Spatio-Temporal Data Warehouses: Current Status and Research Issues

Esteban ZIMÁNYI
Department of Computer & Decision Engineering (CoDE)
Université Libre de Bruxelles
ezimanyi@ulb.ac.be

Invited Talk
19th International Symposium on Temporal Representation and Reasoning
Leicester, UK
12-14 September 2012
Contents

Background
- Data Warehouses and OLAP
- Nested Relational Calculus
- Temporal Aggregation

- Spatio-Temporal Data Warehouses
- From Spatio-Temporal Data to Trajectory Data
- Conclusions & Future Work
Data Warehouses

◆ **Operational databases (OLTP)** are not suitable for data analysis
  - Contain detailed data
  - Do not include historical data
  - Perform poorly for complex queries due to normalization

◆ **Data Warehouses (DW)** address requirements of decision-making users
  ⇒ **Business Intelligence**

◆ A data warehouse is a collection of subject-oriented, integrated, nonvolatile, and time-varying data to support data management decisions

◆ **Online analytical processing (OLAP)** allow decision-making users to perform interactive analysis of data

◆ **Data Warehouses** typically store huge amounts of data
Typical Data Warehouse Architecture

- **Operational databases**
- **External sources**
- **Internal sources**
- **OLAP tools**
- **Reporting tools**
- **Data mining tools**
- **Data staging**
- **Metadata**
- **OLAP server**
- **Data marts**
- **Statistical tools**
- **Data warehouse tier**
- **Back-end tier**
- **Data sources**
- **OLAP tier**
- **OLAP server**
- **Front-end tier**
- **ETL process**
- **Enterprise data warehouse**
- **Metadata**
- **OLAP server**
Conventional Data Warehouses: Example
Nested Relational Calculus

- **Simple query**: Name and population of districts of the Antwerp province
  
  \[
  \{ d.name, d.population \mid \text{District}(d) \land \exists p (\text{Province}(p) \land d.province = p.id \land p.name = 'Antwerp') \}
  \]

- **Aggregation query with group filtering**: Total population by province provided that it is greater than 100,000

  \[
  \{ p.name, \text{totalPop} \mid \text{Province}(p) \land \\
  \text{totalPop} = \sum \{ d.id, d.population \mid \text{District}(d) \land d.province = p.id \} \land \\
  \text{totalPop} > 100,000 \}
  \]

- Corresponds to an SQL query with the **GROUP BY** and **HAVING** clauses
Temporal Aggregation [Gamper et al. 2009]

- **Instantaneous Temporal Aggregation**
  - To each time instant \( t \) is associated an aggregation group valid at \( t \)
  - Aggregation function applied to each group \( \Rightarrow \) a single aggregate value at \( t \)
  - **Span Temporal Aggregation**: Similar, at a coarser granularity

- **Moving Window Temporal Aggregation**
  - To each time instant \( t \) is associated a window \( w = [t-w, t+w'] \) and an aggregation group valid at \( w \)
  - Aggregation function applied to each group \( \Rightarrow \) a single aggregate value at \( t \)
  - **Multi-Dimensional Temporal Aggregation**: Similar, at a coarser granularity

- Equivalent to **temporal group composition** and **temporal partition composition** in [Vega López et al. 2005]
Temporal Aggregation

- **Instantaneous Temporal Aggregation**
  - $f(t) ightarrow v$

- **Span Temporal Aggregation**
  - $f(t_1, t_n) ightarrow v$

- **Moving Window Temporal Aggregation**
  - $f(t-w, t+w') ightarrow v$

- **MultiDimensional Temporal Aggregation**
  - $f(t_1, t_n) ightarrow v$
By pollutant and day count the number of stations that exceeded the load limit

\[
\{ p.\text{name}, t.\text{date}, \text{countExc} \mid \text{Pollutant}(p) \land \text{Time}(t) \land \\
\text{countExc} = \text{count}\_1(\{ w.\text{station} \mid \text{WaterPollution}(w) \land w.\text{pollutant} = p.\text{id} \land \\
w.\text{time} = t.\text{id} \land w.\text{load} > p.\text{loadLimit}\}) \}
\]
By station, pollutant, and day, give the 3-day moving average of load

