sustainability

MDPI

Article

A Taxi Trajectory and Social Media Data Management Platform for Tourist Behavior Analysis

Pattama Krataithong [1,*], Chutiporn Anutariya [1,*] and Marut Buranarach [2]

[1] ICT Department, School of Engineering & Technology, Asian Institute of Technology, Klong Luang 12120, Pathum Thani, Thailand
[2] National Electronics and Computer Technology Center, Klong Luang 12120, Pathum Thani, Thailand; marut.bur@nectec.or.th
* Correspondence: pattama.krataithong@ait.ac.th (P.K.); chutiporn@ait.ac.th (C.A.)

Abstract: Taxis are commonly used by tourists to travel around unfamiliar cities they visit. These taxis today have GPS devices, which can then be used to collect a significant amount of data on the movement of tourists. One problem with this idea, however, is the question of how to extract that movement data from the raw GPS data, which includes a lot of other data, such as vehicle IDs, timestamps, and speeds, etc. The purpose of this research is to propose a data management platform to process heterogeneous data including taxi data, social media data, and place data for tourist behavior analysis. We propose a data pipeline that can be scaled in order to process a significant amount of data regarding taxi trajectory and social media, with two objectives. The first objective is to extract the tourist trajectory data from the raw GPS data and produce a data integration module enriched with a knowledge base of tourist trajectories. This knowledge base is constructed through the extension of semantic trajectory ontology (STO) and mobility behavior ontology (MBO). The second objective is to extract tourist activities/point of interests (POIs) from geo-tagged Twitter data. The results of the data pipeline can readily be used for tourist behavior analysis, such as tourist descriptive analysis, popular tourist destinations/zones, and tourist movement patterns identification. We leverage the study's results to demonstrate the real-life case study in Bangkok during the Songkran Festival in 2019. Thus, we could precisely identify tourist movement during various periods, determine popular destinations/zones, discover high density density of taxi destination points for a given trajectory type, and display the top ten tourist destinations, as well as prominent tourism keywords or trends at the time. This can provide insight to governments and businesses related to tourism regarding the trajectories and activities of tourists, and it will help predict future tourism trends.

Keywords: taxi trajectory data; social media data; tourist behavior analysis; ontology; semantic enrichment process

Citation: Krataithong, P.; Anutariya, C.; Buranarach, M. A Taxi Trajectory and Social Media Data Management Platform for Tourist Behavior Analysis. *Sustainability* **2022**, *14*, 4677. https://doi.org/10.3390/su14084677

Academic Editors: Elena Carvajal-Trujillo and David Castilla-Espino

Received: 26 February 2022
Accepted: 8 April 2022
Published: 13 April 2022

1. Introduction

Tourism is an important economic sector that generates revenue, employment, and foreign-exchange earnings, all of which contribute to economic development. In tourism economics, it has been shown that tourist visits have a major impact on regional economic growth [1]. Tourism involves the movement of people across time and places, either between their accommodations and destinations or inside destination locations. Analyzing and mining tourist movement behavior about the time/space relationships and activities that tourists have with their destinations has important impacts on infrastructure and transportation development, tourism product creation, and the tourism industry's commercial sustainability. The rapid development of technology based on a location has led to a significant amount of data about trajectories generated by devices equipped with GPS. These data are then used to track moving objects in many domains. GPS is a valuable resource for collecting travel survey data because GPS devices can provide data every

second with reasonably high location accuracy [2]. Several papers leverage GPS data to analyze travel patterns [3–8], daily travel characteristics and activities [9–11], and individual route/destination prediction [12,13]. Taxis equipped with GPS are part of an intelligent transport system. This mode of transportation is preferred by tourists for commuting within a city. The data from a taxi's GPS, however, do not usually include specific data on the passengers' trajectories and activities, which makes analyzing the data for use in the tourism sector more difficult. A data management platform that handles significant amounts of data on taxi trajectories and can isolate more meaningful data would thus be very helpful. Previous studies have designed data management platforms that help us to understand human movement in daily life and other circumstances such as trips for work, education, and shopping. These previously developed platforms can be divided into two categories: (i) semantic web technology based platforms, which stores, manages, and enriches data on trajectories with global data [14–17], and (ii) big data technology based platforms, which support large-scale data about trajectories and improve performance regarding movement analysis [18–20].

Most existing mobility data platforms have been created to store, process, and analyze massive amounts of data in order to better understand object/people movement in general purpose. However, there is still a gap in the creation of data management platform to manage a large amount of data about taxi trajectories and social media, specifically for analysis of tourist behaviors. Taxi GPS data provide travel characteristics and taxi origin/destination (OD) points that can be used to discover tourist movement behavior. To predict the visiting places of tourists, only taxi GPS data are insufficient. Social media is another emerging technology that people can use to generate, share, and exchange information and opinions. Recent studies have demonstrated that social media data enable to promote the tourism industry such as discovering popular tourist places [21–25], comparing domestic and foreign tourists [26] and investigating tourist opinions [27–29]. Therefore, social media data are integrated into our data platform to improve the accuracy of inferring taxi trip purpose as well as to uncover popular tourist destinations and tourism trends. Our research questions (RQs) are formulated as follows:

- RQ1: How could we extract tourist trajectories from taxi GPS data?
- RQ2: How could we extract tourist activities/POIs from social media data?
- RQ3: How could we model and construct a knowledge base for enriching tourist trajectories and classify trajectory types?

This paper proposes a platform, which extends from our previous work by employing a scalable document-based database and semantic web technology [30]. The proposed platform incorporates a semantic enrichment data pipeline with following specific objectives: (i) to semantically extract tourist origin/destination (OD) points and tourist trajectory types from taxi GPS data, and (ii) to infer tourist activities or point-of-interests (POIs) from Twitter data. The results of the data pipeline are tourist trajectories extracted from taxi GPS data, and tourist activities/POIs extracted from Twitter data, to be used for tourist behavior analysis. This study has two main contributions: the data pipeline process and the knowledge base construction process. We want to detail the process of integrating multiple sources of place data and building an ontology knowledge base with place information and rules that classify tourist trajectories regarding activities and expenses. In addition, this study demonstrates how to apply the results of the data pipeline to analyze tourist behavior in Bangkok during the Songkran Festival in 2019.

This paper is organized as follows. Section 2 reviews the related work. Section 3 describes the proposed data management platform. Section 4 presents the developed data pipeline. Section 5 introduces the knowledge base construction process. Section 6 demonstrates examples of tourist behavior analysis and visualization. Section 7 concludes.

2. Related Work

This section will review previously developed data management platforms for movement analysis. There are two main categories of platforms: semantic web technology-based and big data technology-based.

Some previous studies integrated data on trajectories and context using ontologies in order to perform movement analysis. Geo-Ontology [14] is an ontology design pattern, used for modeling concepts and properties of semantic trajectory and related data. The main concepts involved are semantic trajectories with segments and fixes. A segment is a entity that begins and ends at a fix. The fixes have locations and times as well as sources that indicate the originating person or device. Baquara2 ontology [16] also produces a conceptual model from data on trajectory and context. A moving object's sequence comprises movement segments, which are represented as tuples, including type, geometry, beginning and ending positions, time, other possible relationships, order, and annotations. Some studies apply ontology and linked data to enrich the trajectory data's semantic information in their proposed process. Segmented Trajectory Ontology [17] integrates trajectory data with global data with linked open data technology. Suitable semantic entities are pulled from linked data sources such as DBpedia and FOAF and used to enrich segments. The mobile object trajectory framework proposed by [15] supports trajectory data mining by enriching trajectory data with OpenStreetMap using linked data technology. This produces a knowledge graph that users can query using SPARQL language.

GPS-embedded devices can accurately track position data of things and people via mobile phones, buses, subways, taxis, etc. [10]. Cloud computing platforms have been applied in many studies to handle large-scale trajectory data for movement analysis. The data management system proposed in [18] used a cloud computing platform to support mobile data analysis. The proposed system had a core system design based on Hadoop and HIVE and included data processing, storage, and visualization. Refs. [19,20,31–33] also focus on big trajectory data management and analysis. The UlTraMan data management platform [20] extended Apache Spark and integrated Chornical Map (a key-value store) in order to make data retrieval, aggregation analyses, and pattern mining of trajectory data more scalable, efficient, and flexible. The trajectory data management platform proposed in [19] employed cloud computing, namely, the Microsoft Azure package including Azure Table, Azure Redis, and Azure Storm, and was used for preprocessing, indexing, and query processing. Similarly, the study [32] processed a large amount of NYC taxi traces on Spark platform to increase the reasonable scheduling of taxis and reduces the waste of resources. Ghosh et al. [31] developed Traj-Cloud, a geographical trajectory data management platform that focuses on storing, retrieving, and processing mobility traces for real-time applications. The platform was built on Google Cloud Platform (GCP) with the study results based on the NYC taxi traces. It consists of three main modules: (i) trajectory-indexing to efficiently handle large real-time trajectory updates; (ii) geo-tagging, map-matching services; and (iii) trajectory-processing to resolve mobility-based spatio-temporal queries. Kong et al. [33] proposed a method and an application of big data mining for massive taxi traces using MapReduce. The MapReduce distributed computing framework in Hadoop was used with the mining approach to extract the feature points of the taxi driving trajectory, and to conduct hotspot analysis. The analytical framework proposed in [7] was developed through the use of big data technology (e.g., MapReduce, HBase, HDFS) and analyzes the movement patterns of tourists in Xi'an, China, from the large-scale mobile tracking data of 12 million users.

