УДК 004.021

СЦЕНАРИЙ ОБРАБОТКИ ТРЕХЭТАПНОЙ СТРУКТУРЫ ОБЪЕДИНЕНИЯ ДАННЫХ TPRUDF

Евдокимова И.А., Андриалович И.В.

Белорусский государственный университет информатики и радиоэлектроники, г. Минск, Республика Беларусь

Научный руководитель: Лихачевский Д.В. – к.т.н., доцент, доцент кафедры ПИКС

Аннотация. Огромные данные, собираемые устройствами Интернет вещей (ИВ) в умных городах, требуют надежного места для обработки и хранения, когда это необходимо, мгновенно на уровне сервера без перегрузки [1]. Следовательно, методы использования ресурсов пользуются большим спросом в эпоху Интернет вещей.

Ключевые слова: объединения данных, слияние, Интернет вещи, алгоритм

Введение. Структуру трехэтапного объединения данных с использованием ресурсов (*TPRUDF*) можно адаптировать к любой вычислительной модели. Его возможности разделены на три основных уровня слияния данных: уровень слияния пространственно-временных данных, который представляет собой уровень слияния входных и выходных данных, управляя различными функциями данных и работая непосредственно с данными ИВ, уровень слияния данных об используемых ресурсах функций, который извлекает некоррелированные функции объединенных данных Интернет вещей, и уровень слияния данных об используемых ресурсах принятых решений, который определяет наилучшее использование ресурсов планирования путем объединения результатов с использованием нескольких методов использования ресурсов на сервере обработки [2].

В данной статье авторами представлен один и возможных сценариев использования трехэтапной структуры объединения данных об использовании ресурсов *TPRUDF*.

Основная часть. В этой статье показан полный сценарий обработки *TPRUDF*, как показано на рисунке 1. Сценарий проходит через три этапа слияния данных. Предположим, что несколько идентификаторов источников данных ИВ SID_1, SID_2, …, SID_n постоянно генерируют огромные объемы данных, связанных с ценными метаданными, и передают их на первую фазу объединения данных через слой объединения пространственно-временных данных с использованием ресурсов вывода данных.

Полученные данные проверяются на достоверность. Например, записи данных о SID_3 игнорируются из-за недействительности этого источника. Затем данные проверяются на достоверность форматов файлов для поддержки различных функций данных ИВ и для обеспечения достоверности сгенерированных данных ИВ в соответствии с доменом ИВ. Записи данных о SID_4 игнорируются из-за недопустимости формата файла.

Принятые записи данных фильтруются на основе их свежести для управления неустойчивыми данными ИВ в соответствии с определенным интервалом времени. Например, интервал времени свежести в домене для умных домов длится секунды, а интервал времени свежести в географическом домене длится месяцами. Следовательно, самые свежие записи данных поддерживаются и предварительно обрабатываются для исправления отсутствующих значений атрибутов и выбросов со средними значениями атрибутов, которые поддерживают функцию неточных данных ИВ [2].

Затем, чтобы поддерживать огромный объем данных ИВ, записи данных *M* (*M* – среднее значение допустимых значений данных) группируются на основе их идентификатора источника, а затем сокращаются с помощью выборки набора *STDF* (*STDF* – это подход к слиянию данных на основе ИВ для низкоуровневого слияния входных и выходных данных для пространственной агрегации источников ИВ в режиме реального времени с использованием аналитики больших данных), которая отбирает все группы с одинаковыми вероятностными выборками размера *N (N* – количество записей данных в наборе данных).