\[
\{s.name, p.name, t.date, 3dMovAvg \mid \text{Station}(s) \land \text{Pollutant}(p) \land \text{Time}(t) \land \\
3dMovAvg = \text{avg}_2(\{w.id, w.load \mid \text{WaterPollution}(w) \land w.station = s.id \land \\
w.pollutant = p.id \land \exists t_1 (\text{Time}(t_1) \land w.time = t_1.id \land \\
0 \leq t.date - t_1.date \land t.date - t_1.date \leq 3) \})\}
\]
Contents

- Background

  ➖ Spatio-Temporal Data Warehouses
    - Spatial Data Warehouses
    - Temporal Data Warehouses
    - Spatio-Temporal Data Warehouses
    - Support of Continuous Fields

- From Spatio-Temporal Data to Trajectory Data

- Conclusions & Future Work
Spatio-Temporal Data Warehouses

- Several proposals aim at extending DW and OLAP with spatial/temporal features
- No commonly agreed definition of what is a spatio-temporal DW and what functionality it should support
- Proposed solutions vary considerably in the kind of data that can be represented and the kind of queries that can be expressed
- [Vaisman & Zimányi 2009] defined
  - Conceptual framework for spatio-temporal DWs using an extensible type system
  - Taxonomy of several classes of queries of increasing expressive power extending the tuple relational calculus with aggregated functions [Klug 1982]
- This provides the underlying basis for implementing spatio-temporal DWs
A Taxonomy for Spatio-Temporal Data Warehouses
Two complementary ways to represent spatial data

- **Object-based** (vector) model: Objects of interest are stored with their spatial extent
- **Space-based** (raster) model: Space is represented as a continuum, to each point in space is associated a value of a phenomenon of interest
Spatial Data Types [Güting et al. 2000]

- **Spatial types**: point, points, line, region

- These types have an associated set of operations

<table>
<thead>
<tr>
<th>Class</th>
<th>Operations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Predicates</td>
<td>isempty, =, ≠, intersects, inside, &lt;, ≤, ≥, &gt;, before,</td>
</tr>
<tr>
<td></td>
<td>touches, attached, overlaps, on_border, in_interior</td>
</tr>
<tr>
<td>Set Operations</td>
<td>intersection, union, minus, crossings,</td>
</tr>
<tr>
<td></td>
<td>touch_points, common_border</td>
</tr>
<tr>
<td>Aggregation</td>
<td>min, max, avg, center, single</td>
</tr>
<tr>
<td>Numeric</td>
<td>no_components, size, duration, length, area, perimeter</td>
</tr>
<tr>
<td>Distance and Direction</td>
<td>distance, direction</td>
</tr>
</tbody>
</table>
For stations located on the Schelde river, give the average content of lead in the last quarter of 2008

\[
\{ s\text{.name, avgLead} \mid \text{Station}(s) \land \exists r, \exists p \text{ (River}(r) \land r\text{.name = ‘Schelde’} \land \text{intersects}(r\text{.geometry}, s\text{.geometry}) \land \text{Pollutant}(p) \land p\text{.name = ‘Lead’} \land \text{avgLead} = \text{avg}_2(\{w\text{.id, w.load} \mid \text{WaterPollution}(w) \land w\text{.station = s.id} \land w\text{.pollutant = p.id} \land \exists t \text{ (Time}(t) \land w\text{.time = t.id} \land t\text{.date} \geq 1/10/2008 \land t\text{.date} \leq 31/12/2008) \}) \}\}
\]
Spatial Aggregation: Space Based

- Local Spatial Aggregation
- Moving Window Spatial Aggregation
- MultiDimensional Spatial Aggregation

