Analyse Formelle de Concepts

Module Ingénierie des Connaissances

Université De Montpellier - Faculté Des Sciences

5 mars 2015

Sommaire

1 Fondements

2 Données complexes

Analyse formelle de concepts (AFC)

- Méthode d'analyse de données
- Application de la théorie des treillis
- Extraction de concepts *Unité de base de la pensée humaine*
- Contextes restreints (monde clos)

Analyse formelle de concepts (AFC)

Applications

- Construction de classifications
- Recherche d'informations (indexation)
- Fouille de données (règles)
- Apprentissage
- Aide à la construction d'ontologies : alignement, explicitation et classification des concepts, etc.

Les planètes du système solaire

Objets étudiés : planètes Attributs : caractéristiques des planètes

	Taille			Dist soleil		Satellite	
	petite	moyenne	grande	proche	loin	oui	non
Mercure	x			x			x
Vénus	x			x			x
Terre	x			x		x	
Mars	x			x		x	
Jupiter			x		x	x	
Saturne			x		x	x	
Uranus		x			x	x	
Neptune		x			x	x	
Pluton	x				x	x	

Contexte binaire et ses applications caractéristiques

Contexte (O, A, R)

- deux ensembles finis O et A
- une relation binaire $R \subseteq 0 \times A$.

Définition (applications caractéristiques d'une relation binaire)

Attributs communs à un ensemble d'objets

$f : \mathcal{P}(O) \to \mathcal{P}(A)$

$X \longmapsto f(X) = \{y \in A \mid \forall x \in X, (x, y) \in \mathcal{R}\} = X'$

Objets partageant un ensemble d'attributs

$g : \mathcal{P}(A) \to \mathcal{P}(O)$

$Y \longmapsto g(Y) = \{x \in O \mid \forall y \in Y, (x, y) \in \mathcal{R}\} = Y'$

Autre notation $'$

Contexte binaire et ses applications caractéristiques

$$f(\{Mercure, Terre\}) = \{Taille : petite, Soleil : proche\}$$

$$g(\{Taille : petite, Soleil : proche\}) = \{Mercure, Venus, Terre, Mars\}$$

Concept

Un concept formel C est un couple (E, I) tel que
$f(E) = I$ (ou de manière équivalente) $E = g(I)$

$$E = \{ \text{e} \in \text{O} \mid \forall \text{ i} \in \text{I}, (\text{e}, \text{i}) \in \text{R}\}$$
est *l'extension* (objets couverts),

$$I = \{ \text{i} \in \text{A} \mid \forall \text{ e} \in \text{E}, (\text{e}, \text{i}) \in \text{R}\}$$
est *l'intension* (caractéristiques partagées).

Concept

	Taille			Dist soleil		Satellite	
	petite	moyenne	grande	proche	loin	oui	non
Mercure	x			x			x
Vénus	x			x			x
Terre	x			x		x	
Mars	x			x		x	
Jupiter			x		x	x	
Saturne			x		x	x	
Uranus		x			x	x	
Neptune		x			x	x	
Pluton	x				x	x	

Exemple

Le concept (E, I) des *petites planètes proches du soleil*

$E = \{Mercure, Venus, Terre, Mars\}$,

$I = \{Taille : petite, Soleil : proche\}$

Un concept est un ensemble maximal d'objets possédant un ensemble maximal d'attributs

Concept

	Taille			Dist soleil		Satellite	
	petite	moyenne	grande	proche	loin	oui	non
Mercure	x			x			x
Vénus	x			x			x
Terre	x			x		x	
Mars	x			x		x	
Jupiter			x		x	x	
Saturne			x		x	x	
Uranus		x			x	x	
Neptune		x			x	x	
Pluton	x				x	x	

Contre-exemple

$$(E = \{Mercure, Terre\}, I = \{Taille : petite, Soleil : proche\})$$

$$E \neq g(I) = \{Mercure, Venus, Terre, Mars\}$$

Un concept est un ensemble maximal d'objets possédant un ensemble maximal d'attributs

