Discret event trace analysis: from an automatical to an interactive process

Céline Robardet

Lundi 12 mai 2014
Data Mining: a KDD step

• Making sense of voluminous data
• Elicit hypothesis on the mechanisms that generate the data
Data mining

It is the use of sophisticated data analysis tools to discover previously unknown, valid patterns and relationships in large data sets

Data mining ingredients:

- a language of patterns/models
- some criteria
- a search strategy
## Pattern example

### Formal concept extraction

<table>
<thead>
<tr>
<th></th>
<th>$p_1$</th>
<th>$p_2$</th>
<th>$p_3$</th>
<th>$p_4$</th>
<th>$p_5$</th>
<th>$p_6$</th>
<th>$p_7$</th>
<th>$p_8$</th>
<th>$p_9$</th>
<th>$p_{10}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$o_1$</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>$o_2$</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>$o_3$</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>$o_4$</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>$o_5$</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
**Pattern example**

**Formal concept extraction**

<table>
<thead>
<tr>
<th></th>
<th>$p_1$</th>
<th>$p_2$</th>
<th>$p_3$</th>
<th>$p_4$</th>
<th>$p_5$</th>
<th>$p_6$</th>
<th>$p_7$</th>
<th>$p_8$</th>
<th>$p_9$</th>
<th>$p_{10}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$o_1$</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>$o_2$</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>$o_3$</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>$o_4$</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>$o_5$</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
Pattern mining

- Patterns are subclasses of directed graphs
- Patterns depend on the data type and the applications needs

- Criteria: Evaluation of constraints over the patterns
- Search strategy: Exact search approaches
Mining Graphs

A timely challenge? Why is it such a keen interest?

- Large networks are now available (social networks, biology, etc.).
- New questions!

Main Challenge: knowledge discovery from large networks.
Discovering useful knowledge (patterns, relations, etc.) in structured data depicted as graph(s).

From data to knowledge.
Different application contexts

Mining traces of use of

a web search engine

the performance of photovoltaic installations

Vélo'V trips
1. Vertex attributes that are "linked" to the graph structure in static graphs: *Mining Graph Topological Patterns: Finding Co-variations Among Vertex Descriptors* (A. Prado, M. Plantevit, C. Robardet, J-F Boulicaut)


3. Causality between attribute value and graph structure: *Mining Triggering Attributes in Dynamic Graphs* (Y. Pitarch, M. Kaytoue, M. Plantevit, C. Robardet)

4. Temporal dependencies between sequences of lasting events: *Mining State Dependencies Between Multiple Sensor Data Sources* (M. Scuturici, C. Robardet, M. Plantevit)
1 Mining static graphs

Study of topological properties of the graphs

- Diameters.
- Identifying power laws.
- Centrality measures.
- A macroscopic view.

Local pattern mining

- Frequent sub-graphs.
- Dense sub-graphs.
- A microscopic view.
Mining co-variations among vertex descriptors

- **Attributed Graphs:**
  - Local attributes provide information on the entities represented as vertices.

- We want to consider the two views:
  - Macro + Micro = Topological patterns

- **Main idea:** discover constrained patterns containing:
  - local properties of vertices,
  - and their topological properties within the graph (given by statistical measures).

Co-authorship network.
Co-authorship network

• Local attributes: number of publications in some conferences series/journal (KDD, ICDM, VLDB, etc.).

Patterns we can discover:

• The higher the number of publications in VLDB and SIGMOD, the higher the centrality (Pagerank) of the related authors within the graph.
• The higher the number of publications in SAC:
  • the lower the number of publications in KDD and ICDM.
  • the lower the centrality in the graph.
How to Discover Such Patterns?

- **Propositionalization**: the topological properties are associated to each vertex.

⇒ Tabular view.