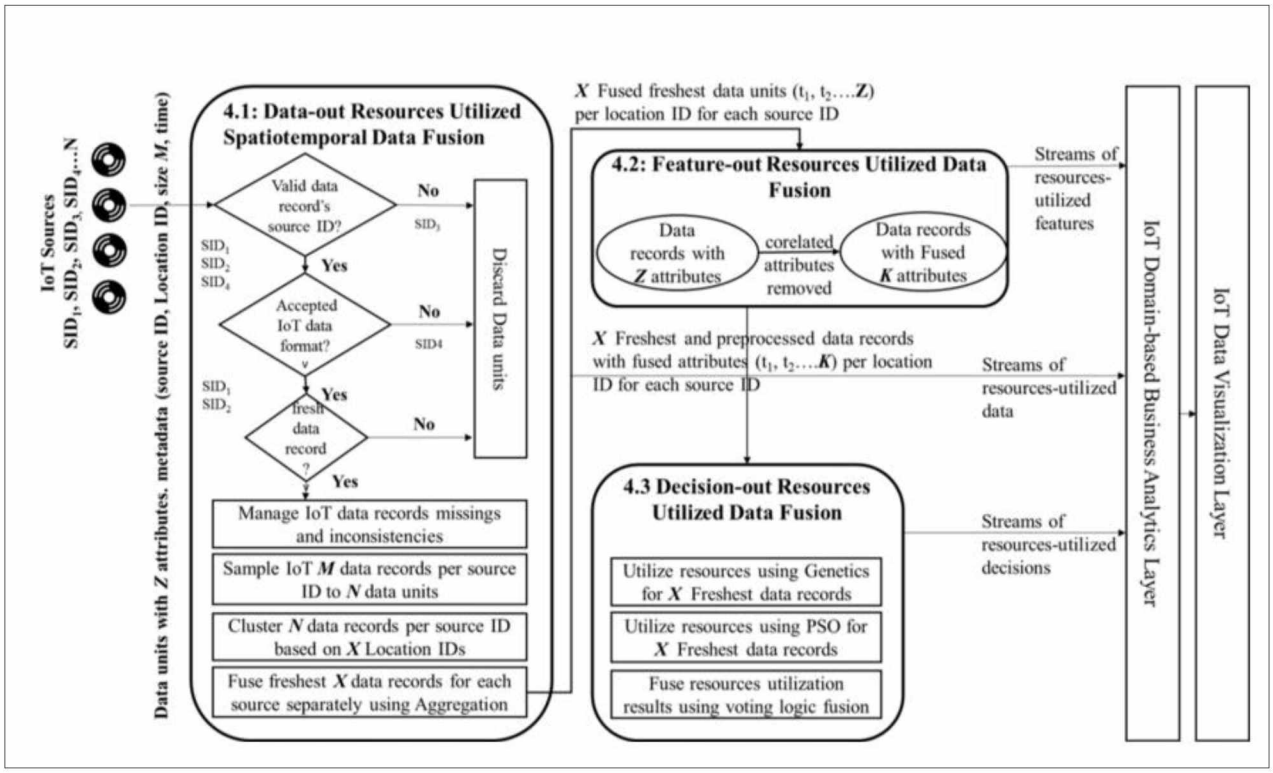

Рисунок 1 – Предлагаемый сценарий обработки TPRUDF

STDF определяет размер выборки *N* с использованием метода выборки «Вероятность, пропорциональная размеру» (*PPS*), при котором вероятность выбранных записей данных пропорциональна размеру набора. После этого *STDF* определяет список *X* идентификаторов местоположения для каждой группы с *N* выборочными записями данных, чтобы сгруппировать их на основе идентификаторов местоположения для создания карты T_1, где ключевой параметр – идентификатор местоположения, значение – матрица *D* записи данных с *Z* атрибутами, как результат алгоритма *K*-средних [3].

Наконец, на этом уровне для каждого идентификатора источника *STDF* сохраняет функцию временных данных ИВ, объединяя все записи данных в результирующую T_1 карту для каждого набора в соответствии с их минимальным временем генерации для создания карты T_2 (ключевой параметр: идентификатор источника, значение: матрица *X* агрегированные записи данных с *Z* атрибутами), которая используется в качестве входных данных для алгоритма анализа основных компонентов (*PCA)* на следующем слое слияния данных.

Второй этап слияния данных – это слой «Выделения ресурсов, используемых для слияния данных». Этот уровень поддерживает массив данных ИВ и обеспечивает точность использования ресурсов за счет извлечения некоррелированных функций и уменьшения количества атрибутов из Z к K. Данный уровень использует PCA для создания $K \times X$ матрицы некоррелированных признаков и обновление T_2 карты (ключевой параметр: идентификатор источника, значение: матрица X агрегированные записи данных с K атрибутами), которая может быть отправлена непосредственно на уровень бизнес-аналитики или на следующий уровень слияния данных, используемых для принятия решений.

