Issue 6 | March 2022

MASTER

Multiple ASpects TrajEctoRy management and analysis

Inside this issue

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant No 777695 Duration: March 1 2018 | December 31 2023

Content

Editorial

Featured

Research Highlights

Next events

Editorial

We are very happy that the MASTER project is officially restarting on January 2nd, 2022 after 8 months of formal suspension due to the pandemic restrictions. Even under a formal interruption, the dissemination activities of MASTER went on since we published papers and we also organised the Dagstuhl Seminar on Mobility and Ethics.

We are very proud to also restart our networking with European funded projects related to MASTER. We are hosting in this issue the project MOBIDATALAB (Labs for prototyping future mobility data sharing solution in the Cloud) whose aim is to foster data sharing in the transport sector, providing mobility organising authorities with recommendations on how to improve the value of their data, contributing to the development of open tools in the cloud. The project also organises hackathons aiming to find innovative solutions to concrete mobility problems.

We report on three articles recently published by our partners. The first two papers are proposing methods to store and compare multiple aspects trajectories in an efficient way. The first one is a collaboration of UFSC and CNR on proposing a new index for storing and comparing multiple aspects trajectories:
MAT-INDEX: an Index for Semantic Trajectory Similarity Measuring by Ana Paula Ramos Souza (UFSC), Chiara, Renso (CNR), Raffaele Perego (CNR) and Vania Bogorny (UFSC)

A second paper is a collaboration between HUA, UFC and UFSC on a compression algorithm for trajectories:
Evaluating the effect of compressing algorithms by Antonios Makris, (HUA), Vania Bogorny (UFSC)

A third paper is again a collaboration between CNR and UFSC on the conceptual modelling of dependency rules in multiple aspects trajectories and extending the MASTER model with rules to express dependency between attributes:
On Modelling Dependency Rules for Multiple Aspects Trajectories by Ronaldo dos Santos Mello (UFSC) and Chiara Renso (CNR)

We hold a Dagstuhl seminar on the encounter of Mobility Data with Ethics. Inspired by our Independent Ethical Advisor, Prof. Berendt, and organised with the support of Prof. Stan Matwin from Dalhousie and Chiara Renso from CNR, we invited researchers from Mobility and Ethics to reflect on the possible ethical issue on analysis of mobility data. We traced a possible path for the creation of a Mobility Data Ethics field.

With the project resumption we also restart our events: the Second MASTER workshop is planned to be held on May 13, 2022 hosted by University Ca' Foscari in Venice. The objective of the workshop is to invite representatives from non academic and industrial environment to present their data and open problems, while discussing MASTER results. We also replanned the Data Science for Mobility summer school to be held in Santorini, in October 2022.

You can download this and previous issues of the newsletter from the MASTER web site:
http://www.master-project-h2020.eu

Stay tuned and happy reading!

Chiara Renso, MASTER Project Coordinator

Mobility Data Analysis: from Technical to Ethical

Dagsthul seminar report

Chiara Renso, ISTI-CNR, Italy

MASTER promoted a Dagstuhl seminar with the title: “Mobility Data Analysis: from Technical to Ethical”, organized by Bettina Berendt (MASTER Independent Ethical Advisor), Stan Matwin (Principal Investigator of Dalhousie University) and Chiara Renso (MASTER Coordinator of CNR).

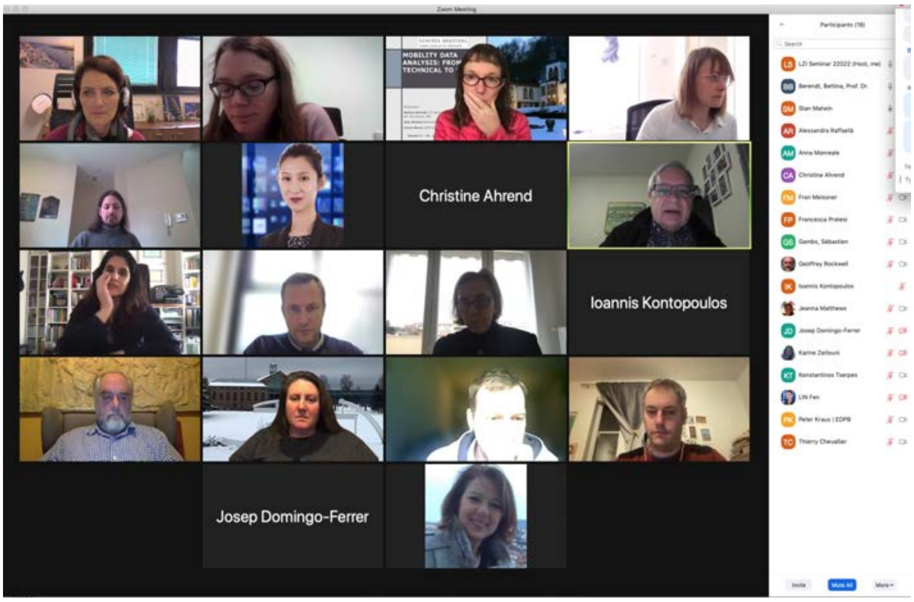
Schloss Dagstuhl – Leibniz-Zentrum für Informatik GmbH (English: Schloss Dagstuhl – Leibniz Center for Informatics) pursues its mission of furthering world class research in computer science by facilitating communication and interaction between researchers.

Schloss Dagstuhl was founded in 1990 as a non-profit organization and quickly became established as one of the world’s premier meeting centers for informatics research. Schloss Dagstuhl is providing the DBLP in bibliographic services.

The seminar has been held fully remote due to the COVID-related restrictions and lasted three full days from January 10 to January 12, 2022. The objective was to start a deep interacting discussion between Mobility Data Analysis researchers and Ethics experts to link these two fields to create the foundations of a new Mobility Data Ethics research field.

The program included three tutorials:
– **Location privacy:** an overview by Sébastien Gambs (Université du Québec à Montréal (UQAM), Canada)
– **Mobility data analysis:** ethical issues by Geoffrey Rockwell (University of Alberta, Canada)
– **Connected vehicles and mobility data** – work done by the EDPB by Peter Kraus (European Data Protection Board, Brussels, Belgium)

Twenty-five participants from the mobility and ethics world – inspired by the three tutorials – discussed some topics:
1) Data utility and data privacy: which trade off?
2) Mobility Data Analysis Ethics beyond the data,
3) Ethics on mobility data: what is unique and which guidelines?
4) Mobile data anonymity: mobile data is never anonymous?
5) Ethics for non-human movements?



