

Discover spatio-temporal cluster from trajectory data enhance by heterogeneous contextual knowledge using FCA and the NextPriorityConcept Algorithm

Jérémy Richard^{1,*}, Guillaume Savarit¹, Salah Eddine Boukhetta¹, Cyril Faucher¹, Karell Bertet¹ and Christophe Demko¹

¹Laboratory L3i, La Rochelle University, La Rochelle, France

Abstract

The rising number of different kinds of data that can be used to describe a human trajectory (Such as GPS Coordinate, GSM, RFID, RSSI...) put in the spotlight the semantically rich trajectory. A semantic trajectory annotates semantic knowledge directly into raw data based on features of the studied area such as point of interest or *weather* conditions. One of the challenges of mobility studies nowadays is to find the right data model to shape all those data coming from different source into a framework flexible enough to multiply the contextual data that can be used; where contextual data are knowledge coming from external data source (public city dataset, web pages, national *weather* services etc...). Such data models are the key component of mobility studies, but oftentimes lose the computational aspect of trajectories. In this paper, we will use the semantically rich trajectory as a way to analyse behavioral data enriched by contextual knowledge as this issue has rarely been addressed in the state of the art. We will study the use of formal concept analysis and pattern mining as a way to compute complex sequential patterns in a dataset of semantic trajectories by using the NEXTPRIORITYCONCEPT algorithm. This kind of formal concept analysis allows an interactive analysis between individuals path and contextual data resulting in a hierarchy of spatio-temporal clusters where each cluster contains a specific pattern depicting the trajectories within.

1. Introduction

A trajectory represents the path of a moving object in space. Most of raw trajectories are formalised as a sequence of time related coordinates (x, y) at a time t , noted $\langle (x_i, y_i), t_i \rangle$. The most common type of data used to reconstruct such paths is GPS coordinates. Yet, with the exponential growth of the internet of Things, we collect more and more data from various ways (such as GSM, radio frequency etc...) that are related to a person's path (or animals, vehicles etc ..). Therefore, making mobility studies more and more complexes taking into account a wide range of data that can be used to reconstitute the movement of an object in space. This question the way we can use and model this information. To address this problem, researchers consider

Published in Pablo Cordero, Ondrej Kridlo (Eds.): *The 16th International Conference on Concept Lattices and Their Applications, CLA 2022, Tallinn, Estonia, June 20–22, 2022, Proceedings*, pp. 159–173.

*Corresponding author.

✉ jeremy.richard2@univ-lr.fr (J. Richard); guillaume.savarit1@univ-lr.fr (G. Savarit); salah.boukhetta@univ-lr.fr (S. Boukhetta); cyril.faucher@univ-lr.fr (C. Faucher); karell.bertet@univ-lr.fr (K. Bertet); christophe.demko@l3i.univ-lr.fr (C. Demko)



© 2022 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).



CEUR Workshop Proceedings (CEUR-WS.org)

the use of semantic trajectories [1] [2]. A semantic trajectory adds contextual knowledge to a path of a mobile object; where contextual knowledge can be defined as every information bringing a better understanding of an event, person or object. This information brings to trajectories another point of view avoiding to rely only on movement data. *Weather* condition or events such as protests or all kinds of “background” information can be used to discover trajectory-shaping decisions. This data needs to be added onto those trajectories and are often time shaped as an episode. An episode is a time interval (A, t, \bar{t}) where t, \bar{t} are timestamps and A a semantic annotation. Using a succession of episodes such as the sequence of “*district*” crossed by a trajectory can be interpreted as a “discretization” of this one. Nevertheless, not only the “context-aware” trajectory has been hard to analyze, but also the temporal aspects of a semantic trajectory are largely ignored and only a few studies address this problem directly. In this paper, we propose a way to analyse such enhanced trajectories by using the similarities between semantic trajectories and temporal sequences. Recent advances in the field of pattern mining and Formal Concept Analysis (FCA) show promising ways to analyze heterogeneous data coming from multiple datasets simultaneously; method that we are going to apply to such semantic trajectories. Processing complex semantic trajectories using the **GALACTIC** framework and the **NEXTPRIORITYCONCEPT**[3] algorithm shows that it is possible to extract such mobility patterns. We also can add any type of contextual data during the process enabling an interactive analysis where we can follow the mobility of an object enriched with all kinds of contextual data evolving on the edges of its trajectory. Also, adding contextual knowledge is a mean to avoid the deluge of pattern by focusing on specific behaviors. We propose a method for movement data analysis focusing on the semantic aspect of trajectories alongside interactive analysis with contextual data. As of today, the number of methods for analysing the relationship between a moving object and background information is relatively low.

The paper is organized as follows; in section 2 we will present related work both from the field of trajectory modelling and formal concept analysis. Section 3 will explain the library **GALACTIC**. In section 4 we will explore the computational properties of such a tool with experimentations on heterogeneous data. We are using real life semantic trajectories captured by our team and composed by touristic path in the city of La Rochelle (France) and enriched with contextual knowledge.

2. From trajectories to temporal sequences: Related work

2.1. Trajectory processing :

In this section, we will present an overview of existing state-of-the-art models for semantic trajectories. First, we will detail models and solutions to add more contextual knowledge into trajectories. Next, we will study methods and analysis techniques for movement data. Based on a dataset of semantic trajectories, we will describe the existing solutions used to retrieve similar sub-trajectory common behaviors or movement patterns.