Ontologies are used to make data on human movement more meaningful and to store trajectory data in an RDF repository. These ontologies can then be used to help analyze human movement behavior. As the volume of available trajectory data increases, however, there are serious performance limitations when using RDF repository, especially regarding querying and analyzing such a massive amount of data. This paper seeks to use big data technology to help improve performance. We propose a data management platform that manages tourist trajectory data by using a scalable document store. We have also developed

an ontological knowledge base that can be used to enrich trajectory data and categorize tourist trajectories to aid further analysis of tourist behaviors.

3. Proposed Data Management Platform for Tourist Trip and Analysis

The data management platform stores and manages data on taxi trajectory, social media, and place, and can be used for analysis of tourist behavior. The data process proposed in this study identifies tourist trajectories from taxi origin and destination (OD) points. Moreover, we propose a data extraction process for tourist activities from social media data, such as Twitter data. The platform thus collects data, prepares data, integrates data with a knowledge base in order to enrich and classify trajectories, and extract and classify social media data. Figure 1 shows the eight components of the developed platform.

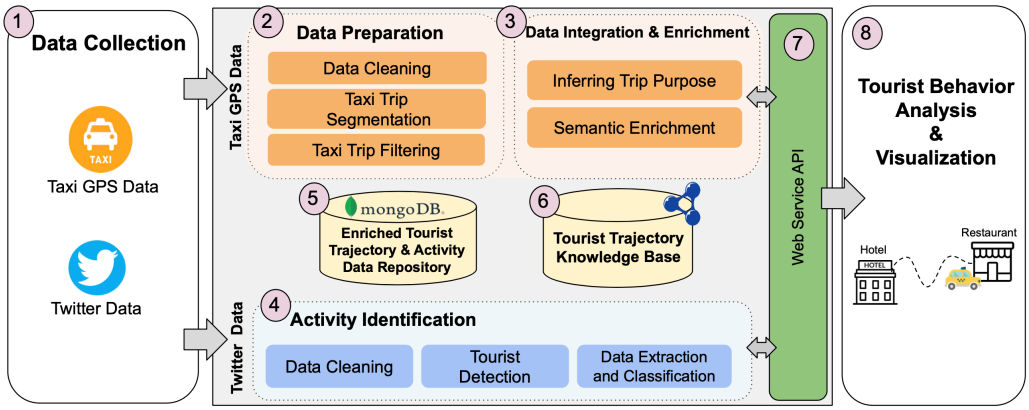

Figure 1. System Architecture.

1. Data Collection Module gathers taxi GPS data and social media data for tourist behavior analysis. The GPS data comes from the ITIC Foundation (https://org.iticfoundation.org/download (accessed on 25 February 2022)), which publishes the data as a CSV file. Each of these files includes a day's worth of data from the GPS instruments of a taxi traveling in Thailand. Second, we crawl Twitter data via Search Tweet: Full-Archive API (https://developer.twitter.com/en/docs/twitter-api/premium/search-api/quick-start/premium-full-archive (accessed on 25 February 2022)) to acquire Twitter data in 2019, then collect data in CSV file format.

2. Data Preparation Module prepares the GPS data by data cleaning, taxi trip segmentation, and taxi trip filtering. This process removes invalid data, separates out taxi trajectories with passengers, and identifies which trajectories can be used for the study area.

3. Data Integration and Enrichment Module has two parts. The first part, inferring trip purpose, aims to predict passengers' destination. The second part, semantic enrichment, enriches the trajectory data by integrating it with information on tourist places, such as name, category/subcategory, location, rating, and opening hours. The trajectories are then classified with a knowledge base about tourist trajectories.

4. Activity Identification Module prepares and extracts tourist activities from social media data consisting of three components, including data cleaning, tourist detection, and data extraction and data classification, which cleans a dataset, identifies tourists from user location, then extracts and classifies tourist activities/POIs from social media messages.

5. Enriched Tourist Trajectory and Activity Data Repository stores data on tourist trajectories and social media with a document-based database called MongoDB. The database supports a significant amount of data and scales it with a cloud computing platform [34,35]. It also includes functions that manage geospatial data, including geospatial indexes and geospatial query.

6. Tourist Trajectory Knowledge Base enriches and classifies the data by connecting with the Data Integration and Enrichment Module, making the trajectory data more meaningful.

7. Web Service APIs exchange data among systems and are used in this work to connect the four modules: Data Integration and Enrichment, Enriched Tourist Trajectory Data Repository, Tourist Trajectory Knowledge Base, and Tourist Behavior Analysis and Visualization. There are seven core APIs for data collection and retrieval across systems:

- *addTouristTrajectory* collects enriched tourist trajectory data, such as taxi OD points, place names, place categories/subcategories at OD points, total distances, and total travel time.
- *addActivityAndPOI* collects enriched twitter text, such as tweet data, tweet data, enriched with information on tourist activities/POIs and the user country.
- *getTouristTrajectory* returns a list of enriched tourist trajectories used for tourist behavior analysis and visualization.
- *getActivityAndPOI* returns a list of enriched twitter text used for tourist behavior analysis and visualization.
- *getPlaceList* returns a list of places after submitting a certain area, such as the Pathum Wan district, which is used for data integration and enrichment.
- *getPlaceInformation* returns a specific place information, such as place name, address, place category/subcategories, rating, and opening hours, which is used for data integration and enrichment.
- *classifyTouristTrajectory* returns tourist trajectory types after submitting passenger's probable place category at taxi OD points, which is used for data integration and enrichment.

8. Tourist Behavior Analysis and Visualization Module analyzes the movement behavior of tourists, such as activities and movement patterns.

4. Data Pipeline for Tourist Trajectory Extraction and Enrichment

Figure 2 shows the data pipeline developed from the proposed system architecture outlined above. To address the defined RQ1 and RQ2, the pipeline includes two main processes: (i) extraction of tourist trajectories from taxi GPS data, and (ii) extraction of tourist activities from social media data.

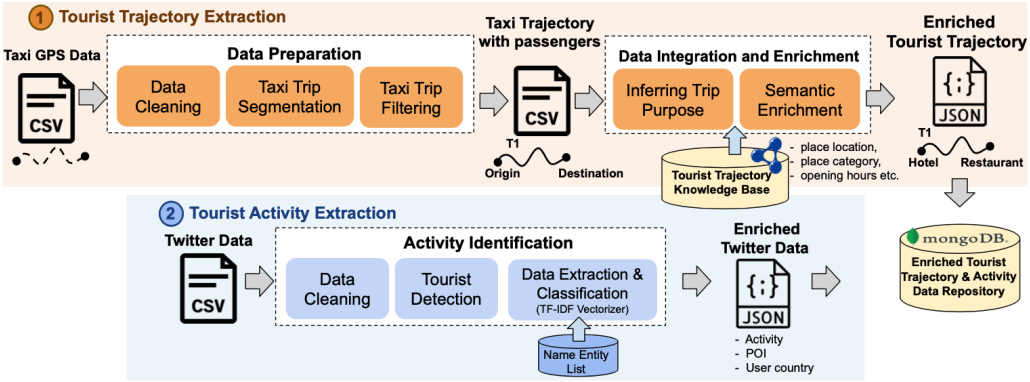

Figure 2. Tourist Trajectory Data Pipeline.

4.1. Tourist Trajectory Extraction from Taxi GPS Data

Taxi GPS Data are inputted into the data pipeline process and tourist trajectories are extracted. Taxi GPS data from the Songkran Festival period (11–17 April 2019) was obtained from the ITIC Foundation (https://org.iticfoundation.org/download (accessed on 25 February 2022)) in CSV file format. Each dataset included GPS data from taxis traveling within Thailand from 00:00 to 23:59 each day. Taxi GPS data are generated every 1–2 min in general. The data includes vehicle ID, location (latitude and longitude), date&time, speed, direction, and hired status, as can be seen in Figure 3a. The datasets we acquired show that

about 4225 taxis in Bangkok have a GPS device installed. The data we obtained includes more than 2.4 million records, with about 97,923 unique trips.

