Knowledge Discovery over Complex Data
Applications in Pharamcogenomics

Adrien Coulet[1,2] and Amedeo Napoli[2]

[1] Inria Paris

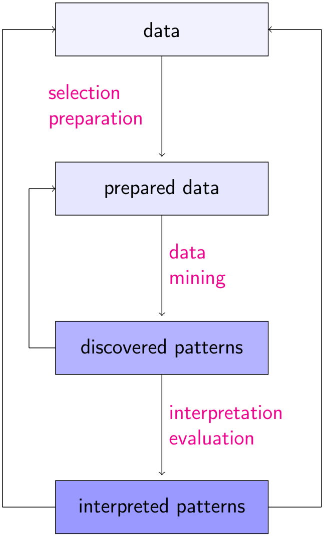
[2] Orpailleur Team, Université de Lorraine, CNRS, Inria, LORIA, 54000 Nancy, France

{Adrien.Coulet,Amedeo.Napoli}@loria.fr

Journée Science Ouverte
Université de Lorraine
8 octobre 2020

Knowledge Discovery in Databases (KDD)

- Knowledge Discovery in Databases (KDD) consists in processing large volumes of data in order to discover "patterns" that are significant, useful, and reusable.
- KDD relies on three main steps: data preparation, data mining, and pattern interpretation.
- KDD is iterative and interactive as it can be replayed and guided by an analyst.



Research Tracks about KDD in the Orpailleur Team

- Knowledge Discovery:
  - pattern mining, rule mining, Formal Concept Analysis (FCA) and extensions, dependencies (functional, approximate)
  - mining complex data: sequences, trees, graphs, linked data, time series...
  - meta-mining: preference and constraint management in mining, dimensionality reduction, production of explanations, fairness of algorithms
  - combining numerical and symbolic data mining methods
  - visualization
- Knowledge Discovery and Knowledge Engineering:
  - mining for ontology engineering, text mining
  - knowledge mining, discovery of link (keys) in linked data
  - mining and decision theory
- Application domains: agronomy, astronomy, biology, chemistry, medicine...

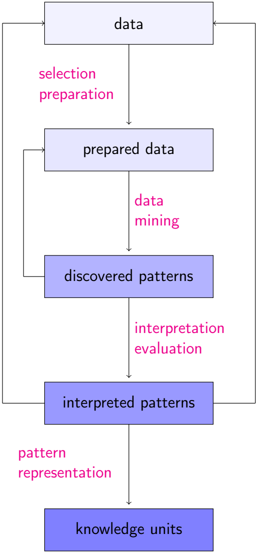
Dimensions in KDD

- A formulation of the KDD problem by Mannila et al.:
- Given a database DB, a finite language L of patterns, an interestingness predicate Q, the mining task amounts to discover a set of patterns α such that: $\{\alpha \in L^* | Q(DB, \alpha) \text{ holds}\}$.

- The data dimension: a database DB and a language L of patterns.
- The knowledge dimension: an interestingness predicate Q.
- The control dimension: find a "mining strategy" for searching the pattern space and discover the "most interesting" patterns.

The Knowledge Dimension in KDD

- Data have a context and KDD is knowledge oriented, depending on domain knowledge, e.g. constraints, preferences...
- At each step, domain knowledge can be embedded to guide KDD, e.g. interestingness measures, preferences...
- The knowledge dimension involves interpretation and the production of actionable knowledge (knowledge construction).



Knowledge Discovery and Knowledge Representation

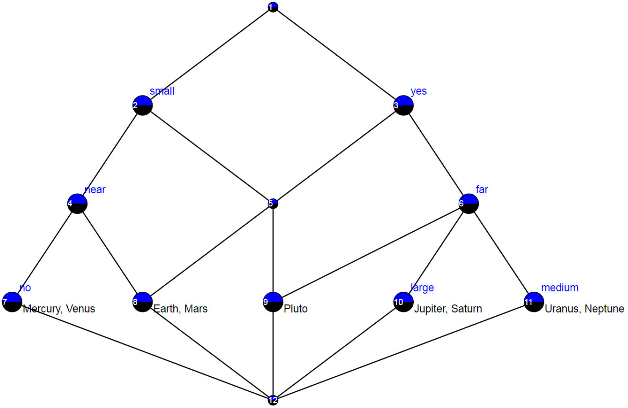
- Knowledge Discovery and Knowledge Engineering are complementary.
- A parallel can be drawn with the "Knowledge Level" (Newell):
- Three Levels: data, information, and knowledge.
- A main idea underlying declarative knowledge representation and reasoning can be reused in KDD, i.e. Describe the problem and the solver will take care of the solution.