<table>
<thead>
<tr>
<th></th>
<th>KDD</th>
<th>PKDD</th>
<th>PODS</th>
<th>...</th>
<th>Page Rank</th>
<th>...</th>
<th>ClustCoef</th>
</tr>
</thead>
<tbody>
<tr>
<td>Author1</td>
<td>14</td>
<td>14</td>
<td>15</td>
<td>...</td>
<td>0.95</td>
<td>...</td>
<td>0.55</td>
</tr>
<tr>
<td>Author2</td>
<td>12</td>
<td>15</td>
<td>14</td>
<td>...</td>
<td>0.85</td>
<td>...</td>
<td>0.65</td>
</tr>
<tr>
<td>Author3</td>
<td>15</td>
<td>13</td>
<td>13</td>
<td>...</td>
<td>0.75</td>
<td>...</td>
<td>0.45</td>
</tr>
<tr>
<td>Author4</td>
<td>11</td>
<td>12</td>
<td>12</td>
<td>...</td>
<td>0.65</td>
<td>...</td>
<td>0.3</td>
</tr>
<tr>
<td>Author5</td>
<td>13</td>
<td>11</td>
<td>11</td>
<td>...</td>
<td>0.55</td>
<td>...</td>
<td>0.1</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

Patterns are extracted after the propositionalization.
• The higher the number of publications in VLDB and SIGMOD, the higher the centrality of the authors within the graph (Pagerank).
  • \{VLDB^+, SIGMOD^+, PageRank^+\}

• The higher the number of publications in SAC, the lower the number of publications in KDD and ICDM:
  • \{SAC^+, KDD^−, ICDM^−\}

• \{SAC^+, Betweenness^−\}
Identification of the representative vertices of the pattern:

- Who are the authors that best represent the pattern \{VLDB^+, SIGMOD^+, PageRank^+\}?
For each vertex, topological properties are computed:

- based on the direct neighborhood:
  - Degree centrality,
  - Clustering coefficient,
  - …

- based on the whole graph:
  - Betweenness centrality,
  - Closeness centrality,
  - PageRank Index,
  - …
Input: an array!
A raw = local attributes + topological properties for a given vertex.

Goal:
Finding patterns among local attributes and topological properties

How?
• Similar to itemset mining:
  • Efficiently enumerate all subsets of signed attributes,
  • Compute the frequency of the patterns within the data, and
  • Return frequent patterns (w.r.t. a threshold).

Question:
What is the frequency of such patterns? How to compute it?
Kendall's \( \tau \) based support definition

- \( P = \{PODS^+, PageRank^+\} \)
  - \( Supp(P) = \) proportion of the number of vertex pairs \((u, v)\) such that: \( u.PODS < v.PODS \) and \( u.PageRank < v.PageRank \)

- \( P = \{KDD^+, ClusCoef^-\} \)
  - \( Supp(P) = \) proportion of the number of vertex pairs \((u, v)\) such that: \( u.KDD < v.KDD \) and \( u.ClusCoef > v.ClusCoef \)
1 Example

\[ P = \{ KDD^+, ClusCoef^- \} \]

- Supported by 5 pairs.
- Ex: (A2,A1)
- \( \text{supp}_T(P) = \frac{5}{\binom{5}{2}} = \frac{1}{2} \)

\[
\begin{array}{|c|c|c|c|c|}
\hline
\text{ } & \text{KDD} & \text{PKDD} & \text{PODS} & \text{Clus Coef} \\
\hline
\text{A1} & 14 & 14 & 15 & 0.55 \\
\hline
\text{A2} & 12 & 15 & 14 & 0.65 \\
\hline
\text{A3} & 15 & 13 & 13 & 0.45 \\
\hline
\text{A4} & 11 & 12 & 12 & 0.3 \\
\hline
\text{A5} & 13 & 11 & 11 & 0.1 \\
\hline
\end{array}
\]
Other constraints?

$A^+B^+$ and $A^+B^-$ can be both frequent. ⇒ Contradiction !!!