2.1.1. The semantic aspect and the knowledge management.

Semantic trajectory is defined by Parent et al. 2013 [4] as : “ ... a [raw] trajectory that has been enhanced with annotations and/or one or several complementary segmentation.”. The addition of semantic knowledge directly into a raw trajectory enables researchers to better interpret such data. A simple GPS coordinate can become “street A”, and a series of radio frequency from multiple sensors in a system of indoor localization can become “room A”. The definition and the formalisation of the semantic trajectory will differ from a data model to another but the main idea is to provide a framework generic enough to be used in many contexts. To do so, such models are usually structured as a sequence of time related episodes where an episode is a group of positions from a raw trajectory over a time period according to a preset defined by the data-scientist [5]. This preset can be a POI (Place Of Interest), a behavior or any other attribute that the researcher chooses to focus on in his model which will be shaped as an alphabet. The first model for semantic trajectories is the “stop and move” model [6]. By semantically annotating when an individual stops and moves again, S. Spaccapietra et al. (along L. O Alvares also based on stop and move [7]) built their semantic trajectories as a succession of *stop* and *move*. A *stop* is a non-empty time interval $[t, \bar{t}]$ where the travelling object does not move and a *move* is a non-empty time interval $[t, \bar{t}]$ where the travelling object is in motion. In this model, *stop* and *move* are episodes. This work made the first step in the field of semantic trajectory inspiring several works such as [8] where a framework for semantic annotation and enhancement is proposed. Shortly after, semantic trajectories began to use ontology-based models such as [9] still based on the first *stop and move* model. Ontology formalism is then used in those models as a mean to convert the raw trajectory into a high-level semantic representation. The “stop” and “move” model can be seen as a discretization of the speed while a “*district*” is a discretization of space. Those ontology-based data models for semantic trajectories are only “queryable” and cannot be used in pattern mining processes. As a matter of fact, the analysis of these data models is a major issue in the field of semantic trajectories.

2.1.2. Trajectory analysis.

Mobility patterns tend to vectorize raw GPS trajectories such as the trajectory clustering patterns where the goal is to determine common trajectories in order to group similar ones. Using a similarity measure ([10] where authors proposed a small overview), several studies proposed clustering algorithms for trajectories such as [11] where authors used those metrics to extract common sub-trajectories. Nevertheless, these studies used raw GPS trajectories and similarity measures can be difficult to adapt to other types of trajectories such as indoor semantic trajectories. In a notable uncommon trajectory clustering study very close to our work [12], J. Nyoman et al. experiment the use of sequential pattern mining for analysing visitor path in a museum. By shaping trajectories as sequences, the article combines sequence mining with algorithms such as MFCS (for “Mining Frequent Contiguous Subsequences”) and MRGS (for “Mining Rare General Subsequences”) with movement data analysis; authors were able to identify four visiting behaviors. The use of pattern mining techniques to process trajectories shows promising results for analysing and clustering movement data with semantic trajectories [13]. These works are closely related to what has been done in sequential pattern

analysis. This is an emerging issue in the field of movement data analysis, but still we can cite [14] where authors proposed Splitter, a framework capable of efficiently mines sequential pattern in semantic trajectories with as a classification approach. In [15], C. Huiping et al. proposed to use an *Apriori*-like algorithm to solve this problem in a dataset of trajectories. Several studies have addressed this problem in the larger field of sequence mining and they will be discussed next section.

2.2. Introduction to FCA, pattern mining and sequence mining:

2.2.1. The basis of pattern mining.

A sequence is defined as a succession of elements from an alphabet Σ , often in the form of $s = \langle A_i \rangle_{i \leq n}$, where $A_i \in \Sigma$. Each element of the dictionary can be text, action, item, protein and so on. The ordering of elements in a sequence is very important and is a key element in this data modelling. This form of representation gained popularity in the 90s, in particular with the work of Agrawal Srikant with the Mining sequential pattern [16]. Given a dataset of customers, they build sequences as a succession of items purchased by an individual and managed to get several lists of items that are frequently bought together through a pattern mining algorithm and more specifically the *Apriori* algorithm. By using pattern mining techniques, we allow a semantic kind of representation to be computed and the goal is then to find common patterns in our dataset. Sequence mining is a subfield of data mining which aims at finding patterns in a dataset of sequences that appear more frequently. Patterns can be subsequences, prefixes, suffixes, subsequences according to a sliding window [17]. The first generation of algorithms emerged in the 90s with the article of Agrawal and Srikant [16] which extends the well-known *Apriori* algorithm to sequence mining. All these algorithms take as input a dataset of sequences and a minimum support threshold, and generate the set of frequent subsequences. For big datasets and a short minimum support, these algorithms take a huge computing time and generate a too large number of patterns that are difficult to interpret and contain redundancy. A second generation of algorithms focuses on maximal patterns also know as closed patterns because they verify a well-known property of closure in order to limit the number of patterns extracted - called the deluge of pattern. Many algorithms directly address this problem (CloSpan [18], ClaSP [19]).

