Sixth Framework Programme
Information Society Technologies
Future Emerging Technologies

Project Acronym:
GeoPKDD

Project full title:
Geographic Privacy-aware Knowledge Discovery and Delivery

Project Number: FP6-014915
Instrument: Specific Target Research Project

Project Deliverable D2.6

Consolidated report on privacy-aware spatio-temporal data mining methods

Date of preparation: M6
Revised: M6
Revision: final

Operative commencement date of Contract: 01/11/2005
Project Coordinator: Fosca Giannotti
Project institute coordinator: Knowledge Discovery and Delivery-LAB / ISTI-CNR

approved by the EC on 25/9/2005
Programme Name: ..................... IST
Project Number: ...................... 014915
Project Title: ......................... GEOPKDD

Document Number: .................... D2.6
Work-Package: .......................... WP2
Document Type: ........................ Document
Contractual Date of Delivery: ........ M6
Actual Date of Delivery: ............... M6
Title of Document: ........................ Deliverable 2.6 – Consolidated report on privacy-aware spatio-temporal data mining methods

Author(s): ............................... Dino Pedreschi (KDD LAB), Mirco Nanni (KDD LAB), Yannis Theodoridis (CTI & Univ. Piraeus), Bart Kuipers (Univ. Hasselt), Yücel Saygin (Sabanci Univ.), Vassilis S. Verykios (CTI & Univ. Thessaly)

Dissemination Level............... This document is public (except for some papers in the appendix that are currently submitted for publication and subject to revision)

GeoPKDD consortium participants

<table>
<thead>
<tr>
<th>Partic. N.</th>
<th>Participant name</th>
<th>Short name</th>
<th>Country</th>
</tr>
</thead>
<tbody>
<tr>
<td>CR 7</td>
<td>Sabanci University, Faculty of Engineering and Natural Sciences. <a href="http://www.sabanciuniv.edu/">http://www.sabanciuniv.edu/</a></td>
<td>UNISAB</td>
<td>TK</td>
</tr>
<tr>
<td>CR 8</td>
<td>WIND Telecomunicazioni SpA, Direzione Reit Wind Progetti Finanziati &amp; Technology Scouting. <a href="http://www.wind.it">http://www.wind.it</a></td>
<td>WIND</td>
<td>I</td>
</tr>
<tr>
<td>CR 9</td>
<td>Polytechnic University of Madrid, Latingeo Laboratory <a href="http://www.latingeo.net">http://www.latingeo.net</a></td>
<td>UPM</td>
<td>ES</td>
</tr>
</tbody>
</table>
Contents

INTRODUCTION ................................................................................................................................. 5

PROGRESSES IN SPATIO-TEMPORAL DATA MINING ................................................................. 5

1 SPATIO-TEMPORAL LOCAL PATTERNS DISCOVERY (TASK 2.1) .................................. 5
2 SPATIO-TEMPORAL GLOBAL MODELS DISCOVERY (TASK 2.2) .................................. 6
3 APPLICATIONS AND PREPROCESSING TOOLS ............................................................... 8

PROGRESSES IN PRIVACY-PRESERVING DATA MINING (TASK 2.3) ................................ 9

BIBLIOGRAPHY .............................................................................................................................. 11

APPENDIX: PAPER ABSTRACTS ................................................................................................ 14
Introduction

This report summarizes the results obtained in Work Package 2 in the third year of activity of the GeoPKDD project. The mission of the work package is to develop a software toolkit, i.e., a repertoire of methods and tools, for privacy-preserving trajectory mining, and therefore: (i) to design, implement and test algorithms for privacy-preserving spatio-temporal data mining, specifically for trajectories of moving objects; and (ii) to provide efficient and scalable versions of the algorithms, which enable their practical application to realistic, massive datasets of mobility data.

The main achievements within the third year of the project, are:

- the consolidation of a comprehensive prototype toolkit for privacy-preserving spatio-temporal data mining, which, on the basis of the work carried out in the previous years, has been extended, validated, and integrated into the coherent framework of the Data mining query language Daedalus, in collaboration with WP3;
- an extensive collaboration with WP4, which brought to the application of the privacy-preserving spatio-temporal data mining methods developed in WP2 to the GeoPKDD demonstrators.

We briefly report the scientific progress occurred within Work Package 2. The discussion is divided in two parts: one related to spatio-temporal data mining (Tasks 2.1 and 2.2), and one related to privacy-preserving data publishing and data mining (Task 2.3).

