jh231004

Representation Learning for Large-scale Geospatial Data Towards Society 5.0

Toyotaro Suzumura (The University of Tokyo)

Abstract

To build robust foundation models in the mobility area, we explored representation learning models primarily aimed at Point-of-Interest (PoI) recommendations this year. The first model, MobGT, utilizes Graph Transformer to process spatial and temporal information from individual movement trajectories as graph structures, significantly enhancing our understanding of mobility patterns. With real-world datasets, MobGT demonstrated an average performance improvement of 24% over existing models. The second model is a semi-multimodal recommendation system that predicts the next PoI a user may visit by integrating textual descriptions, photo data, and user visit history. Our results show that incorporating image descriptions into the recommendation process substantially increases accuracy, offering a robust approach to multimodal data integration for POI recommendations.

1 Basic information

1.1 Collaborating JHPCN centers

- mdx

1.2 Theme area

- Data science/data usage area

1.3 Research area

None

1.4 Project members and their roles

All the following members belong to The University of Tokyo.

- Toyotaro Suzumura (Overall coordination, Model design)
- Hiroki Kanezashi (Model design, Evaluation, Paper writing)
- Chuang Yang (Model design, Evaluation, Paper writing)
- Xiaohang Xu (Model design, Evaluation, Paper writing)
- Matsatoshi Hanai (Model design, Evaluation, Paper writing)

2 Purpose and Significance of the Research

Recently, large-scale movement data collected from smartphone applications and car navigation systems are recorded as sequences of GPS coordinates, representing space-time trajectory data. These data capture the movement behavior of various entities like taxis, pedestrians, and animals with high spatial and temporal precision. In addition to this spatio-temporal information, there is a wealth of text data available, such as place names and spot metadata, which enriches our understanding of regional characteristics and specific locations.

Utilizing representation learning with these diverse data sources, it is feasible to construct foundation models for road traffic that support various machine learning tasks, including POI recommendations and traffic congestion prediction. In recent studies, several representation learning with trajectory data have been propoesd, yet often these efforts have been limited by focusing solely on a single downstream task, thereby constraining the generality and applicability of the models.

In this research, we explore a framework designed for representation learning of foundation models that harness large-scale geospatial data. At the core of this framework lay the integration of a time-series mobility model with a large-scale language model (LLM) that processes textual data related to geographic information. The specific architecture of this framework was illustrated in Figure 1.

In constructing the time-series mobility model (upper part of Figure 1), dynamic data such as vehicle movement trajectories and traffic conditions are inputted, each represented as a graph structure, which is then learned through a Transformer-based spatiotemporal model. Concurrently, in the LLM construction process (middle part of Figure 1), inputs like map data and spot metadata are used to build a knowledge graph of geographic spaces. This process facilitates iterative fine-tuning. Such an infrastructure model is versatile, supporting a wide array of downstream tasks such as estimating travel times, recommending spots to users, and predicting fuel consumption, as depicted in the bottom part of Figure 1.

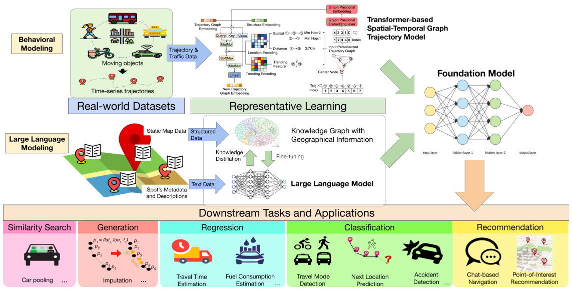

Fig. 1 Large-scale Geospatial Representation Learning Framework

3 Significance as JHPCN Joint Research Project

The final objective of this research is to develop a foundation model that integrates spatiotemporal mobility trajectories of users (vehicles) and textual data extracted from map metadata. Training such a model necessitated significant computational resources, as handling geospatial data with temporal components involved substantial computation. Moreover, processing textual data also requires large computational resources, such as many GPUs, especially when employing Large Language Models (LLMs). With these technical demands, access to robust computational resources was crucial. Utilizing the GPU-equipped mdx environment provided by JHPCN enabled us to further the model' s design and experimentation effectively.

