Data and Resources for Combining Point of Interest Semantics, Locations, and Road Networks

Joseph Zuber
Iowa State University
Ames, Iowa, USA
zubes@iastate.edu

Xu Teng
ESRI
Redlands, California, USA
xteng@esri.com

Andreas Züfle
Emory University
Atlanta, Georgia, USA
azufle@emory.edu

Goce Trajcevski
Iowa State University
Ames, Iowa, USA
gocet25@iastate.edu

ABSTRACT

The advancements in Location Based Services (LBS) and Location Based Social Networks (LBSN) have spurred multiple research efforts in query processing as well as recommendation systems that enable planning trips based on combining location and semantic properties of Points of Interest (POI). However, often times such trips need to involve the reality of existing road networks, for the purpose of obeying constraints such as distance or travel-time. Although there are many publicly available datasets (e.g., Gowalla) that include check-in data at POIs with location, they are often not integrated with existing roads-based data (e.g., Open Street Maps (OSM)) causing researchers to spend extra time and labour to experimentally evaluate their findings. In this paper, we present: (1) methodologies for extracting information regarding POIs from publicly available datasets based on users posting; (2) extracting concise semantic categories for each POI; (3) integrating their location and semantic categories with an existing road network. In addition to the methodologies, we also provide two datasets (based on POIs and road networks in Chicago and New York City) constructed using our methodologies that researchers can readily use for their semantic-aware POIs with location and trip based query processing tasks as well as deep learning tasks.

CCS CONCEPTS

• **Information systems** → **Data extraction and integration**; **Spatial-temporal systems**.

KEYWORDS

Spatial-temporal data, GIS, datasets, data extraction tools, review datasets, semantic data

ACM Reference Format:
Joseph Zuber, Xu Teng, Andreas Züfle, and Goce Trajcevski. 2024. Data and Resources for Combining Point of Interest Semantics, Locations, and Road Networks. In *The 32nd ACM International Conference on Advances in Geographic Information Systems (SIGSPATIAL '24), October 29-November 1, 2024, Atlanta, GA, USA*. ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3678717.3691300

SIGSPATIAL '24, October 29-November 1, 2024, Atlanta, GA, USA

ACM ISBN 979-8-4007-1107-7/24/10.
https://doi.org/10.1145/3678717.3691300

1 INTRODUCTION

The ubiquity of smart mobile devices has stimulated a generation of large geo-textual datasets available on the Web [2]. This data provides users of mobile devices with opportunities to search for content including both keywords and geo-location [6] and to recommend locations and trips, based on particular semantic preferences – core tasks of Location Based Systems (LBS) [17]. Such data also support researchers to improve the understanding of human mobility [14, 15].

Specific downstream applications of LBS enabled by combining semantic and (geo-)location data abound: from targeted advertising [10], through crowdsourcing [11] and route recommendation [13], to group-based POI recommendations [18] and dynamic spectrum allocation [22]. However, from the perspective of research paradigms related to combining spatial and spatio-temporal data with semantic ones, there are two end-points of the spectrum:
(1) Query Processing: Many works have addressed specific objectives such as keyword awareness in optimal routes construction [4], indexing and range/k-NN query processing of semantic trajectories [8], sequential tasks allocation [11], semantic heterogeneity and duration constraints [20], etc. What is common in these works is that data from multiple sources is integrated and indexing structures over combined data are built, that speed up the query processing.
(2) Machine Learning: While differing in the specific application objectives (e.g., specific POI and routes recommendation; Trajectory-User Linking; Anomalies; transportation mode) and architectural models (e.g., Spatio-Temporal (and other variations of) GNN vs. Transformers; Contrastive learning) as well as modeling categories (e.g.., representation learning, meta-learning) [7, 12, 24] – majority of the recent works combine the spatial, temporal and semantic data as features, essential during the training and validation [9]

Regardless of whether a particular problem is based on novel data structures and efficient (exact or approximate) algorithms or on novel (deep) learning models – it is paramount for the development of next generation approaches to be aware of not just diverse time-location information, but also all of the contextual and semantic information surrounding a particular location.