Progresses in spatio-temporal data mining

General goals of the tasks devoted to spatio-temporal data mining are to specify the forms of spatio-temporal local patterns and global models to be mined, and to design and implement efficient algorithms for extracting such patterns, developing the necessary theory and frameworks and investigating the peculiarities of the temporal and spatial dimensions.

1 Spatio-temporal local patterns discovery (Task 2.1)

Task 2.1 has developed in year 1 and 2 the method of T-pattern mining, which extracts a set of frequent spatio-temporal sequential patterns from an input dataset of GPS trajectories. Extracted patterns have the form of sequences of spatial regions with typical transition times for each pair of consecutive regions of a sequence. This tool is based on the notion of trajectory pattern developed during the first year and consolidated in the second, resulting in a solid pattern mining tools made available within the DMQL. During year 3, this method has been applied to many datasets of the project, and was experimented in the demonstrators (see Deliverable D4.6).

Also, novel mining methods, such as the location prediction method [PMPP08] developed in Task 2.2, have been defined on top of T-patterns, coherently with the idea that global models can be built out of a collection of local patterns. In addition, following a similar local-to-global principle, an approach to trajectory classification [NQ08] was proposed, that represents each trajectory in terms of how it interacts with the other trajectories around it, then extracts a set of frequent patterns over such alternative representation, which are later used as discriminant features of each trajectory. Details of
both approaches are provided in the next section, devoted to spatio-temporal global models.

Another line of work related to Tasks 2.1 and 2.2, studied how moving object data analysis can benefit from replacing raw trajectory data by a sequence of stops and moves (a notion studied in the first years of the project, in the context of semantic trajectories) [GKV08b]. A formal model and query language was proposed, to express complex queries involving spatial data stored in a GIS, non-spatial data (stored in a data warehouse) and moving object data. This query language, denoted Lmo also supports different forms of aggregation. Compression of trajectory data produced by moving objects, using the concepts of stops and moves was studied, and it was proved that stops and moves are expressible Lmo and that there exists a fragment of this language (that can be expressed by means of regular expressions) allowing to talk about temporally ordered sequences of stops and moves. This fragment is used to perform data mining over trajectory data.

Finally, a method to perform sequential pattern mining over trajectory data has been developed, revising the classic notion of support in sequential pattern mining, by introducing the concept of temporal support of a sequential expression [GKV09], intuitively defined as the number of sequences satisfying a target pattern, out of the total number of sequences that could have possibly matched such pattern. This is generalized to regular expressions, which encapsulate the definition of a collection of sequential expressions. A theoretical framework for this novel notion of support was presented, as well as an algorithm to compute this.

2 Spatio-temporal global models discovery (Task 2.2)

T-Optics, the trajectory clustering tool, which extracts a set of density-based clusters from an input dataset of GPS-like trajectories, has been re-interpreted as a scalable, incremental method within the visual analytics environment, in order to support a user-driven analysis of massive trajectory dataset [AARNPG09]. The method divides the elaboration of massive trajectory datasets into two steps: first, a density-based trajectory clustering step is applied to a subset of the original dataset that can fit in main memory; then, each cluster is partitioned into a set of center-based sub-clusters; then, the user can visually interact to modify, create and eliminate clusters and sub-clusters; finally, the set of sub-clusters is used as a simple classification model to efficiently label all trajectories left out of the sample. This result has been achieved with a tight collaboration with WP3.

The impact of uncertainty on trajectory data mining and, in particular, on clustering has been studied in to two different directions.

In the first approach [Pel09-b], a new trajectory clustering algorithm has been devised, which finds the centroids of each clusters based on a global uncertainty-supporting similarity function. In a first step, trajectories are modelled as sequences of visited regions accompanied with intuitionistic fuzzy values, i.e. elements of an intuitionistic fuzzy set, which are the mean to quantify uncertainty. Then, a novel distance metric was especially designed to operate on such intuitionistic fuzzy vectors, aimed to incorporate it in a variant of the FCM algorithm that will effectively cluster trajectories under uncertainty. At the second step of the approach, CenTra, a novel density- as well as similarity-based algorithm was introduced to tackle the problem of discovering the centroid of a group of trajectories. Finally, CenTra is used in the centroid update procedure of a new trajectory clustering algorithm, called CenTR-I-FCM, which uses a global uncertainty-supporting similarity function, to group trajectories at a higher level, and iteratively refine the results using local similarity between sub-trajectories. This algorithm has the efficiency advantages of partitioning clustering algorithms in comparison to the
higher processing cost of density-based algorithms, whereas produces non-spherical clusters due to the inclusion of CenTra, that recognises representative movements of any shape.