The participants could interact through Zoom and Gather Town.

Participants:
Darren Abramson (Dalhousie University, CA)
Christine Ahrend (TU Berlin, DE)
Bettina Berendt (TU Berlin, DE)
Florence Chee (Loyola University Chicago, US)
Thiery Chevallier (Akka Technologies, FR)
Maria Luisa Damiani (University of Milan, IT)
Josep Domingo-Ferrer (Universitat Rovira i Virgili – Tarragona, ES)
Sébastien Gambs (University of Montreal, CA)
Ioannis Kontopoulos (Harokopio University – Athens, GR)
Peter Kraus (European Data Protection Board – Brussels, BE)
Fen Lin (City University – Hong Kong, HK)
Jeanna Matthews (Clarkson University – Potsdam, US)
Stan Matwin (Dalhousie University – Halifax, CA)
Fran Meissner (University of Twente, NL)
Anna Monreale (University of Pisa, IT)
Francesca Pratesi (ISTI-CNR – Pisa, IT)
Alessandra Raffaetà (University of Venice, IT)
Chiara Renso (ISTI-CNR – Pisa, IT)
Paula Reyero-Lobo (The Open University – Milton Keynes, GB)
Geoffrey Rockwell (University of Alberta – Edmonton, CA) [dblp]
Yannis Theodoridis (University of Piraeus, GR)
Konstantinos Tserpes (Harokopio University – Athens, GR)
Karine Zeitouni (University of Versailles, FR)

As a first result of the seminar a group of participants published an article at the magazine The Conversation Canada to discuss the (non) anonymity of location data collected in Canada for monitoring pandemic and the related privacy issues. https://theconversation.com/ottawas-use-of-our-location-data-raises-big-surveillance-and-privacy-concerns-175316

MobiDataLab

A solution to make mobility data findable, accessible, interoperable and reusable.

Thierry Chevallier, MobiDataLab project coordinator, AKKA, Toulouse, France | thierry.chevallier@akka.eu

For just over a year, a consortium of 10 European research organisations, including academics, associations and industry, has been working on a mobility data sharing solution, funded by the EU as part of the H2020 research programme for smart, green and integrated transport. The rationale for this project is based on the observation that new mobility services are emerging in the public and private sector, that more and more data is being produced, and that sharing mobility data can lead to more efficient processes and new products.

In a nutshell, MobiDataLab's objective is to foster data sharing in the transport sector. As its name suggest, it is a "lab" for prototyping new mobility data sharing solutions. The MobiDataLab consortium brings a wide range of complementary expertise, in fields as different as legal and governance (KU Leuven), market analysis (AETHON and ICOOR), cloud technologies (AKKA and CNR), data privacy (university Rovira i Virgili), location-based services (HERE), Mobility as a Service (KISIO), and communication (POLIS and F6S). MobiDataLab will produce:

• an open knowledge base, a "Wikipedia of mobility data", with a particular focus on the technical, regulatory and commercial aspects of sharing mobility data. Why this knowledge base? because for Mobility organising authorities and transport operators, sharing best practices is as important as sharing data.

• a prototype of a cloud solution for sharing mobility data, intended for transport actors, which we have called a "transport cloud". This data sharing platform ensures the transnational access to mobility data in a secure, performant and seamless way. It integrates a standardized Open Data catalogue for improving their discoverability, data processors for enriching the data geographically and semantically, and anonymization tools to increase trust in the platform via privacy-preserving techniques.

• an innovation environment (namely living and virtual labs), bringing together data providers and users in hackathons and datathons, and where innovators will use our transport cloud prototype as a tool to address mobility challenges proposed by a reference group of European municipalities (comprising for example the mobility agencies of Rome,

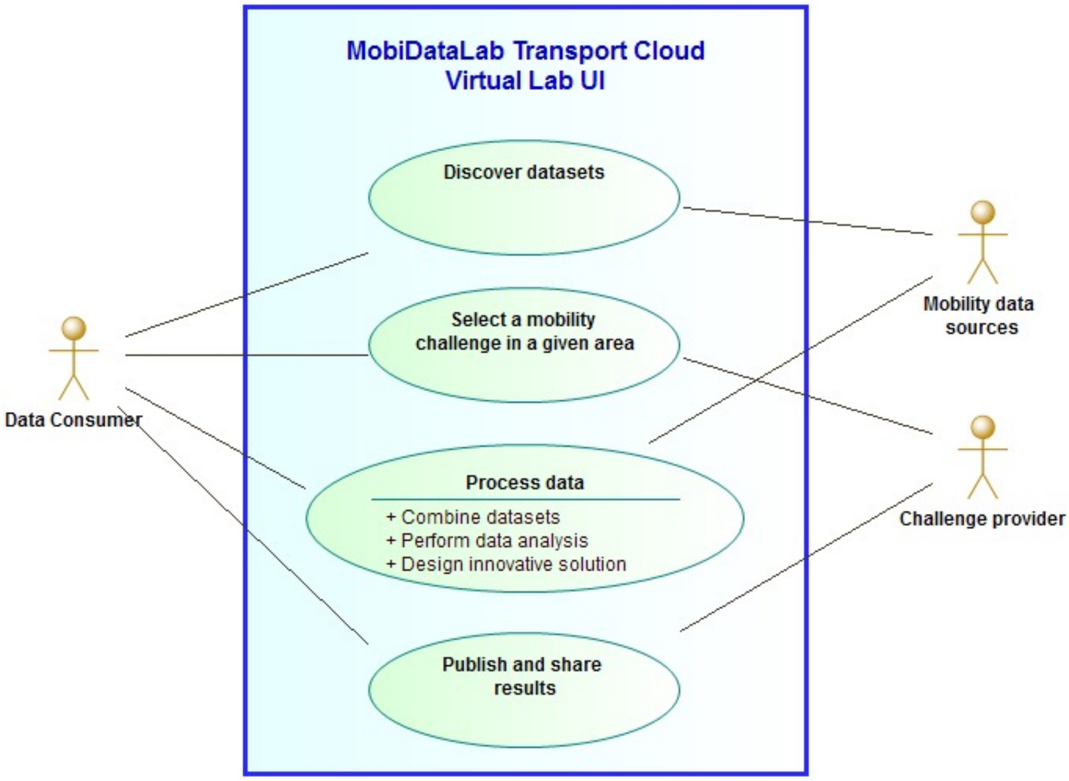


Figure 1: MobiDataLab use case

Milan, the cities of Eindhoven, Leuven, Malaga, Timisoara, etc.)