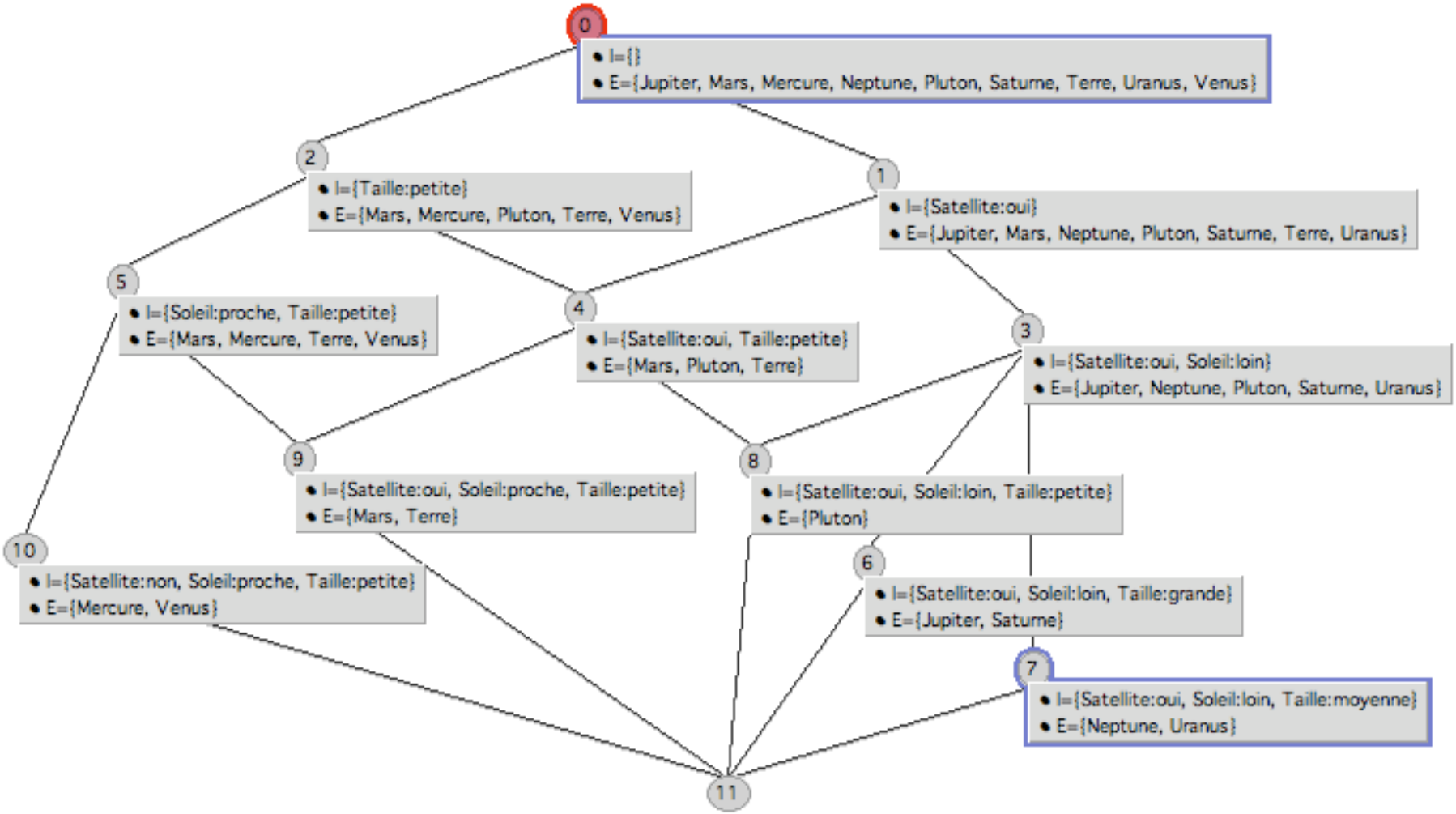
Spécialisation entre concepts

L'ensemble de tous les concepts $\mathcal{C}$ forme un treillis $\mathcal{L}$ lorsqu'il est muni de l'ordre suivant :

$$(E_1, I_1) \leq_{\mathcal{L}} (E_2, I_2) \Leftrightarrow E_1 \subseteq E_2$$
(or de manière équivalente $I_2 \subseteq I_1$).

Le concept des *petites* planètes *proches* du soleil
$E_1 = \{Mercure, Venus, Terre, Mars\}$,
$I_1 = \{Taille : petite, Soleil : proche\}$
spécialise
Le concept des *petites* planètes
$E_2 = \{Mercure, Venus, Terre, Mars, Pluton\}$,
$I_2 = \{Taille : petite\}$

Classification des planètes - Treillis



Galicia : http ://www.iro.umontreal.ca/˜ galicia/

Simplification des étiquettes dans le treillis

La simplification est basée sur l'ordre :

$$(E_1, I_1) \leq_{\mathcal{L}} (E_2, I_2) \Leftrightarrow E_1 \subseteq E_2$$
(ou de manière équivalente $I_2 \subseteq I_1$).

Dans I_1 les attributs de I_2 sont hérités (en descendant),
donc peuvent être retirés

Dans E_2 les objets sont hérités (en montant),
donc peuvent être retirés

Simplification des étiquettes dans le treillis

Dans I_1 les attributs de I_2 hérités (en descendant) sont retirés
Dans E_2 les objets hérités (en montant) sont retirés

Intension simplifiée du concept des *petites* planètes *proches* du soleil :
$I_1 = \{Soleil : proche\}$
car l'attribut *Taille* : *petite* est hérité (concept 2)

Extension simplifiée du concept des planètes *proches* du soleil :
$E_2 = \emptyset$
car les objets sont hérités de concepts plus spécialisés
(concepts 8, 9 et 10)