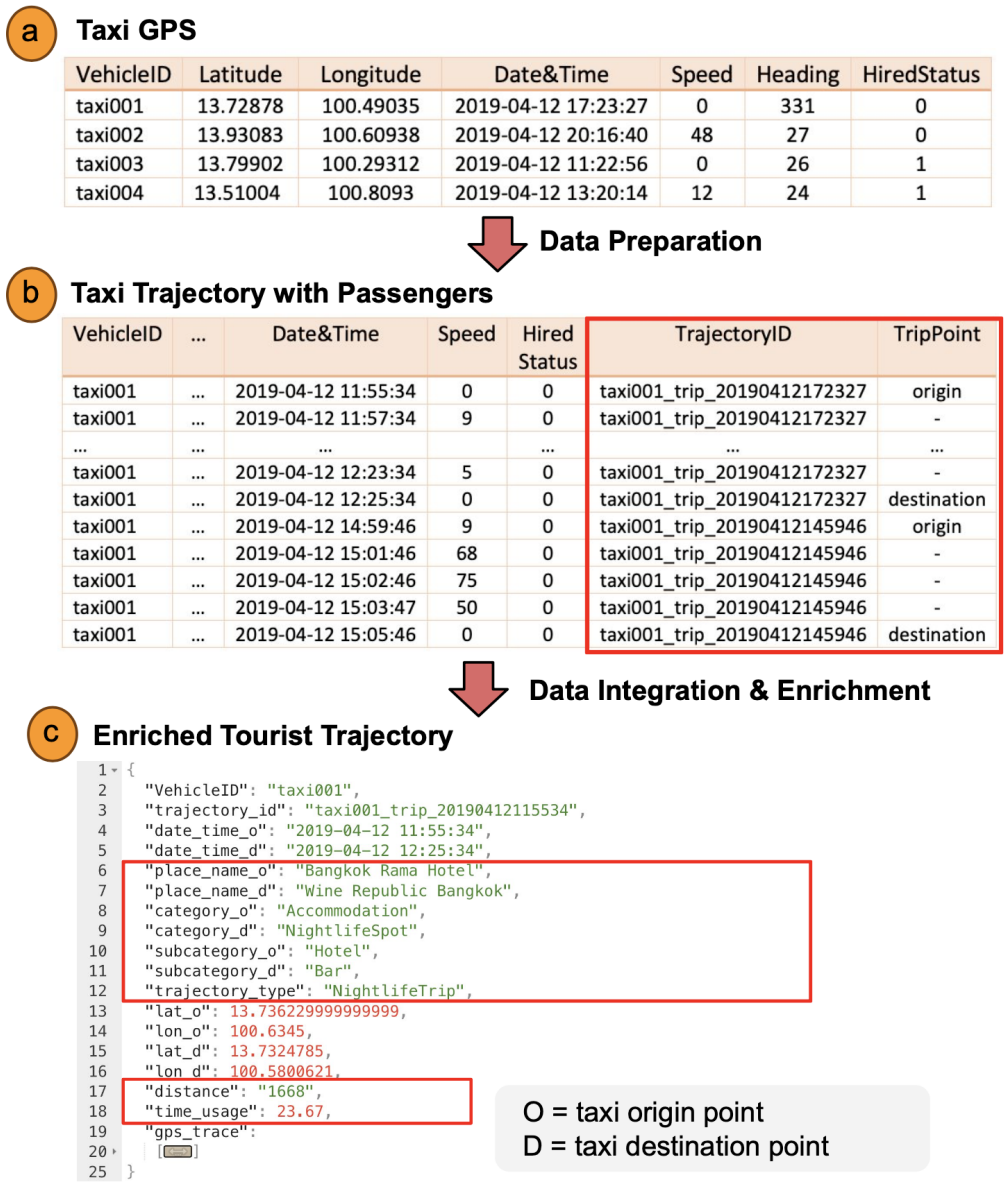

Figure 3. Example of tourist trajectory extraction process: (**a**) Taxi GPS data; (**b**) Taxi trajectory with passengers enriched with trajectory ID, and origin and destination points; and (**c**) Tourist trajectory data enriched with place and trip information.

Data Preparation cleans and extracts segmented taxi trajectories from raw taxi GPS data. There are three sub-processes in this module: data cleaning, taxi trip segmentation, and taxi trip filtering. The data are cleaned to remove taxi GPS data that is invalid because of GPS device malfunction. To do this cleaning, duplicate records are removed from the CSV file of taxi GPS data. After that, the dataset only includes records describing unique taxi trajectories with data that includes vehicle ID, location, timestamp, speed, direction, and hired status. Taxi trajectories are typically a mix of taxis carrying passengers and taxis without passengers. To segment taxi trajectory, we group trajectories by vehicle ID, and then sorted records using timestamp. Taxi trip segmentation refers to identifying the taxi trajectories that have passengers (i.e., hired status = 0). Each segmented trajectory is then given a trajectory ID, and origin and destination points are identified as shown in Figure 3b. Taxi trajectories are removed if they meet the following conditions:

- Only one point (GPS Error),
- Incorrect Date (Ex. 20 July 1963),
- Total distances < 500 m.

Finally, the remaining segmented trajectories are filtered based on particular areas of interest. The results are published and are available via the IEEE data port (http://ieee-dataport.org/9158 (accessed on 25 February 2022)).

Data Integration and Enrichment enhances the segmented taxi trajectories by using related data. This process identifies possible tourist attractions that may have been passenger destinations and then classifies the type of taxi trajectory. There are two sub-processes involved here: inferring the trip purpose and semantic enrichment. Inferring the trip purpose identifies possible tourist places at the taxi's OD points. The study aims at tourist trajectory extraction among foreign tourists traveling by taxi. Hence, one of these trajectories is considered a tourist trajectory if it originates at an accommodation and ends at a tourist place [36]. In this step, we filter the segmented taxi trajectories in Bangkok that have the taxi origin points close to the accommodations by 200 m. Then, we employ the probabilistic model to determine a passenger's possible destination using information from the taxi and nearby places, such as drop-off time, drop-off point, opening hours, and POI rating. Then, the rules defined in the tourist trajectory knowledge base are used to classify tourist trajectory types by submitting the probable OD place category at taxi OD points through the web service API. The result of this API will return a trajectory type to the user. Another sub-process is semantic enrichment, which refers to the process whereby additional related data are integrated into the data about place and trajectory. This process employs the constructed knowledge base described in Section 5.

Enriched Tourist Trajectory Data is produced and stored in the document repository, and is also accessible via the IEEE data port (http://ieee-dataport.org/9175 (accessed on 25 February 2022)). This data can then be used for future analysis of tourist behavior. An example of this trajectory data enriched with additional meaningful data on places and trips (e.g., taxi OD points, place names, place categories/subcategories at OD points, total distances, and total travel time) is shown in Figure 3c.

4.2. Tourist Activity Extraction from Social Media Data

Social Media Data is input into the data pipeline process to extract tourist activities from social media data. In an experiment, we gathered Twitter data during the Songkran Festival between 11–17 April 2019 in Bangkok. The datasets were collected in CSV file format. Typically, Twitter data contains geo-tagged information such as tweet message, tweet location, timestamp, and user location as shown in Figure 4a. According to our dataset, approximately 3500 unique tweets from users who tweeted within the Bangkok area were collected. However, the dataset was rather small because we filtered for only tweet data with user location to identify them as tourists.

