PI-0369 Application of Graph Database in Multisource Transportation Data Integration

Requested by
Zhenyu Zhu, Division of Traffic Operations

Prepared by
Charina Guarino, Division of Research, Innovation and System Information

February 26, 2025

The Caltrans Division of Research, Innovation and System Information (DRISI) receives and evaluates numerous research problem statements for funding every year. DRISI conducts Preliminary Investigations on these problem statements to better scope and prioritize the proposed research in light of existing credible work on the topics nationally and internationally. Online and print sources for Preliminary Investigations include the National Cooperative Highway Research Program (NCHRP) and other Transportation Research Board (TRB) programs, the American Association of State Highway and Transportation Officials (AASHTO), the research and practices of other transportation agencies, and related academic and industry research. The views and conclusions in cited works, while generally peer reviewed or published by authoritative sources, may not be accepted without qualification by all experts in the field. The contents of this document reflect the views of the authors, who are responsible for the facts and accuracy of the data presented herein. The contents do not necessarily reflect the official views or policies of the California Department of Transportation, the State of California, or the Federal Highway Administration. This document does not constitute a standard, specification, or regulation. No part of this publication should be construed as an endorsement for a commercial product, manufacturer, contractor, or consultant. Any trade names or photos of commercial products appearing in this publication are for clarity only.

Table of Contents

Executive Summary

Background

California Department of Transportation (Caltrans) uses and operates a wide variety of traffic data sources which are critical to traffic operation functions. These data sources include:

- Real-time vehicle detector station (VDS) data from Caltrans' Performance Measurement System (PeMS)
- Traffic Census data
- Weigh-in-motion (WIM) data
- California Highway Patrol (CHP) data
- Traffic Incident Management (TIM) data
- Closed-circuit television (CCTV) camera data
- Caltrans Linear Referencing System (LRS)

Each of these datasets plays a crucial role in monitoring and managing California's complex transportation network to ensure smooth flow of daily commuter traffic.

However, the data management is fragmented with isolated data sources. When data sources remain isolated, the capability to comprehensively monitor, predict, and manage transportation events is hindered. This lack of integration not only prevents effective traffic management, but also contributes to broader economic and business challenges. The lack of an integrated data system leads to delayed response times to incidents, inefficient resource allocation and limited ability to make informed decisions.

To address these challenges, the Division of Traffic Operations is seeking information on existing methodologies for integrating diverse transportation datasets by geospatial information, particularly the adoption of graph databases, which excels at unifying these datasets by mapping complex relationships among various data points. This integration enables streamlined incident response and support real-time decision-making.

Summary of Findings

Related Research and Resources

A literature search of recent publicly available domestic and international resources identified research organized into two categories: multi-source transportation data integration and graph database for transportation data use.

Multi-Source Data Integration

This section provides a sampling of studies on fusing transportation data from multiple data sources. The sampling includes a study conducted for the Massachusetts Department of Transportation (MassDOT) which fused available data to develop two models for real-time traffic incident detection. The data used include speed and travel time data through the MassDOT GoTime and Regional Integrated Transportation Information System (RITIS) platforms and Waze reports.

Another study, performed in partnership with Washington State DOT (WSDOT), developed a transportation data-integration framework to support multisource data-based traffic analysis. The data sets used in this study include freeway loop data, NPMRDS data, Verizon speed data, incident data, weather data, and INRIX and HERE data. An international study integrated data from multiple sources to develop a method to model highway traffic flow in Zhejian Province, China.

Graph Database for Transportation Data

A sampling of studies on the application of graph databases in transportation. The sampling includes a study which developed an urban data integration framework using a graph database and used data from Palo Alto, California as a case study. Another study conducted a survey on the integration of advanced technologies, such as Real-Time Databases (RT-DBs), Graph Databases (GDBs), and Artificial Intelligence (AI) to improve Intelligent Transportation Systems (ITS) capabilities. A sampling of research at the international level focuses on applying a graph database schema on multimodal transportation network data.

Gaps in Findings

Limited Scope of Case Studies

While the literature highlighted in this Preliminary Investigation provide useful information and highlight innovative applications, a review of two State DOTs' approach does not provide an exhaustive review on the application of multisource data integration. Even more so, as the search uncovered few instances of the use or application of a graph database structure at state DOTs. There may be state DOTs that have applied a graph database schema to integrate their transportation data which may be of interest to Caltrans.

Lack of Empirical Performance Data

There is a shortage of empirical evidence regarding the impact of integrated multisource data systems – particularly those utilizing graph databases on incident response times, resource allocation, and operational efficiency.