An alternative approach to the same problem was also provided in [KMOV09], by studying how uncertainty affects the results of clustering algorithms like, for instance, k-means or k-medoids. For this, the concept of space-time prisms was applied, replacing a (reconstructed) trajectory by a necklace (intuitively, a chain of space-time prisms), connecting consecutive trajectory sample points, and defining a distance function that accounts for uncertainty. A formal proof was produced, showing that this function is also a metric, and therefore it can be used in clustering. The function basically computes the distance between the chains of prisms corresponding to two trajectory samples based on the proportion of the size of the intersection in the size of the union of these prisms. If the prisms do not intersect, the distance is 1, and if they perfectly overlap, the distance is zero. The approach was validated by clustering a set of trajectories of cars in the center of the city of Milan using this distance function.

Also the problem of trajectory similarity – a key notion in several important task, such as clustering and spatio-temporal queries – was tackled from two different viewpoints. In the first one [Pel09-a], the SIG on Similarity notions for trajectories (a joint effort of WP1, WP2 and WP3) consolidated and extended the repertoire of trajectory similarities (both quantitative and qualitative), and proposed a set of query operators (on top of the Hermes query engine described in WP1 report) for similarity search in TD. Two main types of similarities are defined, namely, spatiotemporal and (temporally-relaxed) spatial similarity, as well as three variations, speed-pattern based, acceleration-pattern based and directional similarity. For each type of similarity query some distance operators and respective query processing algorithms were provided. The complexity of the algorithms proposed is proved to be lower compared to existing approaches. A comprehensive set of experiments was conducted over synthetic and real trajectory datasets, in order to thoroughly evaluate the effectiveness of the approach supporting classification and clustering tasks. The efficiency of the proposed similarity-based query framework has been demonstrated by means of the visual analytics environment, through an iterative knowledge discovery process over a real dataset from the fleet management domain.

From a different viewpoint, the problem of trajectory similarity has been studied also using the concept of the double-cross matrix [KM08,KM09]. This is a well-known formalism used to qualitatively describe polylines in the plane. In a double-cross matrix the relative position of any two-line segments in a polyline is described with respect to a double cross based on their starting points, and two polylines are called DC-similar if their double-cross matrices are identical. A geometric interpretation of the similarity that these matrices express was studied, providing an effective characterization of what DC-similarity means for polylines that are drawn on a grid (e.g., a road network). This was also generalized to the plane.

A location prediction method, named WhereNext, has been developed for estimating the next location of a moving object, on the basis of the T-pattern mining method. The prediction method uses the collection of previously extracted Trajectory Patterns (developed in year 1 and 2) as a concise representation of typical behaviors of moving objects. Thus, the learned predictive rules depend on the movement of all available objects in a certain area, rather than on the individual history of an object. A set of quality measures have been proposed and tuned on a real life case study, and an exhaustive set of experiments and results on the real dataset are presented, which show the high prediction accuracy of the method [TMPG09]. In collaboration with WP3, the location prediction method has been integrated into the Dedalus DMQL.

The problem of assigning the category of a trajectory (e.g., the type of vehicle, such pedestrian, bike,
car, etc.) has been studied, resulting in a **trajectory classification** framework [NQ08] that is based on the analysis and detection of frequent patterns describing how trajectories interact with each other. For each time instant, each trajectory is compared with the others in the database that stay within a spatio-temporal neighborhood, and several descriptors are computed that measure the deviation of its behaviour as compared to other trajectories, for instance the difference between the speed of the single trajectory and the average speed of the neighboring group. This process produces an alternative representation of trajectories, over which frequent sequences of values are extracted and treated as typical profiles, and later used as discriminative features in the classification task, i.e., objects are characterized by the interaction patterns that they follow. This represents another local-to-global approach to the construction of a global model, as in the above-mentioned case of WhereNext for location prediction.

A scalable algorithm that learns Bayesian networks from trajectory data [LKM08] has been designed, which allows for a compact representation of **location dependencies**. The algorithm has been applied to the GSM Milano data set, which contains the movement of cell phone users during their calls on the level of GSM cells. The data has been preprocessed by WIND to contain in addition to the start- and end cell also all cells in between. In total, the data set contains 98994 records of 17948 user over one week. Figure 1 shows the results of the Milan Bayes net for three different queries.