![Spatial Aggregation Diagram](image)
Spatial Aggregation: Object Based

Data from French-speaking countries and their neighbors

Data from Germany-speaking countries

Data from Poland

Data from Serbia and its neighbors

Local Spatial Aggregation

Moving Window Spatial Aggregation

MultiDimensional Spatial Aggregation
Union of the geometries of the districts in which the average content of lead in the last quarter of 2008 was greater than the load limit

union({d.geometry | District(d) ∧ ∃p (Pollutant(p) ∧ p.name = ‘Lead’ ∧ avg₂({w.id, w.load | WaterPollution(w) ∧ w.district = d.id ∧ w.pollutant = p.id ∧ ∃t (Time(t) ∧ w.time = t.id ∧ t.date ≥ 1/10/2008 ∧ t.date ≤ 31/12/2008) }) > p.loadLimit) }
Temporal Data Warehouses

- Arise when evolution of dimension instances is supported
  - Also referred to as slowly changing dimensions [Kimball 96]

- **Temporality types**: Valid time (VT), Transaction Time (TT), Bitemporal Time (BT), Lifespan (LS), DW loading time (DWLT)

- Temporality represented using **moving types** moving($\alpha$) where $\alpha$ is a base type
  - Lifespan of Pollutant is of type moving(bool)
  - Temporal attribute loadLimit is of type moving(real)
  - Temporal relationship between Pollutant and Category is of type moving(id)
Moving Types [Güting et al. 2000]

- Capture the evolution over time of base types and spatial types
- Obtained by applying a constructor \texttt{moving(\(\alpha\))}
  - A value of type \texttt{moving(point)} is a continuous function \(f : \text{instant} \rightarrow \text{point}\)
- Operations on moving types

<table>
<thead>
<tr>
<th>Class</th>
<th>Operations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Projection to Domain/Range</td>
<td>\texttt{deftime}, \texttt{rangevalues}, \texttt{locations}, \texttt{trajectory}, \texttt{routes}, \texttt{traversed}, \texttt{inst}, \texttt{val}</td>
</tr>
<tr>
<td>Interaction with Domain/Range</td>
<td>\texttt{atinstant}, \texttt{atperiods}, \texttt{initial}, \texttt{final}, \texttt{present}, \texttt{at}, \texttt{atmin}, \texttt{atmax}, \texttt{passes}</td>
</tr>
<tr>
<td>Rate of change</td>
<td>\texttt{derivative}, \texttt{speed}, \texttt{turn}, \texttt{velocity}</td>
</tr>
<tr>
<td>Lifting</td>
<td>(all new operations inferred)</td>
</tr>
</tbody>
</table>

- **Lifting**: Operations of moving types generalize those of the nontemporal types
  - A \texttt{distance} function with signature \texttt{moving(point)\times moving(point) \rightarrow moving(real)} calculates the distance between two moving points
- **Semantics**: result is \textbf{computed at each time instant} using the non-lifted operation
By district, pollutant category, and month, give the average load

\[
\{ d.\text{name}, c.\text{name}, m.\text{month}, \text{avgLoad} \mid \text{District}(d) \land \text{Category}(c) \land \text{Month}(m) \land \\
\text{avgLoad} = \text{avg}_2(\{w.\text{id}, w.\text{load} \mid \text{WaterPollution}(w) \land \exists t, \exists p (\text{Time}(t) \land \\
\text{Pollutant}(p) \land w.\text{district} = d.\text{id} \land w.\text{time} = t.\text{id} \land t.\text{month} = m.\text{id} \land \\
w.\text{pollutant} = p.\text{id} \land \text{val}(\text{initial}(\text{atperiods}(p.\text{category}, t))) = c.\text{id} \}))\}
\]

Here we consider the category of a pollutant valid at the day of the measure

Alternative: consider the current category of a pollutant
Spatio-Temporal OLAP (ST-OLAP)

- Arise when **spatial objects evolve over time**
- Evolution captured by **moving types** moving(α) where α is a **spatial type**
Spatio-Temporal OLAP (ST-OLAP) Queries