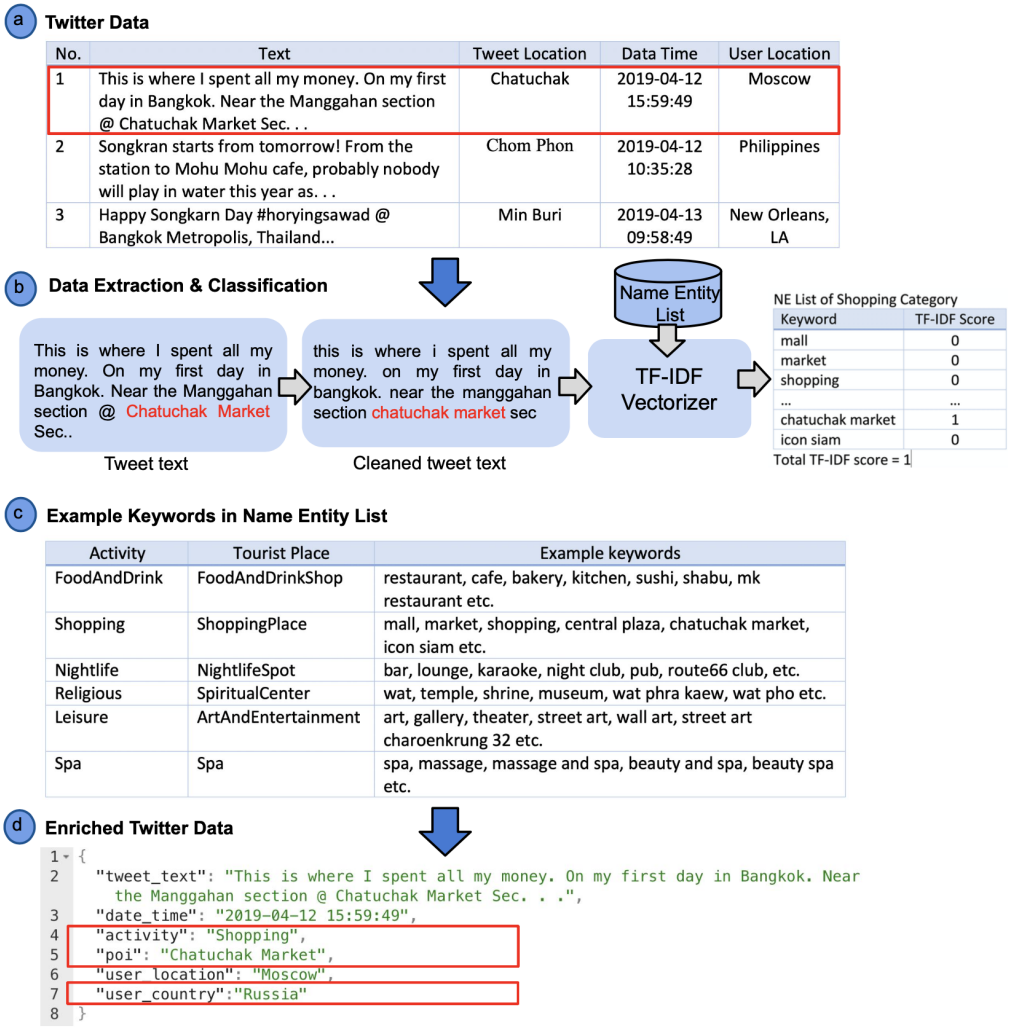

Figure 4. Example of Tourist Activity Extraction Process: (**a**) Twitter data; (**b**) Data Extraction & Classification; (**c**) Example keywords in name entity list; and (**d**) Twitter data enriched with activity, POI, and user country.

Activity Identification extracts and classifies tourist activities/POIs from Twitter data. This module consists of three sub-processes: data cleaning, tourist detection, and data extraction and classification.

- **Data Cleaning** is the process used to remove duplicate records.
- **Tourist Detection** is the process to identify tourists from tweet data. In our study, user location is an important field used to identify tourists [37]. Normally, users can input their locations such as city, country, location (latitude, longitude) into this field. However, users can input any texts, so it is difficult to identify the real locations of users. Here, we use a semi-automatic approach using a string matching between a predefined city/country list and a user location of the tweet data, and then manually verifying the results. For example, for a user location "Moscow", the user country "Russia" is stored. Note that, we filter only oversea tourists.
- **Data Extraction and Classification** is the process to extract and classify tourist activities/POIs from tweet texts. We focus on six tourist activities related tourist trips to be elaborated in Section 5 including FoodAndDrinkTrip, ShoppingTrip, NightlifeTrip, ReligiousTrip, SpaTrip, and LeisureTrip. We first clean tweet texts by removing special characters such as emotion icons and URL, then changing tweet messages to small capital. The TF-IDF score is calculated to describe the similarity between each tweet text and a name entity (NE) list of each activity category. Figure 4c presents sample

words in a NE list of tourist places and activities that match the place category. The results return a frequency score of each document; a document of activity category which has the highest frequency score is the answer. Figure 4b demonstrates an example of tourist activities/POIs extraction and classification from tweet text. Consider as an example, the tweet text: "This is where I spent all my money. On my first day in Bangkok. Near the Manggahan section @ Chatuchak Market Sec…", the process works as follows:

- Remove emotion icons, URLs, and special characters from tweet text. The result of cleaned text is "this is where i spent all my money. on my first day in bangkok. near the manggahan section chatuchak market sec",
- Calculate the TF-IDF score of the relevant document between the above tweet text and a NE list for all categories. Because the tweet text contains the terms "chatuchak market" which is related to the ShoppingPlace category in Figure 4c and has the highest score among the other categories, this twitter text is classified as Shopping activity.
  Note that, if the highest TF-IDF scores of the documents are the same, the answer of such a tweet could be many activities, such as Shopping and FoodAndDrink.

An Enriched Twitter text is then generated and collected in the document repository. Figure 4d presents an example of tweet data, enriched with information on tourist activities/POIs and the user country.

5. Tourist Trajectory Knowledge Base Construction

Construction of tourist trajectory knowledge base has several advantages. Not only can it lead to more meaningful trajectory data with OD point names and categories/subcategories, but it can also be used to explain tourist activities. This allows for a stronger platform that offers many functionalities to answer the defined RQ3: (i) a semantic model of tourist movement behavior can be defined to include places, taxi trajectories, and other contextual data such as events and festivals; (ii) rules classifying tourist trajectories can be formulated, which can help draw inferences about tourist activity; and (iii) Semantic research and data integration. The proposed process for constructing this knowledge base is illustrated in Figure 5. This process has three steps: ontology design and development, tourist place data collection and integration, and contextual data integration. The ensuing subsections will further elaborate on this process.

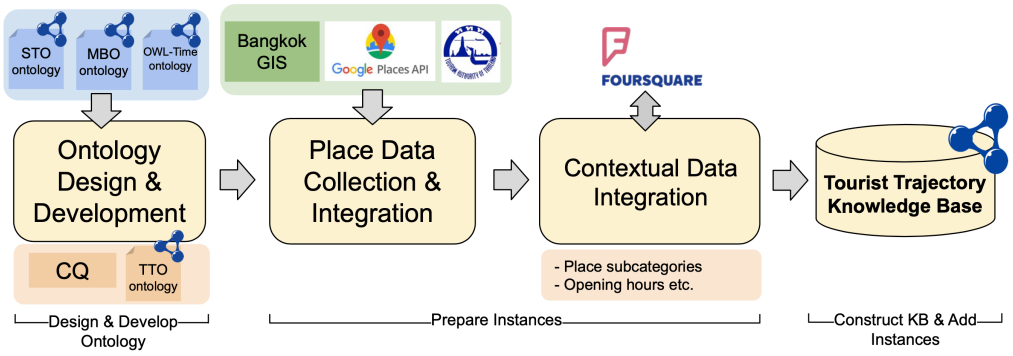

Figure 5. Knowledge Base Construction Process.

5.1. Ontology Design and Development

5.1.1. Ontology Requirements

The proposed ontology aims to model the following: (i) data and attributes of tourist places, divided into categories and subcategories, including accommodation, transportation, and tourist places, (ii) data regarding taxi trajectories, including segments and OD points that connect to tourist places, and (iii) a conceptualization that helps to categorize different types of tourist trajectories based on origins, destinations, and other information such

as visiting date, and time. Importantly, this ontology needs to answer the following competency questions (CQs):

- CQ1: What shopping places are located in Pathum Wan district?
- CQ2: What nightlife venues are still open at the present time?
- CQ3: What possible trajectory types exist for a tourist traveling from an origin point A to a destination point B?

These CQs determine the ability of the ontology to do the following: (i) Model tourist and place data to enrich taxi trajectory data (CQ1–2), and (ii) Categorize tourist trajectories according to OD points (CQ3).

5.1.2. Ontology Development

There are three main parts of the developed ontology, the tourist trajectory ontology (TTO) (https://purl.org/tourist-trajectory (accessed on 25 February 2022)): taxi trajectory data, place data, and tourist movement behavior. To model taxi trajectory data, the ontology extends from the semantic trajectory ontology (STO) [14], which can be used with taxi trajectories since they have segmented trajectories and OD points. STO has three main concepts which are `sto:SemanticTrajectory` , `sto:Segment`, and `sto:Fix`. This is shown in Figure 6a. Regarding place data and tourist movement behavior, the design concept of mobility behavior ontology (MBO) [36] was employed. TTO adds the class `tto:Place` and according subclasses based on Foursquare Categories (https://developer.foursquare.com/docs/build-with-foursquare/categories(accessed on 25 February 2022)), representing tourist destinations that illustrated in Figure 6b. TTO also includes the class `tto:Geometry`, describing a place's geometry type. To model tourist trajectories, TTO has the classes `tto:TouristActivityTrajectory` and `tto:TouristSpendingTrajectory`. The axioms of these two classes define types of tourist trajectories based on the category of place at the OD points. TTO also uses the OWL-Time ontology [38] in order to account for time-related elements such as opening ours and travel time.