![Figure 1: Query results for the Milano Bayes net](image)

The top left picture shows the overall probability to be in some cell during a phone call. Light red indicates a higher probability, dark red a lower probability. The blue cells are not part of the model because of missing input data for these cells. The remaining pictures show the probabilities when the passage of one or more cells is set as certain (green cells). The positive correlation structure of neighboring cells is clearly visible in the pictures. Over the past couple of years, it has become apparent that applications request fine-grained spatial models not only in areas with sufficient mobility data but also in areas with few mobility data for the pursued level of resolution. However, supplementary data sources (e.g., GSM) that provide information on a coarser level of resolution may be available. The advantage of the Bayesian network approach is, that local dependencies can be modeled using artificial trajectories. In combination with global mobility information (e.g., GSM), fine-grained mobility models may be derived for the application.

### 3 Applications and preprocessing tools

In order to exploit at the best the analysis capabilities of the tools developed in the project, particular care has to be paid to the data preparation steps that precede the data analysis. Beside the basic preprocessing functionalities provided by the database and data warehousing systems (see WP 1), a
**map matching** tool for mobility data was also developed, that takes into account the following major challenges for map matching: a) noise and outlier data points, b) warrant of contiguity in street network, c) measurement gaps (e.g., due to tunnels or GPS-warm-up phase) and d) recognition of appropriate street network. The tool provides several possibilities for data pre- and postprocessing to address the above challenges. For data preprocessing, the tool removes invalid GPS points (e.g., points during warm-up phase), extreme outliers (e.g. speed outliers, outliers outside of the map) or carries out any user-specified database query. In a second step, clouds of GPS points that result from stops or very slow motion are detected and replaced by a single center of mass point. Also the street network is preprocessed for performance optimization: it is scaled to the relevant section the street network containing the input GPS trajectory. The map matching algorithm is an extension of Greenfeld’s algorithm [Gre02] and belongs to the class of topological matching algorithms. During postprocessing the trajectory is tested for mismatches of variable length at intersections. Finally, all remaining gaps in the trajectory can be closed using a distance or travel time optimizing routing algorithm.

Finally, the application of the mobility data mining methods developed in the project to a problem of analyzing the **accessibility to public transportation** has also been carried by Fabio Pinelli during his six-month visit to the Senseable City Lab at MIT, Boston, MA, US: the results, achieved applying trajectory clustering techniques to GPS data from public buses in Rome, are documented in [PP09].

**Progresses in privacy-preserving data mining (Task 2.3)**

The general goals of this task are four: (i) to identify anonymity notions for spatio-temporal and trajectory data, (ii) to identify anonymity and privacy notions for spatio-temporal and trajectory patterns, (iii) to design and implement pattern hiding techniques for spatio-temporal and trajectory data, (iv) secure multi-party distributed mining of spatio-temporal and trajectory data.

In the third year of the project, Task 2.3 consolidated, validated and expanded the repertoire of privacy and anonymity tools contributed to the project.

Concerning **trajectory anonymity tools**, both methods TrajKAnon and Never Walk Alone, developed during the first two years of the project, have been thoroughly validated, showing their complementary strong and weak points, so that we now have a methodology to select the most appropriate method depending on the specific GPS trajectory dataset. In particular, we experimented on the Milano data set and evaluated the proposed anonymization techniques in terms of their affect to the data quality. We used clustering and spatial queries on the anonymized data and showed that the anonymized data could still be used for data mining and querying. Details of the comparative experimentation are provided in the WP2 Final Report (deliverable D2.7).

In addition to that, a new and more sophisticated variant of the trajectory anonymization tool Never Walk Alone has been implemented, which improves on the distortion of the underlying trajectory data. The new tool, named Wait for Me [ABN09], adopts a clustering strategy similar to Never Walk Alone, aimed to form groups of trajectories that are anonymous or will be made anonymous through spatial translation of points. The improvements essentially consist in adopting new time-tolerant trajectory distances, and in developing several optimization strategies to make the method highly scalable.