In this paper, we present a methodology for generating a semantic categorization for describing POI locations obtained from publicly available (i.e., users-reported) data and for fusing it with road-network data. In addition to the publicly available source-codes, we also describe (and publicly provide) two readily available datasets that we generated for New York City and Chicago.

Dataset	Attractions	Reviews	Total Words	Earliest Review	Latest Review	Latitudes	Longitudes
Chicago	592	48828	2877118	August 2010[1]	March 2023[1]	41.57 to 42.32	-89.07 to -87.54
New York City	621	32546	1696559	August 2010	May 2020	40.59 to 41.12	-74.15 to -73.77

[1] The Chicago dataset has 1019 reviews without date information.

Table 1: Dataset Details

Dataset	Avg. Reviews per Attraction	Avg. Words per Review	Avg. Reviews per User	Avg. Review Score[2]
Chicago	82.34	58.92	1.38	44.87
New York City	52.24	52.13	1.64	43.03

[2] Allowed scores are 10 to 50, inclusive, in increments of 10

Table 2: Dataset Review Details

2 OVERVIEW OF EXISTING RESOURCES

To position our work in a broader context, we now discuss a few publicly available datasets. We note that, from a broad perspective, there are multiple publicly available geospatial data sources – for example, https://data.gov/ has a dedicated "Geospatial" category.

In the context of problems related to trajectory with semantic awareness of specific locations and with spatio-temporal constraints on the trip, there are several popular public datasets:
– When it comes to planning (predicting and recommending) trajectories with traffic-flow awareness, as well as global/local traffic flow prediction, the Beijing Taxi dataset (https://paperswithcode.com/dataset/taxibj), generated by tracking the motion of taxis in Beijing over a period of time, is quite frequently used. LargeST (https://paperswithcode.com/dataset/largest) is also a popular benchmark for traffic data as it is obtained from 8600 sensors in California (each sensor contains 5 years of data). However, these datasets do not contain much semantic information related to specific locations and do not convey any users experience.
– An orthogonal category of sources stems from individual users' motions as well as reviews that they provide regarding specific locations. One popular source is OpenStreetMap (OSM – https://www.openstreetmap.org/) containing both GPS traces and User Diaries. While often used in research literature, OSM does not provide categorization of semantic attributes and may often be confined to a particular region with no road network. Complementary to OSM, the popular time-at-location (i.e., "check-in") datasets such as Gowalla (https://snap.stanford.edu/data/loc-Gowalla.html), while useful for social network location communication, often lack critical contextual information about a person's visit to a particular location (including the trip taken there).
– Some works, such as [1] and [23], use fine-grained data generators or machine learning models to produce datasets of trajectory and check-in locations, but simulated data like this is not always representative of real-world scenarios and does not offer rich interaction information between agents and the places that they visit.
– Researchers have targeted the problem of providing a semantic annotation for OSM data [16] and, more recently, geo-semantic matching of events and POIs in mobility datasets [2] as well as fusing spatio-temporal features for recommendations based on reinforcement learning [9].

We note also that constructing a dataset using the APIs of popular services can be prohibitively cost-ineffective.