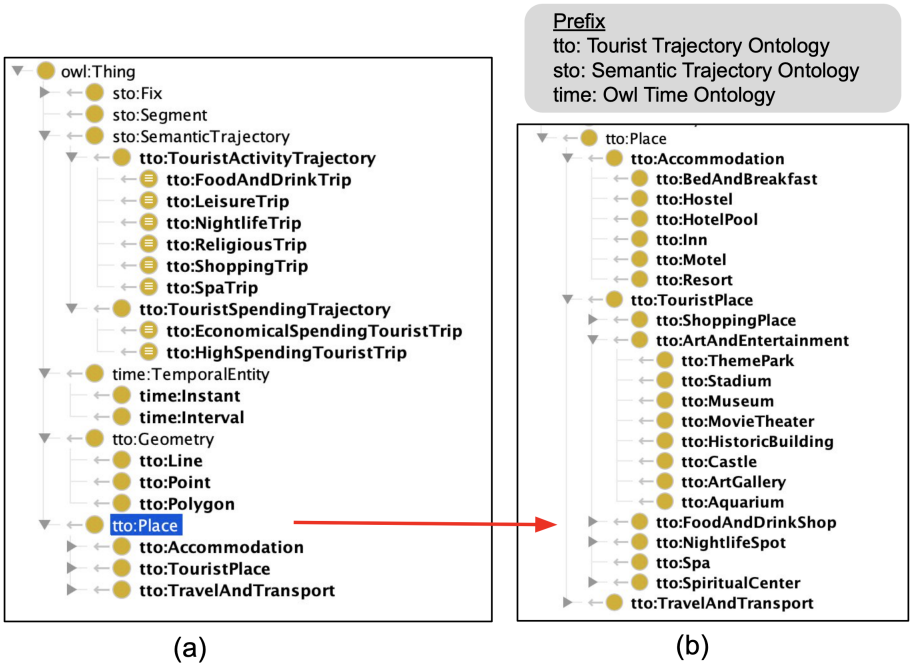

Figure 6. (**a**) Class hierarchy of the Tourist Trajectory Ontology (TTO); (**b**) Class tto:Place and its subclasses.

Concept Hierarchy—To meet the requirements of ontology, the proposed concepts are divided into six top-level classes together with their corresponding subclasses as elaborated by Table 1.

Table 1. Class/Subclass and Description of TTO.

Class/Subclass	Description
sto:Segment	A segment of a taxi trajectory
sto:Fix	The origin and destination of a taxi trajectory. It also connects with tto:Place class
time:TemporalEntity	Time at taxi OD points and opening hours of tourist places. It also connects with sto:Fix and tto:Place class.
tto:Place	The place data such as tto:TouristPlace, tto:Accommadation.
tto:Geometry	The geometry type (Point, Polygon, and Line) of a place.
sto:SemanticTrajectory	Taxi trajectory data with passenger, consisting of two subclasses: tto:TouristActivityTrajectory and tto:TouristSpendingTrajectory
tto:TouristActivity Trajectory	Tourist trajectories based on tourist activities, and consisting of six defined classes including tto:ReligiousTrip, tto:FoodAndDrinkTrip, tto:NightlifeTrip, tto:ShoppingTrip, tto:SpaTrip, and tto:LeisureTrip
tto:ReligiousTrip	A taxi trip that begins at an accommodation and ends at a spiritual center such as temple, shrine, church, and mosque. (hasStartPlace some (tto:BedAndBreakfast and tto:Hostel and tto:HotelPool and tto:Inn and tto:Motel and tto:Resort)) and (hasEndPlace some tto:SpiritualCenter)
tto:FoodAndDrinkTrip	A taxi trip that begins at an accommodation and ends at a food/drink shop such as restaurant, cafe, coffee shop, and dessert shop. (hasStartPlace some (tto:BedAndBreakfast and tto:Hostel and tto:HotelPool and tto:Inn and tto:Motel and tto:Resort)) and (hasEndPlace some tto:FoodAndDrinkShop)
tto:NightlifeTrip	A taxi trip that begins at an accommodation and ends at a nightlife spot such as bar, nightclub, pub, and lounge. (hasStartPlace some (tto:BedAndBreakfast and tto:Hostel and tto:HotelPool and tto:Inn and tto:Motel and tto:Resort)) and (hasEndPlace some tto:NightlifeSpot)
tto:ShoppingTrip	A taxi trip that begins at an accommodation and ends at a shopping place such as shopping mall, super market, and market. (hasStartPlace some (tto:BedAndBreakfast and tto:Hostel and tto:HotelPool and tto:Inn and tto:Motel and tto:Resort)) and (hasEndPlace some tto:ShoppingPlace)
tto:SpaTrip	A taxi trip that begins at an accommodation and ends at a spa. (hasStartPlace some (tto:BedAndBreakfast and tto:Hostel and tto:HotelPool and tto:Inn and tto:Motel and tto:Resort)) and (hasEndPlace some tto:Spa)
tto:LeisureTrip	A taxi trip that begins at an accommodation and ends at a leisure place such as museum, art gallery, and movie theater. (hasStartPlace some (tto:BedAndBreakfast and tto:Hostel and tto:HotelPool and tto:Inn and tto:Motel and tto:Resort)) and (hasEndPlace some tto:ArtAndEntertainment)
tto:TouristSpending Trajectory	Tourist trajectories based on tourist spending, and consisting of two defined classes including tto:HighSpendingTouristTrip and tto:EconomicalSpendingTouristTrip
tto:HighSpending TouristTrip	A taxi trip that begins at a hotel with pool or a resort and ends at a tourist place. (hasStartPlace some (tto:HotelPool and tto:Resort)) and (hasEndPlace some (tto:ArtAndEntertainment and tto:FoodAndDrinkShop and tto:NightlifeSpot and tto:ShoppingPlace and tto:Spa and tto:SpiritualCenter))

Table 1. *Cont.*

Class/Subclass	Description
tto:EconomicalSpending TouristTrip	A taxi trip that begins at an hostel or a motel or an inn or a bed and breakfast and ends at a tourist place. (hasStartPlace some (tto:Hostel and tto:Motel and tto:Inn and tto:BedAndBreakfast)) and (hasEndPlace some (tto:ArtAndEntertainment and tto:FoodAndDrinkShop and tto:NightlifeSpot and tto:ShoppingPlace and tto:Spa and tto:SpiritualCenter))

Object Properties—The eight object properties that describe the trajectory data are defined, as shown by Table 2.

Table 2. Object Properties and Description of TTO.

Object Property	Description
sto:hasSegment	A relation from sto:SemanticTrajectory to sto:Segment.
sto:startsFrom	A relation from sto:Fix to sto:Segment. It specifies the relation "from point" of a segment.
sto:endsAt	A relation from sto:Fix to sto:Segment. It specifies the relation "to point" of a segment.
sto:hasPlace	A relation from sto:Fix to tto:Place.
tto:hasOpeningHours	A relation from tto:Place to time:Interval. It specifies the opening hours of a tourist place.
tto:hasGeometry	A relation from tto:Place to tto:Geometry. It specifies the geometry type of a place.
time:hasBeginning	A relation from time:Interval to time:Instant. It specifies the opening time of a tourist place.
time:hasEnd	A relation from time:Interval to time:Instant. It specifies the closed time of a tourist place.

5.2. Tourist Place Data Collection and Integration

In this step, data from numerous sources and with various data schemas were collected and merged. Our study collected data from Bangkok GIS (http://www.bangkokgis.com/ (accessed on 25 February 2022)), Google Place (https://cloud.google.com/maps-platform/places (accessed on 25 February 2022)), and the Tourism Authority of Thailand (TAT) (http://developers.tourismthailand.org/console/ (accessed on 25 February 2022)). To account for data duplicates, the fuzzywuzzy package, a python library, was used. We refer to the algorithm for checking similar place names in different sources using fuzzy string-matching technique in [30]. This algorithm compares POIs from two different sources with the *checkSimilarity* function, and returns a list of place names with a similarity score, some of which may be duplicated. The next step was to conduct a manual review and remove duplicate place names, maintaining only unique ones.

Merging place data from multiple sources required designing a core schema with important attributes such as place ID, place name, place category, address, location (latitude, longitude), geometry type, opening hours, contact, and source. Table 3 shows the list places left in the data integrated from three different data sources.