To achieve its objective, the MobiDataLab approach is to make mobility data FAIR (Findable, Accessible, Interoperable, Reusable) in the territorial context of our partner municipalities, and for particular use cases. With technology providers such as KISIO and HERE integrating open data transport datasets into their products, and startups participating in our hackathons, a particular emphasis is placed on reusability.

MobiDataLab use cases basically combine real-time public transport, road & vehicle data and allow data consumer to solve specific mobility challenges faced by transport stakeholders (cf. Figure 1). Use cases for Operations include the optimization of transport flows, emission reporting, analytics and Learnings, re-use of transport data for journey planners and Mobility-as-a-Service. Use cases for Research correspond to the geographical enrichment (with e.g. OpenStreetMap and environmental data) and the semantic enrichment of mobility data, detailed later in this article.

Thanks to our colleagues from the CNR, MobiDataLab came into contact with the MASTER project, and we quickly found insightful synergies and fruitful collaboration opportunities. In particular the works of MASTER on holistic trajectories are very useful for the MobiDataLab use cases related to the semantic enrichment of mobility data.

Semantic enrichment of mobility data

Information available from Linked Open Data (LOD) or extracted from data sources on the web like e.g. bus/train companies web sites, social networks, etc. can be used to enrich movement data with timetables, weather and public opinions on the quality of the provided service.

An interesting application is the mobility of tourists. In this example, trajectories of tourists visiting a city (sightseeing and using different transport modes) could be enriched with the features of the attractions they visit. Trajectories can be individual GPS tracks (difficult to obtain due privacy issues) or theoretical trajectories suggested by journey planners (such as the Navitia API from our partner KISIO). These multimodal trajectories are segmented by transport mode and greatly benefit from being semantically annotated. Indeed the choice between different routes does not necessarily depend on the best travel time, but can be based on the carbon footprint of the corresponding transport mode, on the safety of the route, or on the presence of points of interest.

The Virtual Lab, using transport open data as a tool

The works on MobiDataLab, can also be useful to the MASTER project, especially in the implementation of application scenarios related to public transportation. Researchers working on "Analytics and Learnings" of mobility data will be able to use an instance of the MobiDataLab virtual labs. Virtual Labs provide a technological gateway to the Transport Cloud as well as social and collaborative tools necessary for the interaction between data consumers and data providers, and for the execution of co-creation activities.

What's next for MobiDataLab ?

The next step of the project is the first version of the prototype available by the end of 2022. Then the consortium will be able to organise the innovation sessions bringing together data providers and users in hackathons and datathons, and where innovators will use the Transport Cloud prototype as a tool to address mobility challenges proposed by the reference group.

For more info:
https://www.mobidatalab.eu

MOBIDATALAB

Labs for prototyping future mobility data sharing solutions in the cloud

MAT-INDEX

an Index for Semantic Trajectory Similarity Measuring

Ana Paula Ramos de Souza, Santa Catarina Federal University (UFSC) Florianópolis, Brazil

Trajectories are sequences of points located in space and time that can describe any movement behavior. The trajectory definition has evolved from raw, before 2007 representing the spatial and temporal components, to semantic in 2008 representing the semantic label of trajectory points and then to the very recent concept of multiple aspect trajectory [2] where the semantic dimensions become large in numbers and heterogeneous in the form.

The problem of finding when two trajectories are similar is well-known in the literature since several approaches have been proposed [5].

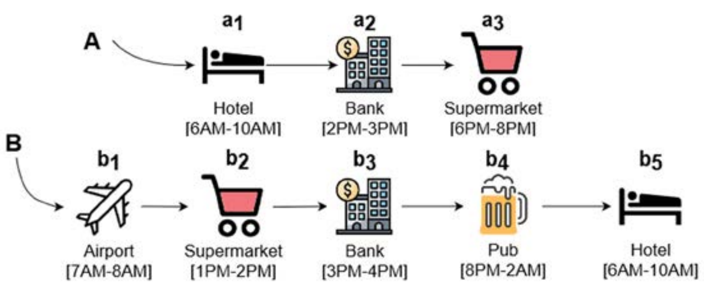


Figure 1. Two Multiple Aspects Trajectories

Figure 1 illustrates two multiple aspects trajectories, A and B having as semantic dimensions the Point of Interest, the category and the time. How to compare these trajectories in all dimensions in an efficient and effective way?

State-of-the-art methods MSM [1] and MUITAS [3] pairwise compare all points including the semantic dimensions of trajectories A and B, therefore they are quite inefficient having high complexity. It is required a smart indexing structure that supports all dimensions -spatial, temporal and semantic- for fast similarity search in real datasets.

In the literature there is a limited number of works that propose to index spatial, temporal and semantic data, but they do not fully index all three trajectory dimensions. To the best of our knowledge, none of these works were developed for trajectory similarity purposes.

We propose a new index called MAT-INDEX that optimizes the storage of the three dimensions by significantly reducing (up to 98.1%) the processing time on evaluations for measuring the multidimensional similarity of trajectories. We experimented the index with two state-of-the-art similarity measures for multiple aspects trajectories MSM [1] and MUITAS[3]. The MAT-index construction avoids the need for point-to-point comparison, consolidating all different dimensions in a single data structure containing the matching scores between trajectories. All details can be found in the full paper [4].

Multiple Aspect Trajectory Index.

MAT-Index focuses on indexing all three multiple aspect trajectory dimensions in a single data structure, eliminating redundant operations. It is divided in six steps: (1) load, (2) spatial computation, (3) temporal computation, (4) semantic combine, (5) semantic compress, and (6) dimensions integration.

(1) Load Step: The Load Stage saves each trajectory dimension into individual intermediate data structures. Figure 2 shows a toy dataset (a) and its corresponding data dimensions after loaded.