2.2.2. The basis of formal concept analysis.

Some algorithms also appear within Formal Concept Analysis (FCA) frameworks [20] and their extensions to pattern structures[21], where the lattice property of closed patterns is promoted. We can mention an article for mining medical care trajectories using pattern structures [22]. Pattern concepts are built as maximal sets of individuals with their maximal common subsequences, the whole set of concepts is equipped with a specialization/generalization relation from a partially ordered set with the lattice property. This lattice represents the initial data where concepts are clusters of “similar” sequences, and the concept lattice is a hierarchy of clusters (regrouping objects with their associated common patterns). The use of the Formal

Concept Analysis while working with semantic sequences is a promising way to find common patterns of behavior [23].

2.2.3. Temporal sequence mining.

To the best of our knowledge, Yoshida et al. [24] were the firsts to introduce the notion of temporal pattern mining, called “Delta patterns” where a pattern $(a, [0, 3], b)$ is a sequential pattern (a, b) that frequently occurs in the dataset and has a transition time from a to b of $[0, 3]$, a time interval. This transition time between two elements has known a notable extension in work such as [25]. In [26], researchers focus on mining chronicles, where a chronicle is a couple $(\mathcal{E}, \mathcal{F})$ with \mathcal{E} is a set of temporal events and \mathcal{F} a set of temporal constraints on the set \mathcal{E} . A temporal constraint noted $e_1[t^-, t^+]e_2$ represent the time gap between two events e_1 and e_2 .

Thus, a parallel can be made between semantic trajectories and temporal sequences. By extracting the semantic annotations and the time information of all episodes of the semantic trajectories, we end up with a sequence of time intervals denoted by $S = \langle A_i, \underline{t}_i, \bar{t}_i \rangle$. As defined in “Extracting temporal patterns from Interval-Based Sequences” [27], T. Guyet and R. Quiniou defined a way of representing temporal sequences “A temporal sequence S is an ordered set of events, where an event $\mathcal{A} = (A, [l, u])$ is composed of a symbol A and a nonempty interval $[l, u]$, where l and u are dates.” Based on this definition, semantic annotations of episodes can then be considered as temporal events, making the whole semantic trajectory a temporal sequence. The goal of temporal pattern mining is then to find sequential patterns from a dataset of temporal sequences, such as $(\{a, b\}, [1, 5])$, where the event a and b are sharing a common time interval. Several temporal pattern mining algorithms showed good results in extracting sequential patterns, such as NeGPSan [28] to extract negative sequential patterns. Nevertheless, these approaches cause a deluge of patterns and remain dedicated to a sequence dataset on a same alphabet.

3. Description of the NEXTPRIORITYCONCEPT algorithm

The NEXTPRIORITYCONCEPT algorithm [3] computes concepts from complex and heterogeneous data for a set of objects G . We first introduce the notion *description* δ , which is an application to provide the smallest set of predicates describing a set of objects $A \subseteq G$, based on their characteristics. A characteristic describes our objects, such as numerical, discrete, semantic, temporal etc ... A concept $(A, \delta(A))$ is composed of a subset of A and a set of predicates $\delta(A)$ describing the objects it contains. An example of predicate can be “less than c ” or “match subsequence s ” where c is the max of the numerical values and s is a maximal common subsequence of the sequence. Depending on the data type, generation algorithms will differ. Each predicate is specific to one type of characteristics. The final *description* δ is the union of all those predicates.

$$\delta(A) = \bigcup_{i \in 1..n} \delta_i(A) \quad (1)$$

Where $(\delta_i(A))_{i \in 1..n}$ is a family of descriptions. This is the exact principle that allows us to perform heterogeneous analyses. The generation of a lattice with NEXTPRIORITYCONCEPT is inspired by Bordat’s algorithm [29]. It recursively computes the immediate successors of a

concept, starting at the bottom concept. The NEXTPRIORITYCONCEPT focuses on objects G and starts the computation at the top concept, $(G, \delta(G))$ containing the whole set of objects G and their common description by predicates $\delta(G)$ until no more concepts can be generated. These concepts are computed on demand and do not need any kind of preprocessing.

In order to generate the immediate predecessors of a concept $(A, \delta(A))$, the algorithm introduces *strategies* σ that select predecessors of such a concept based on each characteristic. The strategy refines the description $\delta(A)$ to a reduced set $A' \subset A$. We call cardinality the number of individuals within A' . σ is an application such that $\sigma : 2^G \rightarrow 2^P$ where P is the set of all possible predicates. Several strategies are available to generate predecessors of a concept such as the naive strategy (considering all possible predecessors of a concept), or strategies reducing the number of predecessors.

The use of predicates regardless of the data allows a generic implementation of algorithms which mixed with a system of plugins, enables for an easy integration of new data types (description and strategies), hence, taking into account a wide range of data types. **GALACTIC**¹ (**G**alois **L**attices, **C**oncept **T**heory, **I**mplicational systems and **C**losures) is a development platform for our algorithm allowing easy integration of new plugins for characteristics, descriptions and strategies.

NEXTPRIORITYCONCEPT maintains the lattice structure using queue for a generation level by level, and a mechanism of propagation of constraints to ensure the meet and join will be generated.

With **GALACTIC** and the NEXTPRIORITYCONCEPT algorithm, we contribute to the field of temporal sequences by creating two descriptions δ and a set of strategies σ where we consider temporal sequences [30] [31] as:

- A temporal sequence such that $S = (\langle a, t_i \rangle)_{i \in 1..n}$ with $a \in \Sigma$ is an item from a dictionary Σ and t is a timestamp. Then, we compute the distance (either transition time or duration) between two items of S .
- A temporal sequence such that $S = (\langle a, [t_s, t_e] \rangle)_{i \in 1..n}$ where $a \in \Sigma$ is an item from a dictionary Σ and $[t_s, t_e]$ the interval of time during the element a occurs. We will then compute common sub-sequences of items during a maximal similar interval of time.