Also, new **anonymization methods for sequential data and GSM cell sequences** have been studied and thoroughly validated over real GSM data provided by WIND. This work has studied an extension of the notion of k-anonymity to sequential data, such that GSM data, consisting of sequences of time-
stamped cell locations corresponding to calls from the same user. An efficient algorithm for producing a k-anonymous version of a sequence dataset has been developed, with the aim of (i) guaranteeing a formal lower bound to the probability of re-identifying a k-anonymized sequence, and (ii) preserving some specified analytical properties, such as sequential patterns and clustering. An extensive experimentation has been carried out, which shows that the proposed method produces efficiently high quality results [PMPP08]. Also, a variant of the method dealing with a simple temporal extension of the sequential data has been designed and applied to GSM data, in collaboration with WIND: the results show that massive GSM data from operational billing systems of telecom operators are amenable to k-anonymization, also for reasonable high (and hence protective) values of k.

Also, further study has concentrated in clarifying how certain trajectory pre-processing methods leave doors open to subtle re-identification attacks [TBSS08, KBSS09], thus bringing further evidence to the fact that a more robust notion of k-anonymity is needed, such as those delivered by this task. The aim of the study was to challenge the state-of-the-art data transformation techniques proposed for time-series data. We particularly investigated time-series data since it is the closest data type to spatio-temporal trajectories. We first considered that distance-preserving transformation is applied to trajectory data, and only the distances among trajectories are released in the form of a distance matrix. Together with the distance matrix, we modeled an adversary who knows some existing trajectories as background information. This is a reasonable assumption since a group of people can work together by recording their own trajectories and collaboratively attack the released distance matrix. We experimented on the real data set obtained from Milano Municipality in the context of the project, and showed that the rest of the trajectories could be recovered when the adversary has such feasible background information. This showed that such distance preserving transformations are not reliable as tools for preprocessing data and further anonymization techniques are needed.

Concerning privacy-preserving data mining, efficient protocols for distributed privacy preserving clustering of trajectories have been consolidated and validated, which are based on secret sharing: remarkably, the involved parties do not need to use expensive cryptographic tools.

Finally, two methodologies for the automatic identification of sensitive routes (i.e. parts of trajectories) in a collection of users’ trajectories have been developed [GVM09]. Knowledge of these routes by an adversary can easily lead to the identification of specific individuals in the population of users, thus breach their location privacy, even when no other identifying information is disclosed. Two settings were considered: (i) unconstrained (free-terrain) user movement, where each user is allowed to move freely within a bounded area (e.g. a football field) [GV08], and (ii) network-constrained user movement, where we assume the existence of an underlying network topology (such as a road network) that defines the alternative paths in the network that one can follow to move from one location to another [GVB09]. For each of these contexts, an algorithm was proposed for the automatic identification of the users’ sensitive routes for their collected trajectories and for the subsequent protection of the users when requesting Location Based Services (LBSs) from within any of these routes. The solutions provided were tested both on real and on synthetic movement data, and were integrated both methodologies into a single toolbox [GVE09].
Bibliography


Appendix: Paper abstracts

In this appendix we collect the abstracts of the papers produced during the third year of activity of the project. The full papers are provided separately, in electronic form (PDF).


One of the most common operations in exploration and analysis of various kinds of data is clustering, i.e. discovery and interpretation of groups of objects having similar properties and/or behaviors. In clustering, objects are often treated as points in multi-dimensional space of properties. However, structurally complex objects, such as trajectories of moving entities and other kinds of spatiotemporal data, cannot be adequately represented in this manner. Such data require sophisticated and computationally intensive clustering algorithms, which are very hard to scale effectively to large datasets not fitting in the computer main memory. We propose an approach to extracting meaningful clusters from large databases by combining clustering and classification, which are driven by a human analyst through an interactive visual interface.


