Data4UrbanMobility: Towards Holistic Data Analytics for Mobility Applications in Urban Regions

Nicolas Tempelmeier
L3S Research Center, Leibniz Universität Hannover
tempelmeier@L3S.de

Yannick Rietz
PROJEKTIONISTEN GmbH
rietz@projektionisten.de

Iryna Lishchuk
Institute for Legal Informatics, Leibniz Universität Hannover
iryna.lishchuk@iri.uni-hannover.de

Tina Kruegel
Institute for Legal Informatics, Leibniz Universität Hannover
kruegel@iri.uni-hannover.de

Olaf Mumm
Institute for Sustainable Urbanism, TU Braunschweig
o.mumm@tu-braunschweig.de

Vanessa Miriam Carlow
Institute for Sustainable Urbanism, TU Braunschweig
v.carlow@tu-braunschweig.de

Stefan Dietze
GESIS - Leibniz Institute for the Social Sciences
stefan.dietze@gesis.org

Elena Demidova
L3S Research Center, Leibniz Universität Hannover
demidova@L3S.de

ABSTRACT

With the increasing availability of mobility-related data, such as GPS-traces, Web queries and climate conditions, there is a growing demand to utilize this data to better understand and support urban mobility needs. However, data available from the individual actors, such as providers of information, navigation and transportation systems, is mostly restricted to isolated mobility modes, whereas holistic data analytics over integrated data sources is not sufficiently supported. In this paper we present our ongoing research in the context of holistic data analytics to support urban mobility applications in the Data4UrbanMobility (D4UM) project. First, we discuss challenges in urban mobility analytics and present the D4UM platform we are currently developing to facilitate holistic urban data analytics over integrated heterogeneous data sources along with the available data sources. Second, we present the MiC app - a tool we developed to complement available datasets with intermodal mobility data (i.e. data about journeys that involve more than one mode of mobility) using a citizen science approach. Finally, we present selected use cases and discuss our future work.

ACM Reference Format:
Nicolas Tempelmeier, Yannick Rietz, Iryna Lishchuk, Tina Kruegel, Olaf Mumm, Vanessa Miriam Carlow, Stefan Dietze, and Elena Demidova. 2019. Data4UrbanMobility: Towards Holistic Data Analytics for Mobility Applications in Urban Regions. In *Companion Proceedings of the 2019 World Wide Web Conference (WWW '19 Companion), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3308560.3317055

WWW '19 Companion, May 13–17, 2019, San Francisco, CA, USA

ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3317055

1 INTRODUCTION

Contemporary urban mobility behavior is undergoing a rapid transition and paradigm shift and also is affected by a wide range of short-, medium- as well as long-term factors. These vary from aspects such as immediate climate conditions, regional construction sites or ephemeral events to long-term trends such as increasing negative environmental impacts and the widespread adoption of intermodal mobility chains (i.e. journeys involving several means of transportation and mobility such as walking, cycling, public transportation, etc.). Moreover, the lifestyle- and population-induced new demand for mobility leads to region-specific, ecological and traffic related problems in growing metropolitan areas and is a limiting factor for urban development.

Traditional urban traffic planning relies on complex models, combining trip generation, trip distribution, mode choice and route choice [21] to simulate travel demands. Data foundations of these models essentially consist of traffic census, socio-demographic data as well as infrastructural and service data, where large-scale and accurate data about actual mobility behavior, in particular in the context of intermodality, is costly to obtain and the aforementioned contextual factors are largely ignored.

Throughout the last decade, the widespread use of Web and mobile applications has led to an increasing availability of data, which captures both actual mobility needs and usage as well as contextual information about, for instance, traffic incidents, city events and weather conditions. This data has the potential to complement existing data sources and information systems currently used for handling urban mobility processes. In particular, within densely populated areas, the correlation of mobility behavior with data obtainable from mobility apps, public transportation websites and social media streams can uncover more complex dependencies and aid the development of supervised models which consider a wide variety of features and enable predictions of future needs.

Despite an overall increasing availability of datasets related to mobility behaviour, this data typically focuses on one mode of transportation, most prominently covering individual traffic and, to

some extent, public transportation services. Other modes of mobility that become particularly important in modern cities, e.g. cycling and walking, are typically not captured at scale. Furthermore, data sources coming from particular mobility services and transportation providers do not adequately capture intermodal mobility sequences. However, such data is of utmost importance to better understand and predict the actual demand in mobility and associated services.

The challenges related to collection and analysis of such data are manifold. They include the provision of tools and methods to capture and analyse intermodal mobility, designing incentives for city inhabitants to share their mobility data, protection of personal data to be collected in accordance to the legal framework as well as data analytics methods and models to provide added value for the individual participants, mobility service providers and city authorities.

In this work, we present our ongoing research within the *Data4UrbanMobility* project (D4UM)[1]. This research aims at gathering, augmenting and analysing mobility data in urban regions to address the problems presented above. We introduce the D4UM platform built to facilitate collection and analysis of mobility-related data from a variety of sources on a long term. In the project we in particular focus on the datasets available in the urban regions of Hanover, Wolfsburg and Brunswick (Germany), whereas the results will be transferable to other urban regions. The overall contributions of the project include a large annotated data catalogue including regionally and globally relevant mobility-related data sources, models built upon these data sources and analytic results in the particular regions.