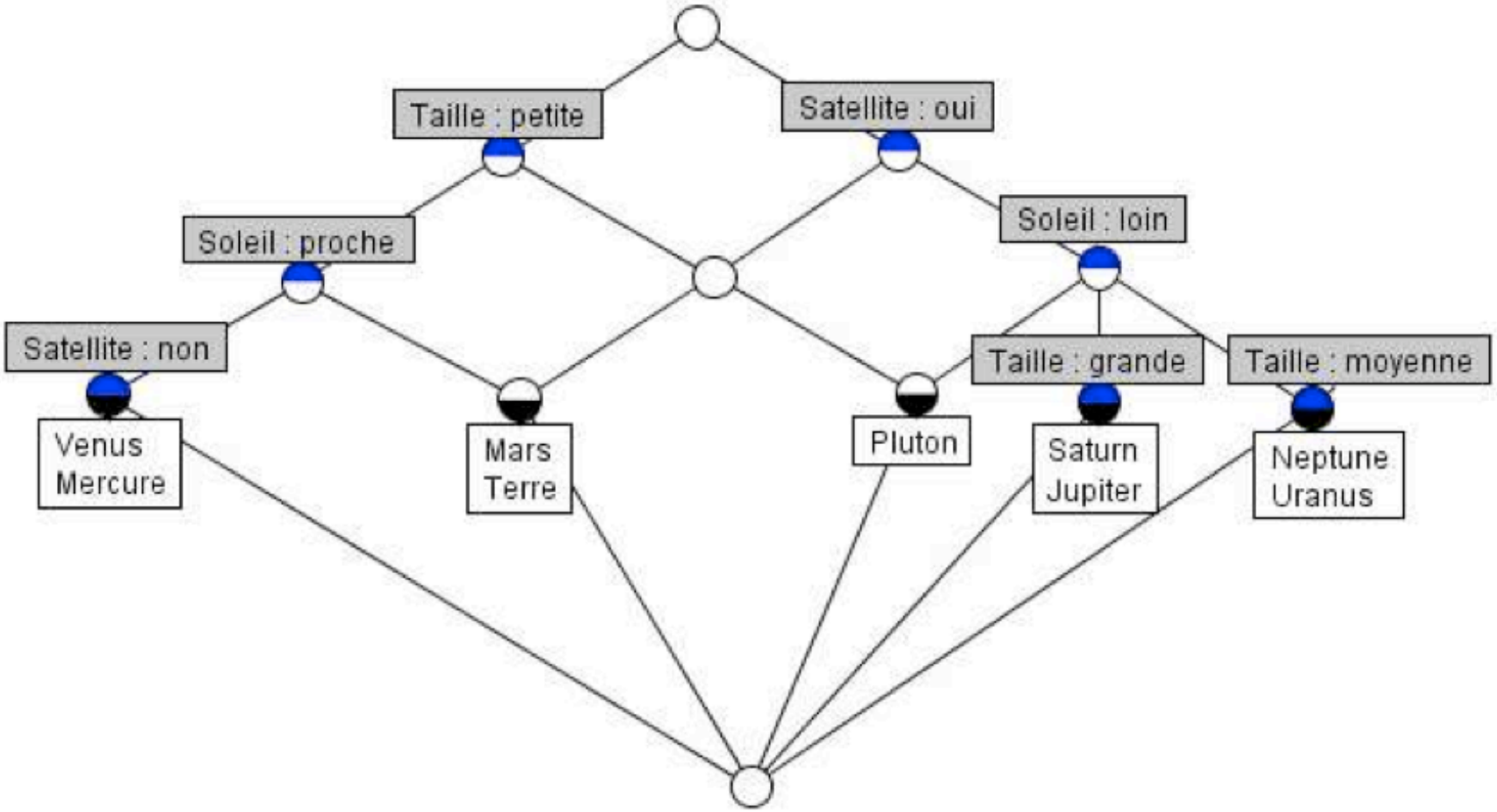
Simplification des étiquettes dans le treillis

Le concept 5 *déclare* ou *introduit* l'attribut *Soleil* : *proche*
C'est un *concept-attribut*

Le concept 8 *déclare* ou *introduit* l'objet *Pluton*
C'est un *concept-objet*

Le concept 7 est un *concept-attribut* et un *concept-objet*
Il introduit *Taille* : *moyenne* et *Neptune*

Classification des planètes - Treillis, étiquettes simplifiées



Conexp : http ://conexp.sourceforge.net/index.html

Sous-hiérarchie de Galois / AOC-poset : Simplification du treillis

Retrait des concepts qui ne sont :