Methodology Limitations

The current investigation exhibits no production level graph database across different state DOTs. In addition, available studies vary widely in their methodological approaches, with some providing high-level conceptual frameworks and others offering more detailed technical designs.

Next Steps

Moving forward, Caltrans could consider:

- Contacting any identified State DOTs in this Preliminary Investigation to inquire on the methodologies to integrate multisource data.
- Seeking information from other state DOTs and agencies on their application of a graph database structure and if/how this is applied to their data.
- Reviewing the findings of the literature search for information on multisource data integration methods and graph database schema application that can be applied to Caltrans' existing data source and their needs.

Detailed Findings

Background

California Department of Transportation (Caltrans) uses and operates a wide variety of traffic data sources which are critical to traffic operation functions. These data sources include:

- Real-time vehicle detector station (VDS) data from Caltrans' Performance Measurement System (PeMS)
- Traffic Census data
- Weigh-in-motion (WIM) data
- California Highway Patrol (CHP) data
- Traffic Incident Management (TIM) data
- Closed-circuit television (CCTV) camera data
- Caltrans Linear Referencing System (LRS)

These data sources, however, are isolated and lack integration, making it difficult to efficiently manage traffic incidents or improve operational safety. The lack of an integrated data system leads to delayed response times to incidents, inefficient resource allocation and limited ability to make informed decisions. To overcome these challenges, the Division of Traffic Operations is seeking information on existing methodologies for integrating diverse transportation datasets by geospatial information, particularly the adoption of graph databases, which excels at unifying these datasets by mapping complex relationships among various data points. This method offers a streamlined incident response, and support real-time decision-making.

A graph database is a database management system based on graph theory which uses nodes as entities or discrete objects, edges as relationships or connections between nodes and properties as key-value pairs used for storing data on nodes and edges. Traditionally, most transportation network data is managed in relational databases, which organizes data into tables with rows representing records and columns representing attributes. Graph databases, however, offer a flexible approach to modeling the intricate and connected relationships inherent between physical objects, such as stations, camera, WIM, etc.

Related Research and Resources

A literature search of recent publicly available domestic and international resources identified research organized into two categories: multi-source transportation data integration and graph database for transportation data use.

Multi-Source Data Integration

Multisource Data Fusion for Real-Time and Accurate Traffic Incident Detection via Predictive Analytics, Polichronis Stamatiadis, Nathan H. Gartner, Yuanchang Xie and Ruifeng Liu, University of Massachusetts Lowell and Massachusetts Department of Transportation and Federal Highway Administration, 23-040, April 2023.
https://rosap.ntl.bts.gov/view/dot/73043/dot_73043_DS1.pdf
From the abstract: The objectives of this study are to (1) identify data sets available to MassDOT that can be used for real-time incident detection; (2) investigate how data from different sources can be

integrated to improve incident detection; and (3) develop guidance for establishing trigger points to alert Highway Operations Center (HOC) operators about incidents on the road. Speed data available through the Regional Integrated Transportation Information (RITIS) platform are used for developing two alternative strategies: (a) an Artificial Intelligence (AI) model using supervised learning based on Long Short-Term Memory (LSTM) and Variational Autoencoders (VAE) layers for classifying records as normal events or incidents, and (b) an empirical rule-based method using historical speeds to establish threshold values, below which an alarm is issued requiring the HOC operator's attention. Results on the AI model and a verified incident data set indicate a False Alarm Rate (FAR) of 0.0069% and a detection rate of 91.70%. For the empirical rule-based model, a 30-day off-line "field-test" was conducted for June 2021. Most of the events recorded by the MassDOT HOC were detected, and for most of these events the detection time was well before the "SENT-ON" time recorded in the HOC incident database.

"Establishing Multisource Data-Integration Framework for Transportation Data Analytics," Zhiyong Cui, Kristian Henrickson, Salvatore Antonio Biancardo, Ziyuan Pu and Yinhai Wang, *Journal of Transportation Engineering, Part A: Systems*, Volume 146, Issue 5, February 2020.
https://ascelibrary.org/doi/10.1061/JTEPBS.0000331
From the abstract: In recent years, with the advancement in traffic sensing, data storage, and communication technologies, the availability and diversity of transportation data have increased substantially. When the volume and variety of traffic data increase dramatically, integrating multisource traffic data to conduct traffic analysis is becoming a challenging task. The heterogeneous spatiotemporal resolutions of traffic data and the lack of standard geospatial representations of multisource data are the main hurdles for solving the traffic data-integration problem. In this study, to overcome these challenges, a transportation data-integration framework based on a uniform geospatial roadway referencing layer is proposed. In the framework, on the basis of traffic sensors' locations and sensing areas, transportation-related data are classified into four categories, including on-road segment-based data, off-road segment-based data, on-road point-based data, and off-road point-based data. Four data-integration solutions are proposed accordingly. An iterative map conflation algorithm as a core component of the framework is proposed for integrating the on-road segment-based data. The overall integration performance of the four types of data and the efficiency of the iterative map conflation algorithm in terms of percentage of integrated roadway segments and computation time are analyzed. To produce efficient transportation analytics, the proposed framework is implemented on an interactive data-driven transportation analytics platform. Based on the implemented framework, several case studies of real-world transportation data analytics are presented and discuss.

