Dependencies and Formal Concept Analysis

Jaume Baixeries\textsuperscript{1}, Mehdi Kaytoue\textsuperscript{2}, and Amedeo Napoli\textsuperscript{3}

\textsuperscript{1} Universitat Politècnica de Catalunya, 08032, Barcelona, Catalonia (jbaixer@lsi.upc.edu).

\textsuperscript{2} Université de Lyon (CNRS, INSA-Lyon, LIRIS), F-69621 Villeurbanne, France (mehdi.kaytoue@insa-lyon.fr).

\textsuperscript{3} LORIA (CNRS – INRIA Nancy Grand-Est – Université de Lorraine), B.P. 239, 54506 Vandoeuvre les Nancy, France (amedeo.napoli@loria.fr).

17 de novembre de 2014
Outline of This Talk

1. Dependencies.
2. Theoretical Aspects.
3. Computational Aspects.
Our Research

What we do is:

- Data Mining and Knowledge Discovery (we are finding patterns).
- Big Data (we are dealing with large amounts of data).
Our Research

How We Do It:

- We use different kinds of dependencies to analyze data.
- We use different techniques to compute them.
From here, we have two different (yet complementary) perspectives

- A *theoretical* perspective: formalize the relationship between these dependencies.
- A *computational* perspective: algorithms and data structures to compute these dependencies.
A functional dependency (FD) $X \rightarrow Y$ holds in $T$ if:

$$\forall t_i, t_j \in T : t_i(X) = t_j(X) \Rightarrow t_i(Y) = t_j(Y)$$

For example, the functional dependencies $a \rightarrow d$ and $d \rightarrow a$ hold whereas $a \rightarrow c$ does not hold since $t_2(a) = t_4(a)$ but $t_2(c) \neq t_4(c)$.
More Dependencies

In a table, there may be some tuples that prevent a functional dependency from holding.

Removing such tuples allows the dependency to exist: a threshold can be set to define a set of approximate dependencies holding in a table.

The dependency $Month \rightarrow Av.\ Temp$ holds if 6 tuples are removed: e.g. keeping only one tuple for $Month\ 1$ and one tuple for $Month\ 5$.

Then, if the threshold $\geq 75\%$ then $Month \rightarrow Av.\ Temp$ is a valid “approximate dependency”.

<table>
<thead>
<tr>
<th>id</th>
<th>Month</th>
<th>Year</th>
<th>Av. Temp.</th>
<th>City</th>
</tr>
</thead>
<tbody>
<tr>
<td>$t_1$</td>
<td>1</td>
<td>1995</td>
<td>36.4</td>
<td>Milan</td>
</tr>
<tr>
<td>$t_2$</td>
<td>1</td>
<td>1996</td>
<td>33.8</td>
<td>Milan</td>
</tr>
<tr>
<td>$t_3$</td>
<td>5</td>
<td>1996</td>
<td>63.1</td>
<td>Rome</td>
</tr>
<tr>
<td>$t_4$</td>
<td>5</td>
<td>1997</td>
<td>59.6</td>
<td>Rome</td>
</tr>
<tr>
<td>$t_5$</td>
<td>1</td>
<td>1998</td>
<td>41.4</td>
<td>Dallas</td>
</tr>
<tr>
<td>$t_6$</td>
<td>1</td>
<td>1999</td>
<td>46.8</td>
<td>Dallas</td>
</tr>
<tr>
<td>$t_7$</td>
<td>5</td>
<td>1996</td>
<td>84.5</td>
<td>Houston</td>
</tr>
<tr>
<td>$t_8$</td>
<td>5</td>
<td>1998</td>
<td>80.2</td>
<td>Houston</td>
</tr>
</tbody>
</table>
More Dependencies

- There is still a more sophisticated kind of Functional Dependencies: **Impurity Dependencies**.
- These dependencies rely on an *impurity measure* that describes how bad a functional dependency holds.
- They seem to generalize approximate dependencies.
Instead of considering measures that deal with the sets of tuples as a whole, dependencies could be directly related with the notion of “similarity”.

If two tuples have “similar values” for Month and City, then they should have a “similar value” for Av.Temp.

If two cities are “close enough” and two months are also “close enough”, then the average temperature in the cities should be “close enough” or “similar” as well.

This leads to the notion of Similarity Dependencies.
More Dependencies

- There is a more general kind of Similarity Dependencies: **Matching Dependencies**.
- They not only take into account a similarity measure for each attribute, but also, the distribution of values among the values of each attribute.
There are different kinds of dependencies that seem to be related:

- Functional Dependencies.
- Approximate Dependencies.
- Purity Dependencies.
- Similarity Dependencies.
- Matching Dependencies.
Dependencies

Why are we interested in Dependencies?

- Because they provide useful information in table datasets.
- They are easy to interpret and work with.
- They can have flexible semantics.
- They fit in various frameworks: the Relational Database Model, Formal Concept Analysis, Formal Logics, ...
TASK (1): Theoretical Aspects of Dependencies

- Approximate/Impurity/Similarity/Matching Dependencies generalize Functional Dependencies.
- It seems that Impurity Dependencies generalize Approximate Dependencies (?).
- What about the relationship between Similarity/Matching Dependencies and the rest?
TASK (2): Computational Aspects of Dependencies

- FD/Approximate/Impurity/Similarity/Matching Dependencies can be computed using the TANE algorithms.
- FD/Similarity Dependencies can be computed using FCA and pattern structures.
- We need to compare the performance of TANE vs FCA and Pattern Structures!
- We need to study the performance of FCA and pattern structures using different datasets.
Dependencies

This is an ongoing research!


If you are interested, please:

1. Google:
   - Jaume Baixeries dblp.
   - Mehdi Kaytoue dblp.
   - Amedeo Napoli dblp.

2. Send me an e-mail to jbaixer@cs.upc.edu
Questions, comments and discussions are welcome!
Thanks!!

Merci!