The academic significance of this work lies in its ability to facilitate various downstream tasks, such as estimating travel times, recommending spots to users, and predicting fuel consumption, through the foundation model. This approach eliminates the need to design and pre-train task or dataset-specific models from scratch. Instead, modest fine-tuning adjustments suffice, thereby conserving both computational and operational resources. This advancement aligns with the goals of Japan's Society 5.0 - transforming from "human-based information analysis" to "AI-based machine intelligence." and significantly contributes to the broader application and efficiency of AI technologies in real-world scenarios.

4 Outline of Research Achievements up to FY2022 (Only for continuous projects)

None

5 Details of FY2023 Research Achievements

We have two achievements in this research: (1) the Transformer-based Spatial Temporal Graph Trajectory Model, which utilizes spatial and temporal graphs combined with Transformer mechanisms for POI recommendations, and (2) the Multimodal POI recommendation model that integrates textual and visual information.

5.1 Transformer-based Spatial Temporal Graph Trajectory Model

The foundation model we introduced integrates two key components: a behavioral modeling from time-series trajectory data and a large language model from textual data, which work together to enhance Point of Interest (POI) recommendation systems. First, we proposed the former model, named Mobility Graph Transformer (MobGT), leverages both temporal and spatial data to predict user movement and recommend the next spot a user may visit. This model uniquely captures the dynamic movement trajectories of users, spatial information from maps, and meta-information of spots to provide tailored recommendations.

Although numerous graph neural network (GNN)-based POI recommendation models have been previously proposed, none have effectively combined temporal and spatial information within a GNN framework to capture localized user behavior comprehensively. Our approach with MobGT introduces applications of graph-based methods to represent the temporal and spatial relationships of global POIs and local user mobility patterns as distinct graphs.

The MobGT model utilizes a Graph Transformer-based architecture to process this information, effectively modeling user movement patterns by integrating spatial, temporal, and POI category information. This integration is designed to offer a detailed representation of user behaviors and their interactions with various locations. Additionally, MobGT addresses the challenge of the long-tail distribution in recommendation systems by introducing a new loss function called "Tail Loss," which improves the prediction accuracy for less frequently visited spots.

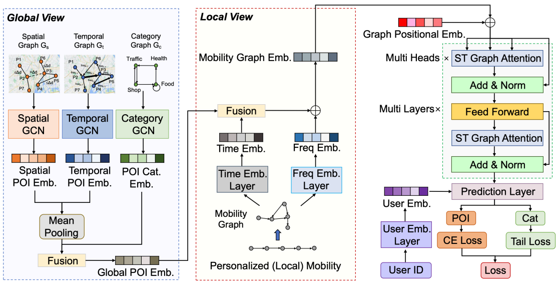

Fig. 2 Overview architecture of MobGT

The structure of the MobGT model is detailed in Figure 2, where it shows how global POI information is synthesized alongside individual user data to generate both a global representation of POIs and a local representation based on each user's movement trajectory.

To validate the effectiveness of MobGT, we conducted performance evaluation experiments using open data of vehicle movement trajectories and data provided by Toyota Motor Corporation. These experiments focused on predicting the next spot a user may visit, and results indicated that MobGT achieved an average improvement of 24% in NDCG and MRR indicators compared to existing POI recommendation models.

This research fills a gap in the existing literature by providing a sophisticated model that not only incorporates but also effectively utilizes temporal and spatial data within a GNN framework to enhance the accuracy and personalization of POI recommendations.

5.2 Multimodal POI Recommendation Model

In our research on representation models utilizing textual data, we investigated POI recommendation models that employ multimodal language representations and large language models (LLMs). These models integrate both textual and image data to enhance the accuracy and relevance of recommendations across various POIs, including restaurants and other venues.

We developed a framework that combines visual data with conventional text-based attributes, such as venue names and locations, as depicted in Figure 3. This framework incorporates LLaVA, a multimodal model that generates textual descriptions from images, and Recformer, a sequential recommendation framework. By transforming visual data into textual descriptions and integrating them with traditional text attributes, the model facilitates enhanced and personalized recommendations.