"A Modeling Method for Complex Traffic Flow on Highway Based on the Fusion of Heterogenous Data from Multiple Sensors," Shaoweihua Liu, Yunyan Tang, Yiliu He, Junyi Ren, Yujie Zhang, Xi Luo and Hongyun Yang, *Journal of Transportation Engineering, Part A: Systems*, Volume 150, Issue 6, March 2024.
https://ascelibrary.org/doi/10.1061/JTEPBS.TEENG-8207
From the abstract: Improving the efficiency and safety of highway traffic relies heavily on accurately modeling the complex dynamics of traffic flow. This study aims to develop a novel method for modeling highway traffic flows by integrating data from multiple sources, such as roadside cameras, gantry cameras, and communication devices. The method leverages heterogeneous sensor data characteristics, temporal information, and spatial structures to achieve deep fusion of latent features at the sensor level. To minimize human intervention, a multistage training approach is employed, combining large-scale self-supervised learning with supervised fine-tuning, leveraging abundant unlabeled unstructured data such as monitoring videos recorded by roadside cameras and snapshots captured by gantry cameras, alongside limited accurate structured traffic flow data aggregated by communication devices from gantries and toll stations. We demonstrate the effectiveness and stability of the proposed method

on a case study, the G92 ring highway around the Hangzhou Bay in Zhejiang Province, China, achieving the mean absolute percentage error of traffic flow within 7.5% and 8.3% for fixed and variable highway sensor networks, respectively. Ablation studies further demonstrate the significant improvement in predictive accuracy achieved by the designed self-supervised pretraining task. To summarize, our approach provides a promising solution for efficient and safe management of highway traffic flow, with potential applicability to real-world scenarios.

Graph Database for Transportation Data

"Urban Data Integration Using Proximity Relationship Learning for Design, Management, and Operations of Sustainable Urban Systems," Karan Gupta, Zheng Yang and Rishee K. Jain, Journal of Computing in Civil Engineering, Volume 33, Issue 2, 2019.
https://ascelibrary.org/doi/abs/10.1061/%28ASCE%29CP.1943-5487.0000806
From the abstract: The world is rapidly urbanizing, with 66% of the world's population expected to reside in cities by 2050. This massive influx of new urban citizens is putting enormous pressure on city systems and bringing forth challenges at the intersection of urban infrastructure, governance, and the environment. As a result, researchers and practitioners have turned to new advanced sensing and data analytics developed under the burgeoning smart city movement to improve the design, management, and operations of urban systems. However, it has been challenging to integrate, organize, and analyze the data emerging from urban systems due to their natural spatial, temporal and typological heterogeneity. This paper introduces an urban data integration (UDI) framework that is capable of integrating heterogeneous urban data. The proposed UDI framework is extensible to multiple types of urban systems, scalable to the growing volume of data streams (as a result of increasing geographical areas, higher sampling frequencies, and so on), and interpretable enough to help inform municipal decision-making. The UDI framework uses a series of proximity relationship learning algorithms to reconstruct urban data in a graph database. The merits, applicability, and efficacy of the proposed framework is demonstrated by validating and testing it on data from a midsize city in the United States and by benchmarking its interpretability and computational performance for a typical urban analytics scenario against current practice (i.e., a relational database). Results indicate that the UDI framework provides easier and more computationally efficient exploration and querying of urban data, and in turn can enable new computational approaches to urban system design, management, and operations.

"Intelligent Transportation Systems: A Survey on Data Engineering," Safa Batita, Achraf Makni and Ikram Amous, 13th International Conference on Data Science, Technology and Applications - DATA, pages 169-179, 2024.
https://www.scitepress.org/Link.aspx?doi=10.5220/0012857300003756
From the abstract: This paper presents an examination of data engineering within Intelligent Transportation Systems (ITS), focusing on integrating advanced technologies such as Real-Time Databases (RT-DBs), Graph Databases (GDBs), and Artificial Intelligence (AI) to improve ITS capabilities. The decision to focus on database systems and AI in this paper is based on their crucial roles in shaping modern transportation systems and offers a comprehensive view o the technological framework influencing ITS. Through an extensive review of existing literature, the paper explores how these solutions synergistically contribute to data collection, organization, processing, and extraction of value from various ITS data. The paper analyzes the transformative impact of real-time data management in connected vehicle systems and the efficacy of GDBs in capturing complex relationships within intelligent transportation networks. Additionally, it assesses the adaptability of AI in various ITS applications, including traffic prediction, driver assistance, and accident analysis. Despite their benefits, the paper discusses persistent challenges related to system complexity, interoperability, data management, and model accuracy, which impact the widespread deployment of ITS. Furthermore, the paper presents

recommendations for addressing these challenges and emphasizes research directions that require further exploration, underscoring the importance of intelligent and efficient transportation worldwide.