- **Loading the Spatial Dimension:** MAT-Index allocates the spatial (b) dimension using an inverted list, simulating a (logical) grid data structure. This logical grid only stores the cells with at least one trajectory point, therefore in the worst-case, the number of cells will be equal to the number of trajectory points in the dataset. The well-dimensioned cell size ensures that all points belonging to the same cell automatically match.
- **Loading the Temporal Dimension:** Analogously, MAT-Index also uses an inverted list data structure to allocate the temporal (c) dimension. However, considering the very limited range of possibilities and unidimensionality of the time expressing a day (eg., hours, minutes, etc.), the index stores the temporal data using the same unit of the similarity measure threshold to facilitate its computation.
- **Loading the Semantic Dimension:** The load step preliminary stores the semantic content into two structures: a list (d) with the distinct semantic composite keys occurred in the dataset, and a two level index (e). Concerning the second structure, the first dataset row provides the attribute names price, poi, and weather that composes its first level keys and their corresponding attribute values, preserving the context of the values. MAT-Index merges both semantic structures in the Semantic Combine Step then optimizes it in the Semantic Compress Step. The semantic content preliminary saved in the load step into two structures

After the first step (load) stores the intermediate structures, each dimension is processed in the following five steps considering its particularities and then the results are consolidated in the final step in a single structure that contains the number of matches between pairs of trajectories. The basics are briefly explained in the follows:

(2) Spatial Computation: The well-dimensioned cell size of the logical grid

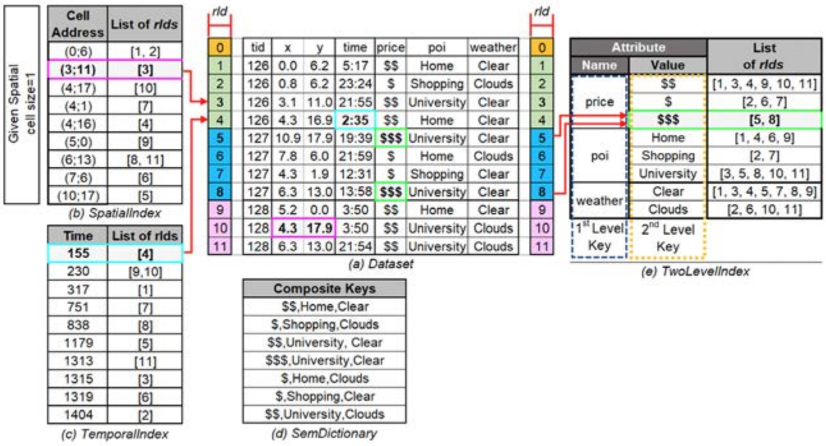


Figure 2. The dataset and the intermediate structures after the load computation

assures that the maximum distance between two points in the same cell never exceeds the threshold, thus only points in neighbor cells demand to pairwise calculate the distance to check if they match. The pairs of matches are used to update the final MAT-index score, in the Index Integration Step.

(3) Temporal Computation: For each cell in Figure 2 (c), the step retrieves a list of trajectory points allocated in the threshold admitted that automatically matches with all points regarding the cell in processing, thus it is computed and associated only once.

(4) Semantic Combine: MAT-Index gets each valid combination of attribute values in Figure 2 (d), and consulting the list of occurrences by attribute, for each attribute value in Figure 2 (e), it creates a list of matches that the combination have with each trajectory point in the dataset. The step explores the transitive property that if a = b and b = c, thus a = c to avoids pairwise comparisons. At this point, a single direct access may retrieve the number of semantic features that match between trajectory points, although the points were never pairwise compared.

(5) Semantic Compress: Both MSM and MUITAS retrieve the best semantic match of a point when compared to a trajectory. In this case, we can further compress the index by keeping the trajectory maximum score of each semantic composite key. Figure 3 shows the scores of the example before (b) and after (a) the compression.

(6) Dimensions Integration: Indexing the trajectory dimensions in a single data structure avoids redundant comparisons, speeding up the trajectory similarity analysis. Therefore, this phase consolidates the matching scores into a single data structure, finalizing the MAT-index construction. It updates each trajectory point with the spatial and/or temporal matches computed in the previous steps, resulting in a single data structure containing only the top scores (number of matches) between trajectories. The Mat-Index depicted in Figure 3 (c) is ready to be used.

We have tested MSM [1] and MUITAS [3] with and without MAT-Index support using FourSquare1 and BerlinMOD2 datasets considering three perspectives: running time, number of semantic composite keys and number of attributes.

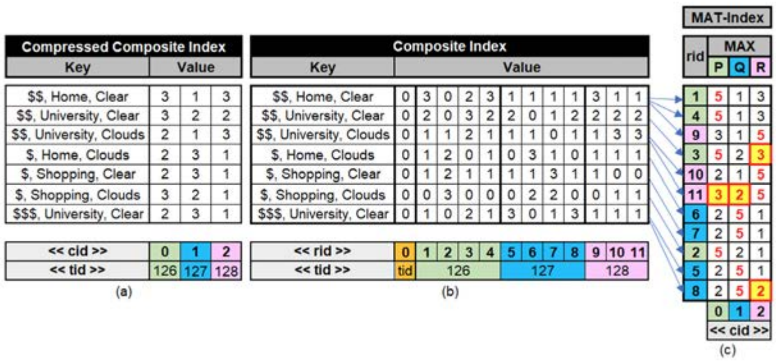


Figure 3. Composite Index after (a) and before (b) the compression, and the final MAT-Index (c) after Integration

MAT-Index is stable and faster in all tested scenarios of both datasets. While the original implementation has its performance strictly related to the number of attributes, the results using MAT-Index support tend to proportionally improve as the number of trajectory aspects/attributes increases. This characteristic is essential for multiple aspect trajectories since they tend to have many attributes.

We also notice that the MAT-Index performance is more associated with the number of processed semantic composite keys.

The proposed MAT-Index intends to fill this gap by indexing the trajectory dimensions (including a semantic content with multiple attributes) into a compact data structure, assuring efficiency in similarity computation while reducing data redundancy. The MAT-Index support for MSM and MUITAS drastically reduces the needed comparisons. Experiments show an improvement of up to 98.1%. Future works include: (1) experi- ment MAT-Index with additional semantically enriched trajectories datasets; (2) perform a complexity analysis with average and worst-case scenarios to study the index limits; (3) develop an efficient index update to avoid reprocessing.