- ni *concept-attribut*
- ni *concept-objet*

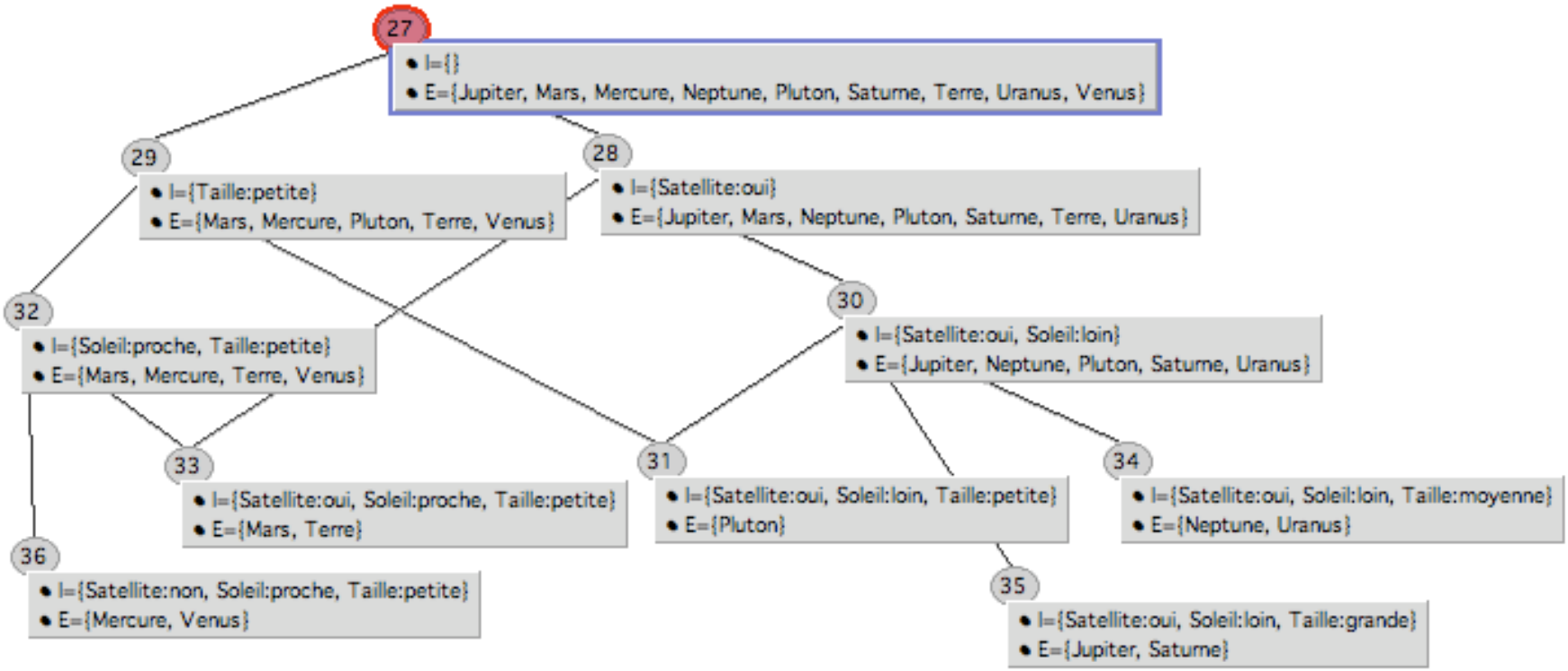
Exemple

concept 4

$\{\{Mars, Pluton, Terre\}, \{Satellite : oui, Taille : petite\}\}$

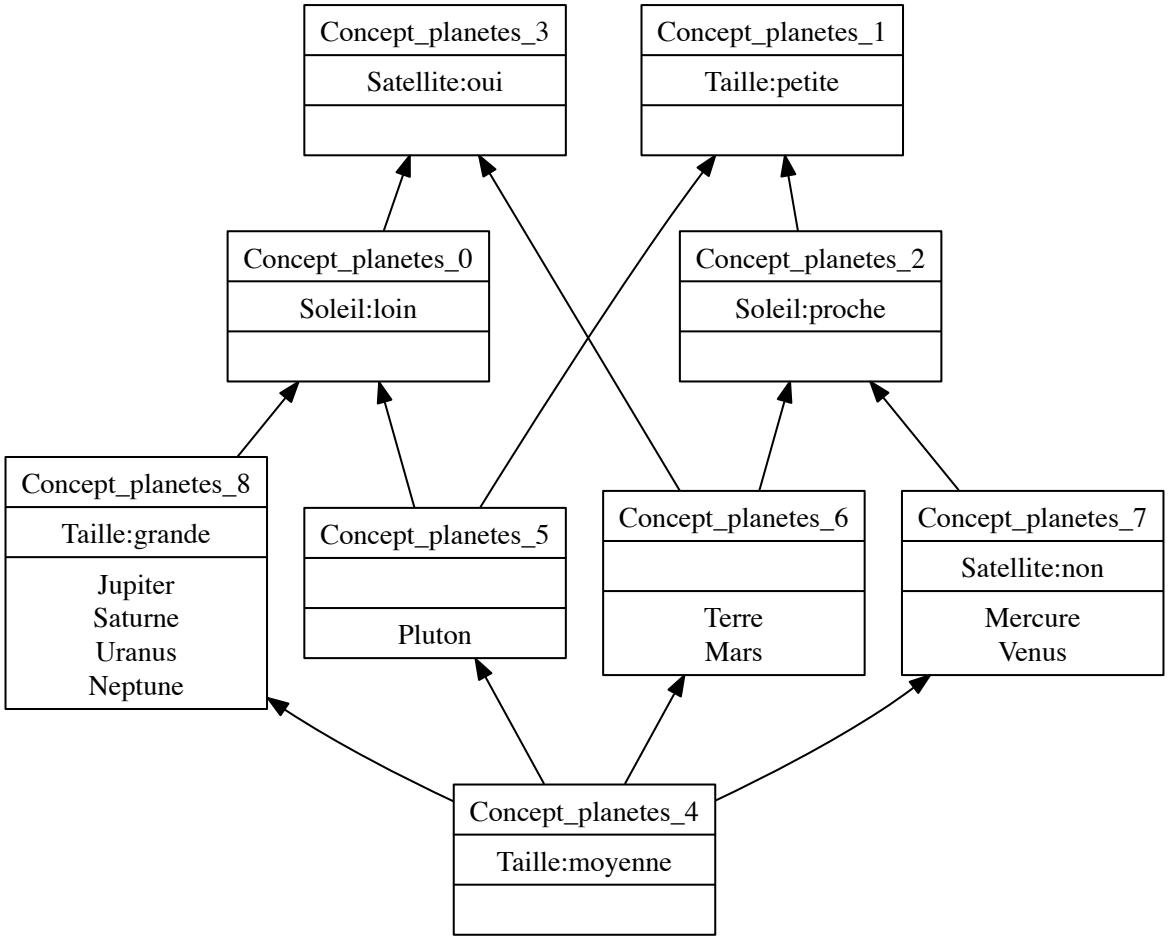
La structure obtenue n'est plus un treillis, mais un ordre partiel quelconque

Classification des planètes - AOC-poset



Galicia : http ://www.iro.umontreal.ca/˜ galicia/

Classification des planètes - AOC-poset



AOC-poset builder : http ://www.lirmm.fr/AOC-poset-Builder/

Sous-hiérarchie de Galois / AOC-poset : Simplification du treillis

La structure obtenue n'est plus un treillis,
mais un ordre partiel quelconque

Les concepts 31 et 33 ont deux plus petits majorants incomparables

concepts 29 et 28

Règles d'implication d'un contexte formel

A_1 et A_2 deux ensembles d'attributs

$A_1 \Rightarrow A_2$ ssi $A_1' \subseteq A_2'$

(ou encore $f(A_1) \subseteq f(A_2)$)

une planète sans satellite est de petite taille et proche du soleil

$\{Satellite : non\} \Rightarrow \{Soleil : proche, Taille : petite\}$

$$f(\{Satellite : non\}) = \{Mercure, Venus\}$$
$$\subseteq$$
$$f(\{Soleil : proche, Taille : petite\}) = \{Mars, Mercure, Terre, Venus\}$$

Règles d'implication d'un contexte formel

$A_1 \Rightarrow A_2$ se ramène à :
$A_1 \Rightarrow a, \forall a \in A_2$

une planète sans satellite est de petite taille et proche du soleil
$\{Satellite : non\} \Rightarrow \{Soleil : proche, Taille : petite\}$
se ramène à

$\{Satellite : non\} \Rightarrow \{Taille : petite\}$
et
$\{Satellite : non\} \Rightarrow \{Soleil : proche\}$