By district and month, give the total number of persons affected by polluting clouds

\[
\{d.\text{name}, m.\text{month}, \text{totalNo} \mid \text{District}(d) \land \text{Month}(m) \land \\
\text{totalNo} = \text{area}(\text{union}(\{\text{traversed}(p.\text{commonArea}) \mid \text{AirPollution}(p) \land \\
p.\text{district} = d.\text{id} \land \exists t (\text{Time}(t) \land t.\text{month} = m.\text{id}))\}))/\\
\text{area}(d.\text{geometry}) \times d.\text{population}\}
\]
**Spatial TOLAP (S-TOLAP)**

<table>
<thead>
<tr>
<th>Month</th>
<th>District</th>
<th>LS</th>
<th>Pollutant</th>
</tr>
</thead>
<tbody>
<tr>
<td>month</td>
<td>name</td>
<td>...</td>
<td>VT loadLimit</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Calendar</th>
<th>Water Pollution</th>
<th>Station</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time</td>
<td>load</td>
<td>name</td>
</tr>
</tbody>
</table>

- Arise when there are **spatial objects/attributes and temporal dimensions**

- By pollutant and month, give the average load in stations of the Namur district, if it is larger than the load limit during the reported month

\[
\{p.\text{name}, m.\text{month}, \text{avgLoad} \mid \text{Pollutant}(p) \land \text{Month}(m) \land \exists d, \exists s (\text{District}(d) \land d.\text{name} = \text{‘Namur’} \land \text{Station}(s) \land \text{inside}(s.\text{geometry}, d.\text{geometry}) \land \\
\text{avgLoad} = \text{avg}_2(\{w.\text{id}, w.\text{load} \mid \text{WaterPollution}(w) \land \exists t (\text{Time}(t) \land \\
\text{w.station} = s.\text{id} \land w.\text{time} = t.\text{id} \land t.\text{month} = m.\text{id} \land w.\text{pollutant} = p.\text{id})\}) \land \\
\text{avgLoad} > \text{val}(\text{initial}(\text{atperiods}(p.\text{loadLimit}, m.\text{month})))} \}
\]
**Spatio-Temporal TOLAP (ST-TOLAP)**

- **Most general case**: there are moving geometries and temporal dimensions

- Number of days when the Gent district was under at least one cloud of carbon monoxide (CO) with a load larger than the load limit valid when the cloud appeared

\[
\text{duration}(\text{union}(\{ t.\text{date} \mid \text{Time}(t) \land \exists a, \exists d, \exists c, \exists p \ (\text{AirPollution}(a) \land \text{District}(d) \land \text{Cloud}(c) \land \text{Pollutant}(p) \land a.\text{time} = t.\text{id} \land a.\text{district} = d.\text{id} \land d.\text{name} = \text{‘Gent’} \land a.\text{cloud} = c.\text{id} \land c.\text{pollutant} = p.\text{id} \land p.\text{name} = \text{‘CO’} \land a.\text{load} > \text{val(\text{atinstant}(p.\text{loadLimit}, \text{inst(\text{initial(\text{at}(c.\text{lifespan, true))))}))})})})
\]

27
Spatio-Temporal Aggregation: Space Based (1)
Spatio-Temporal Aggregation: Space Based (2)

Local Spatial Aggregation

Moving Window Spatial Aggregation

Moving Window Temporal Aggregation

Local Spatial Aggregation
Continuous Fields

- **Non-temporal** fields describe phenomena that change continuously in space
  - land elevation, soil type, ... 
- Represented by **field types** field\((\alpha)\) where \(\alpha\) is a base type
- **Temporal** fields describe phenomena that change continuously in space and time
  - temperature, precipitation, ... 
- Represented by **field types** moving(field\((\alpha))\) where \(\alpha\) is a base type
  - moving(field(real)) defines a continuous function \(f : \text{instant} \rightarrow (\text{point} \rightarrow \text{real})\)
- Operators for moving fields as before
- Field levels have a **geometry** attribute of type field\((\alpha)\) or moving(field\((\alpha))\)
- Field dimensions are **not connected to a fact relationship**
Spa\(\text{tial Data Warehouses with Continuous Fields: Example}\)