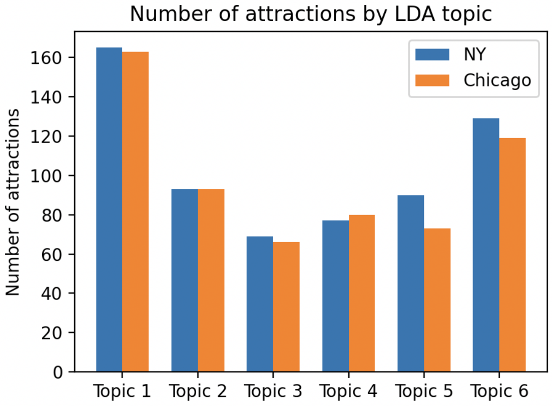


Figure 1: Results of Applying Latent Dirichlet Allocation

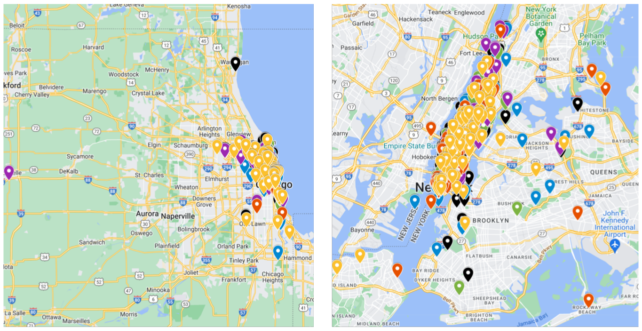
3 DATA SOURCES AND METHODOLOGY

We now discuss the details of our methodology and describe the steps taken to generate some readily available datasets (cf. Sec.4). We note that all the source codes used, as well as the (partial and) integrated datasets are publicly available at https://github.com/Zubes01/POISemanticsLocationAndRoadNetworks.

Firstly, to collect our datasets, we utilized publicly available information from the popular TripAdvisor website. Using a simple webscraper, we collected a list of attractions and their addresses for a desired city by browsing TripAdvisor's "Top things to do in [city]" pages. Review information was collected by navigating to the page associated with each attraction and collecting as much information from each review as possible to enhance the richness of the dataset. The associated attraction, the username of the reviewer, the date the review was posted, the title of the review, the review itself, and the rating given by the reviewer were all extracted by our script.

Table 1 shows the details of the two datasets (Chicago and NYC) for which we ran our scripts, and Table 2 provides more focused statistics of the reviews in each dataset.

To extract the latent topics (and subsequently annotate the sites network accordingly) we used Latent Dirichlet Allocation (LDA) [3]. A vector α of length K is used to parameterize the a priori distribution of topics. The parameter K corresponds to the number of latent topics and for each topic the prior parameter β is used to generate the distribution of words within a topic (for details, see [21]). Fig. 1 shows the distribution of the six topics thus extracted. Since LDA



(a) POIs in Chicago

(b) POIs in New York

Figure 2: POIs from different topics on maps

is unsupervised, one cannot know the exact categories our topics represent. However, they can be inferred based on the highest probability keywords for each topic. In the case of Chicago, the topics have the following components: Topic#1 – monuments and memorials; Topic#2 – parks and beaches; Topic#3 – theaters and shows; Topic#4 – shops and restaurants; Topic#5 – bars; and Topic#6 – Museums. We note that, as a consequence of LDA all the keywords in a particular topic are semantically related. For example, 'monument' and 'war' in Topic#1, as well as 'art' and 'history' in Topic#6. A detailed table with the 10 keywords with highest probabilities in each of the topics is available at [19]. More details about the LDA topics are described at https://github.com/Zubes01/POISemanticsLocationAndRoadNetworks/tree/main/LDA_Model_6.

The next part was to then translate the address information for each attraction into latitude and longitude format. We relied on the Nominatim and ArcGIS Geocoders provided in GeoPy. Upon completion of this translation, the final phase is to snap the attractions to an existing road network. To achieve this we used the road networks provided by Open Street Map. To speed up the procedure, we used an R-tree to obtain the nearest edge (the nearest road) and then used the shortest distance to that edge to place the POI along the road network – however, any map-matching algorithm can be used [5]. Finally, we attached the relevant category label to the POI thus aligned within the road network.

4 OVERVIEW OF INTEGRATED DATASETS

We now briefly describe the integrated datasets for Chicago and New York, along with an application to a specific query (cf. [20]).

Firstly, we note that python scripts were created to take in coordinates, placenames, and assigned LDA Topic #s to create a KML file which can be imported into GoogleMyMaps. Using the scripts one can see the POIs color-coded by their respective LDA Topic # and their locations within the each of the two cities. Fig. 2 provides an illustration (cf. https://www.google.com/maps/d/edit?mid=1nW63RpMGl27QCxE4x6DFDoY4dxmJZ-U&usp=sharing and https://www.google.com/maps/d/edit?mid=1IA36U2EQGXVkHoMgUtO0t8qNq6MknVc&usp=sharing) and using the leftmost part, one can see the list of terms in each of the topics.