Furthermore, the paper introduces the MiC-App - a human-centric data tool developed in the project framework to capture individual movements – tracks and modes – to complement existing sources in the context of intermodal urban mobility. Combined with other available data sources, the data collected through the MiC-App and the underlying D4UM platform facilitates the capturing and the analyses of data to better understand the growing demand in intermodal mobility in urban regions.

2 PROBLEM DESCRIPTION

The aim of the D4UM project is to facilitate efficient analytics and data-driven estimates of mobility demand and evaluation of mobility network quality, addressing the needs of different stakeholders. The stakeholders of the project include individual citizens, city administrations, mobility service providers and urban traffic planners. The aims of these stakeholders can be categorised with respect to the different time horizons, as follows:

- Short Term: Facilitate efficient access to mobility services, taking into account temporary high load (citizens).
- Medium Term: Provide new and optimise existing mobility services (mobility service providers).
- Long Term: Facilitate data-driven integrated urban planning processes (urban traffic planers, city administrations).

To address these problems the project develops the D4UM platform that facilitates collection and analytics of relevant data sources. The research questions that can be addressed using this platform include e.g.:

- Which external conditions such as climate or spatial structures influence the typical movement patterns of city inhabitants?
- How are the urban movement patterns influenced by the external factors, such as e.g. weather conditions, given specific urban contexts and mobility options?
- How can an increased demand in mobility services be determined and addressed?

3 D4UM PLATFORM

This section describes the architecture of the D4UM platform. This platform enables integrated analytics of heterogeneous data sources for different stakeholders in the context of urban mobility. Figure 1 provides an overview of the platform layers and its individual components.

The input layer of the D4UM platform consists of *heterogeneous data sources* that cover various aspects of urban mobility. These data sources are described in Section 4 in more detail.

The *data aggregation and integration* layer conducts all necessary pre-processing and transforms these data sources to comply with the D4UM data model. This data model formally specifies an integrated schema and establishes spatial, temporal and contextual connections across these sources, thus facilitating integrated analytics going across the dataset boundaries. For example, traffic speed records are aligned with the street segments obtained from the OpenStreetMap data.

The D4UM platform conducts long-term collection of mobility-related data from data streams to create an overview of relevant mobility data over longer time periods (i.e. months or years) within the *long-term data collection* component. In particular, recording of dedicated streaming sources such as traffic warnings published as RSS feeds or data extracted from social media channels (e.g. police channels on Twitter) allows for long-term analytics to observe patterns and temporal fluctuations, which would otherwise not be possible. *Data enrichment* facilitates collection of supplemental data by employing Web mining (e.g. by identifying events in social media streams) and citizen science tools (e.g. by collecting data with the MiC App presented in Section 5).

The *data analytics* layer introduces models that build on the integrated data. For example, analysis of long-term data collections can be employed to identify typical urban movement patterns. For instance, spatio-temporal dependencies between traffic conditions on different roads can reveal structural problems within the road network [6, 19]. Another example is the analysis of the impact of external factors (e.g. heavy rain or snowfall) on mobility behaviour [29]. Data-based forecasts can make use of long-term data collections to predict urban mobility patterns in the future. For example, in the presence of planned special events (e.g. football matches or concerts), an increased load on the mobility infrastructure such as roads and public transportation capacity can be estimated [27, 39].

Dedicated *interfaces* such as APIs and graphical user interfaces grant access to the data analytics results. APIs serve as a conceptual abstraction from the complex models and provide information in

[1] http://data4urbanmobility.l3s.uni-hannover.de/

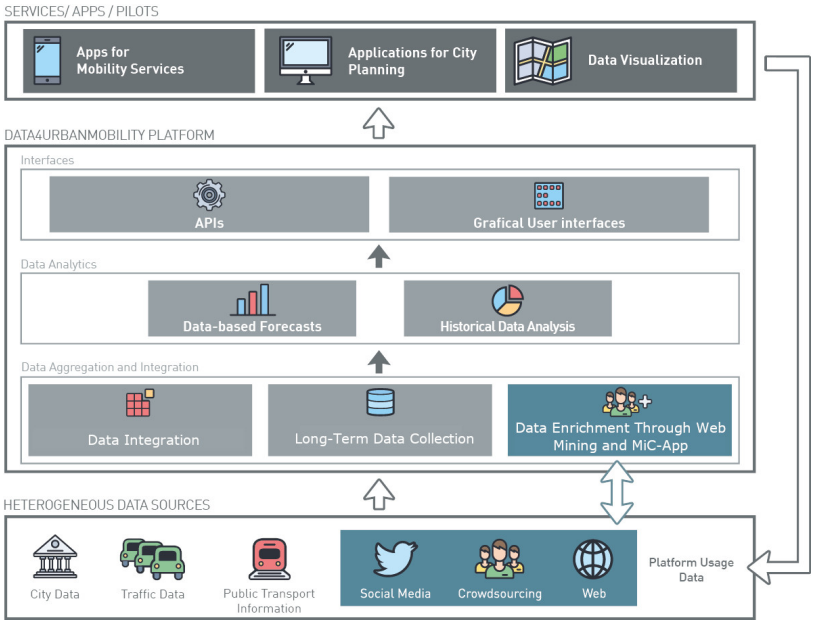


Figure 1: Overview of the Data4UrbanMobility platform architecture

a machine readable form, such as the GeoJSON[2] format. Graphical interfaces such as map layers graphically present the analysis results and allow for embedding a map into Web pages.