4. Experimentations

4.1. The Geoluciole dataset

The Geoluciole dataset is composed of GPS trajectories in the city of La Rochelle France. This dataset was collected during the summer of 2019, from tourists visiting the city during their vacation. It contains 192 trajectories of different individuals. Their positions were obtained by an application we developed, giving their position every 30 seconds during their stay. Before activating their application, we also had them fill out a survey, giving contextual

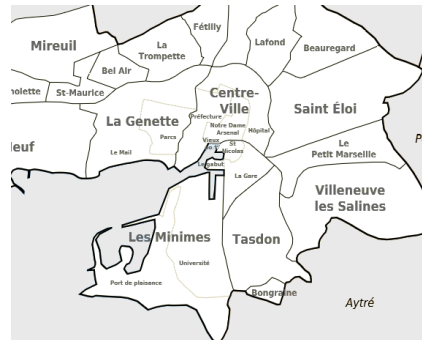


Figure 1: Segmentation of the city of La Rochelle into *district*

¹<https://galactic.univ-lr.fr>

knowledge, such as the *staying status* or the *arrival means*.

By matching the GPS coordinates to *districts* of the city such as shown in figure 1, we can then transform these raw data into semantic trajectories. The segmentation of the trajectory into a sequence of districts has been done by using a GIS (Geographic Information Systems). We can also add all kinds of knowledge we have webscrapped, such as the *beach* they were and parks which is spatial information and *weather* and *tide* information which is more of a temporal information (figure 2). This data representation is very similar to the one we can find in [32] a “Multi-Level and Multiple Aspect Semantic Trajectory Model”.

4.2. The semantic trajectory model

In Table 2, each “aspect” (or side) of those semantic trajectories keeps the specificity of the data within. The *district*, *weather*, *tide*, *green space* (parks) and *beach* are sequences of temporal intervals. In each “temporal” aspect of a sequence, the temporal information represents the hour of the day. While the *district*, *green space* and *beach* depends on the localisation of the user between t_k and \bar{t}_k , the *weather* and *tide* depends on the temporal information and are issued from webscraping.

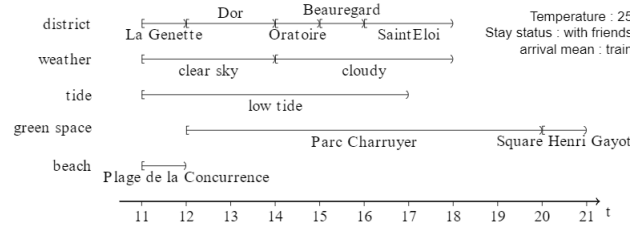


Figure 2: Semantic trajectory of a tourist in La Rochelle

We formalise the semantic trajectory model for a trajectory T as follows : $T : \langle \alpha_1, \alpha_2 \dots \alpha_m \rangle$ where T is described by m aspects and each aspects α_i , with $i \leq m$ can either be a temporal sequence $\alpha_i = (\langle a_k, t_k, \bar{t}_k \rangle)_{k \leq n}$, on an alphabet Σ_i and $a_k \in \Sigma_i$, a semantic value $\alpha_i = X$ on an alphabet Σ_i and $X \in \Sigma_i$, or a numeric value $\alpha_i = x$. Our dataset of trajectories \mathcal{D} is given by $T_j : \langle \alpha_{1_j}, \alpha_{2_j} \dots \alpha_{m_j} \rangle$. The description space used with **GALACTIC** for any trajectory $T_j \in \mathcal{D}$ is a set of predicate defined by :

$$\delta(T_j) : \langle \delta_1(\alpha_{1_j}), \delta_2(\alpha_{2_j}) \dots \delta_m(\alpha_{m_j}) \rangle \quad (2)$$

4.3. Experiment description

In this section, we use **GALACTIC** to extract concepts from the Geoluciole dataset. First, we will study the impact of a heterogeneous analysis on the number of concepts (ie. clusters) generated. Then, we will focus on specific parts of the dataset with selected semantic information in order to better define touristic behaviors. Table 1 shows the descriptions and strategies used during this experimentation. Table 2 show a description of each dataset.

	δ	σ
Temporal Sequence	maximal common interval	minimal cardinality
Chain	chain matching	complete match
Numeric	simple numerical	quantile

Table 1
Descriptions and strategies associated with datatype

NEXTPRIORITYCONCEPT uses descriptions δ to define a set of predicates describing the attributes. As predicates, the descriptions can then be seen as binary table. We also makes use of strategies to choose next predecessors at each iteration.