\begin{itemize}
  \item Time
    \begin{itemize}
    \item date
    \item season
    \end{itemize}
  \item Calendar
  \item Month
    \begin{itemize}
    \item month
    \end{itemize}
  \item Quarter
    \begin{itemize}
    \item quarter
    \end{itemize}
  \item Year
    \begin{itemize}
    \item year
    \end{itemize}
  \item District
    \begin{itemize}
    \item name
    \item population area
    \end{itemize}
  \item Geo location
  \item Province
    \begin{itemize}
    \item name
    \item major activity capital
    \end{itemize}
  \item Water Pollution
    \begin{itemize}
    \item common Area
    \item load
    \end{itemize}
  \item Categories
  \item Pollutant
    \begin{itemize}
    \item name
    \item type
    \item load limit
    \end{itemize}
  \item River
    \begin{itemize}
    \item name
    \item flow
    \end{itemize}
  \item Station
    \begin{itemize}
    \item name
    \end{itemize}
  \item Category
    \begin{itemize}
    \item name
    \item description
    \end{itemize}
  \item SoilType
    \begin{itemize}
    \item \(f(\text{\#})\)
    \item classifSystem
    \item date
    \item characteristics
    \end{itemize}
  \item Elevation
    \begin{itemize}
    \item \(f(\text{\#})\)
    \item units
    \item minValue
    \item maxValue
    \end{itemize}
  \item Temp
    \begin{itemize}
    \item \(f(\text{\#},\text{\#})\)
    \item units
    \item startDate
    \item endDate
    \end{itemize}
\end{itemize}
Field Types [Vaisman & Zimányi 2009]

- Capture the variation in space of base types
- Obtained applying a constructor \( \text{field}(\alpha) \)
  - A value of type \( \text{field}(\text{real}) \) is a continuous function \( f : \text{point} \rightarrow \text{real} \)
- Operations on field types

<table>
<thead>
<tr>
<th>Class</th>
<th>Operations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Projection to Domain/Range</td>
<td>( \text{defspace}, \text{rangevalues}, \text{point}, \text{val} )</td>
</tr>
<tr>
<td>Interaction with Domain/Range</td>
<td>( \text{atpoint}, \text{atpoints}, \text{atline}, \text{atregion}, \text{at}, \text{atmin}, \text{atmax}, \text{defined}, \text{takes}, \text{concave}, \text{convex}, \text{flex} )</td>
</tr>
<tr>
<td>Lifting</td>
<td>(all new operations inferred)</td>
</tr>
<tr>
<td>Rate of change</td>
<td>( \text{partialder}_x, \text{partialder}_y )</td>
</tr>
<tr>
<td>Aggregation operators</td>
<td>( \text{integral}, \text{area}, \text{surface}, \text{favg}, \text{fvariance}, \text{fstdev} )</td>
</tr>
</tbody>
</table>

- **Lifting** applies to fields
  - The + operator with signature \( \alpha \times \alpha \rightarrow \alpha \) generalized by allowing any argument to be a field as in \( \text{field}(\alpha) \times \text{field}(\alpha) \rightarrow \text{field}(\alpha) \)
- **Semantics**: result is **computed at each point in space** using the non-lifted operation
For districts having at least 30% of their surface of clay soil, give the average load of lead on February 1st, 2009

\[
\{d.name, \text{avgLead} \mid \text{District}(d) \land \exists p, \exists s (\text{Pollutant}(p) \land p.name = \text{‘Lead’} \land \\
\text{SoilType}(s) \land \text{area(defspace(atregion(at(s.geometry, ‘Clay’), d.geometry))))/area(d.geometry) \geq 0.3 \\
\text{avgLead} = \text{avg}_2(\{w.id, w.load \mid \text{WaterPollution}(w) \land \exists t (\text{Time}(t) \land \\
w.district = d.id \land w.time = t.id \land t.date = 1/2/2009 \land w.pollutant = p.id\})\})\}
\]
For each river and month, give a field computing the average temperature of the month at each point in the river