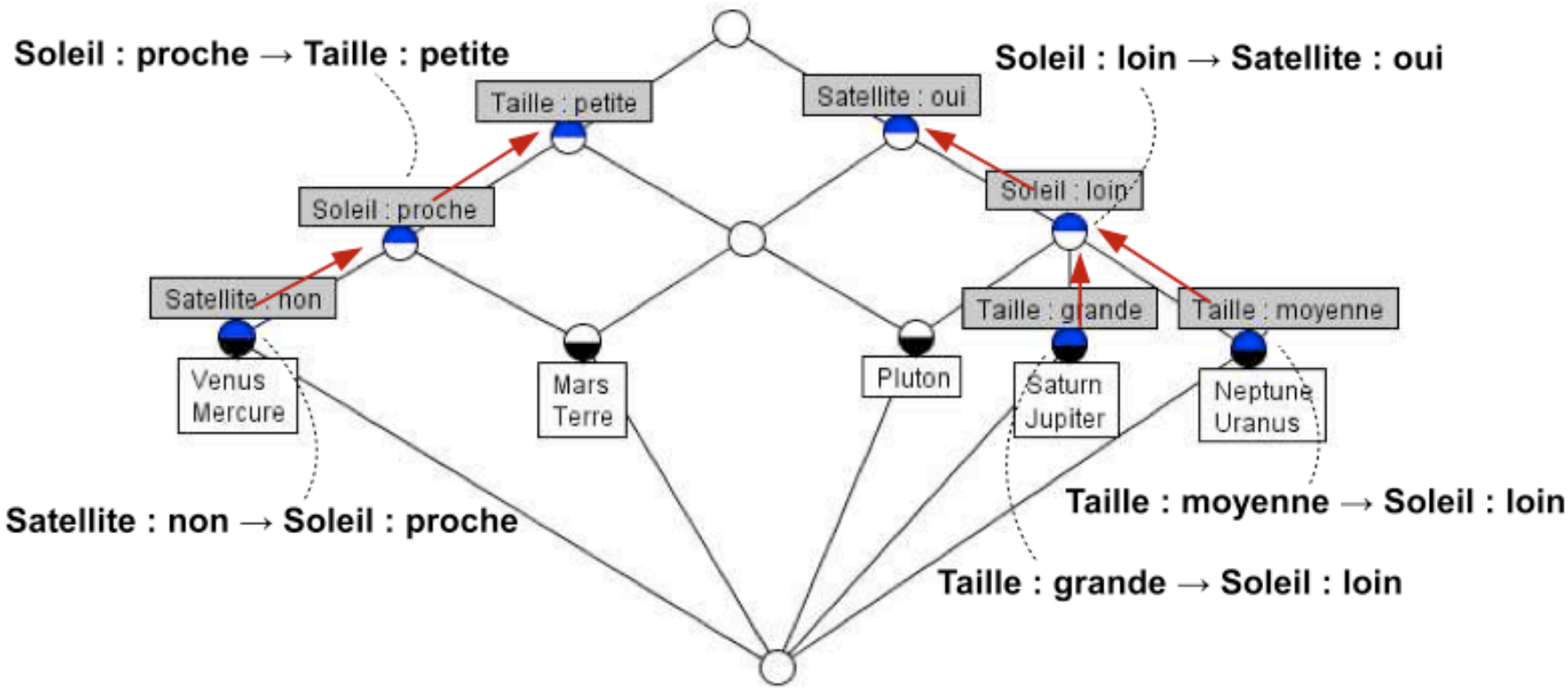
Règles d'implication d'un contexte formel

Certains ensembles de règles sont redondants

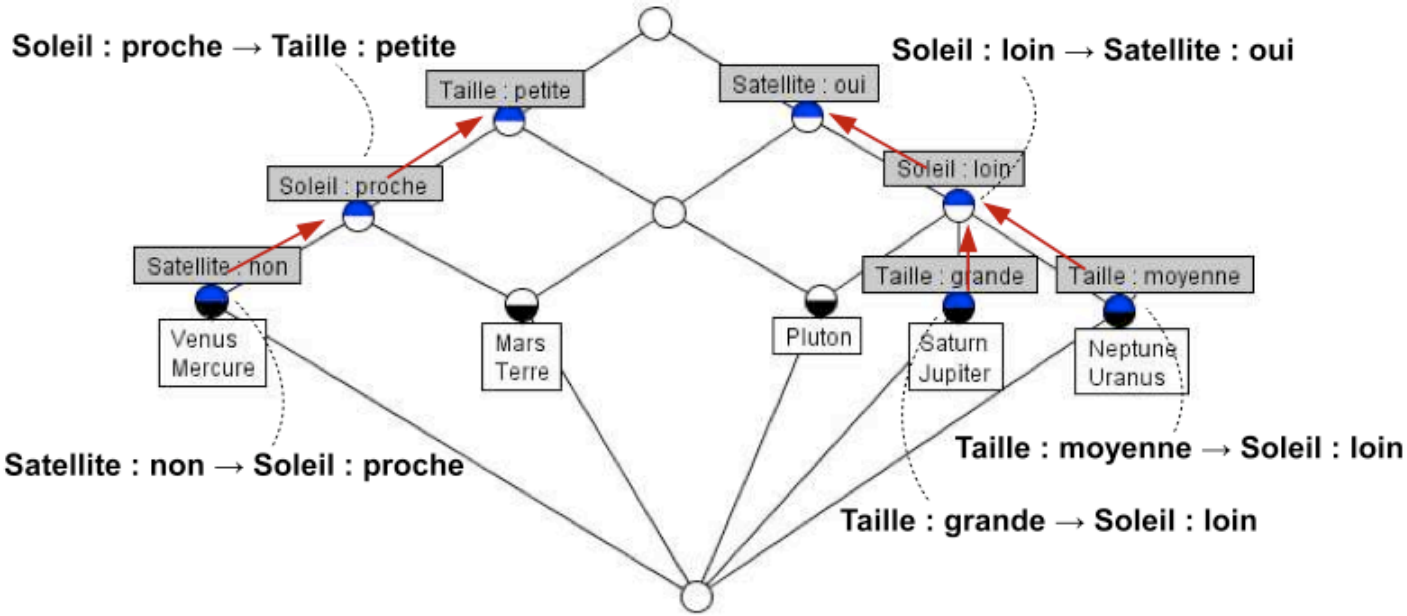
$$\{Satellite : non\} \Rightarrow \{Soleil : proche, Taille : petite\}$$
$$\{Satellite : non\} \Rightarrow \{Taille : petite\}$$
$$\{Satellite : non\} \Rightarrow \{Soleil : proche\}$$
$$\{Soleil : proche\} \Rightarrow \{Taille : petite\}$$

On peut déduire de la réduction transitive du treillis l'ensemble minimal non redondant des implications du contexte qui ont un support non nul (il existe des objets vérifiant l'implication).