1)https://sites.google.com/site/yangdingqi/home/foursquare-dataset/
2) https://secondo-database.github.io/BerlinMOD/BerlinMOD.html

References

[1] FURTADO, A. S., KOPANAKI, D., ALVARES, L. O., AND BOGORNY, V. Multidimen- sional similarity measuring for semantic trajectories. TGIS (2016).
[2] MELLO, R. D. S., BOGORNY, V., ALVARES, L. O., SANTANA, L. H. Z., FERRERO, C. A., FROZZA, A. A., SCHREINER, G. A., AND RENSO, C. MASTER: A multiple aspect view on trajectories. TGIS (2019).
[3] PETRY, L., FERRERO, C., ALVARES, L. O., RENSO, C., AND BOGORNY, V. Towards semantic-aware multiple-aspect trajectory similarity measuring. TGIS (2019).
[4] SOUZA, A. P. R. D., RENSO, C., PEREGO, R., AND BOGORNY, V. MAT-index: An index for fast multiple aspect trajectory similarity measuring. Transactions in GIS (2022). First published: 20 January 2022.
[5] SU, H., LIU, S., ZHENG, B., ZHOU, X., AND ZHENG, K. A survey of trajectory distance measures and performance evaluation. The VLDB Journal 29, 1 (jan 2020), 3–32.

Evaluating the Effect of Compressing Algorithms

for trajectory similarity and classification problems

Antonios Makris, Department of Informatics, Telematics, Harokopio University, Athens, Greece | amakris@hua.gr
Vania Bogorny, Programa de Pós-Graduação em Ciência da Computação, Universidade Federal de Santa Catarina (UFSC), Florianópolis, Brazil | vania.bogorny@ufsc.br

During the last few years the volumes of the data that synthesize trajectories have expanded to unparalleled quantities. This growth is challenging traditional trajectory analysis approaches and solutions are sought in other domains. We focus on data compression techniques with the intention to minimize the size of trajectory data, while, at the same time, minimizing the impact on the trajectory analysis methods. The effectiveness of the compression is evaluated using classification techniques and similarity measures. As the results showed, there is a tradeoff between the compression rate and the achieved quality. There is no "best algorithm" for every case and the choice of the proper compression algorithm is an application-dependent process.

Massive volumes of moving object trajectories are being generated by a range of devices, services and systems, such as GPS devices, RFID sensors, location-based services, satellites and wireless communication technologies. In essence, these data are spatio-temporal in nature, represented as chronologically ordered points consisting of the position of a moving object, along with the time it was collected. The more points a device collects, the more accurate a trajectory representation becomes. However, the increasing amount of data poses new challenges related to storing, transmitting, processing, and analyzing these data [1-2].

Storing the sheer volumes of trajectory data can quickly overwhelm available data storage systems. For instance, one AIS-receiving station receives 6 AIS messages, or 1800 bytes, per second. Considering that a ship tracking company may operate thousands of such receivers, we end up with terabytes of data to deal with each month. For example, MarineTraffic's [L2] on-line monitoring service collects 60MB of AIS and weather data every second. Additionally NASA's Earth Observing System produces 1 TB of data per day while Hubble Telescope [L1] generates 140 GB of raw spatial data every week. These facts demonstrate that the transmission of large trajectory data can be proven prohibitively expensive and problematic [3]. In addition, these data often contain large amounts of redundant information which in turn can overwhelm human analysis.
A typical approach towards these challenges is to reduce the size of trajectory data by employing compression techniques. The latter aims at substantial reductions in the amount of data while preserving the quality of information, and thus ensuring the accuracy of the trajectory. Most works on trajectory compression neglect to evaluate the impact of compression techniques over their intended task.

The objective of this work is to move one step forward by investigating and quantifying the relationship as well as the impact of trajectory compression techniques over trajectory classification and similarity analysis. In particular, five top-down compression algorithms were compared, namely Douglas-Peucker (DP), Time-Ratio (TR), Speed-Based (SP), Time Ratio-Speed Based (TR_SP) and Speed Based-Time Ratio (SP_TR) over similarity and classification. The comparison is based on the compression rate achieved and the execution time across four real-world dataset namely, GeoLife, Hurricane, Animals and UFC.

The datasets used for evaluating the compression trajectory algorithms were selected according to some criteria: i) the data are categorized in classes, so they can be used to evaluate both classification problems and similarity measures., ii) they contain trajectories of different objects (people, vehicles, hurricanes, and animals) and iii) they present variations in their sampling rates. The performance characteristics of the compared algorithms were evaluated against six different dynamically defined thresholds. This practically means that in every trajectory of each dataset, a different threshold is applied in the corresponding algorithm, which depends on the actual features and peculiarities of this trajectory. Thus, we have eliminated the need of arbitrary user-defined thresholds. The examined algorithms are lossy which means that they present information loss and therefore it is important to measure the quality of the compressed data. To assess the quality of the compression algorithms, data mining techniques were employed in order to evaluate the capability of extracting meaningful conclusions from the trajectories even after their compression. These techniques include four trajectory classification methods namely Dodge, Zheng, Xiao and Movelets and two trajectory similarity measures namely EDR and UMS. More specifically, the classification techniques aim to identify the classes of the trajectories by extracting features that are capable of distinguishing the classes while the purpose of the similarity measures is to obtain a quantitative measure between any two trajectories, thus to identify to what extent two objects are similar.

Figure 1 demonstrates the obtained results by using the Random Forest classification algorithm for the different classification tasks. The higher values in each line are presented in bold, while the second highest is underlined. The accuracy results are presented for each classification technique and dataset, by using the original dataset in the first line, and the maximum and minimum accuracy achieved by any of the compression techniques in the second and third lines, respectively. This is the best way for presenting the accuracy without any compression, the best result after compression, and also how "bad" in terms of quality the accuracy can be after compression.