In this paper, we will use the following descriptions and strategies :

- **Numeric** : We describe a set $A = \{\alpha_1, ..\alpha_k\}$ of numerical values by the “simple numerical” description δ_S :

$$\delta_S(A) = \{\text{is greater than } \min(A), \text{ is lesser than } \max(A)\}$$

The strategy is the quantile description $\sigma_Q(A, k)$ where k is the number of quantiles:

$$\sigma(A) = \{\text{is greater than } q_j, \text{ is lesser than } q_j : q_j \text{ is a k-quantile } \} \quad (3)$$

- **Chain** : For a set $A = \{< s_i >\}$ of sequences, we use the “chain matching” description $\delta_{CM}(A, k)$ that compute the maximal number of subsequences of size k, with \sqsubseteq_s is the subsequence relation:

$$\delta_{CM}(A, k) = \{x \in \Sigma^* : \forall a \in A, x \sqsubseteq_s a \text{ and } |x| = k\} \quad (4)$$

We use the “complete match” strategy σ_{CM} that computes all the possibles subsets $A' \subset A$:

$$\sigma_{CM}(A) = \{x : x \in \delta_{CM}(A'), A' = A \setminus \{a\}, \text{ for all } a \in A\} \quad (5)$$

- **Temporal sequence** : We consider a set of temporal sequences $A = \{< X_i, T_i >\}$ where $T_i = [t_i, \bar{t}_i]$ is an interval and X_i an itemset. For a sequence $a \in A$, the projection $\phi_T(a)$ selects all the itemsets of a included in the interval. We use the “maximal common interval” description that computes the set of maximal common sub-intervals:

$$\delta_{MCTF}(A) = \{\langle (T, X) \rangle : \forall a \in A, X \subseteq \Phi_T(a)\} \quad (6)$$

The “minimal cardinality” strategy $\sigma_{AMC}(A)$ adds element of minimal cardinality to the subsequences of the description:

$$\sigma_{AMC}(A) = \{\langle (T, X) \rangle : \forall a \in A, \phi_T(a) \subseteq X \text{ and } \forall x \in X \quad (7)$$

$$nb(A, T, x) = |A| \text{ and } nb(A, T, X) = nb_{\min}(A, T)\} \quad (8)$$

Aspects	Type	Dictionary	Source
<i>district</i>	Temporal sequence	$\Sigma = \text{set of city } district$	Trajectory
<i>weather</i>	Temporal sequence	$\Sigma = \{\text{"raining"}, \text{"moderate raining"}, \text{"cloudy"}, \text{"sunny"}\}$	Web scrapping
<i>beach</i>	Temporal sequence	$\Sigma = \{\text{"no beach"}, \text{"Plage de la Concurrence"}, \text{"Plage des Minimes"}\}$	Trajectory
<i>green space</i>	Temporal sequence	$\Sigma = \text{set of city parks}$	Trajectory
<i>tide</i>	Temporal sequence	$\Sigma = \{\text{"high tide"}, \text{"low tide"}\}$	Web scrapping
<i>average temperature</i>	Numeric		Web scrapping
<i>stay status</i>	Chain	$\Sigma = \{\text{"with family"}, \text{"with friends"}, \text{"couple"}, \text{"alone"}\}$	questionnaire
<i>arrival means</i>	Chain	$\Sigma = \{\text{"train"}, \text{"car"}, \text{"bus"}, \text{"velo"}\}$	questionnaire

Table 2
Dataset heterogeneous aspects' description and source

4.4. Impact of heterogeneous data on the patterns

Running a heterogeneous analysis with multiple description and strategy types can reduce the number of concepts in the final lattice. While a "naive" strategy selects all predicates within a description δ , such as σ_{CM} , other strategies such as σ_{AMC} will select specific predicates and skip other to refine the analysis.

Example: Table 3 shows a dataset D of five individuals, characterized by a numeric value and a sequence of intervals. In this example, we will be using the simple numerical description and quantile strategy for the numerical aspects and the maximal common interval description alongside the minimal cardinality strategy for the sequence of intervals.

Individual	Numeric (x)	Interval
a	1	(8.3, 11): "P", (13, 15): "M", "H"
b	2	(10, 12): "P", "H", (14, 16): "M"
c	3	(8.3, 12): "P", (14, 16): "M", "C"
d	4	(7, 9): "P", "H", (12, 13): "M"
e	5	(10, 11): "P", (12, 12): "M", "C"

Table 3
Dataset D

Let L be the lattice shown in Figure 3 as the result of a numeric description and quantile strategy, taking into account only the numeric characteristics. By adding an interval description and strategy, it results in the lattice L' with a fewer number of concepts. Figure 4 shows that concept \$1 ($x \leq 4$) and \$2 ($x \geq 2$) from L do not appear in the lattice L' as they have not been selected by the minimal cardinality strategy. The concepts contained in L' are the most representative of the two description spaces.

4.4.1. Naive analysis using time interval.

first, we will only consider *districts* since they are a geographical segmentation of trajectories into *districts* of the city. By doing so, we obtain 416 concepts with a low cardinality value. The huge number of concepts generated involving only one or two trajectories is an indicator that there are no significant places where multiple individuals were at a same time of the day. To sharpen our research, we then choosed to add other aspects of semantic trajectories in order to discover new significant behaviors depending on external factors. The next section will present

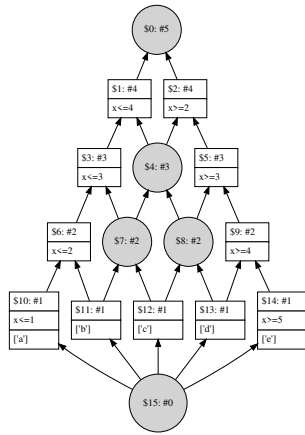


Figure 3: Lattice L with quantile strategy

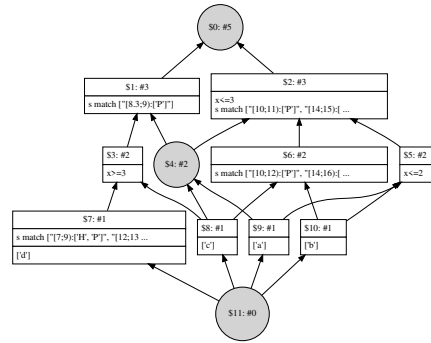


Figure 4: Lattice L' by adding interval description and strategy

experimentations by taking into account multiple aspects of semantic trajectories, such as the *weather* or the *staying status*. With only the *weather* and the same approach, we obtain 379 concepts.