Règles d'implications sur les planètes



Règles d'implications sur les planètes



$$\{Satellite : non\} \Rightarrow \{Soleil : proche\}$$
$$\{Soleil : proche\} \Rightarrow \{Taille : petite\}$$
$$\{Taille : grande\} \Rightarrow \{Soleil : loin\}$$
$$\{Taille : moyenne\} \Rightarrow \{Soleil : loin\}$$
$$\{Soleil : loin\} \Rightarrow \{Satellite : oui\}$$

Sommaire

1 Fondements

2 Données complexes

Modélisation de données complexes

Sans prétendre à l'exhaustivité :

- Attributs valués (entier, réels, termes, structures, objets symboliques, etc.) (Ganter et Wille, Polaillon, ...)
- Description floue (Yahia et al., Belohlavek, ...)
- Hiérarchisation des valeurs (Godin et al., Carpineto et Romano, ...)
- Relations entre objets (Priss, Rouane et al., ...)
- Description logique (Chaudron et al., Ferré et al., ...)
- Graphes (Liquière, Prediger et Wille, ...)
- Pattern Structures (Kuznetsov)
- ....

Contextes multi-valués

Définition

Un contexte multi-valué est un quadruplet $K = (O, A, V, J)$ où O et A sont les ensembles d'objets et d'attributs respectivement et $J \subseteq O \times A \times V$ représente la valuation des attributs. $(o, a, v) \in J$ signifie que l'objet o a la valeur v pour l'attribut a.

	Taille	Dist soleil	Satellite
	(km)	(10^6 km)	(nombre)
Mercure	4 878	58	0
Vénus	12 400	108	0
Terre	12 756	150	1
Mars	6 800	228	2
Jupiter	142 800	778	16
Saturne	120 800	1 427	19
Uranus	47 600	2 870	5
Neptune	44 600	4 500	8
Pluton	2 320	9 950	1

Echelonnage (échantillonage) conceptuel

Transformation du contexte K en contexte binaire K^d

- transformer chaque attribut multivalué en plusieurs attributs binaires
- la partie du contexte correspondante est un contexte d'échelle (noté K_a), qui donne naissance à un treillis ($\mathcal{L}_a$)
- réassembler les contextes d'échelle et leurs treillis pour former K^d

Echelle conceptuelle

Les échelles représentent des descriptions séparées des attributs

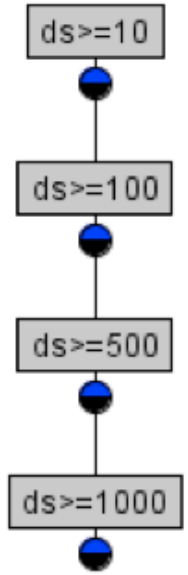
Définition (interprétation d'un contexte d'échelle)

Soit un contexte multi-valué $K = (O, A, V, J)$ et $a \in A$. Le contexte d'échelle de a, noté $K_a = (V_a, P_a, J_a)$, s'interprète :

- $V_a \subseteq V$ est l'ensemble des valeurs de a
- P_a contient un ensemble de propriétés des valeurs de a
- $J_a \subseteq V_a \times P_a \times V_a$ associe à une valeur ses propriétés

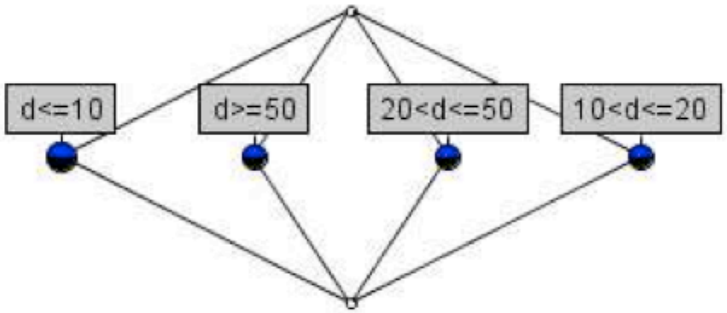
Echelle conceptuelle de l'attribut distance au soleil

	$ds \geq 10$	$ds \geq 100$	$ds \geq 500$	$ds \geq 1000$
58	x			
108	x	x		
150	x	x		
228	x	x		
778	x	x	x	
1 427	x	x	x	x
2 870	x	x	x	x
4 500	x	x	x	x
9 950	x	x	x	x



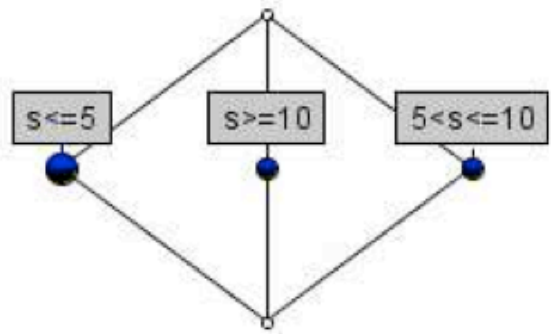
Echelle conceptuelle de l'attribut diamètre

	$d \leq 10$	$10 < d \leq 20$	$20 < d \leq 50$	$d \geq 50$
4 878	x			
12 400		x		
12 756		x		
6 800	x			
142 800				x
120 800				x
47 600			x	
44 600			x	
2.320	x			