Finally the *services, apps and pilots* layer makes the analysis results available for end users. In particular, we develop a dashboard Web application that can be used by city planers to analyse traffic patterns on the long term, while citizens can use the MiC app to derive insights about their own mobility behaviour and voluntarily contribute data to the D4UM system.

4 INTEGRATION OF WEB-BASED MOBILITY DATA

An overview of existing data sources from which we currently extract regional and other relevant information in the context of urban mobility is presented in Table 1. The majority of these data sources (except the traffic flow information) are Web-based. Traffic Data information sources include traffic flow information and traffic feeds. Traffic flow data, provided as aggregated Floating Car Data (FCD), reflects the average speed of the road traffic with respect to the individual road segments, whereas traffic feeds contain traffic warnings and accident notifications provided as RSS feeds. Public Transport Information includes public transportation query logs and GTFS data. We obtain query log data from the *EFA*-system[3]. *EFA* is the official Web service for routing and time table information of public transportation for the region of Hannover. In this context, a query is a request for a public transportation route with specified origin, destination and departure time, issued via a Web-interface or a mobile application. Moreover, we consider Global Transit Feed Specification (GTFS)[4] data which provides timetable information for public transportation such as departure times, routes, etc. This data is complemented with regional weather conditions, which can potentially impact the selection of the transportation mode and routes.

Next to such directly related mobility data, we consider additional information obtained from social media and the Web. We consider social media data obtained from the Twitter streaming API[5], which potentially contains information about regional events as well as traffic incidents. Web data includes event-centric Web markup, which is prevalent in Web pages through standards such as RDFa[6], and Microdata[7]. We currently investigate in particular data complying with schema.org as the most established markup vocabulary on the Web so far [23]. In our previous work we developed methods to infer missing categorical information in noisy and sparse Web markup data [31], increasing usefulness of this data for event-centric applications. Furthermore, we consider event-centric focused crawls from the Web [7] and Web archives [8], [9], as well as Twitter data regarding events and traffic. Another source of event-centric information is the recently proposed EventKG knowledge graph [10, 11]. Finally, this information is complemented with geographic data (e.g. street networks, locations of event venues) obtained from *OpenStreetMap*[8]. The information contained in these sources is highly complementary.

These data sources can be used to address parts of the afore introduced research questions. For instance, in the context of events, the traffic flow and public transportation query data can be used to determine typical patterns while Web, Web archives, knowledge graphs and social media sources can provide information about

[2]https://tools.ietf.org/html/rfc7946
[3]https://www.efa.de
[4]https://developers.google.com/transit/gtfs/
[5]https://developer.twitter.com/en/docs/tweets/filter-realtime/overview
[6]RDFa W3C recommendation: http://www.w3.org/TR/xhtml-rdfa-primer/
[7]http://www.w3.org/TR/microdata
[8]http://www.openstreetmap.org/

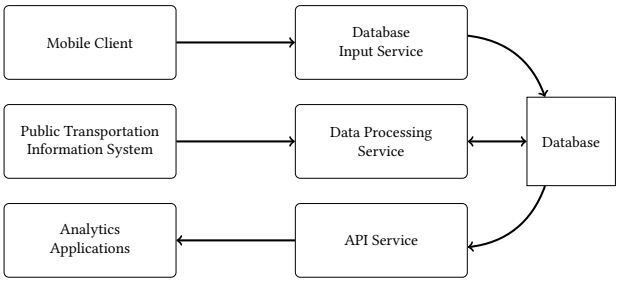


Figure 2: Overview of the MiC app data flow.

planned special events such as concerts or football matches. Another example is the computation of the average load on the roads with respect to the external factors, e.g. weather. These approaches require integration of data originating at several data sources, where semantic data descriptions, methods of dataset profiling [5], [4] and data quality analytics play an important role.

However, for comprehensive analysis of urban mobility some data is still missing. In particular foot walks, bicycle rides and intermodal data is not captured by any of the existing data sources. On the contrary, these modes become increasingly popular and thus important for urban mobility applications. To this end, the project introduces the MiC App, which can be used to collect the required data.

5 MIC APP

To address the lack of data regarding foot walks, bicycle rides and intermodal trips, the project introduces the *Move in the City* (MiC) app. This app poses a citizen science inspired method to gather movement data. Users of the MiC app can capture GPS traces of their trips with their mobile devices and voluntarily contribute the data to the D4UM project. In return, individual movement statistics are provided to the user. Moreover, the increased amount of available mobility data enables novel analytics of mobility behaviour. Based on the analytics, previously unseen problems can be identified and addressed, creating further benefits for the user and for the mobility ecosystem as a whole.

Movement data captured by the MiC app undergoes a preprocessing routine tailored to the specific needs of urban mobility analytics. Different modes of mobility such as walking, cycling, driving and the use of public transportation are automatically distinguished. Furthermore, if public transportation is used, the data is enriched with additional information such as entry stop, exit stop and the public transportation line used. This way, the data captured by the MiC app represents a valuable data source for further mobility analytics approaches.

5.1 Architecture

The architecture of the MiC app aims at keeping a low computation effort on the user device to reduce the energy consumption of the app. Furthermore, the architecture is platform independent to enable as many users as possible to use the app. To this end, MiC makes use of established, flexible Web technologies such as Web-based user interfaces and the MQTT protocol[12]. To further ensure platform independence, the app does not directly access sensor data of the mobile devices. Instead, the relatively new, built-in activity recognition APIs of the mobile operating systems are used. In addition to the activity recognition data, the app records fine grained location data.