4.4.2. Analysis with heterogeneous data

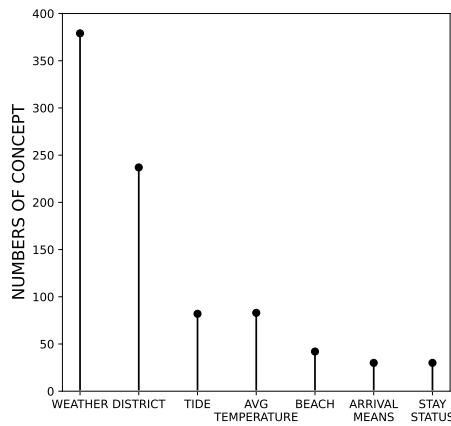


Figure 5: Number of concept by attributes successively added

In this experiment, we exploit the ability of **GALACTIC** to deal with heterogeneous data by adding successively the other attributes to the “*weather*” attributes (the ones that generate the largest number of concepts). Hence it is possible to mix several sequences (such as shown in Equation 1 and 2) together with other data in a heterogeneous analysis. With nearly all attributes added successively, we diminish the number of concepts to only 36. Figure 5 gives

the number of concepts obtained for each added attributes that decreases from 379 (only the *weather*) to 36 (with all our attributes).

The table 6 in appendix gives the descriptions by common predicates of three concepts (concept number 0 is the top concept for all attributes). We can observe that the addition of more knowledge gives a more precise description of concepts, and a reduction of the number of patterns.

4.5. Focus on some particular behaviours when reducing pattern

One way for the data analyst to analyse the dataset is to focus on particular behaviors, where a behavior can be a subset of objects (people going to the *beach*) or a part of the description (an action during the night).

4.5.1. The detection of *beach* trip related to the *weather*.

In order to focus on visitors going to the *beach*, we choose to analyse *beach* with the *weather* that is semantically meaningful for this experiment. The dataset is composed of 108 trajectories, and generates 93 concepts as a result of the formal concept analysis. The Table 7 (found in *Appendix*) shows some concepts containing valuable information, such as a strong correlation between “raining” and “no beach”, meaning that people won’t go to the *beach* if it’s raining .

In the dataset, 31 trajectories contain rain.

On that part of the trajectory we can see that 8 of them can be described with a predicate which matches “raining” between 11 and 12 am and “no beach”. Other predicates are also generated which also describe that bad *weather* has an impact on a *beach* trip and this predicate seems to focus on bad *weather* in the morning.

4.5.2. The detection of sleeping habits during touristic stay.

With the time information, it is also possible to analyse any part of the day. One possible user driven analysis could be to see if sleeping habits are dependent on the *staying status* of individuals. To do so, we will make an analysis with two attributes, the *staying status*, which is a chain and locations which are temporal. With 138 trajectories, we have in total 340 concepts, and 108 for the sleeping period of time (between 22 and 10).

Some of the results can be seen in Table 8 (see *Appendix*). As we can see, predicates are generated and are supported by nearly 10% of “family” trajectories and more than 10% of “with friends” trajectories to locations subsequences, which by comparing it with other predicates of supersequence seem to represent sleeping places (no movement during a long period of time between 0 and 10 am).

5. Conclusion and future works

In this paper, we propose a way to process heterogeneous data using the NEXTPRIORITYCONCEPT algorithm alongside the GALACTIC platform. In the experiments, we illustrate the advantage of mixing heterogeneous data while keeping their semantic and readability with two examples.

First, by reducing the number of patterns by adding contextual knowledge onto trajectories and afterwards by focusing on specific aspects of the data to detect and identify distinctive behaviors. The specificity of this approach is to keep the raw data within and avoid any kind of vectorization technique. Also, it allows us to enrich the dataset with semantic information from the space and/or temporal knowledge directly from the trajectory (such as *districts* in this paper), surveys filled out by individuals (*staying status, arrival means*) or even web scrapping (*weather, tide*). The plugin system of **GALACTIC** allows to easily integrate descriptions and strategies for heterogeneous data and offering a wide range of data types to work with. The **NEXTPRIORITYCONCEPT** and its ability to run interactive, heterogeneous analysis is a big step in data science and offers new ways to deal with complex data structures such as semantic trajectories.

For future works, we would like to measure the quality of generated concepts to better represent and sort the relevance of their predicates (such as the *stability* measure). These results should become available very soon. Also, we are currently working on an incremental tool in order to reinforce the interactivity with a user-driven tool where the data-analyst can select data to analyse it in an interactive way. By doing so, we allow the data scientist to select which side of the data he wants to explore, because only he knows best the semantic of the data he manipulates. In particular, it gives the possibility to change the strategy or description during every step of the process (where new concepts are generated), in an intuitive and simple way. For example *weather* and *tide* information are not interesting during the night. We wish that by giving the opportunity to the data-scientist to interactively explore the data in a user-driven approach, we will be able to optimize data mining processes.