Echelle conceptuelle de l'attribut Satellite

	$s \leq 5$	$5 < s \leq 10$	$s \geq 10$
0	x		
0	x		
1	x		
2	x		
16			x
19			x
5	x		
8		x	
1	x		



Calcul de K^d

Définition

Le scaling du contexte K produit le contexte binaire K^d tel que :

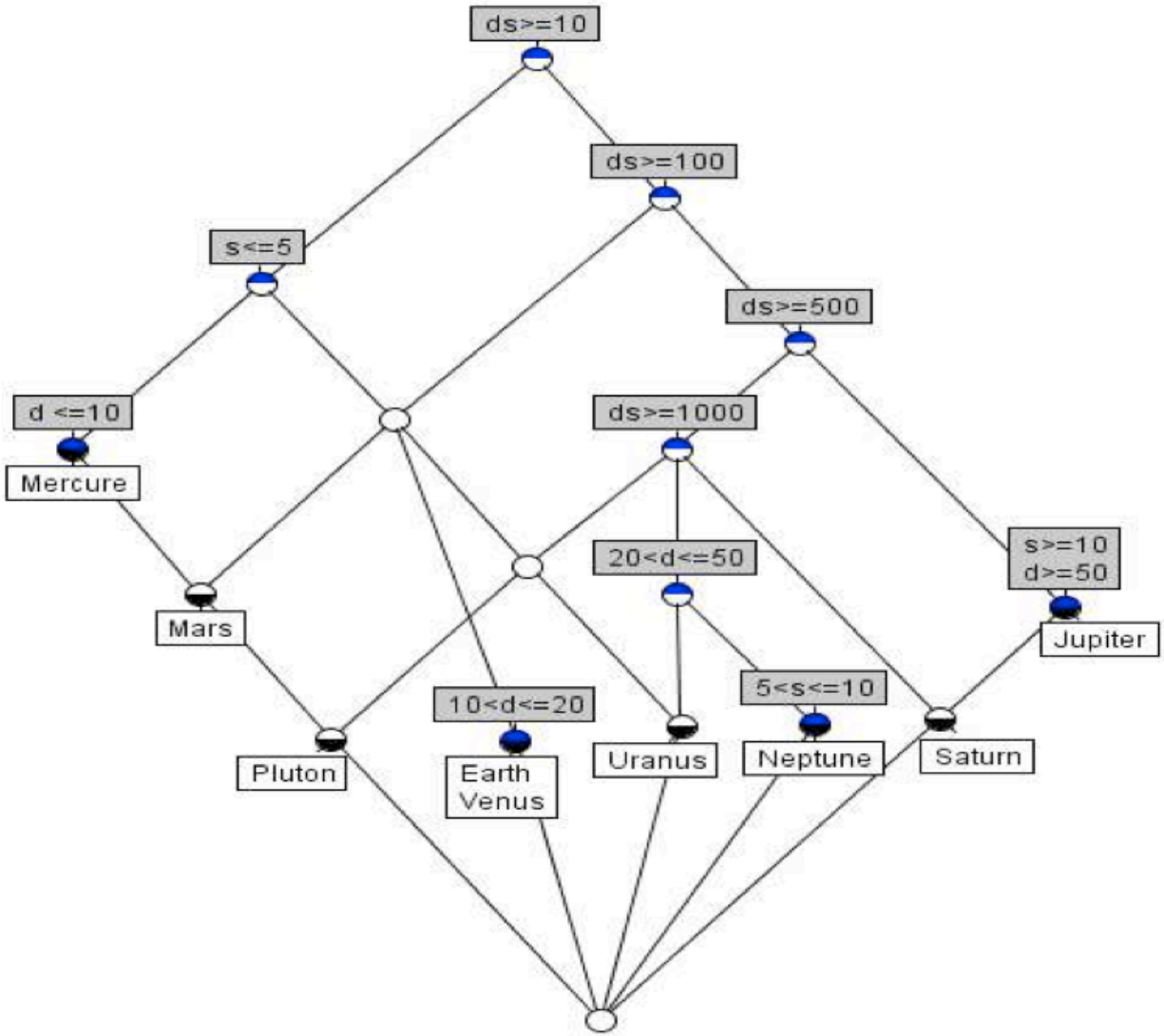
- $O^d = O$
- $A^d = \cup_{a \in A}$ {concepts du treillis $\mathcal{L}_a$}
- $J^d = \{(o, a_s) | (o, a, v) \in J \ et \ v \in extent(a_s)\}$

Nota : on pourrait utiliser directement les descriptions des valeurs mais dans certains cas, de nouveaux concepts apparaissent

Contexte après échantillonage

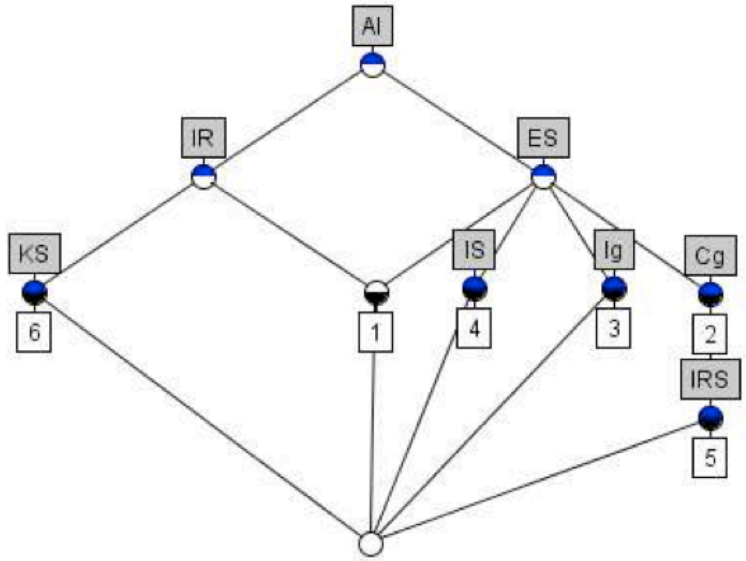
	Diamètre (km)				Distance au Soleil (10^6km)				Satellite		
	$d \leq 10$	$10 < d \leq 20$	$20 < d \leq 50$	$d \geq 50$	$ds \geq 10$	$ds \geq 100$	$ds \geq 500$	$ds \geq 1000$	$s \leq 5$	$5 < s \leq 10$	$s \geq 10$
Mercure	x				x				x		
Venus		x			x	x			x		
Terre		x			x	x			x		
Mars	x				x	x			x		
Jupiter				x	x	x	x				x
Saturne				x	x	x	x	x			x
Uranus			x		x	x	x	x	x		
Neptune			x		x	x	x	x		x	
Pluton	x				x	x	x	x	x		