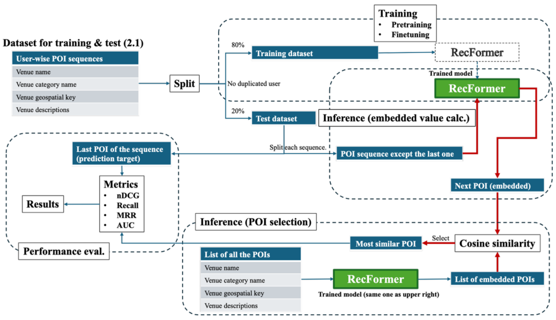

Fig. 3 Flow of Multi-modal POI Recommendation

By incorporating of image data as part of the sequential recommendation input allows for the generation of personalized recommendations that are informed by the user' s visit history and preferences. This approach is particularly effective in categories where visual information plays a crucial role, such as restaurants and tourist attractions. Furthermore, we introduced a novel loss function designed to optimize the interaction between image and text data, thereby improving both the efficiency of model training and the performance of recommendations.

In our evaluation experiments, we utilized a combined dataset from Foursquare for POI check-ins and the FoodX-251 dataset for food images. The proposed semi-multimodal model, which incorporates image descriptions, demonstrated superior performance compared to a baseline model that did not include image descriptions. The inclusion of visual contexts in the recommendation process allows for a more precise capture of user preferences and behaviors, resulting in more accurate and relevant recommendations. This work has been under review by a top-tier recommendation-related international conference called ACM RecSys 2024.

This detailed approach underscores the significant advancements made in leveraging multimodal data for enhancing POI recommendations, bridging the gap between traditional text-based recommendation systems and the dynamic, visually enriched preferences of modern users.

6 Self-review of Current Progress and Future Prospects

In our research on the Transformer-based Spatial Temporal Graph Trajectory Model, we introduced the MobGT model, which leverages spatial and temporal attention mechanisms along with graph structures to recommend Points of Interest (POIs). MobGT integrates information about space, time, and POI categories to construct a global graph, enhancing the learning of higher-order relationships among POIs through the use of a local mobility graph for individual users. A new loss function, Tail Loss, was also developed to address the long-tail distribution problem in POI recommendations. Extensive experimentation on real-world datasets yielded an average improvement of 24% over existing state-of-the-art models. The results of this research were accepted for presentation at the ACM SIGSPATIAL 2023.

For future work on MobGT, it is necessary to test the effectiveness of the current model in different geographic and social contexts, as its performance has so far been validated only on specific datasets. Moreover, exploring methods to more accurately predict users' long-term preferences and behavior patterns remains a critical next step. Addressing these challenges will help expand the model's applicability and refine it into a more versatile and practical POI recommendation system.

In the area of multi-modal POI recommendation, we proposed a new approach that effectively learns from textual attributes of venues and descriptions generated from visual information under geographical constraints. Including image data as part of the textual descriptions improved model performance.

Overall, our research proposed two models: a Graph Transformer-based trajectory model and a multimodal LLM-based POI recommendation model that incorporates visual information. However, to achieve the initial goal of constructing a foundation model, it is essential to develop a unified model that demonstrates applicability across a broader range of downstream tasks and datasets.

As future work, we plan to integrate these proposed models to create a comprehensive foundation model that can be adapted to various mobility-related downstream tasks. This model will be validated and optimized across diverse datasets and applications to ensure its effectiveness and practical utility in real-world scenarios.

7 List of publications and presentations

Journal Papers (Refereed)

Proceedings of International Conference Papers (Refereed)

Xu, X., Suzumura, T., Yong, J., Hanai, M., Yang, C., Kanezashi, H., Jiang, R. and Fukushima, S., 2023, November. Revisiting Mobility Modeling with Graph: A Graph Transformer Model for Next Point-of-Interest Recommendation. In Proceedings of the 31st ACM International Conference on Advances in Geographic Information Systems (pp. 1-10), November, 2023 (Top 10%)

Presentations at International conference (Non-refereed)

Presentations at domestic conference (Non-refereed)

Published open software library and so on

Other (patents, press releases, books and so on)