References

- [1] A. Kontarinis, K. Zeitouni, C. Marinica, D. Vodislav, D. Kotzinos, Towards a Semantic Indoor Trajectory Model, in: 2nd International Workshop on "Big Mobility Data Analytics" (BMDA) with EDBT 2019, Lisbon, Portugal, 2019. URL: <https://hal.archives-ouvertes.fr/hal-02314572>.
- [2] S. Ilarri, D. Stojanovic, C. Ray, Semantic management of moving objects: A vision towards smart mobility, *Expert Systems with Applications* 42 (2015) 1418 – 1435. doi:10.1016/j.eswa.2014.08.057.
- [3] C. Demko, K. Bertet, C. Faucher, J. Viaud, S. O. Kuznetsov, Nextpriorityconcept: A new and generic algorithm computing concepts from complex and heterogeneous data, *Theoretical Computer Science* 845 (2020) 1–20.
- [4] C. Parent, S. Spaccapietra, Semantic trajectories modeling and analysis, *ACM Comput. Surv.* 45 (2013) 1–32. doi:10.1145/2501654.2501656.
- [5] D. Mountain, J. Raper, Modelling human spatio-temporal behaviour: A challenge for location-based services, 2001.
- [6] S. Spaccapietra, C. Parent, M. L. Damiani, J. A. de Macedo, F. Porto, C. Vangenot, A conceptual view on trajectories, *Data Knowledge Engineering* 65 (2008) 126–146.
- [7] L. O. Alvares, V. Bogorny, B. Kuijpers, J. A. F. de Macedo, B. Moelans, A. Vaisman, A model for enriching trajectories with semantic geographical information, *Proc. of the 15th*

- Annual ACM International Symposium on Advances in Geographic Information Systems. ser. GIS '07. ACM, 2007 (2007) 22:1–22:8.
- [8] V. Bogorny, C. Renso, A. R. de Aquino, F. de Lucca Siqueira, L. O. Alvares, Constant – a conceptual data model for semantic trajectories of moving objects, *Transactions in GIS* 18 (2014) 66–88. doi:10.1111/tgis.12011.
 - [9] Z. Yan, J. Macedo, C. Parent, S. Spaccapietra, Trajectory ontologies and queries, *Transactions in GIS* 12 (2008) 75–91. doi:10.1111/j.1467-9671.2008.01137.x.
 - [10] K. Toohey, M. Duckham, Trajectory similarity measures, *Sigspatial Special* 7 (2015) 43–50.
 - [11] S. J. Camargo, A. W. Robertson, S. J. Gaffney, P. Smyth, M. Ghil, Cluster analysis of typhoon tracks. part i: General properties, *Journal of Climate* 20 (2007) 3635–3653.
 - [12] N. Juniarta, M. Couceiro, A. Napoli, C. Raïssi, Sequential pattern mining using fca and pattern structures for analyzing visitor trajectories in a museum, in: *CLA 2018-The 14th International Conference on Concept Lattices and Their Applications*, 2018.
 - [13] Y. Chen, P. Yuan, M. Qiu, D. Pi, An indoor trajectory frequent pattern mining algorithm based on vague grid sequence, *Expert Systems with Applications* 118 (2019) 614–624.
 - [14] C. Zhang, J. Han, L. Shou, J. Lu, T. La Porta, Splitter: Mining fine-grained sequential patterns in semantic trajectories, *Proceedings of the VLDB Endowment* 7 (2014) 769–780.
 - [15] H. Cao, N. Mamoulis, D. W. Cheung, Mining frequent spatio-temporal sequential patterns, in: *Fifth IEEE International Conference on Data Mining (ICDM'05)*, IEEE, 2005, pp. 8–pp.
 - [16] R. Agrawal, R. Srikant, Mining sequential patterns, in: *Proceedings of the Eleventh International Conference on Data Engineering*, 1995, pp. 3–14.
 - [17] H. Mannila, H. Toivonen, A. Inkeri V., Discovery of frequent episodes in event sequences, *Data mining and knowledge discovery* 1 (1997) 259–289.
 - [18] X. Yan, J. Han, R. Afshar, Clospan: Mining: Closed sequential patterns in large datasets, in: *Proceedings of the 2003 SIAM international conference on data mining*, SIAM, 2003, pp. 166–177.
 - [19] A. Gomariz, M. Campos, R. Marin, B. Goethals, Clasp: An efficient algorithm for mining frequent closed sequences, in: *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, Springer, 2013, pp. 50–61.
 - [20] B. Ganter, Attribute exploration with background knowledge, *Theoretical Computer Science* 217 (1999) 215–234.
 - [21] B. Ganter, S. O. Kuznetsov, Pattern structures and their projections, in: *International conference on conceptual structures*, Springer, 2001, pp. 129–142.
 - [22] A. Buzmakov, E. Egho, N. Jay, S. O. Kuznetsov, A. Napoli, C. Raïssi, On projections of sequential pattern structures (with an application on care trajectories), in: *The Tenth International Conference on Concept Lattices and Their Applications-CLA'13*, 2013.
 - [23] N. Juniarta, M. Couceiro, A. Napoli, C. Raïssi, Sequential Pattern Mining using FCA and Pattern Structures for Analyzing Visitor Trajectories in a Museum, in: *CLA 2018 - The 14th International Conference on Concept Lattices and Their Applications*, Olomouc, Czech Republic, 2018. URL: <https://hal.inria.fr/hal-01887914>.
 - [24] M. Yoshida, T. Iizuka, H. Shiohara, M. Ishiguro, Mining sequential patterns including time intervals, in: *Data Mining and Knowledge Discovery: Theory, Tools, and Technology II*, volume 4057, SPIE, 2000, pp. 213–220.
 - [25] S. Yen, Y. Lee, Mining non-redundant time-gap sequential patterns, *Applied Intelligence*