Preserving individual privacy when publishing data is a problem that is receiving increasing attention. Thanks to its simplicity the concept of k-anonymity, introduced by Samarati and Sweeney, has established as one fundamental principle for privacy preserving data publishing. According to the k-anonymity principle, each release of data must be such that each individual is indistinguishable from at least k – 1 other individuals. In this article we study the problem of anonymity preserving data publishing in moving objects databases. We propose a novel concept of k-anonymity based on co-localization that exploits the inherent uncertainty of the moving object’s whereabouts. Due to sampling and imprecision of the positioning systems (e.g., GPS), the trajectory of a moving object is no longer a polyline in a three-dimensional space, instead it is a cylindrical volume, where its radius delta represents the possible location imprecision: we know that the trajectory of the moving object is within this cylinder, but we do not know exactly where. If another object moves within the same cylinder they are indistinguishable from each other. This leads to the definition of (k, delta)-anonymity for moving objects databases. In this article we first characterize the (k, delta)-anonymity problem, then we recall NWA (Never Walk Alone) a method that we introduced in previous work, based on clustering and spatial perturbation, and enhanced with ad hoc preprocessing and outlier removal techniques. Starting by a discussion on the limits of NWA we develop a novel clustering method that, being based on EDR distance, it has the important feature of being time-tolerant. As a consequence the method perturbs trajectories both in space and time. The novel method, named W4M (Wait for Me) is empirically evaluated in terms of data quality and efficiency, and deeply compared to its predecessor NWA. Data quality is assessed both by means of objective measures of information distortion, and by more usability oriented measure, i.e., by
comparing the results of the same spatio-temporal range queries, and mining-like analyses, executed on the original database and on the (k, delta)-anonymized one. Experimental results show that for a wide range of values of delta and k, over both real-world and synthetic mobility data, the relative distortion introduced by our methods is kept low, with W4M producing higher quality (k, delta)-anonymized data than NWA.

[DBSSL08] Mahir Can Doganay, Thomas Brochmann Pedersen, Yücel Saygin, Erkay Savas, Albert Levi: Distributed privacy preserving k-means clustering with additive secret sharing. PAIS 2008: 3-11

Recent concerns about privacy issues motivated data mining researchers to develop methods for performing data mining while preserving the privacy of individuals. However, the current techniques for privacy preserving data mining suffer from high communication and computation overheads which is prohibitive considering even a modest database size. Furthermore, the proposed techniques have strict assumptions on the involved parties which need to be relaxed in order to re

ect the real-world. In this paper we concentrate on a distributed scenario where the data is partitioned vertically over multiple sites and the involved sites would like to perform clustering without revealing their local databases. For this setting, we propose a new protocol for privacy preserving k-means clustering based on additive secret sharing. We show that the new protocol is more secure than the state of the art. Experiments conducted on real and synthetic data sets show that, in realistic scenarios, the communication and computation cost of our protocol is considerably less than the state of the art which is crucial for data mining applications.


The study of moving objects has been capturing the attention of Geographic Information System (GIS) researchers. Moving objects, carrying location-aware devices, produce trajectory data in the form of a sample of (Oid; t; x; y)-tuples, that contain object identifier and time-space information. Recently, the notion of stops and moves was introduced. Intuitively, if a moving object spends a sufficient amount of time in a certain geographic place (which we denote a place of interest of an application), this place is considered a stop of the object's trajectory. In-between stops, a trajectory has moves. In this paper we study how moving object data analysis can benefit from replacing raw trajectory data by a sequence of stops and moves. We first propose a formal model and query language (denoted Lmo) to express complex queries involving spatial data stored in a GIS, non-spatial data (stored in a data warehouse) and moving object data. This query language also supports different forms of aggregation. We then study the compression of trajectory data produced by moving objects, using the concepts of stops and moves. We show that stops and moves are expressible in Lmo and that there exists a fragment of this language (that can be expressed by means of regular expressions) allowing to talk about temporally ordered sequences of stops and moves. We use this fragment to perform data mining over trajectory data. We present an implementation and a case study, and discuss different applications of our approach.
We address aggregate queries over GIS data and moving object data, where non-spatial information is stored in a data warehouse. We propose a formal data model and query language to express complex aggregate queries. Next, we study the compression of trajectory data, produced by moving objects, using the notions of stops and moves. We show that stops and moves are expressible in our query language and we consider a fragment of this language, consisting of regular expressions to talk about temporally ordered sequences of stops and moves. This fragment can be used not only for querying, but also for expressing data mining and pattern matching tasks over trajectory data.

In sequential pattern discovery, the support of a sequence is computed as the number of data-sequences satisfying a pattern with respect to the total number of data-sequences in the database. When the items are frequently updated, the traditional way of counting support in sequential pattern mining may lead to incorrect (or, at least incomplete), conclusions. For example, if we are looking for the support of the sequence A.B, where A and B are two items such that A was created after B, all sequences in the database that were completed before A was created, can never produce a match. Therefore, accounting for them would underestimate the support of the sequence A.B. In this paper we propose to revise the classic notion of support in sequential pattern mining, introducing the concept of temporal support of a sequential expression (SE), intuitively defined as the number of sequences satisfying a target pattern, out of the total number of sequences that could have possibly matched such pattern. We then generalize this notion to regular expressions (RE) which encapsulate the definition of a collection of SEs. We present and discuss a theoretical framework for these novel notion of support, and present an algorithm to compute it.