Figure 2 provides an overview on the data flow. Data recorded by the mobile client is sent to a Web-based endpoint which stores the recorded data in a database. The data is then enriched with information about public transportation, obtained by querying the local information system for public transportation. Finally, a Web-based endpoint provides an API that makes the data accessible for data analytics applications.

5.2 App Design

MiC aims at motivating the users to use the app by providing an appealing and easy to use interface. Figure 3 presents two exemplary screenshots of the MiC app. The start view can be seen in Figure 3a. We achieve a lightweight user interaction by only requiring users to press one button to start or stop the recording of their movement. Figure 3b depicts the view presenting statistics of the individual user. The view presents the fraction of the user transportation modes by providing absolute and relative numbers as well as a graphical visualisation. Ideas for further user statistics include visualisation of the environmental impact of the users mobility behavior or the overall contribution to the urban mobility network. Figure 3c presents the map view, where the user can visualise captured trip data, i.e. GPS traces, on a map.

5.3 Piloting

The first closed test of the system was conducted by five participants who made 41 trips over the duration of five days, where the median duration of the recorded trips took 24 minutes. The participants wrote detailed travel diaries for these trips that were then used as a ground truth to assess the accuracy of the mode of travel classification. Due to limited GPS signals paths in underground trains, they were excluded from the study.

Table 2 presents the median duration and accuracy of transportation mode recognition of the recorded trips with respect to mode of transportation, where the total row summarises all recorded trips. Trips that include multiple modes of transportation were divided into individual trips with only one mode of transportation. In addition to classifying the correct transportation mode, we only considered tram and bus trips to be recognised correctly, if the correct entry stop, exit stop and public transportation line was identified by the system.

In total, 86.2% percent of the trips were recognized correctly by the system. The highest accuracy was achieved for the recognition of bicycle trips (100%). This is due to the relatively clear movement signature of riding a bicycle, which is well recognised by the activity recognition of the smartphones. The second highest accuracy is achieved for cars. Even though trams and busses achieve the lowest accuracy, the absolute accuracy is at least 73%. The reduced accuracy for the two classes is likely to be caused by the additional

[12] http://mqtt.org/

Table 1: An overview of regional and general mobility-relevant data sources for Lower Saxony, Germany.

Source	Description	Granularity / Size	Timespan	License	Format
Traffic Data					
Traffic flow	Average car traffic speed records in Lower Saxony, Germany.	Average traffic speed per road segment and time interval (15 min)	September 2017 - January 2019	Commercial	CSV
Traffic feeds	Traffic warnings and incidents[9].	~ 25000 notifications	June 2017-January 2019	Provider-specific	RSS feeds / XML
Public Transport Information					
Public transportation query logs	Query logs for public transportation routes and timetables.	$\sim 4 \cdot 10^6$ queries per month	October 2016-January 2019	Commercial	CSV
Global Transit Feed Specification (GTFS) data	Public transportation timetable information for Lower Saxony, Germany.	8800 stops, 2600 routes, $2 \cdot 10^6$ stop times	until January 2019	Open Data	CSV
City Data					
Rainfall data	Volume of rain in a region.	1 Record / hour / km^2	2005-2019	Open data	CSV
Social Media					
Twitter	Event- and location-centric tweets from Twitter API. Traffic information channels[10].	German tweets from Twitter API	May 2017-January 2019	Twitter API license	JSON
Web					
Event-centric Web markup	Annotated Web pages, e.g. using schema.org.	Web Data Commons event subset: 263×10^6 facts	until November 2017	Common Crawl ToU	RDFa, Micro-Data
Focused crawls	Event-centric crawls, news[11].	~ 22000 events located in Hannover, Germany	Crawl-specific	Provider-specific	HTML
EventKG	Multilingual event-centric temporal knowledge graph	$690 \cdot 10^3$ events and $2.3 \cdot 10^6$ temporal relations (V1.1)	until today	Creative Commons Attribution Share Alike 4.0	RDF
OpenStreetMap	Geometries (points, lines, polygons) annotated with properties. Subset for parts of Lower Saxony.	$2.6 \cdot 10^6$ facts	until January 2019	ODbL	XML

Table 2: Median duration and accuracy of mode of travel recognition with respect to transportation mode. Trips including multiple modes of transportation are divided into individual trips with only one mode of transportation.

Mode	Number	Median Duration	Accuracy
Bicycle	16	10 min.	100%
Car	14	9.5 min.	92.9%
Tram	13	7 min.	76.9%
Bus	15	12 min.	73.3%
Total	58	11 min.	86.2%

constraints, i.e. the recognition of start stop, entry stop and public transportation line.

The first public test phase of the system is currently ongoing. The application beta testing platforms by Apple and Google are being used to carry out the application. Users were recruited at universities and expositions. Table 3 presents statistics about data currently captured by the MiC app. To today, 78 participants were found to test the system under real-world conditions.

Table 3: Statistics of data currently captured by the MiC app.

Property	Value
Number of Users	78
Number of Recorded Trips	218
Average Trip Duration	69 min.
Number of Captured GPS Points	92.550

5.4 Data Protection

To ensure data protection compliant to latest regulations (e.g. the *General Data Protection Regulation* by the European Union), a concept was developed in joint work with researchers from legal informatics that includes the following components:

- The consent of app users (Article 6 GDPR);
- A privacy policy (Article 13 GDPR);
- Data pseudonymisation (Article 89 GDPR);
- The agreement between joint controllers (Article 26 GDPR).