- 39 (2013) 727–738.
- [26] D. Cram, B. Mathern, A. Mille, A complete chronicle discovery approach: application to activity analysis, *Expert Systems* 29 (2012) 321–346.
- [27] T. Guyet, R. Quiniou, Extracting temporal patterns from interval-based sequences, in: *International Joint Conference on Artificial Intelligence (IJCAI)*, Barcelone, Spain, 2011. URL: <https://hal.inria.fr/inria-00618444>.
- [28] T. Guyet, R. Quiniou, NeGPSpan: efficient extraction of negative sequential patterns with embedding constraints, *Data Mining and Knowledge Discovery* 34 (2020) 563–609.
- [29] J. Bordat, Calcul pratique du treillis de galois d’une correspondance, *Mathématiques et Sciences humaines* 96 (1986) 31–47.
- [30] S. E. Boukhetta, C. Demko, K. Bertet, J. Richard, C. Cayère, Temporal sequence mining using fca and galactic, in: *International Conference on Conceptual Structures*, Springer, 2021, pp. 185–199.
- [31] S. E. Boukhetta, J. Richard, C. Demko, K. Bertet, Interval-based sequence mining using fca and the nextpriorityconcept algorithm., in: *FCA4AI@ ECAI, 2020*, pp. 91–102.
- [32] C. Cayère, C. Sallaberry, C. Faucher, M. Bessagnet, P. Roose, M. Masson, J. Richard, Multi-level and multiple aspect semantic trajectory model: Application to the tourism domain, *ISPRS International Journal of Geo-Information* 10 (2021) 592.

Appendix

Concepts	Patterns (common predicates)	Individuals
0	\emptyset	All
1	$temperature = 20.75$ $stay_status: 'family'$ $arrival_mean: 'personal\ car'$ <i>district</i> $\xrightarrow{\text{Dor}}$ <i>weather</i> $\xrightarrow{\text{cloudy}}$ <i>tide</i> $\xrightarrow{\text{low\ tide}}$ <i>green\ space</i> $\xrightarrow{\text{Square\ Valin}}$ $\xrightarrow{\text{Parc\ Charruyer}}$ <i>beach</i> $\xrightarrow{\text{No\ beach}}$ 12 13 14 15 16 17 18 19 20 t	2
2	$25.0417 \geq temperature \geq 20.75$ $stay_status: 'family'$ <i>district</i> $\xrightarrow{\text{Dor}}$ <i>weather</i> $\xrightarrow{\text{cloudy}}$ <i>tide</i> $\xrightarrow{\text{low\ tide}}$ <i>green\ space</i> $\xrightarrow{\text{Parc\ Charruyer}}$ <i>beach</i> $\xrightarrow{\text{No\ beach}}$ 12 13 14 15 16 17 18 t	2

Figure 6: Examples of heterogeneous predicates

Patterns (common predicates)	Support	Support "raining"	Individuals
<i>weather match</i> ["[11;12):['raining']"] <i>beach match</i> ["[0;24):['No beach']"]	0.074	0.260	8
<i>weather match</i> ["[1;2):['raining']"] <i>beach match</i> ["[0;24):['No beach']"]	0.037	0.130	4
<i>weather match</i> ["[8;9):['raining']"] <i>beach match</i> ["[0;24):['No beach']"]	0.028	0.100	3
<i>weather match</i> ["[1;4):['raining']"] <i>beach match</i> ["[0;24):['No beach']"]	0.019	0.065	2

Figure 7: Sample of predicates showing the impact of *weather* on *beach* trip

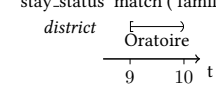
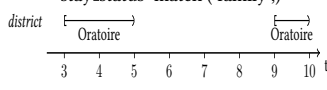
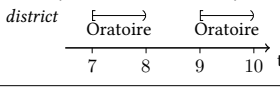
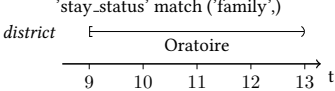
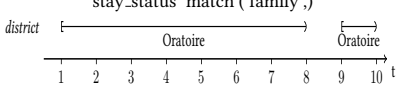
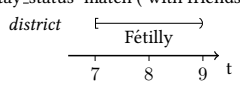
Patterns (common predicates)	Support	Support by status	Individuals
'stay_status' match ('family',) 	0.060	0.098	8
'stay_status' match ('family',) 	0.043	0.073	6
'stay_status' match ('family',) 	0.036	0.061	5
'stay_status' match ('family',) 	0.036	0.061	5
'stay_status' match ('family',) 	0.029	0.049	4
'stay_status' match ('with friends',) 	0.049	0.108	4

Figure 8: Sleeping habits by status