The main contribution of this work is that it provides evidence using some proof concept scenarios that, in terms of data points representing a trajectory, only a small portion of carefully selected points are needed in order to conduct the analysis task without significant loss in its performance metrics (e.g. accuracy, precision, recall). We demonstrate that there is a trade-off between the compression rate and the achieved quality and that choosing a proper compression algorithm is not an easy task: efficiency analysis demonstrates that there is no "best algorithm" and the selection is application-dependent.

Our future plan is to expand the comparison with online compression algorithms that are extensively used, such as STTrace, Dead-Reckoning and SQUISH.

Links:

[L1]: https://hubblesite.org/
[L2]: https://www.marinetraffic.com/

References:

[1] Langran G. Issues of implementing a spatiotemporal system. International Journal of Geographical Information Science. 1993 Jul 1;7(4):305-14.
[2] Potamias M, Patroumpas K, Sellis T. Sampling trajectory streams with spatiotemporal criteria. In18th International Conference on Scientific and Statistical Database Management (SSDBM'06) 2006 Jul 3 (pp. 275-284). IEEE.
[3] Makris A, Tserpes K, Spiliopoulos G et al. MongoDB Vs PostgreSQL: A comparative study on performance aspects. GeoInformatica. 2021 Apr;25(2):243-68.

Dataset	Metric	Techniques			
		Zheng	Dodge	Xiao	Movelets
Animals	original	87.17%	82.05%	85.89%	**91.02%**
	max. ACC	**91.02%** (*SP_TR*)	85.89% (*TR_SP*)	**91.02%** (*SP_TR*)	**91.02%** (*SP_TR*)
	min. ACC	71.79% (*DP*)	65.38% (*TR*)	73.07% (*DP*)	**74.35%** (*DP*)
UFC	original	19.81%	18.22%	26.02%	**89.9%**
	max. ACC	19.65% (SP_TR)	17,39% (*SP_TR*)	24.43% (*SP_TR*)	**89.64%** (*SP_TR*)
	min. ACC	5.57% (*DP*)	3.45% (*DP*)	6.24% (*DP*)	**83.13%** (*TR*)
Hurricane	original	58.22%	60.52%	60.19%	**61.84%**
	max. ACC	60.85% (*DP*)	59.86% (*SP*)	60.85% (*TR*)	**62.82%** (*TR_SP*)
	min. ACC	54.60% (*SP*)	53.28% (*TR*)	55.59% (*DP*)	**56.56%** (*DP*)
GeoLife	original	82.33%	81.01%	83.77%	**83.98%**
	max. ACC	80.26% (*SP_TR*)	79.69% (*SP*)	**82.14%** (*SP_TR*)	80.26% (*SP_TR*)
	min. ACC	69.73% (*SP_TR*)	56.01% (*SP*)	67.10% (*SP*)	**72.74%** (*TR_SP*)

Figure 1. Summary of the classification results with the accuracy obtained from the original dataset without compression and the maximum and minimum accuracy achieved from the compressed datasets.

On Modeling Dependency Rules for Multiple Aspects Trajectories

Ronaldo dos Santos Mello, Universidade Federal de Santa Catarina, Florianópolis, Brazil | r.mello@ufsc.br
Chiara Renso, ISTI-CNR, Pisa, Italy | chiara.renso@isti.cnr.it

This article introduces dependency rules to represent patterns discovered from the analysis of trajectories with multiple aspects. These rules are conceptually represented as an extension of a conceptual model for multiple aspects trajectories called MASTER, allowing trajectories and their patterns to be queried by any data scientist working on trajectory data management and analysis.

Trajectories of moving objects are usually modeled as sequences of space-time points or, in case of semantic trajectories, as labelled stops and moves. Data analytics methods on these kinds of trajectories to discover geometrical and temporal patterns, or simple semantic patterns based on the labels of stops and moves. A recent extension of semantic trajectories is called Multiple Aspects Trajectory (MAT), i.e., a trajectory associated to different semantic dimensions called aspects. Our research group had proposed a conceptual model for representing MAT data (the MASTER data model [mello 2019]). It includes the modeling of MATs, their points, the moving object that owns the MAT, the aspects related to them and their attributes, as well as aspect types. Figure 1 (yellow entities) gives a MASTER overview.

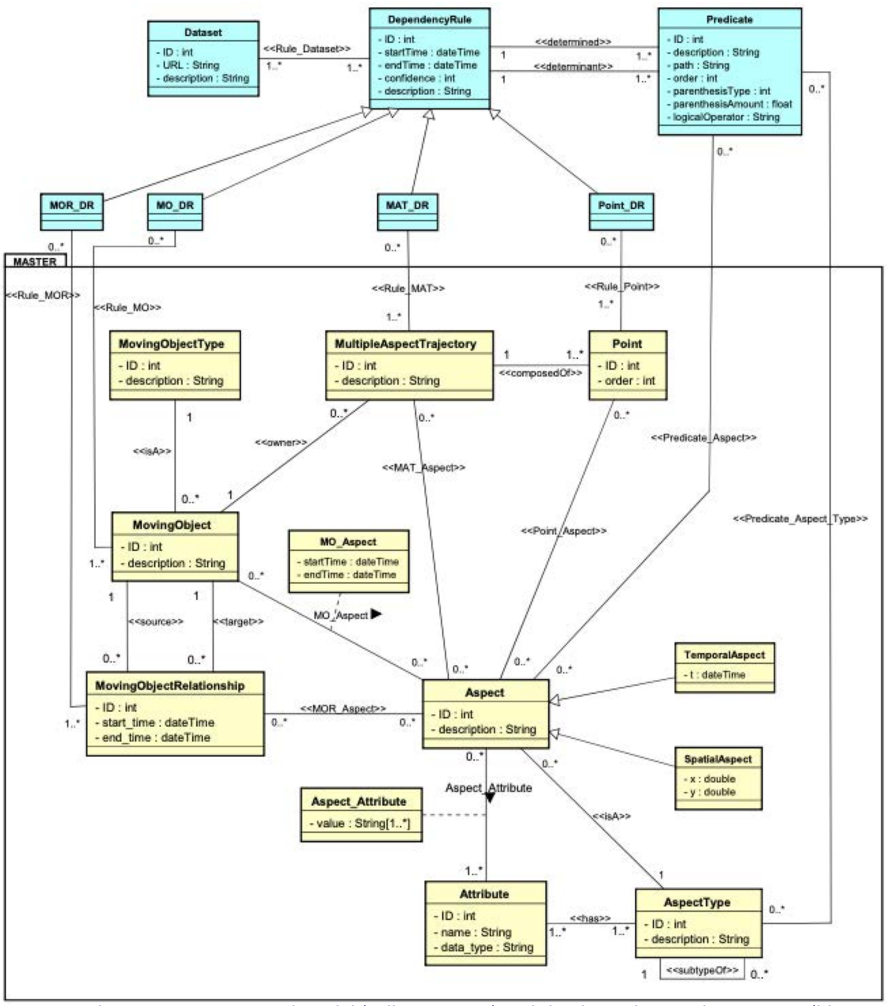


Figure 1. The MASTER conceptual model (yellow entities) and the dependency rule extension (blue entities).