The MiC app has an integrated privacy policy and consent form providing app users with the information about the data procession, data rights and possibility to consent. It contains details about the types of data the MiC app collects, the collection procedure, purposes of research, data storage and data sharing within the D4UM

(a) Start View **(b) User Statistics View** **(c) Map View**

Figure 3: Screenshots taken from the MiC app. Subfigure (a) presents the start view. Users can start recording their movement by pressing a single button. Subfigure (b) presents the user statistics view. It visualises details about the user specific mobility behaviour, e.g. the user specific fraction of transportation modes. Subfigure (c) presents the map view that projects captured trip data onto a map.

project, data rights of the users (i.a. access, deletion, withdrawal of consent) and contact details of the app providers.

The geo-referenced data collected by the MIC app is processed in pseudonymised form. By this, the data is purified from all personal identifiers, such as name, e-mail, IMEI number (a cellphone serial number), etc. Pseudonymisation is chosen as a de-identification measure (in contrast to anonymisation or irreversible de-identification) to enable enforcement of the data subjects' rights to access and deletion of the data, as may be requested by the users. This is part of an integrated and innovative citizen science approach which enables everyone to actively participate in research and planning processes with a strong effort to establish new standards for data sovereignty of citizens and cities.

6 USE CASES

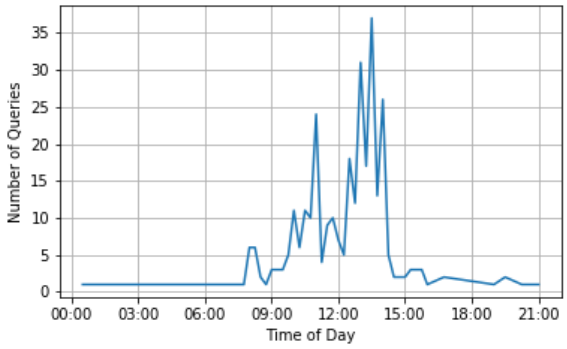
In this section we discuss two exemplary use cases for the D4UM platform and the MiC app.

6.1 Urban Mobility in Presence of Planned Special Events

Large-scale planned special events such as football games, concerts, etc. are known to have impact on both road traffic [15] and public transportation services [27]. Analysing mobility behaviour in such situations is particularly challenging due to a great number of influence factors (e.g. target audience, weather conditions, public transportation infrastructure). Moreover, large-scale events typically have an impact on all modes of transportation, which are mutually dependent. Thus, a holistic approach, which considers data about all modes of transportation as well as external factors, is required to gain a better understanding of the mobility behaviour in presence of planned special events.

We illustrate the complexity of the problem at the example of a football game that took place in the city of Hannover (Germany) on December 17th, 2017 at 3:30pm. Figure 4 presents the conditions around the football stadium in Hannover for both road traffic and public transportation services. Figure 4a depicts the road traffic conditions around the stadium half an hour before the start of the game. As we can observe, the roads with a high load (red color) as well as the roads with good traffic conditions (green color) are present nearby the stadium. This illustrates that the impact of planned special events might be complex and such impact does not necessarily evenly spread around the event venue, but might be subject to other factors such as road topology or availability of parking spaces. Figure 4b presents the number of queries to the public transportation information system that were issued for the bus stop near the stadium during the day of the football game. We observe a relatively low number of queries during the morning hours. At 9am, the number of queries starts to rise and reaches their maximum at 2:30pm, one hour before the game starts. We assume that an increased number of queries in the temporal proximity of the event is likely related to the football game. Overall, we observe,

(a) Traffic conditions on the major roads located nearby the football stadium half an hour before the game starts. Red color indicates heavy load, green color indicates low load.



(b) Number of public transportation queries for the bus stop nearest to the football stadium with respect to the time of day on the day of the football game.

Figure 4: Illustration of an impact of a football game on road traffic and public transportation. The football game took place on December 17th, 2017 at 3:30pm in Hannover, Germany.

that a single factor, i.e. the taking place of a large-scale event, can impact both road traffic and public transportation simultaneously. Therefore, analysis of these effects should take both mobility modes into account. To this extent, the D4UM platform presented in this paper aims at holistic data analysis of urban mobility data.

6.2 Analytics of Urban Mobility Infrastructure

The rapid growth of cities has lead to an increased demand of suitable urban mobility infrastructure. New services such as car sharing, (e-)bike sharing and ride sharing have emerged and begin to gain importance within the mobility services ecosystem. However, data about these modes of transportation is typically only sporadically available. Moreover, intermodal data which captures different modes of transportation on a single journey is evenly rare.

The interdependencies between individual and public transportation increase the complexity of the problem. For instance, a poor coverage of public transportation services might lead to an increased load on the road infrastructure. In turn, an overloaded road system can lead to an increased demand of public transportation services or bicycle tracks.

Therefore it is crucial for city planers and mobility service providers to conduct holistic analysis, which take all modes of transportation into account. Such analysis should include the identification of:

- typical mobility patterns
- demand of mobility services
- interdependencies between transportation modes
- coverage of public transportation services

The D4UM platform enables the holistic analysis of urban mobility data in these scenarios. Furthermore, the MiC app presented in this paper is a valuable approach to capture rarely available data about urban mobility behaviour, including bicycle rides and intermodal data and complement available data sources.