Trajectories are spatio-temporal traces of moving objects which contain valuable information to be harvested by spatio-temporal data mining techniques. Applications like city traffic planning, identification of evacuation routes, trend detection, and many more can benefit from trajectory mining. On the other hand, special care needs to be taken if trajectories of individuals are going to be disclosed for analysis due to private information contained in those trajectories. Removing identifiers from trajectories before the release is not effective against linkage type attacks, and rich sources of background information make it even worse. An alternative is to apply transformation techniques to map the given set trajectories into another set where the distances are preserved. This way, the actual trajectories are not released and the distance information can be used for data mining techniques such as clustering. In this paper, we show that an unknown private trajectory can be re-constructed using the available background information together with the mutual distances released for data mining purposes. The background knowledge is in the form of
known trajectories and extra information such as the speed limit. We provide analytical results which bound the number of the known trajectories needed to reconstruct private trajectories. Experiments performed on real trajectory data sets show that the number of known samples is surprisingly smaller than the actual theoretical bounds.


One of the formalisms to qualitatively describe polylines in the plane are double-cross matrices. In a double-cross matrix the relative position of any two line segments in a polyline is described with respect to a double cross based on their start points. Two polylines are called DC-similar if their double-cross matrices are identical. Although double-cross matrices have been widely applied, a geometric interpretation of the similarity they express is still lacking. In this paper, we provide a first step in the geometric interpretation of this qualitative definition of similarity. In particular, we give an effective characterization of what DC-similarity means for polylines that are drawn on a grid. We also provide algorithms that, given a DC-matrix, check whether it is realizable by a polyline on a grid and that construct, if possible, in quadratic time example polylines that satisfy this matrix. We also describe algorithms to reconstruct polylines, satisfying a given double-cross matrix, in the two-dimensional plane, that is, not necessarily on a grid.


One of the formalisms to qualitatively describe polylines in the plane are double-cross matrices. In a double-cross matrix the relative position of any two line segments in a polyline is described with respect to a double cross based on their start points. Two polylines are called DC-similar if their double-cross matrices are identical. Although double-cross matrices have been widely applied, a geometric interpretation of the similarity they express was still lacking. In this paper, we provide a geometric interpretation of this qualitative definition of similarity. In particular, we give an effective characterization of what DC-similarity means for polylines in terms of the new notion of local carrier order of a polyline. We also discuss the realizability of a DC-matrix and if it is, how to construct example polylines that satisfy it.


A key issue in clustering data, regardless the algorithm used, is the definition of a distance function. In the case of trajectory data, different distance functions have been proposed, with different degrees of complexity. All these measures assume that trajectories are error-free, which is essentially not true. Uncertainty is present in trajectory data, which is usually obtained through a series of GPS of GSM observations. Trajectories are then reconstructed, typically using linear interpolation. A first source of error in trajectory are the GPS observations themselves, since many reported points lie outside the road network. Thus, the
users position must be matched to a map, leading to the problem of map matching. A well-known model to deal with uncertainty in a trajectory sample, uses the notion of space-time prisms (also called beads), to estimate the positions where the object could have been, given a maximum speed. Thus, we can replace a (reconstructed) trajectory by a necklace (intuitively, a chain of prisms), connecting consecutive trajectory sample points. When it comes to clustering, the notion of uncertainty requires appropriate distance functions. The main contribution of this paper is the definition of a distance function that accounts for uncertainty, together with the proof that this function is also a metric, and therefore it can be used in clustering. We also present an algorithm that computes the distance between the chains of prisms corresponding to two trajectory samples. Finally, we discuss some preliminary results, obtained clustering a set of trajectories of cars in the center of the city of Milan, using the distance function introduced in this paper.


Traffic routes through a street network contain patterns and are no random walks. Such patterns exist for instance along streets or between neighbouring street segments. The extraction of these patterns is a challenging task due to the enormous size of city street networks, the large number of required training data and the unknown distribution of the latter. We apply Bayesian Networks to model the correlations between the locations in space-time trajectories and address the following tasks. We introduce and examine a Bayesian Network Learning algorithm enabling us to handle the complexity and performance requirements of the spatial context. Furthermore, we apply our method to German cities, evaluate the accuracy and analyse the runtime behaviour for different parameter settings.