Третий этап объединения данных получает обновленную T_2 карту для обработки на стороне сервера. Он применяет два метода использования ресурсов: генетические алгоритмы (GA) и оптимизация множества частиц (PSO) с использованием списка доступных виртуальных машин и идентификатора каждого источника $K \times X$ некоррелированной матрицы признаков для выбора наиболее оптимальных виртуальных машин, необходимых для обработки матрицы признаков.

Наконец, $TPRUDF$ предоставляет наиболее точные результаты использования ресурсов, выбирая лучшие результирующие виртуальные машины из обоих методов, используя метод слияния логики, чтобы вывести решения об использовании ресурсов, которые отправляются на уровни бизнес-аналитики и визуализации данных для дальнейшего анализа.

Заключение. В этом иллюстративном примере подчеркивается, что $TPRUDF$ может поддерживать различные функции данных ИВ, где каждая функция должна обрабатываться отдельно от других, при этом учитываются все параметры использования ресурсов. Например, рассматривая отношениям между функциями данных ИВ и параметрами использования ресурсов, можно сделать следующие выводы [4]:

1 Пропускная способность обеспечивается за счет обработки массивных данных ИВ и за счет сокращения как записей данных ИВ с использованием метода выборки, так и атрибутов данных ИВ с использованием PCA при обработке быстро генерируемых данных ИВ за счет поддержки реальных данных

2 Чтобы гарантировать параметр надежности, $TPRUDF$ поддерживает *частные данные*, устанавливая порог надежности для источников данных ИВ для получения данных; *разнообразные данные*, поддерживая различные форматы файлов; *изменчивые данные*, устанавливая интервал времени свежести для принятия записей данных; *неточные данные* за счет сохранения ошибок данных и шумов, *информативные данные* за счет учета метаданных во время анализа, таких как идентификатор источника, время генерации и т. д.

3 Параметр доступности предоставляется за счет поддержания как временных, так и быстрых данных, сгенерированных путем агрегирования записей данных до их минимального времени генерации с обработкой в режиме реального времени.

4 Наконец, параметр задержки сохраняется за счет сокращения данных ИВ в двух измерениях с использованием двух методов: выборки и PCA.

Список литературы

1. *Deadlock free resource management technique for IoT-based post disaster recovery systems S. Agrawal, R.R. Rao. – Scalable Comput. Pract. Exp. 21 (2020) 391–406.*
2. *Zhang T. Collaborative algorithms that combine AI with IoT towards monitoring and control system /T. Zhang, Y. Zhao, W. Jia, M.Y. Chen// Futur. Gener. Comput. Syst.– 2021.– 125. – pp 677–686.*
3. *Lv Z. Intelligent edge computing based on machine learning for smart city/ Z. Lv, D. Chen, R. Lou, Q. Wang// Futur. Gener. Comput. Syst.– 2021.– 1215. – pp 90-99.*
4. *Fawzy D. The spatiotemporal data reduction (STDR): an adaptive IoT-based data reduction approach/ D. Fawzy, S. Moussa, N. Badr // in: Proceedings of the 10th International Conference on Intelligent Computing Information System [Electronic resource] – https://doi.org/10.1109/ICICIS52592.2021.9694199.*

UDC 004.021

PROCESSING SCRIPTION OF THE THREE STEP DATA COMBINATION STRUCTURE TPRUDF

I.A. Evdokimova, I.V.Andryalovich

Belarusian State University of Informatics and Radioelectronics, Minsk, Republic of Belarus

D.V. Likhachevsky– Cand. of Sci., associate professor, associate professor of the department of ICSD

Annotation. The huge data collected by Internet of Things (IoT) devices in smart cities requires a reliable place to process and store when needed, instantly at the server level without overloading. Hence, resource utilization techniques are in great demand in the era of Internet of Things.

Keywords: data merging, fusion, internet of things, algorithm