7 RELATED WORK

In this section we discuss related work in the area of smart city mobility systems as well as approaches to predict individual aspects of urban mobility.

7.1 Smart City Mobility Systems

[24] conducted a systematic review of smart city data analysis approaches and provide taxonomies for data sources, data analytic methods and smart city services. Urban CPS [38] is a cyber physical system that integrates floating car data, cellphone data and public transportation data to make prediction about real time traffic speeds. [17] employ semantic technologies to develop STAR-CITY, a system for traffic prediction and reasoning, used for spatio-temporal analysis of the traffic status as well as for the exploration of contextual information such as nearby events. [3] presents a system that makes use of cellphone data to analyse the demand of public transportation services within a city. While these approaches focus on predictions for one mode of mobility only and mainly use a single dataset as their primarily data source, we consider the other modes of transportation as well, e.g. bicycle rides or public transportation. [1] provides a summary of common enterprise, logical and physical architectures for digital smart city applications. While the authors consider architectures for general smart city applications, we focus on the mobility domain. [33] proposed the Compressed Start-End Tree (CSE-tree), a spatio-temporal index structure that can be used to effectively index and retrieve temporal GPS data which is in particular relevant for smart city mobility applications.

7.2 Urban Mobility Analytics

Recently a number of studies addressed several individual prediction tasks at the interface of the urban infrastructure and mobility. We consider these approaches as potential use cases for the D4UM platform.

Traffic Forecasting: Short-term forecasting of urban road traffic has been the focus of numerous studies where data sparsity is a particular challenge. [35] tackles the problem by using sparse FCD (floating car data) and a context-aware tensor decomposition approach to estimate travel times for road segments for which no FCD is available. The information is then used to estimate the required travel time for a given route. [22] proposed a framework for the city-wide inference of traffic volume. They make use of a semi-supervised learning algorithm that can be used with sparse loop detector data as well as taxi GPS data. Similar, [34] employs a hidden Markov model to estimate traffic speeds of a road network based on sparse FCD where the speed to be estimated on a single road is considered as a hidden state.

Social Network Data: Further approaches utilized location-based social network (LBSN) data. [28] leverages LBSN data to identify functional urban regions. They employ latent Dirichlet allocation and unsupervised machine learning algorithms to determine the regions. [18] investigated the general predictability of LBSN data. They conducted a case study on Foursquare datasets, where users indicate their geographic location, i.e. the users can indicate that they are at a certain event venue. The authors do not focus on a specialised prediction task, but provide general insights on working with the aforementioned data. [36] make use of LBSN data to infer boundaries of functional regions in urban environments. They construct a mobility network from spatial user interactions and delineate boundaries by identifying strongly connected communities within the network space. [25] detects events from social media by employing a hashtag-based algorithm. The event information extracted from the social media is then used for prediction of the public transportation flow.

Structural Analysis: Another class of approaches focuses on discovering structural dependencies within cities by analysing urban mobility data. [2] propose a method to keep track of the congestions in urban road networks to identify unstable road segments. [13] make use of context-aware tensor decomposition to identify so called *urban black holes*, i.e. traffic anomalies with a greater inflow than outflow. [12] detect urban black holes by using a grid-based index-structure that is build on top of a spatio-temporal graph representing the road network. They extract candidate cells from the index which then are used to determine the exact subgraphs that are urban black holes. [19] identify cascading patterns on congested roads. They propose a generative probabilistic model that maximises the likelihood of a cascade to be present with respect to the observed traffic data. [14] employs topic modelling to analyse urban street works. They proposed the concept of interactional regions, i.e. regions that commonly bound routes within the street networks. [32] leverage taxi flow data to learn vector representation of city regions. The representations are then used to make predictions about the regions such, e.g. crime rate, average income or average house prices. [20] employ deep learning techniques, i.e. a combination of restricted Boltzmann machines and recurrent neural networks, to learn high-dimensional congestion patterns from taxi GPS data. [26] make use of taxi GPS-trajectories to classify the land use of urban areas. They propose an iterative DBSCAN algorithm to cluster regions with respect to the frequency with which passengers are picked up or set down. They make use of the same information to classify the land use of regions, e.g. the land use for hospitals or commercial districts. [37] proposes the infinite urbanization process model that employs a topic modelling approach to simultaneously discover the function of an urban region (i.e. the distribution of present shops, restaurants, etc.) and to estimate the region popularity (e.g. in terms of real estate prices). Similar, [30] extracts urban regions of interest from online map search queries. They propose a spatio-temporal latent factor model which identifies travel patterns that influence points of interest.

Special Traffic Conditions Several approaches target the identification of problematic segments and areas of urban networks under specific conditions (e.g. planed special events). [16] propose the use of an artificial neuronal network to identify road segments that are typically affected by planned special events that take place at a particular venue. [39] proposed an approach to detect events from traffic data. They make use of a two-dimensional grid to partition the space. The authors then infer a graph-based representation of the grid that captures the flow of vehicles between the individual cells of the grid where the root of the graph is located at the events location. Finally, [27] investigate the effect of public events on the public transportation network. They propose a Bayesian additive model that can be employed to gain an understanding of public transportation demand in the presence of events. I.e. the model is able to predict the number of public transportation trips to the venues where the respective event takes place.