\[
\{ r\text{.name}, m\text{.month}, \text{temp} \mid \text{River}(r) \land \text{Month}(m) \land \\
\text{first} = \min(\{ t\text{.date} \mid \text{Time}(t) \land t\text{.month} = m\text{.id} \} \land \\
\text{last} = \max(\{ t\text{.date} \mid \text{Time}(t) \land t\text{.month} = m\text{.id} \} \land \\
\text{temp} = \text{avg}(\{ \text{atperiods(atregion}(f\text{.geometry}, r\text{.geometry}), \text{range}(\text{first}, \text{last})) \mid \\
\text{Temp}(f) \}) \}) \}
\]
Spatio-Temporal Data Warehouses: Conclusions

- Spatio-temporal DWs result from combining GIS, OLAP, and temporal data types
  - Temporal data types model geometries that evolve over time (moving objects) and evolving (slowly changing) dimensions
  - Field data types model continuous fields that change in space
  - Temporal fields obtained by composing field and temporal data types

- Our taxonomy for spatio-temporal OLAP queries
  - characterizes features required by spatio-temporal DWs
  - allows to classify different work in the literature

- Our framework is at a conceptual level: implementation issues were omitted
  - From abstract to concrete models: e.g., grid and TIN models for continuous fields
  - Optimization issues: index structures, pre-aggregation, query optimization, ...
Contents

- Background
- Spatio-Temporal Data Warehouses

- From Spatio-Temporal Data to Trajectory Data
  - Trajectory Construction
  - Visual Analytics
  - Trajectory Mining

- Conclusions & Future Work
Typical Data Warehouse Architecture (Reminder)
Movement Data

- Typically, a **temporal sequence** of position records \((id, x, y, t)\)

- Collected nowadays in **rapidly growing amounts** due to the development of tracking technologies
  - GPS, RFID, WiFi, mobile phones, banking transactions, sensor networks, ...

- Complexities
  - **Huge amounts**: number of moving entities, number of records
  - **Geographic space** with its structure and complexity
  - **Time domain** at multiple granularities, linear and also multiple nested and overlapping cycles
  - **Data properties**: imprecision (errors in location, time, attributes), irregular and/or sparse sampling, missing values, ...
  - Large **diversity of types of movement**: constrained vs. unconstrained, vehicles/persons/animals, 2D vs 3D ...
  - **Real-world**: ill-defined, application-dependent problems
Examples of Movement Data: Migration of White Storks

Tracks of 35 storks during 8 years, about 2,000 positions ⇒ Animals
Examples of Movement Data: Cars in Milan

2,075,216 position records of 17,241 cars during 1 week ⇒ Vehicles, network-constrained
Examples of Movement Data: Children in Amsterdam

GPS tracks of 303 school children playing an educational game in Amsterdam, about 57,000 points ⇒ Pedestrians

Slide kindly lend by Gennady Andrienko
Examples of Movement Data: Air Traffic

427,652 position records of 17,851 planes during 1 day ⇒ 3D movement
From Movement to Trajectories [Parent et al. 2013]

- Movement is continuous and never-ending

- A trajectory is a finite meaningful movement segment

- Trajectory data is semantically richer than spatio-temporal data
Cleaning Raw Data

- Input: Raw data

- Output: Cleaned raw data

- Methods: filtering, smoothing, outlier removal, missing point interpolation, map-matching, data compression, etc.
Segmentation into Trajectories

- **Input:** Cleaned raw data

- **Output:** Trajectories

- **Methods:** various segmentation algorithms, based on spatial gaps, temporal gaps, time intervals, time series, etc.
Trajectory Structuring: Stop & Move Episodes

- **Input:** Trajectories

- **Output:** Structured trajectories (Stop, Move)

- **Methods:** various stop identification algorithms, based on velocity, density, etc.
Velocity-Based Stop Identification
Semantic Enrichment

- **Input:** Structured trajectories

- **Output:** Semantic trajectories

- **Methods:** relate structured components (begin, end, stops, moves, ... ) to application knowledge (i.e., meaningful objects)
Visual Analytics [Andrienko & Andrienko 2012]