$A^+B^+$ seems more interesting if the support of $A^+$ is significantly larger in the "class" $B^+$

We define emerging patterns in this context

$$Gr(P, C^*) = \frac{\text{Supp}(P \cup C^*)}{\text{Supp}(P \cup C^*)} \cdot \frac{\text{Supp}(C^*)}{\text{Supp}(C^*)}$$

where the attribute $C$ partitions the set of vertex pairs in three classes:

- $C^+ = \{(u, v) \mid (u.C < v.C) \wedge (u \neq v)\}$
- $C^- = \{(u, v) \mid (u.C > v.C) \wedge (u \neq v)\}$
- $C^- = \{(u, v) \mid (u.C = v.C) \wedge (u \neq v)\}$
Patterns correlated with the graph

Only consider pairs that are edges of $G = (V, E)$ in the support counting:

$$C_E = \{(u, v) \in V^2 \mid \{u, v\} \in E\}$$

As before, to evaluate the impact of $E$ on the support of $P$, we consider the growth rate of the support of $P$ over the partition of vertex pairs $\{C_E, C_{\bar{E}}\}$:

$$Gr(P, E) = \frac{\text{Supp}_E(P)}{\text{Supp}_{\bar{E}}(P)}.$$
1 Example

What about $h^+, t^+$?

Growth rate w.r.t. $t$

- $Gr(\{h^+\}, t^+) = 2.13$

Growth-rate w.r.t. the structure

- $Gr(\{h^+, t^+\}, E) = 1.23$
- $Gr(\{h^+, t^-\}, E) = 0.59$
Experimentation on Movie Network

- Netflix and IMDb (1998-2005)
- Users rate movies (1 to 5 stars)
- 5,972 vertices (films)
- 64,338 edges (one common actor)
- 5 local attributes:
  - Release year, num_customers (number of users that rate the movie), avg_rating, stdev_rating, et num_actors
- 9 topological properties
Some Results

\{\text{avg\_rating}^+, \text{num\_customers}^+\}\n
"People tend to rate movies they like"

\{\text{num\_customers}^+, \text{Degree}^+\}\n
"People tend to rate movie with major actors." (e.g., R de Niro, S. Connery, et T. Hanks)

\{\text{stdev\_rating}^+, \text{PageRank}^-\}\n
Controversial movies are isolated
What are the sub-graphs whose vertex attributes evolve similarly?

Mining attributed sub-graphs (Nodes, Times, Trends) that satisfy some constraints on the graph topology and on the attribute values

- The connectivity of the dynamic subgraphs is constrained by a maximum diameter value.
- The vertices (Nodes) evolve similarly (Trends) through time (Times)
Cohesive Co-evolution Patterns

Maximum diameter constraint

Additional interestingness measures to assess the patterns and guide their search by user-parametrized constraints:

- What about the vertices that do no belong to the trend dynamic sub-graph? Do they behave similarly?
- Are trends specific to the vertices of the pattern?
- How is characterized the dynamic? Does the pattern appear suddenly or continuously?
US domestic flights database
All the US domestic flights over the last 20 years. About 370 airports (vertices) described with 8 attributes (number of departures/arrivals, the number of canceled flights, the number of flights whose destination airport has been diverted)

What are the vertices whose vertex attributes evolve similarly?
Top pattern w.r.t. time specificity (in red)

- 71 airports whose arrival delays increase over 3 weeks.
- Arrival delays never increased in these airports during another week.
- The hurricane strongly influenced the domestic flight punctuality.

Top pattern w.r.t. trend specificity (Yellow)

- 5 airports whose number of flights increased over 3 weeks.
- Substitutions flights were provided from these airports during this period.
- This behavior is rather rare in the rest of the graph.
Result on Brazil landslides

Two satellite images taken before and after huge landslides in Brazil (394,885 vertices -- segmented areas, 11 attributes -- colors).

Goal: identify regions in which a landslide appears

- 69% of the true landslide regions identifies.
- 54% of false positive that are either
  - (1) regions nearby true landslides (border effect),
  - (2) deforested area (e.g., human activity),
  - (3) misalignment of the segmentation technique
  - (4) cities and human activity footprints.
• Discovering (temporal) combinations of attribute changes that are strongly correlated with significant structural changes within the graphs.