8 CONCLUSIONS & FUTURE WORK

In this paper we presented our ongoing work towards holistic urban data analytics conducted in the context of the Data4UrbanMobility project. We presented the D4UM platform that facilitates seamless long-term analytics of heterogeneous mobility-related data sources, including but not limited to floating car data, weather conditions, traffic warnings and Web queries. Furthermore, we presented the MiC app - a citizen science application that facilitates complementing this data with intermodal mobility patterns of city inhabitants. In our future work we intend to further increase the MiC app user base, and implement further use cases on top of the integrated D4UM platform.

ACKNOWLEDGMENTS

This work was partially funded by the Federal Ministry of Education and Research (BMBF), Germany, "Data4UrbanMobility" project, grant ID 02K15A040.

REFERENCES

[1] L. Anthopoulos and P. Fitsilis. 2010. From Digital to Ubiquitous Cities: Defining a Common Architecture for Urban Development. In *Proceedings of the 2010 Sixth International Conference on Intelligent Environments*. 301–306.
[2] Tarique Anwar, Chengfei Liu, Hai L. Vu, and Md. Saiful Islam. 2016. Tracking the Evolution of Congestion in Dynamic Urban Road Networks. In *Proceedings of the ACM CIKM 2016*.
[3] Michele Berlingerio, Francesco Calabrese, Giusy Di Lorenzo, Rahul Nair, Fabio Pinelli, and Marco Luca Sbodio. 2013. AllAboard: A System for Exploring Urban Mobility and Optimizing Public Transport Using Cellphone Data. In *Machine Learning and Knowledge Discovery in Databases*, Hendrik Blockeel, Kristian Kersting, Siegfried Nijssen, and Filip Železný (Eds.). 663–666.
[4] Stefan Dietze, Elena Demidova, and Konstantin Todorov. 2019. RDF Dataset Profiling. In *Encyclopedia of Big Data Technologies*.
[5] Mohamed Ben Ellefi, Zohra Bellahsene, John G. Breslin, Elena Demidova, Stefan Dietze, Julian Szymanski, and Konstantin Todorov. 2018. RDF dataset profiling - a survey of features, methods, vocabularies and applications. *Semantic Web* 9, 5 (2018), 677–705.

[6] U. Feuerhake, O. Wage, M. Sester, N. Tempelmeier, W. Nejdl, and E. Demidova. 2018. Identification of similarities and prediction of unknown features in an urban street network. *ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences* XLII-4 (2018), 185–192.

[7] Gerhard Gossen, Elena Demidova, and Thomas Risse. 2015. iCrawl: Improving the Freshness of Web Collections by Integrating Social Web and Focused Web Crawling. In *Proceedings of the JCDL'15*.

[8] Gerhard Gossen, Elena Demidova, and Thomas Risse. 2017. Extracting Event-Centric Document Collections from Large-Scale Web Archives. In *Proceedings of the 21st International Conference on Theory and Practice of Digital Libraries, TPDL 2017*. 116–127.

[9] Gerhard Gossen, Thomas Risse, and Elena Demidova. 2018. Towards extracting event-centric collections from Web archives. *International Journal on Digital Libraries* (27 Oct 2018).

[10] Simon Gottschalk and Elena Demidova. 2018. EventKG: A Multilingual Event-Centric Temporal Knowledge Graph. In *Proceedings of the 15th International Conference, ESWC 2018*. 272–287.

[11] Simon Gottschalk and Elena Demidova. 2019. EventKG - the Hub of Event Knowledge on the Web - and Biographical Timeline Generation. *Semantic Web* (2019).

[12] Liang Hong, Yu Zheng, Duncan Yung, Jingbo Shang, and Lei Zou. 2015. Detecting Urban Black Holes Based on Human Mobility Data. In *Proceedings of the 23rd SIGSPATIAL International Conference on Advances in Geographic Information Systems (SIGSPATIAL '15)*. ACM, Article 35, 10 pages.

[13] Li Jin, Zhuonan Feng, and Ling Feng. 2016. A Context-aware Collaborative Filtering Approach for Urban Black Holes Detection. In *Proceedings of the ACM CIKM 2016*.

[14] Kira Kempinska, Paul Longley, and John Shawe-Taylor. 2018. Interactional regions in cities: making sense of flows across networked systems. *International Journal of Geographical Information Science* 32, 7 (2018), 1348–1367.

[15] Simon Kwoczek, Sergio Di Martino, and Wolfgang Nejdl. 2014. Predicting and visualizing traffic congestion in the presence of planned special events. *Journal of Visual Languages & Computing* 25, 6 (2014), 973 – 980.

[16] Simon Kwoczek, Sergio Di Martino, and Wolfgang Nejdl. 2015. Stuck Around the Stadium? An Approach to Identify Road Segments Affected by Planned Special Events. In *IEEE ITSC 2015*.

[17] Freddy Lécué, Simone Tallevi-Diotallevi, Jer Hayes, Robert Tucker, Veli Bicer, Marco Luca Sbodio, and Pierpaolo Tommasi. 2014. STAR-CITY: Semantic Traffic Analytics and Reasoning for CITY. In *Proceedings of the IUI '14*.

[18] Ming Li, René Westerholt, Hongchao Fan, and Alexander Zipf. 2018. Assessing spatiotemporal predictability of LBSN: a case study of three Foursquare datasets. *GeoInformatica* 22, 3 (2018), 541–561.

