И.* Направления развития подготовки бакалавров и магистров в области автоматизации, цифровизации и управления // Информатизация образования и науки. – 2023. – № 4(60). – С. 58-68.

6. *Сидоренко В.Г., Сафронов А.И.* Применение генетических алгоритмов при решении задач планирования перевозочного процесса городской рельсовой транспортной системы // Автоматика на транспорте. – 2023. – Т. 9. № 1. – С. 49-62. – DOI: 10.20295/2412-9186-2023-9-01-49-62.

7. *Сидоренко В.Г., Логинова Л.Н., Сафронов А.И.* Разработка информационного обеспечения для интеллектуального управления городскими рельсовыми транспортными системами // Автоматика на транспорте. – 2023. – Т. 9. № 2. – С. 178-192. – DOI: 10.20295/2412-9186-2023-9-02-178-192.