"The Shortest Path Algorithm Performance Comparison in Graph and Relational Database on a Transportation Network," Mario Miler, Dražen Odobašić and Damir Medak, Promet-Traffic & Transportation, Volume 26, pages 75-82, 2014.
https://doi.org/10.7307/ptt.v26i1.1268
From the abstract: In the field of geoinformation and transportation science, the shortest path is calculated on graph data mostly found in road and transportation networks. This data is often stored in various database systems. Many applications dealing with transportation network require calculation of the shortest path. The objective of this research is to compare the performance of Dijkstra shortest path calculation in PostgreSQL (with pgRouting) and Neo4j graph database for the purpose of determining if there is any difference regarding the speed of the calculation. Benchmarking was done on commodity hardware using OpenStreetMap road network. The first assumption is that Neo4j graph database would be well suited for the shortest path calculation on transportation networks but this does not come without some cost. Memory proved to be an issue in Neo4j setup when dealing with larger transportation networks.
From the conclusion: The graph database management systems are not routing engines and are not suitable for full graph traversal, which is used in the shortest path calculations. …Although in most cases, Neo4j outperforms pgRouting, the Neo4j "greed" for memory has to be considered. This is especially important for large transportation networks. If memory is not an issue, then graph database is the right choice for the shortest path calculation.

"Framework for constructing multimodal transport network and routing using a graph database: A case study in London," Seula Park and Tao Cheng, Transactions in GIS, Volume 27, Issue 5, pages 1391-1417, 2023.
https://doi.org/10.1111/tgis.13071
From the abstract: Most prior multimodal transport networks have been organized as relational databases with multilayer structures to support transport management and routing; however, database expandability and update efficiency in new networks and timetables are low due to the strict database schemas. This study aimed to develop multimodal transport networks using a graph database that can accommodate efficient updates and extensions, high relation-based query performance, and flexible integration in multimodal routing. As a case study, a database was constructed for London transport networks, and routing tests were performed under various conditions. The constructed multimodal graph database showed stable performance in processing iterative queries, and efficient multi-stop routing was particularly enhanced. By applying the proposed framework, databases for multimodal routing can be readily constructed for other regions, while enabling responses to diversified routings, such as personalized routing through integration with various unstructured information, due to the flexible schema of the graph database.
From the conclusion: Compared to previous transport networks with a traditional relational database for multimodal routing applications, a graph database is suitable for managing multimodal networks more effectively regarding data integration, expansion, and updates, based on a flexible schema. To evolve existing routing services to personalized and context-aware routing, it is necessary to establish a database combining different typed contextual data with spatial data.

"Graph Database Schema for Multimodal Transportation in Semarang," Panji Wirawan, Djalal Riyanto, Dinar Nugraheni, and Yasmin Yasmin, Journal of Information Systems Engineering and Business Intelligence, Volume 5, page 163, 2019.
https://doi.org/10.20473/jisebi.5.2.163-170

From the abstract: **Background:** Semarang has broad area that cannot be covered entirely by single transportation mode. To reach a specific location, people often use more than one public transportation mode. Apart from Bus Rapid Transit, another exist namely Angkot or city transportation. Multimodal traveler information is then required to help passenger searching for a route. Several studies of multimodal traveler information systems have been conducted, however the data model for multimodal transportation did not conceived in detail.
Objective: Proposes a database of multimodal transportation design using graph data model by taking Semarang as a case study.
Method: We create our model in oriented entity-relationship diagram (O-ERD) and map this O-ERD to the graph database schema.
Result: We develop our data model in graph database schema and we implement the model using Neo4J graph database for validation purpose. Our model consists of three graph node label namely Shelter, Angkot Stopper, and Closer Place. To validate our model, we execute a search query using the Cypher query to look for location with closer place to it.
Conclusion: Our data model was successfully developed and implemented. Searching transportation route in the implementation of our model has been conducted using cypher query. It can successfully display all possible paths and routes. Our query can distinguish between one mode of transportation with another.