[19] Yuxuan Liang, Zhongyuan Jiang, and Yu Zheng. 2017. Inferring Traffic Cascading Patterns. In *Proceedings of the 25th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, GIS 2017, Redondo Beach, CA, USA, November 7-10, 2017*. 2:1–2:10.

[20] Xiaolei Ma, Haiyang Yu, Yunpeng Wang, and Yinhai Wang. 2015. Large-Scale Transportation Network Congestion Evolution Prediction Using Deep Learning Theory. *PLOS ONE* 10, 3 (03 2015), 1–17.

[21] Michael McNally. 2008. The four step model. In *Handbook of Transport Modeling*, David A. Hensher and Kenneth J. Button (Eds.). Emerald, Bingley, 35–53.

[22] Chuishi Meng, Xiuwen Yi, Lu Su, Jing Gao, and Yu Zheng. 2017. City-wide Traffic Volume Inference with Loop Detector Data and Taxi Trajectories. In *Proceedings of the 25th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, GIS 2017*. 1:1–1:10.

[23] Robert Meusel, Petar Petrovski, and Christian Bizer. 2014. The WebDataCommons Microdata, RDFa and Microformat Dataset Series. In *Proceedings of the ISWC'14*.

[24] Vaia Moustaka, Athena Vakali, and Leonidas G. Anthopoulos. 2018. A Systematic Review for Smart City Data Analytics. *ACM Comput. Surv.* 51, 5, Article 103 (Dec. 2018), 41 pages. https://doi.org/10.1145/3239566

[25] Ming Ni, Qing He, and Jing Gao. 2017. Forecasting the Subway Passenger Flow Under Event Occurrences With Social Media. *IEEE Trans. Intelligent Transportation Systems* 18, 6 (2017), 1623–1632.

[26] G. Pan, G. Qi, Z. Wu, D. Zhang, and S. Li. 2013. Land-Use Classification Using Taxi GPS Traces. *IEEE Transactions on Intelligent Transportation Systems* 14, 1 (2013), 113–123.

[27] F. Rodrigues, S. S. Borysov, B. Ribeiro, and F. C. Pereira. 2017. A Bayesian Additive Model for Understanding Public Transport Usage in Special Events. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39, 11 (2017).

[28] Gao Song, Janowicz Krzysztof, and Couclelis Helen. 2017. Extracting urban functional regions from points of interest and human activities on location-based social networks. *Transactions in GIS* 21, 3 (2017), 446–467.

[29] R. Soua, A. Koesdwiady, and F. Karray. 2016. Big-data-generated traffic flow prediction using deep learning and dempster-shafer theory. In *Proceedings of the 2016 International Joint Conference on Neural Networks (IJCNN)*. 3195–3202.

[30] Ying Sun, Hengshu Zhu, Fuzhen Zhuang, Jingjing Gu, and Qing He. 2018. Exploring the Urban Region-of-Interest Through the Analysis of Online Map Search Queries. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD '18)*. 2269–2278.

[31] Nicolas Tempelmeier, Elena Demidova, and Stefan Dietze. 2018. Inferring Missing Categorical Information in Noisy and Sparse Web Markup. In *Proceedings of The Web Confonference 2018*.

[32] Hongjian Wang and Zhenhui Li. 2017. Region Representation Learning via Mobility Flow. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management (CIKM '17)*. ACM, 237–246.

[33] L. Wang, Y. Zheng, X. Xie, and W. Ma. 2008. A Flexible Spatio-Temporal Indexing Scheme for Large-Scale GPS Track Retrieval. In *Proceedings of the Ninth International Conference on Mobile Data Management (mdm 2008)*. 1–8. https://doi.org/10.1109/MDM.2008.24

[34] Xiaomeng Wang, Ling Peng, Tianhe Chi, Mengzhu Li, Xiaojing Yao, and Jing Shao. 2016. A Hidden Markov Model for Urban-Scale Traffic Estimation Using Floating Car Data. *PLOS ONE* 10, 12 (12 2016), 1–20.

[35] Yilun Wang, Yu Zheng, and Yexiang Xue. 2014. Travel Time Estimation of a Path Using Sparse Trajectories. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '14)*. ACM, 25–34.

[36] Junjun Yin, Aiman Soliman, Dandong Yin, and Shaowen Wang. 2017. Depicting urban boundaries from a mobility network of spatial interactions: a case study of Great Britain with geo-located Twitter data. *International Journal of Geographical Information Science* 31, 7 (2017), 1293–1313.

[37] Bang Zhang, Lelin Zhang, Ting Guo, Yang Wang, and Fang Chen. 2018. Simultaneous Urban Region Function Discovery and Popularity Estimation via an Infinite Urbanization Process Model. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD '18)*. 2692–2700.

[38] Desheng Zhang, Juanjuan Zhao, Fan Zhang, and Tian He. 2015. UrbanCPS: A Cyber-physical System Based on Multi-source Big Infrastructure Data for Heterogeneous Model Integration. In *Proceedings of the ACM/IEEE Sixth International Conference on Cyber-Physical Systems (ICCPS '15)*. 238–247.

[39] Xun Zhou, Amin Vahedian Khezerlou, Alex X. Liu, M. Zubair Shafiq, and Fan Zhang. 2016. A traffic flow approach to early detection of gathering events. In *Proceedings of the 24th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, GIS 2016*. 4:1–4:10.