The potential benefits of the integrated datasets are for both researchers and practitioners who would like to combine semantic-aware location information along with constraints on a trip, to recommend a specific motion plan. However, the results of multiple queries can also be combined to generate a larger dataset that can subsequently be used for various tasks (e.g., semantic and constraint-aware clustering). As a specific example, Fig. 3 shows a route generated as an answer for a query posed by a user who would like to: (1) Select a starting location for the trip; (2) See as many *diverse* POIs as possible; and (3) Complete the trip within a desired travel-distance limit (based on [20]).

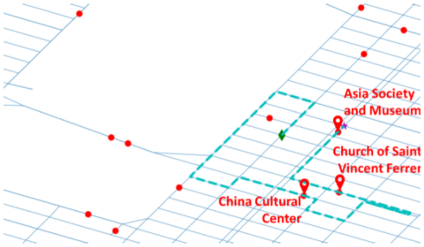


Figure 3: Route in NYC with constraints on origin, distance and semantic diversity of categories for visited POIs

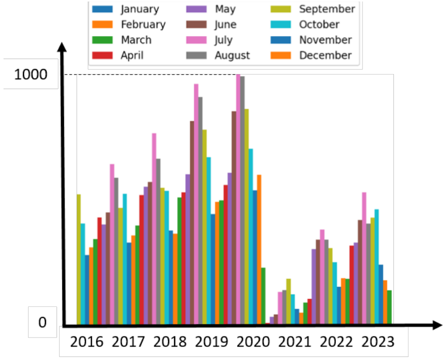


Figure 4: Monthly review counts (subset of Chicago data)

We close this section with a discussion regarding the positioning of our dataset with respect to existing ones. More specifically, since our datasets consist of dated reviews, we can make a comparison to the Gowalla check-in dataset, noting several key differences. Clearly, the Gowalla dataset has much more precise check-in information, with the exact date and time recorded, compared to our date information which is simply a month and a year. Gowalla also has a slightly denser collection of records within the area of our dataset, which we can observe by filtering Gowalla's check-ins by the longitude and latitude ranges we have in our dataset. In New York City, Gowalla includes 136,205 check-ins which is much larger than our 35,546 reviews as seen in Table 1; in Chicago, Gowalla includes 95,117 check-ins which is almost twice as large as our 48,826 reviews as seen in Table 1.

However, (1) our data is much richer in interaction information than Gowalla, including full text reviews and review scores (2) Our data contains much more recent reviews than any Gowalla check-ins, the most recent of which is from October 2010 (3) Our datasets are composed of reviews from a much larger period of time, collected over 116 months for New York City and 150 months for Chicago instead of just 21 months in the Gowalla dataset. For

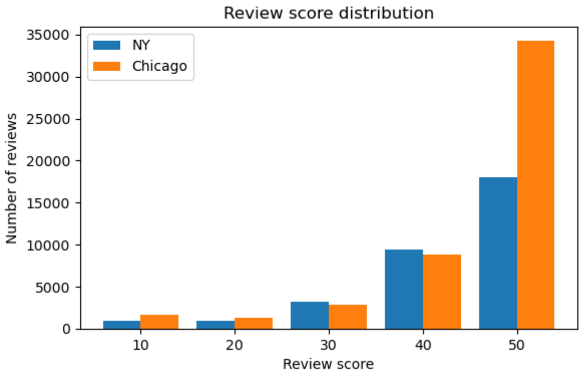


Figure 5: Distribution of Review Scores

an illustration, Fig. 4 shows a monthly distribution of reviews for Chicago (2016-2023 only). We note that both sets show an upward trend until early 2020, when they both fall dramatically due to the pandemic. Fig. 5 shows the distribution of review scores in both New York City and Chicago. Both bar charts show an upward trend, with more reviews at progressively higher scores.