• The so-called triggering attributes thus bring new insights on the structure evolution

In a social network, whose users are linked if they mutually follow their blogs, we may observe that updating his status more often while giving positive opinions about others and then receiving less negative opinions from the others is often followed by an increase of the user popularity.
A dynamic attributed graph with 6 timestamps entailing the triggering attributes pattern $\langle \{a^+, b^+\}, \{c^-\}, \{\text{deg}^+\} \rangle$. 

SILEX seminar
Triggering Attributes in Dynamic Graphs
30 / 43
Example

In a co-citation network
Publication to Nature and/or science lead to a high increase of the inner degree and the centrality.

In social network
- Like Logic leads to a sparsification of your neighbourhood.
- Example with Lance Armstrong, ...
Two steps:

Identification of the changes on the graph structure and on the vertex attributes

Transformation into a problem of sequence mining guided by a graph measure

- A sequence depicts vertex history
- Coverage measure on an aggregated graph
- Mining covering emerging sequences
- No support measure but we can take advantage of closed-based pruning techniques while pushing the coverage (AM)
4 Temporal dependencies between lasting events

Identifying temporal dependencies between sequence events
• Events last for a period of time
• What is the time lag between dependent events?

Measuring event dependency

Event shift
4 Our Proposal

- **Main idea**: discover temporal dependencies and their associated time-delay intervals
  - robust to the temporal variability of events
  - characterizes the time intervals during which the events are related.
  - without user-determined threshold
4 Measuring the dependency confidence

Confidence of state dependency $A \xrightarrow{[\alpha,\beta]} B$

The proportion of time where the two states are active over the active time period of $A$:

$$\text{conf}(A \xrightarrow{[\alpha,\beta]} B) = \frac{\text{len}(A \cap B^{[\alpha,\beta]})}{\text{len}(A)}$$

Pearson’s chi-squared test of independence

Are the occurrences of $A$ and $B^{[\alpha,\beta]}$ statistically independent?

<table>
<thead>
<tr>
<th></th>
<th>$B^{[\alpha,\beta]}$</th>
<th>$\overline{B}^{[\alpha,\beta]}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$A$</td>
<td>$P(A \cap B^{[\alpha,\beta]})$</td>
<td>$P(A \cap \overline{B}^{[\alpha,\beta]})$</td>
</tr>
<tr>
<td>$\overline{A}$</td>
<td>$P(\overline{A} \cap B^{[\alpha,\beta]})$</td>
<td>$P(\overline{A} \cap \overline{B}^{[\alpha,\beta]})$</td>
</tr>
</tbody>
</table>

$$\chi^2 \geq \chi^2_{0.05}$$
Video surveillance cameras dependencies

Temporal dependencies between lasting events
Video surveillance cameras dependencies
4 Triggering road deicing operations

Organizing the deicing operations on the basis of freezing alert prediction

Temporal dependencies between lasting events
Conclusion

Several methods developed to discover new insights in Dynamic Graphs

• Constraint based pattern mining view...
• ... without support threshold (as much as possible) ...
• in a generic way (DBLP, US Flights, Brazil landslides, social TVs, Velo'V, orphadata, IMDB, etc.)
• Preliminary results are promising

Future work

• Yet designing sophisticated approach to describe graphs.
• Actionnable insights in spatio-temporal data.
Trace mining of Vélov movement

• Vélo'V: the bicycle sharing system of Lyon
• large body of data including all trips over several years, the socio-demographic characteristics of stations and related information of system users

Goal:
• disclose the characteristics of the Vélo’V usages at certain times of the week according to socio-economic characteristics of the station areas.
• the usage characteristics encoded by the graph topology will be considered in conjunction with the values of the attributes of vertices and edges.
Trace mining of urban sensors

Two applications are considered:

• The study of habits related to urban mobility
• The analysis of the degradation of performance of PV systems

Goal:

• identification of recurrent phenomena
• or exceptional events
Trace mining of social media

Exploration of geo-tagged and contextualized traces of users of social media
Interactive Knowledge Discovery

Thèse de Damien Cram