Exploratory Knowledge discovery based on FCA

- Formal Concept Analysis (FCA) is a mathematical formalism based on lattice theory, classification and concept discovery providing a generic framework for KDD.
- Moreover, FCA follows a human centered approach and supports exploration operations through the concept lattice.
  - Discovery of concepts, i.e. classes of individuals with a description.
  - Organization of concepts into a poset based on a subsumption relation.
  - The poset supports exploration, e.g. information retrieval, visualization...
- FCA can be a "Discovery Engine for Exploratory KDD" provided that data are not too big, but Small is Beautiful...

Planets	Size			Distance to Sun		Moon(s)	
	small	medium	large	near	far	yes	no
Jupiter			x		x	x	
Mars	x			x		x	
Mercury	x			x			x
Neptune		x			x	x	
Pluto	x				x	x	
Saturn			x		x	x	
Earth	x			x		x	
Uranus		x			x	x	
Venus	x			x			x

- Exploration and Visualization (LatViz)
- Navigation and Information Retrieval
- Interpretation of concepts and rules



- Rules: "far $\longrightarrow$ medium" (confidence 2/5), "small $\longrightarrow$ near" (confidence 4/5).
- Implications: "no $\Longrightarrow$ near" and "near $\Longrightarrow$ small" (confidence 1).

Mining Definitions in the Web of Data

- DBpedia is the largest reservoir of Linked Data with more than 6 million entities and 9.5 billion RDF triples.
- The content of DBpedia is obtained from semi-structured sources of information in Wikipedia, namely infoboxes and categories.
- In Wikipedia, infoboxes are used to standardize entries of a given type. Categories are another important tool used to –manually– organize information.
- Can we use categorical information in DBpedia as a "definition of a class of documents", as it could be expected if DBpedia was an ontology?

- Mehwish Alam, Aleksey Buzmakov, Víctor Codocedo and Amedeo Napoli. Mining Definitions from RDF Annotations Using Formal Concept Analysis, in Proceedings of IJCAI 2015 (Buenos Aires, Argentina), AAAI Press, pages 823–829, 2015.
- Mehwish Alam, Aleksey Buzmakov and Amedeo Napoli. Exploratory Knowledge Discovery over Web of Data, Discrete Applied Mathematics, 249:2–17, 2018.

Discovery of Definitions in RDF data

- For being significant for a software agent, information should be expressed through definitions.
- Accordingly, we propose a formalism relating the syntactic nature of categorical annotations with a semantic counterpart, yielding a concept definition.
- Given a set of RDF data of interest, a concept lattice is built after a suitable transformation of the data.
- Then, mining implications provides a basis for "subject definitions" in terms of necessary and sufficient conditions.
- If $X \Longrightarrow Y$ and $Y \Longrightarrow X$, then $X \equiv Y$ is a definition.
- If $X \Longrightarrow Y$ and $Y \rightarrow X$ has a high confidence, then $X \cong Y$ is a quasi-definition and can be interpreted as a marker of "data incompleteness".
- An interaction with an analyst is used to check whether a quasi-definition should or not be completed into a definition.

RDF triples

```
<Person1,dc:subject,dbpc:Computer_Scientists>
<Person1,dc:subject,dbpc:Turing_Award_Laureates>
<Person1,dbp:field,dbp:Computer Sciences>
<Person1,rdf:type,dbo:Scientists>
...
```

Predicates		Objects	
Index	URI	Index	URI
A	dc:subject	a	dbpc:Computer_Scientists
		b	dbpc:Turing_Award_Laureates
B	dbp:award	c	dbp:TuringAward
C	rdf:type	d	dbo:Scientist
D	dbp:field	e	dbp:Computer Sciences
E	dbp:birthPlace	f	dbo:UnitedStates
		g	dbo:UnitedKingdom

	A		B	C	D		E
	a	b	c	d	e	f	g
Person1	×	×	×	×	×	×	
Person2	×	×	×	×	×		
Person3	×	×	×	×			×
Person4	×	×	×	×			
Person5	×	×	×	×			
Person6	×	×					
Person7	×	×					

- $c, d \Rightarrow a, b$ but $conf(\{a, b\} \rightarrow \{c, d\}) = 0.71$
- A definition may exist provided that data are completed: $a, b \equiv c, d$ i.e., $a, b \Longrightarrow c, d$ and $c, d \Longrightarrow a, b$