tend

A MAT increases in a large scale the number of discovered patterns, as the behaviour of a moving object may involve several aspects and, additionally, some aspects may be strongly correlated (or dependent) and may not be analyzed separately. For instance, suppose a restaurant (an aspect), visited by trajectories of a person, with some reviews, average price and rating. These attributes may hold a dependency stating that ratings equal to 10 have average price higher than U$ 100 and excellent or good reviews. We call these dependency relationships between attributes a Dependency Rule (DR) [mello 2021]. A DR may be learned through data mining or machine learning methods, or it may be predefined by the user.

Finding, representing and storing these dependencies is therefore an essential step when analysing MATs. Although there are advances in trajectory data modeling and mining, there is no con-

```
MAT | owner.is-a[description ='Person'] AND
      MAT_Aspect[description ='retired'].is-a[description ='occupation'] AND
      MAT_Aspect[NOT(description ='rainy')].is-a[description ='weather']
      ⇒
      MAT_Aspect[description = 'foot'].is-a[description = 'transportation']
```

Figure 2. Pattern that involves three aspects (occupation, transportation and weather)

sensus among approaches for modeling discovered patterns from trajectory data, and the existing ones have limitations. One example is the work of Bogorny [bogorny 2010], which models patterns for trajectories only in terms of stops and moves, and considers a few/fixed aspects. Different from the related work, our DR formalism is able to represent complex dependencies involving trajectory aspects.

Specifically, a DR allows the definition of determinant and determined attribute sets, complex predicates involving these attribute sets, and the real-world entity type on which the DR holds (a whole MAT, MAT points, or the MAT moving object). Our contribution with the DR modeling on MASTER is to represent discovered patterns for the main real-world entities based mainly on the analysis of the aspects that surround them, as the aspects represent the relevant features of the trajectories, including spatial and temporal information. As a DR example, suppose we had discovered a pattern in a MATs dataset stating that retired people in a small city usually move on foot when it is not raining. This pattern involves three aspects (occupation, transportation and weather) and could be specified as *Fig 2.*

This DR is a pattern for whole MATs (the entity type). It also holds a pre-condition (before the implication – the determinant part) and a discovered data behaviour based on this pre-condition (after the implication – the determined part). Figure 1 also shows our MASTER extension to provide the representation of DRs and related concepts (the blue entities). The DR entity has some attributes, like its lifetime (start and end times) and confidence. We also indicate the data sets on which it was discovered (the Dataset entity). Besides, a DR has specialized entities that hold specific relationships with the original MASTER entities that own the DR.

As shown in Figure 1, a DR is composed of the determinant and determined parts, which are sets of predicates. For sake of understanding of the Predicate Entity, suppose the predicates in *Fig 3* that could be part of a DR determinant or determined.

```
...((a.b.c[x = 1] OR d.e[y = 'q']) AND ...) ...
```

Figure 3. Predicates that could be part of a DR determinant or determined

We have two predicates (a.b.c[x = 1] and d.e[y = 'q']). The first one is preceded by two open parenthesis (parenthesisType = 'open' and parenthesisAmount = 2), and the second one is preceded by the OR logical operator and succeeded by one close parenthesis (logicalOperator = 'OR', parenthesisType = 'close' and parenthesisAmount = 1). This strategy to model predicates allows the representation of a DR composed of an arbitrary number of logical operators and parenthesis levels. We also associate a predicate with an aspect and/or an aspect type in order to allow queries like "what MATs have patterns that enclose the aspect X?" and "what aspect types are more frequent in patterns for the moving object Y?".

A validation of our extended model considering real trajectory data from the SoBigData Consortium repository (https://sobigdata.d4science.org/web/cityofcitizens/catalogue) was accomplished as a proof-of-concept. SoBigData is an European Union funding project related to Big Data social mining. Two repository datasets with trajectories from different means of transportation were mined in order to find out patterns with diffferent levels of complexity and high confidence. These datasets have several aspects associated to the trajectories, like average speed, day of the week, goal and day period. Satisfactorily, all of these found patterns were able to be represented by our model.

We also analyze query performance over our extended model designed over relational-based database technologies: a traditional database (PostgreSQL), and a NewSQL database (VoltDB). We define five queries covering the main model entities and relationships, with different levels of complexity (from 3 to 10 table joins), and emulate 20 users randomly run each one of the queries over synthetic data in a database with three sizes for each table: 10K, 50K and 100K rows. In short, the performance sounded good for both databases, as they were able to consume hundreds of requests even for the largest database size, with a light advantage for VoltDB.

As a conclusion, our proposed model combines simplicity by adding few entities to the original MASTER, as well as the capability to represent patterns with different levels of complexity. Our evaluations demonstrated that our model is useful and practicable. Future works include the analysis of other MAT datasets to better evaluate performance, expressiveness and limitations of the model. We also intend to simulate data insertions and updates over several database technologies.

References

[bogorny 2010] Bogorny, V., et. al.: A Conceptual Data Model for Trajectory Data Mining. In: GIS Conference. p. 1–15 (2010).

[mello 2019] Mello, R., et. al.: MASTER: A Multiple Aspect View on Trajectories. Trans. GIS 23(4). p. 805–822 (2019).

[mello 2021] Mello, R., et. al.: Dependency Rule Modeling for Multiple Aspects Trajectories. In: ER Conference. p. 123-132 (2021).