By integrating full text reviews and review scores with date information to the road network, our datasets provide unparalleled opportunities for Machine Learning. For example, sentiment analysis can be trained on the review datasets, while also integrating information about the location, category, and nearby PoI. A Graph Neural Network could also be trained on the dataset to predict the PoI category by utilizing the reviews of nearby PoI and its own.

From a broader perspective, our scripts can be used for other cities, thereby enabling a generation of customized richer datasets than the ones publicly available (e.g., OSM).

5 CONCLUSIONS

In this work, we presented new methodologies for extracting large, information-rich POI datasets from publicly available sources. We explained the process for extracting semantic categories for these POI using Latent Dirichlet Allocation (LDA), making these datasets useful for research in Query Processing and Machine Learning. For use in routing, we also elaborated on the process of snapping locations to road networks using Open Street Map (OSM).

Finally, we introduced two novel datasets that we created using our described procedures: New York City, collected in May 2020; and Chicago, collected in March 2023. These datasets exhibit a density competitive with popular check-in datasets, such as Gowalla, but also offer dramatically more related information such as review scores and review comments which can be used to perform sentiment analysis or semantic categorization as we did using LDA. Our code and datasets are publicly available on GitHub.

REFERENCES

[1] Hossein Amiri, Shiyang Ruan, Joon-Seok Kim, et al. 2023. Massive Trajectory Data Based on Patterns of Life. In *SIGSPATIAL'24*. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3589132.3625592

[2] Ndiouma Bame., Ibrahima Gueye., and Hubert Naacke. 2023. Geo-Semantic Event-POI Matching of Large Mobility Datasets. In *Proceedings of the 12th International Conference on Data Science, Technology and Applications - DATA*. INSTICC, SciTePress, 496–503. https://doi.org/10.5220/0012132700003541

[3] David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *JMLR* 3, Jan (2003), 993–1022.

[4] Xin Cao, Lisi Chen, Gao Cong, Jihong Guan, Nhan-Tue Phan, and Xiaokui Xiao. 2013. KORS: Keyword-aware Optimal Route Search System. In *29th IEEE International Conference on Data Engineering*, Christian S. Jensen, Christopher M. Jermaine, and Xiaofang Zhou (Eds.). 1340–1343.

[5] Pingfu Chao, Yehong Xu, Wen Hua, and Xiaofang Zhou. 2020. A Survey on Map-Matching Algorithms. In *ADC'20*. 121–133. https://doi.org/10.1007/978-3-030-39469-1_10

[6] Zhida Chen, Lisi Chen, Gao Cong, and Christian S. Jensen. 2021. Location- and keyword-based querying of geo-textual data: a survey. *VLDB J.* 30, 4 (2021), 603–640.

[7] Qiang Gao, Fan Zhou, Kunpeng Zhang, Fengli Zhang, and Goce Trajcevski. 2023. Adversarial Human Trajectory Learning for Trip Recommendation. *IEEE Trans. Neural Networks Learn. Syst.* 34, 4 (2023), 1764–1776. https://doi.org/10.1109/TNNLS.2021.3058102

[8] Anasthasia Agnes Haryanto, Md. Saiful Islam, David Taniar, and Muhammad Aamir Cheema. 2019. IG-Tree: an efficient spatial keyword index for planning best path queries on road networks. *World Wide Web* 22, 4 (2019), 1359–1399.

[9] Shenggong Ji, Zhaoyuan Wang, Tianrui Li, and Yu Zheng. 2020. Spatio-temporal feature fusion for dynamic taxi route recommendation via deep reinforcement learning. *Knowl. Based Syst.* 205 (2020), 106302. https://doi.org/10.1016/J.KNOSYS.2020.106302

[10] Kai Li and Timon C. Du. 2012. Building a targeted mobile advertising system for location-based services. *Decis. Support Syst.* 54, 1 (2012), 1–8. https://doi.org/10.1016/J.DSS.2012.02.002

[11] Kun Li, Shengling Wang, Hongwei Shi, Xiuzhen Cheng, and Minghui Xu. 2023. Spatial Crowdsourcing Task Allocation Scheme for Massive Data with Spatial Heterogeneity. *CoRR* abs/2310.12433 (2023). arXiv:2310.12433 https://doi.org/10.48550/arXiv.2310.12433