Table 3. Tourist place data collection.

No.	Place Category	Number of Places in the Collection
1	Accommodation	1784
2	FoodAndDrinkShop	3846
3	Spa	2187
4	NightlifeSpot	489
5	ShoppingPlace	419
6	SpiritualCenter	88
7	ArtAndEntertainment	24
8	Transportation	82
	Total	8919

5.3. Contextual Data Integration

Once place data are collected, it can be enriched with contextual data. The contextual data used in this study included related place data and other contextual data. The related place data included information such as opening hours and place subcategory. These details help us to determine trip purpose and provide insight into tourist movement patterns and activities. Opening hours is an important field that can help us to determine trip purpose [9,39]. We determined common business hours for different place categories (e.g., shopping malls are open 10:00–21:00 every day, museums are open 9:00–16:00 but are closed on Mondays). This information can help infer the type of tourist with a given trajectory. A tourist staying in a motel, inn, or hostel, for example, is likely an economically spending tourist. Developing an understanding of the behavior of different tourist groups can be used to help businesses or organization to market to or fill the needs of specific groups of tourists. Foursquare API, a social media platform with which users share locations, can be used for a more detailed categorization of places. The subcategories identified with Foursquare API are more detailed than those from Google Place, TAT, and Bangkok GIS. A place such as 'VX The Fifty' is categorized as an accommodation, for example, but Foursquare would add the subcategory of hostel. This step generates tourist place data divided into detailed subcategories. Other contextual data that may impact tourist travel includes events, festivals, tourism campaigns, or environmental conditions. Future work should also include other contextual data in the analysis.

Once the data was merged, the tourist place data was mapped as concept instances into our ontology (TTO) using the OAM Framework [40]. The OAM Framework provides database-ontology mapping modules and publishes an RDF dataset. An example (tto:Iconsiam) is shown in Figure 7. This data point is in the class `tto:TouristPlace`, subclass `tto:ShoppingPlace`. The other descriptive details include `tto:hasPlaceName`, `tto:hasSubdistrict`, `tto:hasDistrict`, `tto:hasProvince`, and `tto:hasRating`. The specific geometric coordinates of the place as well as its opening times are also defined.

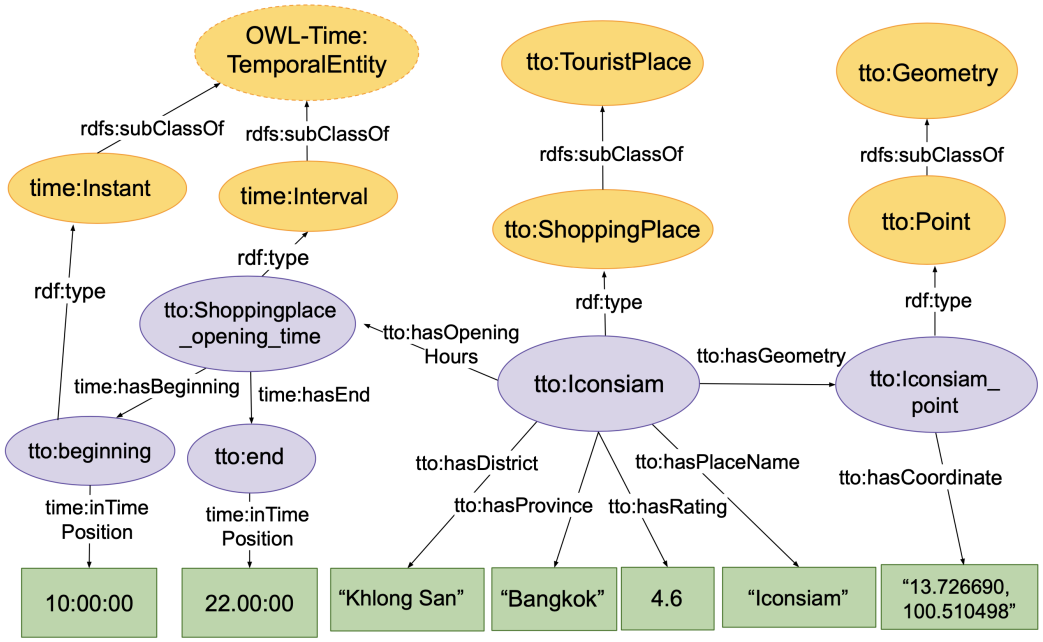

Figure 7. Example of Place Data Instance.

5.4. Example Usage and Queries

Examples of SPARQL query are shown in Figure 8a,b. In Figure 8a, the query asks to retrieve all shopping places in Pathum Wan district (per CQ1). This query results in a list of all instances of ShoppingPlace subclass with place names, locations, and opening and

closing times. Another example, this time per CQ2, is shown in Figure 8b. This query asks to retrieve all the nightlife spots open at the present time.

Figure 8. Examples of queries and results of CQ1-CQ3: (**a**) a SPARQL query for CQ1 to find shopping places in Pathum Wan district, (**b**) a SPARQL query for CQ2 to find nightlife venues that are still open at the present time, and (**c**) a DL query for CQ3 to find the possible trajectory types that have a taxi origin point at a hostel and a taxi destination point at an art gallery.

An example query per CQ3 is shown in Figure 8c. This query classifies the trajectory of a taxi that starts at a hostel and ends at an art gallery. The superclass of tto:ArtGallery is tto:ArtAndEntertainment. A DL query can be used to find the trajectory type, which the results classify as tto:EconomicalSpendingTouristTrip and tto:LeisureTrip.

6. Tourist Behavior Analysis and Visualization Implementation

The results of the data pipeline in Section 4 contain tourist trajectories from taxi GPS data and tourist activities from Twitter data, which can be used for deep insights into tourist behavior analysis. This study applies tourist descriptive analysis and Origin/Destination (OD) analysis to demonstrate examples of tourist behavior analysis and visualization in Bangkok during the Songkran Festival in 2019.

In Figure 9, we utilize tourist descriptive analysis to demonstrate an overview of tourist trajectories in terms of the number of trajectories, travel time, and distance, which helps to understand the amount of travel demand at different times of the day. It also represents the general traffic conditions, as measured by travel time and distance. Figure 9a depicts the number of tourist trajectories grouped by category from 11 April to 17 April 2019. The number of trajectories is pretty small at 121 April, then increase and remains consistent from 13 April to 17 April. Figure 9b depicts the number of tourist trajectories grouped by category at various times. The number trajectories of FoodAndDrink category are peak in the afternoon time and evening time. Next, the number trajectories of Nightclub/Bar category are peak at nighttime and drop in the afternoon time. Last, the number of trajectories in the Spa and Shopping categories are trending in the same direction. It reaches its peak about 16:00–18:00. According to Figure 9c, most tourist trajectories travel for short distances of 0–5 km, followed by 5–10 km, and 10–15 km, respectively. Most tourist trajectories have short travel times of 0–10 min, 10–20 min, and 20–30 min, as illustrated

in Figure 9d. Figure 10 applies OD analysis to reveal popular tourist destinations and the high density of taxi pick-up points. Figure 10a represents the most popular Nightclub/Bar. To identify popular tourist places, we first filter the NightlifeTrip with a starting point at a hotel and a destination point at a NightlifeSpot. The DBSCAN clustering technique is then used to discover the high density of taxi destination points for the NightlifeTrip. The most popular Nightclub/Bar zones in Bangkok are Sukumvit Road in Wattana District, followed by Si Lom Road in Bang Rak District and Khaosarn Road in Phra Nakhon District respectively. Similarly, in Figure 10b, we use NightlifeTrips to discover the most popular taxi pick-up points, which infers the places where tourists stay. The green color denotes the taxi pick-up points and the red color denotes the taxi drop-off points in Sukumvit Road, Bangkok's most popular Nightclub/Bar zone. This figure shows that tourists prefer to stay near tourist attractions.