◆ Aims at helping people in
  ● distilling relevant nuggets of information from large amounts of data
  ● understanding the connections among relevant information
  ● gaining insight from data

◆ Focuses on the division of labour between humans and machines

◆ Goal: computational power amplifies human perceptual and cognitive capabilities

◆ Visual representations: most effective means to convey information to human’s mind, prompt human cognition and reasoning

◆ Combines interactive visualizations with automated analysis techniques such as
  ● database processing
  ● data mining algorithms
  ● statistics
  ● geographical analysis methods
Interactive Transformation of Time

- Space-time cube on the right uses time transformation in respect to daily cycle
- This technique enables interpretation of repeated trajectories
Trajectory Summarization

- Left: major flows of tourists in Germany, according to panoramio.com photos
- Right: major traffic flows in Milano, based on trajectories of about 20,000 cars
Scalable Trajectory Clustering

- Major clusters of trajectories extracted from the same data set (a week of 20K cars in Milano) presented by the trajectory summarization method
Similarity of Situations

- Hourly traffic situations in Milano (spatial distributions of counts of cars) are clustered; colors are assigned to clusters according to the similarity of situations.
- The colors are projected to 2D time plot (bottom right) showing similarities of situations over 7 days x 24 hours.
For a cluster of trajectories, attribute values (speeds) are displayed on the trajectory wall display.

This enables investigation of traffic jams.
Trajectory Data Mining [Giannotti et al. 2011]

- The process of analyzing large amounts of trajectory data to **identify unsuspected or unknown patterns** that might be of value to an application
- A particular step in the knowledge discovery process
- Some key research questions
  - Which spatio-temporal patterns are **useful abstractions** of mobility data?
  - How to **classify trajectories** according to specific behavior?
  - How to **interpret** in a meaningful way the discovered patterns?
  - How to make such analysis **privacy preserving** in a measurable way?
What are Trajectory Patterns (1)?

- Frequent sets/sequences of places visited
What are Trajectory Patterns (2) ?

- Groups of objects moving together
Trajectory Clustering

- **Cluster analysis**: Find groups where objects in a group are *similar* (or near) to one another and *different* from (or distant from) the objects in other groups.

- Some research questions:
  - Which distance between trajectories?
  - Which kind of clustering?
  - What does a cluster mean?
    - A representative trajectory?
The M-Atlas tool

- A knowledge discovery support environment for trajectory data
- Data, patterns, and background knowledge need to be progressively combined
- T-patterns, T-clusters, etc. are the basic primitives within a Data Mining Query Language (DMQL), supporting the entire knowledge discovery process
- T-patterns, T-clusters, etc., once mined from a trajectory dataset, can be stored and later used for query, mining, and interpreting in a progressive way
Looking for Movement Patterns: From Where to Where

Select trajectories exiting the city from the centre towards North-West
Discovering Typical Routes

T-clustering divides trips based on route similarity
Focus on the Three Larger Clusters

One group (red) goes straight to NW, other follow alternative routes
Temporal Analysis on Each Group

A small group of commuters in the morning and a larger one in the afternoon
Temporal Analysis on Each Group

A group of commuters in the afternoon follows an alternative route: are they smart?
Conclusions

◆ A large number of applications in a variety of domains are interested in analyzing movement of some type of objects or phenomena

◆ Nowadays, a huge amount of movement data that is being continuously captured

◆ Current approaches to process these massive data sets are inappropriate

◆ A complex and pipelined process is needed for transforming raw samples to insightful knowledge

◆ This process is by definition iterative, semi-automatic, application-dependent, and multi-disciplinary

◆ This raises many theoretical and implementation issues
  ● A forthcoming book [Renso et al. 2013] surveys some of them

◆ Solutions proposed so far are just the first steps in the direction of mobility data understanding

◆ Exciting research domain, huge potential implications in our lives and in our planet
References (1)

References (2)


Spatio-Temporal Data Warehouses: Current Status and Research Issues

Esteban ZIMÁNYI

Questions ?