[12] Hao Liu, Jindong Han, Yanjie Fu, Yanyan Li, Kai Chen, and Hui Xiong. 2023. Unified route representation learning for multi-modal transportation recommendation with spatiotemporal pre-training. *VLDB J.* 32, 2 (2023), 325–342. https://doi.org/10.1007/S00778-022-00748-Y

[13] Aqsa Ashraf Makhdomi and Iqra Altaf Gillani. 2024. Towards a Greener and Fairer Transportation System: A Survey of Route Recommendation Techniques. *ACM Trans. Intell. Syst. Technol.* 15, 1 (2024), 1:1–1:57. https://doi.org/10.1145/3627825

[14] Mohamed Mokbel, Mahmoud Sakr, Li Xiong, Andreas Züfle, et al. 2022. Mobility data science (dagstuhl seminar 22021). In *Dagstuhl reports*, Vol. 12. Schloss Dagstuhl-Leibniz-Zentrum für Informatik.

[15] Mohamed Mokbel, Mahmoud Sakr, Li Xiong, Andreas Züfle, et al. 2024. Mobility Data Science: Perspectives and Challenges. *ACM Transactions on Spatial Algorithms and Systems* (2024).

[16] Michele Ruta, Floriano Scioscia, Saverio Ieva, Giuseppe Loseto, and Eugenio Di Sciascio. 2012. Semantic Annotation of OpenStreetMap Points of Interest for Mobile Discovery and Navigation. In *2012 IEEE First International Conference on Mobile Services, MS*. 33–39. https://doi.org/10.1109/MOBSERV.2012.17

[17] Pablo Sánchez and Alejandro Bellogín. 2022. Point-of-Interest Recommender Systems Based on Location-Based Social Networks: A Survey from an Experimental Perspective. *ACM Comput. Surv.* 54, 11s (2022), 223:1–223:37. https://doi.org/10.1145/3510409

[18] Mayank Singhal and Suman Banerjee. 2021. Group Trip Planning Queries on Road Networks Using Geo-Tagged Textual Information. In *Advanced Data Mining and Applications - 17th International Conference, ADMA (LNCS, Vol. 13087)*. Springer, 243–257. https://doi.org/10.1007/978-3-030-95405-5_18

[19] Xu Teng. 2022. *Semanticaly Diverse and Spatially Constrained Queries*. Ph. D. Dissertation. Iowa State University.

[20] Xu Teng, Goce Trajcevski, and Andreas Züfle. 2023. Distance, Origin and Category Constrained Paths. *ACM Trans. Spatial Algorithms Syst.* 9, 3 (2023), 18:1–18:27. https://doi.org/10.1145/3596601

[21] Xu Teng, Jingchao Yang, Joon-Seok Kim, Goce Trajcevski, Andreas Züfle, and Mario A. Nascimento. 2019. Fine-Grained Diversification of Proximity Constrained Queries on Road Networks. In *SSTD'19*. ACM, 51–60. https://doi.org/10.1145/3340964.3340970

[22] Yonghua Wang, Zifeng Ye, Pin Wan, and Jiajun Zhao. 2019. A survey of dynamic spectrum allocation based on reinforcement learning algorithms in cognitive radio networks. *Artif. Intell. Rev.* 51, 3 (2019), 493–506. https://doi.org/10.1007/s10462-018-9639-x

[23] Ali Zarezade, Sina Jafarzadeh, and Hamid R. Rabiee. 2018. Recurrent spatio-temporal modeling of check-ins in location-based social networks. *PLOS ONE* 13, 5 (05 2018), 1–20. https://doi.org/10.1371/journal.pone.0197683

[24] Fan Zhou, Pengyu Wang, Xovee Xu, Wenxin Tai, and Goce Trajcevski. 2022. Contrastive Trajectory Learning for Tour Recommendation. *ACM Trans. Intell. Syst. Technol.* 13, 1 (2022), 4:1–4:25. https://doi.org/10.1145/3462331