Trajectory datasets are becoming popular due to the massive usage of GPS and location-based services. In this paper, we address privacy issues regarding the identification of individuals in static trajectory datasets. We first adopt the notion of k-anonymity to trajectories and propose a novel generalization-based approach for anonymization of trajectories. We further show that releasing anonymized trajectories may still have some privacy leaks. Therefore we propose a randomization based reconstruction algorithm for releasing anonymized trajectory data and also present how the underlying techniques can be adapted to other anonymity standards. The experimental results on real and synthetic trajectory datasets show the effectiveness of the proposed techniques.


The increasing pervasiveness of location-acquisition technologies (GPS, GSMnetworks, etc.) is leading to the collection of large spatio-temporal datasets and to the opportunity of discovering usable knowledge about movement behavior, which fosters novel applications and services. In this paper, we move towards this direction and develop an extension of the
supervised learning paradigm that analyzes the trajectories of moving objects. We introduce trajectory interaction patterns as concise descriptors of regions, in terms of both space (i.e., the regions of space visited during movements) and relations of similarity between trajectories, to extract semantics from them. In this paper, we provide a general statement of the novel approach, describe its main aspects, both in methodological and logical components, and validate it over a real dataset of moving vehicles.


Nowadays, the study of the city connectivity is very appealing for many end user analysts such as urban planners, public trasportation companies, and so on. In many recent research works, focused on this topic tries to measure different connectivity indexes over the network using GPS data for few bus routes or using the timetable over all network measuring all the connectivity city. In this paper, we propose a new complete methodological process to analyze and measure the connectivity of a city using only the GPS data. Firstly, the process include the application of recent results in spatio-temporal data mining in order to detect bus lines and bus stops in the city and some number of the real network allow us to validate the process. Finally, a spanning tree algorithm is used to simulate human behaviors in the network studying how the connectivity of the city changes in different hours of the day and in different days of the week. Moreover, by means of a real case of study both the detection process and the connectivity study are deeply investigated.


Sequential pattern mining is a major research field in knowledge discovery and data mining. Thanks to the increasing availability of transaction data, it is now possible to provide new and improved services based on users' and customers' behavior. However, this puts the citizen's privacy at risk. Thus, it is important to develop new privacy-preserving data mining techniques that do not alter the analysis results significantly. In this paper we propose a new approach for anonymizing sequential data by hiding infrequent, and thus potentially sensible, subsequences. Our approach guarantees that the disclosed data are k-anonymous and preserve the quality of extracted patterns. An application to a real-world moving object database is presented, which shows the effectiveness of our approach also in complex contexts.


One of the fundamental challenges that the data mining community faces today is privacy. The question "How are we going to do data mining without violating the privacy of individuals?" is still on the table, and research is being conducted to find efficient methods to do that. Data transformation was previously proposed as one efficient method for privacy preserving data mining when a party needs to out-source the data mining task, or when distributed data mining needs to be performed among multiple parties without each party disclosing its actual data. In this paper we study the safety of distance preserving data transformations proposed for privacy preserving data mining. We show that an adversary
can recover the original data values with very high confidence via knowledge of mutual distances between data objects together with the probability distribution from which they are drawn. Experiments conducted on real and synthetic data sets demonstrate the effectiveness of the theoretical results.


The pervasiveness of mobile devices and location based services is leading to an increasing volume of mobility data. This side effect provides the opportunity for innovative methods that analyse the behaviors of movements. In this paper we propose WhereNext, which is a method aimed at predicting with a certain level of accuracy the next location of a moving object. The prediction uses previously extracted movement patterns named Trajectory Patterns, which are a concise representation of behaviors of moving objects as sequences of regions frequently visited with a typical travel time. A decision tree, named T-pattern Tree, is built and evaluated with a formal training and test process. The tree is learned from the Trajectory Patterns that hold a certain area and it may be used as a predictor of the next location of a new trajectory finding the best matching path in the tree. Three different best matching methods to classify a new moving object are proposed and their impact on the quality of prediction is studied extensively. Using Trajectory Patterns as predictive rules has the following implications: (I) the learning depends on the movement of all available objects in a certain area instead of on the individual history of an object; (II) the prediction tree intrinsically contains the spatio-temporal properties that have emerged from the data and this allows us to define matching methods that strictly depend on the properties of such movements. In addition, we propose a set of other measures, that evaluate a priori the predictive power of a set of Trajectory Patterns. This measures were tuned on a real life case study. Finally, an exhaustive set of experiments and results on the real dataset are presented.