Next Events

SECOND MASTER WORKSHOP

MAY 13 2022

VENICE, ITALY

The objective of this workshop is to strengthen the networking with the non academic sector and specifically to industries on mobility data.
We will have invited speakers from AKKA (France) and Marine Traffic (Greece), plus other mobility related companies located in Venice. The project will present some highlights of the recent research results that could be of interest for companies and start new collaborations. The objective is to collect needs and open problems from industries and discuss the possible links.

Check the MASTER website for the latest news!

http://www.master-project-h2020.eu/second-master-workshop/

Aula Trentin, University Ca'Foscari, Venice, Italy. The room where the workshop will be held.

FIRST INTERNATIONAL SUMMER SCHOOL ON DATA SCIENCE FOR MOBILITY

AUTUMN 2022

SANTORINI, GREECE

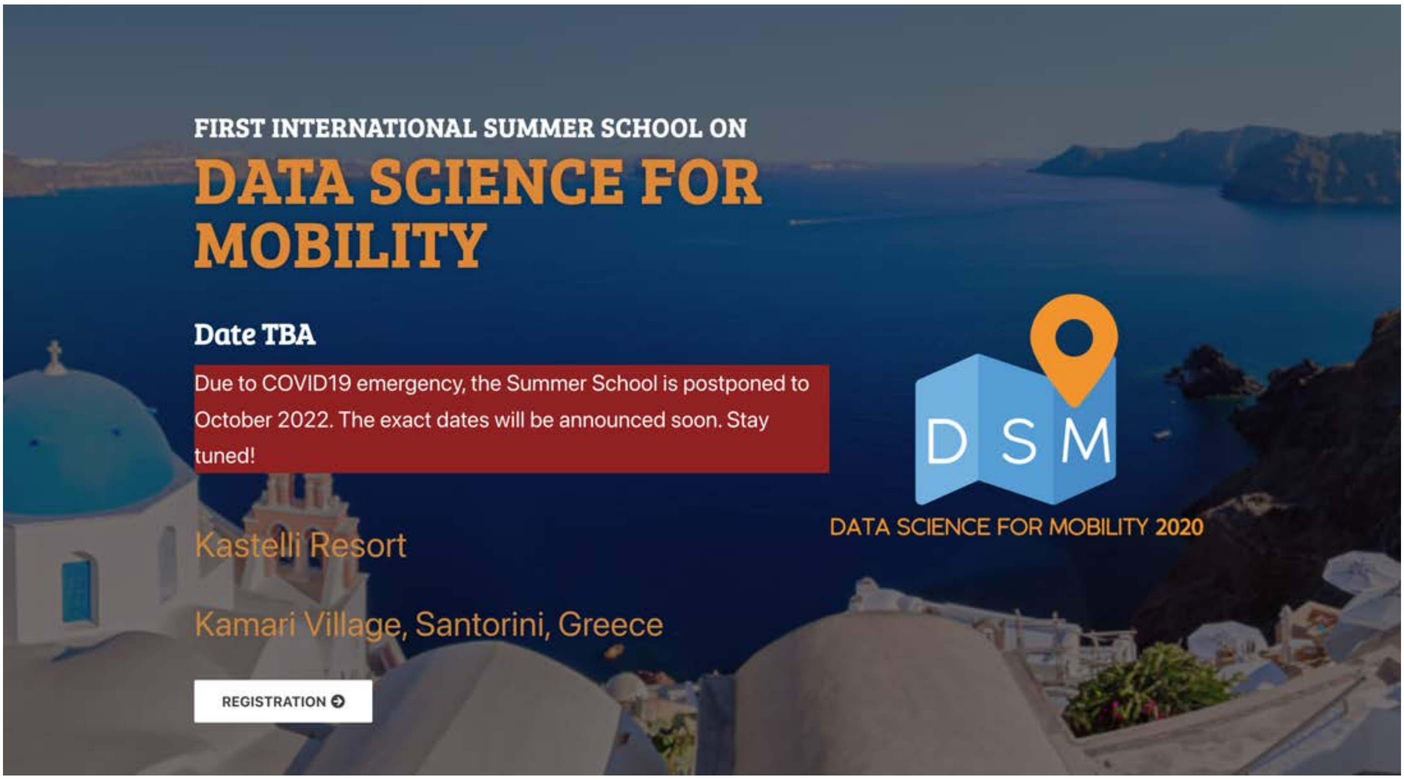
Massive amounts of spatio-temporal data representing trajectories of moving objects are produced by an ever-increasing number of diverse, real-life applications, ranging from mobile to social media apps and surveillance systems, from vehicle tracking systems to IoT mobile sensors. Such mobility-aware traces come in huge numbers and very complex forms, and can be enriched with multiple different semantic dimensions. These semantically enriched trajectories have the potential to unveil novel challenges in several domains, such as urban, maritime and aviation.

The explosion in Data Science is happening now. The Big Data technological infrastructure has reached maturity. Significant interest from the research community is being shown towards the Big Data Value Analytics reference model: data management, data processing, data analytics, data visualization. The time is right for the field of Mobility Data Science to follow the trend!

Our **First International Summer School on Data Science for Mobility** offers participants both visionary keynote speeches and hands-on mini courses held by leading experts in AI and Data Analytics for Mobility from Canada, Greece & Italy. The keynotes speeches will explore the challenges faced due to the voluminous and complex mobility data generated every day in maritime and aviation domains. The hands-on mini courses complement the keynotes by giving practical experience in the usage of analysis tools on real mobility datasets.

The Summer School was supposed to take place in April 27-May 1, 2020, but due to the coronavirus outbreak the school has been postponed to October 2022.

For updates, please visit:
http://master-school.isti.cnr.it/



This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant No 777695
Duration: March 1 2018 | December 31 2023

Coordinator of the project: Chiara Renso | chiara.renso@isti.cnr.it

MASTER Newsletter is not for sale but it is distributed for purposes of study and research and published online at
http://www.master-project-h2020.eu

Consiglio Nazionale delle Ricerche

1868 2018 Università Ca'Foscari Venezia

Dalhousie University Inspiring Minds

Université de Versailles St-Quentin-en-Yvelines

université Paris-Saclay

Municipality of Thira

University of Piraeus Research Center

ΧΑΡΟΚΟΠΕΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ
Harokopio University

Universidade Federal do Ceará

Universidade Federal de Santa Catarina

PUC Rio