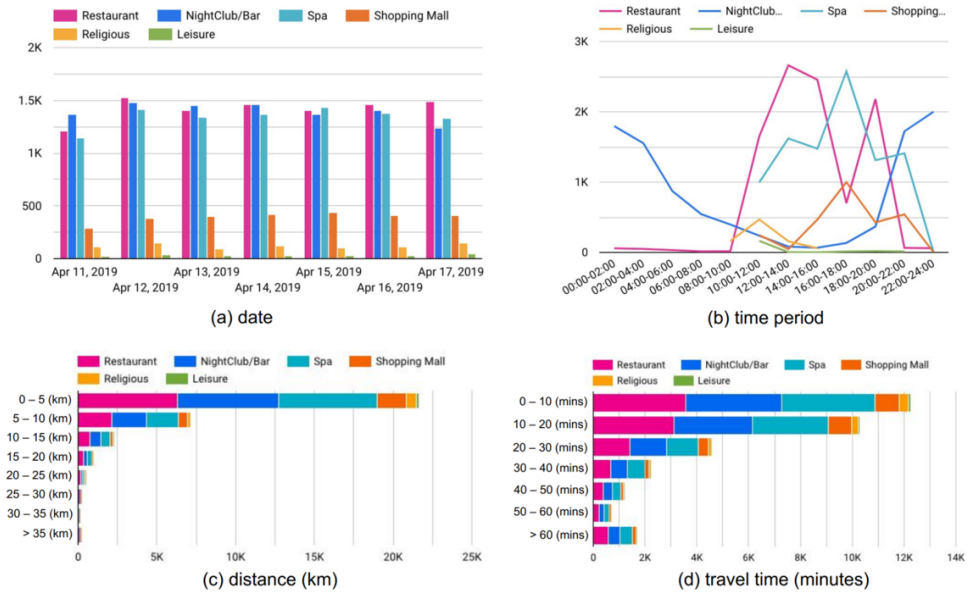

Figure 9. (**a**,**b**) The number of tourist trips by date and time period, (**c**,**d**) Statistics on distance and travel time of tourist trips.

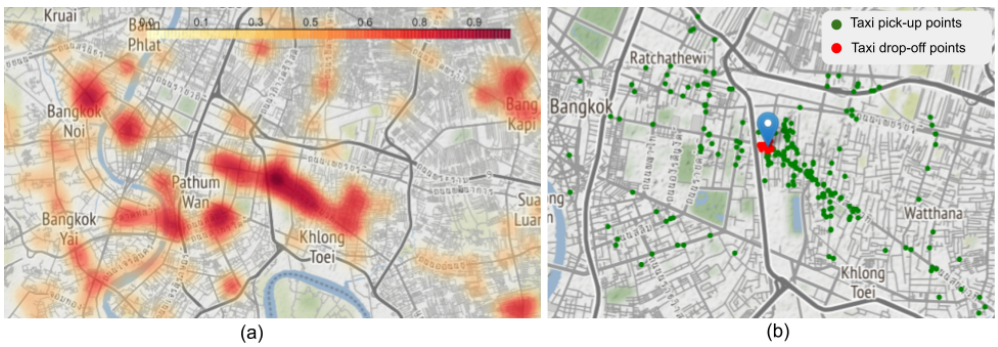

Figure 10. (**a**) Popular Nightclub/Bar in Bangkok area, (**b**) taxi origin points of Nightlife trips.

Figure 11a shows the top ten most popular tourist destinations in Bangkok based on Twitter data. Figure 11b illustrates popular tourism-related keywords, such as tourist activities, tourist locations, events, and festivals, which were popular on Twitter data during Songkran Festival in 2019.

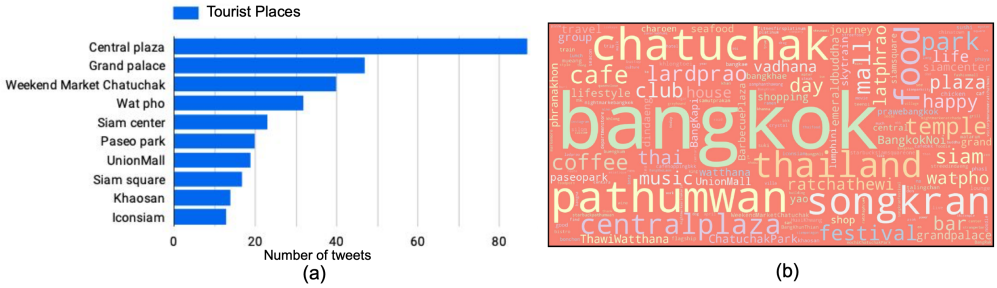

Figure 11. (**a**) Top 10 tourist destinations; (**b**) popular keywords related tourism domain.

7. Conclusions and Discussions

This study proposes a data management platform that manages data on taxi trajectories and social media in order to be used for tourist behavior analysis. The goal of such a platform is to manage significant amounts of trajectory data and incorporate meaningful semantic information. A document-based database with geo-spatial analysis support and scalable performance was used to process the trajectory data. Because taxi trajectories do not generally provide information about individual activities and movement behavior, the data management platform we developed helps to enrich the trajectory data with important semantic information that can elucidate notable tourist movement behavior trends. We also propose a data pipeline to extract tourist trajectories from taxi GPS data, and to extract tourist activities/POIs from Twitter data. Further, we develop a tourist trajectory knowledge base, which provides place information and axioms to classify tourist trajectory types for two aspects: (i) based on tourist activity including ReligiousTrip, FoodAndDrinkTrip, NightLifeTrip, ShoppingTrip, LeisureTrip, and SpaTrip; (ii) based on tourist spending including HighSpendingTouristTrip and EconomicalSpendingTouristTrip.

Finally, we demonstrate how to better understand tourist behavior using our data platform to analyze data in Bangkok during the Songkran Festival in 2019. First, tourist trajectories were extracted from taxi GPS data and contained trip information, place information at taxi OD points, and tourist trajectory types. We utilized tourist descriptive analysis to show the distribution of tourist trajectories at various time, as well as a clustering technique to identify famous tourist destinations/zones. Second, tourist activities/POIs were extracted from social media data and contained check-in information, tourist activities and POIs. We visualized the top ten well-known tourist attractions with high check-in rates and significant tourism keywords or trends at the time.

Although the proposed data platform is constructed for the tourism industry, it can be applied to other domains to process and analyze different data sources. For example, to understand human movement behavior during the COVID-19 pandemic from mobile GPS data, the following modules in the data pipeline need to be modified: (i) Data Preparation—trips must be cleaned and segmented based on mobile GPS data characteristics, (ii) Data Integration and Enrichment—the process for inferring trip purposes must be updated to identify trips with an origin point from a residential area and a destination point at a hospital/vaccination location, (iii) Knowledge Base—add more rules to classify trajectory types related to COVID-19, and (iv) Activity Identification—TF-IDF algorithm and name-entity list need to be updated to extract activities/POIs related the public health and COVID-19. The results of this example will be useful for urban planning purposes, such as detecting the high density of people at hospitals or vaccination locations, and indicating the demand for public transportation.

Further research should seek to (i) enhance and complete the TTO ontology, and (ii) link the TTO ontology and place instances to other published data such as DBpedia and Geonames and use data mining, machine learning, and data visualization techniques to further investigate and elucidate tourist movement trends and popular tourist destinations.

Author Contributions: Conceptualization, P.K. and C.A.; Data curation, P.K.; Methodology, P.K.; Supervision, C.A. and M.B.; Visualization, P.K.; Writing—original draft, P.K.; Writing—review & editing, C.A. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the grants from Asian Institute of Technology (AIT) and National Science and Technology Development Agency (NSTDA) according to the Thailand Graduate Institute of Science and Technology (TGIST) scholarship agreement no. SCA-CO-2562-9683-TH.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Kostakis, I.; Theodoropoulou, E. Spatial analysis of the nexus between tourism–human capital–economic growth: Evidence for the period 2000–2014 among NUTS II Southern European regions. *Tour. Econ.* **2017**, *23*, 1523–1534. [CrossRef]
2. Cui, Y.; Meng, C.; He, Q.; Gao, J. Forecasting current and next trip purpose with social media data and Google Places. *Transp. Res. Part C Emerg. Technol.* **2018**, *97*, 159–174. [CrossRef]
3. Liu, X.; Gong, L.; Gong, Y.; Liu, Y. Revealing travel patterns and city structure with taxi trip data. *J. Transp. Geogr.* **2015**, *43*, 78–90. [CrossRef]
4. Acheampong, R.A. Spatial structure, intra-urban commuting patterns and travel mode choice: Analyses of relationships in the Kumasi Metropolis, Ghana. *Cities* **2020**, *96*, 102432. [CrossRef]
5. Cai, H.; Zhan, X.; Zhu, J.; Jia, X.; Chiu, A.S.; Xu, M. Understanding taxi travel patterns. *Phys. A Stat. Mech. Its Appl.* **2016**, *457*, 590–597. [CrossRef]
6. Zhang, H.; Shi, B.; Zhuge, C.; Wang, W. Detecting taxi travel patterns using GPS trajectory data: A case study of Beijing. *KSCE J. Civ. Eng.* **2019**, *23*, 1797–1805. [CrossRef]
7. Zhao, X.; Lu, X.; Liu, Y.; Lin, J.; An, J. Tourist movement patterns understanding from the perspective of travel party size using mobile tracking data: A case study of Xi'an, China. *Tour. Manag.* **2018**, *69*, 368–383. [CrossRef]
8. Huang, X.; Li, M.; Zhang, J.; Zhang, L.; Zhang, H.; Yan, S. Tourists' spatial-temporal behavior patterns in theme parks: A case study of Ocean Park Hong Kong. *J. Destin. Mark. Manag.* **2020**, *15*, 100411. [CrossRef]
9. Furletti, B.; Cintia, P.; Renso, C.; Spinsanti, L. Inferring human activities from GPS tracks. In Proceedings of the 2nd ACM SIGKDD International Workshop on Urban Computing, Chicago, IL, USA, 11 August 2013; pp. 1–8.
10. Gong, L.; Liu, X.; Wu, L.; Liu, Y. Inferring trip purposes and uncovering travel patterns from taxi trajectory data. *Cartogr. Geogr. Inf. Sci.* **2016**, *43*, 103–114. [CrossRef]
11. Gong, S.; Cartlidge, J.; Bai, R.; Yue, Y.; Li, Q.; Qiu, G. Extracting activity patterns from taxi trajectory data: A two-layer framework using spatio-temporal clustering, Bayesian probability and Monte Carlo simulation. *Int. J. Geogr. Inf. Sci.* **2020**, *34*, 1210–1234. [CrossRef]
12. Chen, L.; Lv, M.; Ye, Q.; Chen, G.; Woodward, J. A personal route prediction system based on trajectory data mining. *Inf. Sci.* **2011**, *181*, 1264–1284. [CrossRef]
13. Zheng, W.; Huang, X.; Li, Y. Understanding the tourist mobility using GPS: Where is the next place? *Tour. Manag.* **2017**, *59*, 267–280. [CrossRef]
14. Hu, Y.; Janowicz, K.; Carral, D.; Scheider, S.; Kuhn, W.; Berg-Cross, G.; Hitzler, P.; Dean, M.; Kolas, D. A geo-ontology design pattern for semantic trajectories. In Proceedings of the International Conference on Spatial Information Theory, Scarborough, UK, 2–6 September 2013; pp. 438–456.
15. Nogueira, T.P.; Braga, R.B.; Martin, H. An ontology-based approach to represent trajectory characteristics. In Proceedings of the 2014 Fifth International Conference on Computing for Geospatial Research and Application, Washington, DC, USA, 4–6 August 2014; pp. 102–107.
16. Fileto, R.; May, C.; Renso, C.; Pelekis, N.; Klein, D.; Theodoridis, Y. The Baquara2 knowledge-based framework for semantic enrichment and analysis of movement data. *Data Knowl. Eng.* **2015**, *98*, 104–122. [CrossRef]
17. Ruback, L.; Casanova, M.A.; Raffaetà, A.; Renso, C.; Vidal, V. Enriching mobility data with linked open data. In Proceedings of the 20th International Database Engineering & Applications Symposium, Montreal, QC, Canada, 11–13 July 2016; pp. 173–182.
18. Witayangkurn, A.; Horanont, T.; Shibasaki, R. The design of large scale data management for spatial analysis on mobile phone dataset. *Asian J. Geoinform.* **2013**, *13*, 17–24.
19. Li, R.; Ruan, S.; Bao, J.; Zheng, Y. A cloud-based trajectory data management system. In Proceedings of the 25th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, Redondo Beach, CA, USA, 7–10 November 2017; pp. 1–4.
20. Ding, X.; Chen, L.; Gao, Y.; Jensen, C.S.; Bao, H. Ultraman: A unified platform for big trajectory data management and analytics. *Proc. VLDB Endow.* **2018**, *11*, 787–799. [CrossRef]

21. Chen, M.; Arribas-Bel, D.; Singleton, A. Understanding the dynamics of urban areas of interest through volunteered geographic information. *J. Geogr. Syst.* **2019**, *21*, 89–109. [CrossRef]
22. García-Palomares, J.C.; Gutiérrez, J.; Mínguez, C. Identification of tourist hot spots based on social networks: A comparative analysis of European metropolises using photo-sharing services and GIS. *Appl. Geogr.* **2015**, *63*, 408–417. [CrossRef]
23. Peng, X.; Huang, Z. A novel popular tourist attraction discovering approach based on geo-tagged social media big data. *ISPRS Int. J. Geo-Inf.* **2017**, *6*, 216. [CrossRef]
24. Zhou, X.; Xu, C.; Kimmons, B. Detecting tourism destinations using scalable geospatial analysis based on cloud computing platform. *Comput. Environ. Urban Syst.* **2015**, *54*, 144–153. [CrossRef]
25. Devkota, B.; Miyazaki, H.; Witayangkurn, A.; Kim, S.M. Using volunteered geographic information and nighttime light remote sensing data to identify tourism areas of interest. *Sustainability* **2019**, *11*, 4718. [CrossRef]
26. Maeda, T.N.; Yoshida, M.; Toriumi, F.; Ohashi, H. Extraction of tourist destinations and comparative analysis of preferences between foreign tourists and domestic tourists on the basis of geotagged social media data. *ISPRS Int. J. Geo-Inf.* **2018**, *7*, 99. [CrossRef]
27. Philander, K.; Zhong, Y.; others. Twitter sentiment analysis: Capturing sentiment from integrated resort tweets. *Int. J. Hosp. Manag.* **2016**, *55*, 16–24. [CrossRef]
28. Hu, Y.H.; Chen, Y.L.; Chou, H.L. Opinion mining from online hotel reviews–A text summarization approach. *Inf. Process. Manag.* **2017**, *53*, 436–449. [CrossRef]
29. Padilla, J.J.; Kavak, H.; Lynch, C.J.; Gore, R.J.; Diallo, S.Y. Temporal and spatiotemporal investigation of tourist attraction visit sentiment on Twitter. *PLoS ONE* **2018**, *13*, e0198857. [CrossRef]
30. Krataithong, P.; Anutariya, C.; Buranarach, M. A Data Management Platform for Taxi Trajectory-based Tourist Behavior Analysis. In Proceedings of the 13th International Conference on Management of Digital EcoSystems, Virtual Event, 1–3 November 2021; pp. 15–21.
31. Ghosh, S.; Ghosh, S.K. Traj-Cloud: A Trajectory Cloud for enabling Efficient Mobility Services. In Proceedings of the 2019 11th International Conference on Communication Systems & Networks (COMSNETS), Bengaluru, India, 7–11 January 2019; pp. 765–770.
32. Li, C.; Liu, Y.; Zhang, H. Analysis of taxi track data based on spark platform. In Proceedings of the 2019 IEEE 4th Advanced Information Technology, Electronic and Automation Control Conference (IAEAC), Chengdu, China, 20–22 December 2019; Volume 1, pp. 2357–2361.
33. Kong, F.; Lin, X. The method and application of big data mining for mobile trajectory of taxi based on MapReduce. *Clust. Comput.* **2019**, *22*, 11435–11442. [CrossRef]
34. Putri, F.K.; Song, G.; Kwon, J.; Rao, P. DISPAQ: Distributed profitable-area query from big taxi trip data. *Sensors* **2017**, *17*, 2201. [CrossRef]
35. Jiang, Y.; Cao, J.; Liu, Y.; Fan, J. West Lake Tourist: A Visual Analysis System Based on Taxi Data. *Smart Cities* **2019**, *2*, 345–358. [CrossRef]
36. Renso, C.; Baglioni, M.; de Macedo, J.A.F.; Trasarti, R.; Wachowicz, M. How you move reveals who you are: Understanding human behavior by analyzing trajectory data. *Knowl. Inf. Syst.* **2013**, *37*, 331–362. [CrossRef]
37. Chua, A.; Servillo, L.; Marcheggiani, E.; Moere, A.V. Mapping Cilento: Using geotagged social media data to characterize tourist flows in southern Italy. *Tour. Manag.* **2016**, *57*, 295–310. [CrossRef]
38. Cox, S.; Little, C. The OWL-Time Ontology. Available online: https://www.w3.org/TR/owl-time/ (accessed on 8 May 2012).
39. Meng, C.; Cui, Y.; He, Q.; Su, L.; Gao, J. Travel purpose inference with GPS trajectories, POIs, and geo-tagged social media data. In Proceedings of the 2017 IEEE International Conference on Big Data (Big Data), Boston, MA, USA, 11–14 December 2017; pp. 1319–1324. [CrossRef]
40. Buranarach, M.; Thein, Y.M.; Supnithi, T. A community-driven approach to development of an ontology-based application management framework. In Proceedings of the Joint International Semantic Technology Conference, Nara, Japan, 2–4 December 2012; pp. 306–312.