Treillis après échantillonage



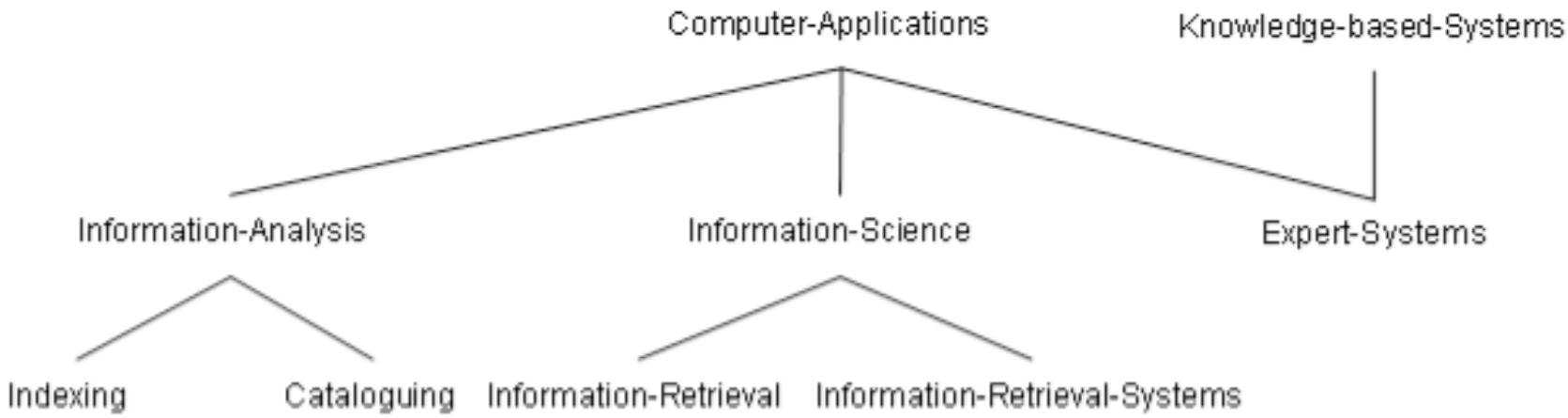
Exemple de données complexes : Contexte avec taxonomies sur les attributs

	Artificial Intelligence (AI)	Expert Systems (ES)	Information Retrieval (IR)	Cataloguing (Cg)	Indexing (Ig)	Information Science (IS)	Information Retrieval Systems (IRS)	Knowledge-based Systems (KS)
1	x	x	x					
2	x	x		x				
3	x	x			x			
4	x	x				x		
5	x	x		x			x	
6	x		x					x



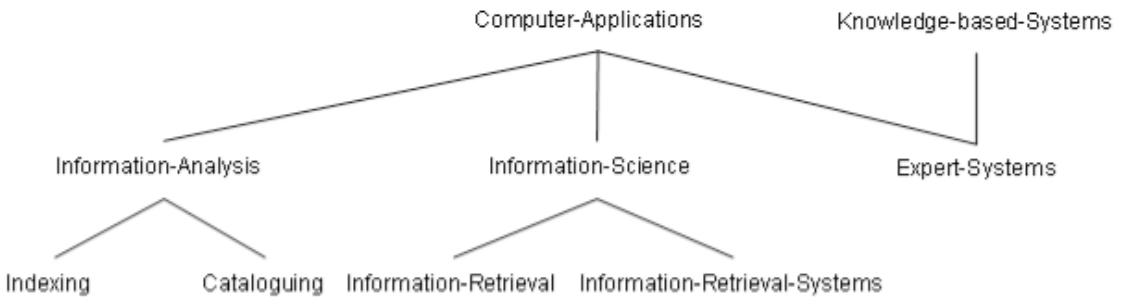
Documents décrits par des domaines de l'informatique

Contexte avec taxonomies sur les attributs

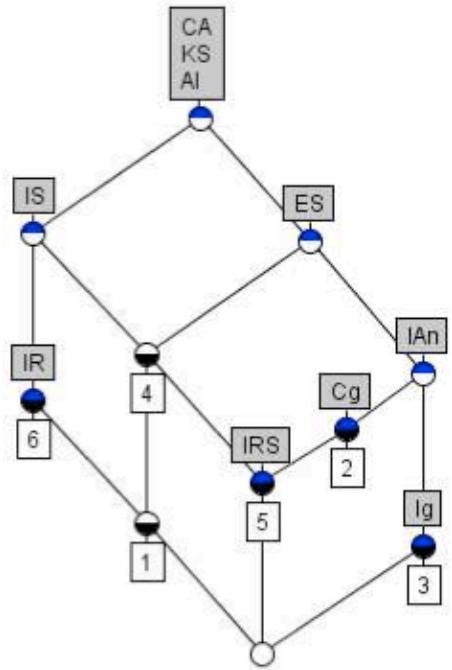


Connaissances taxonomiques sur les domaines de l'informatique

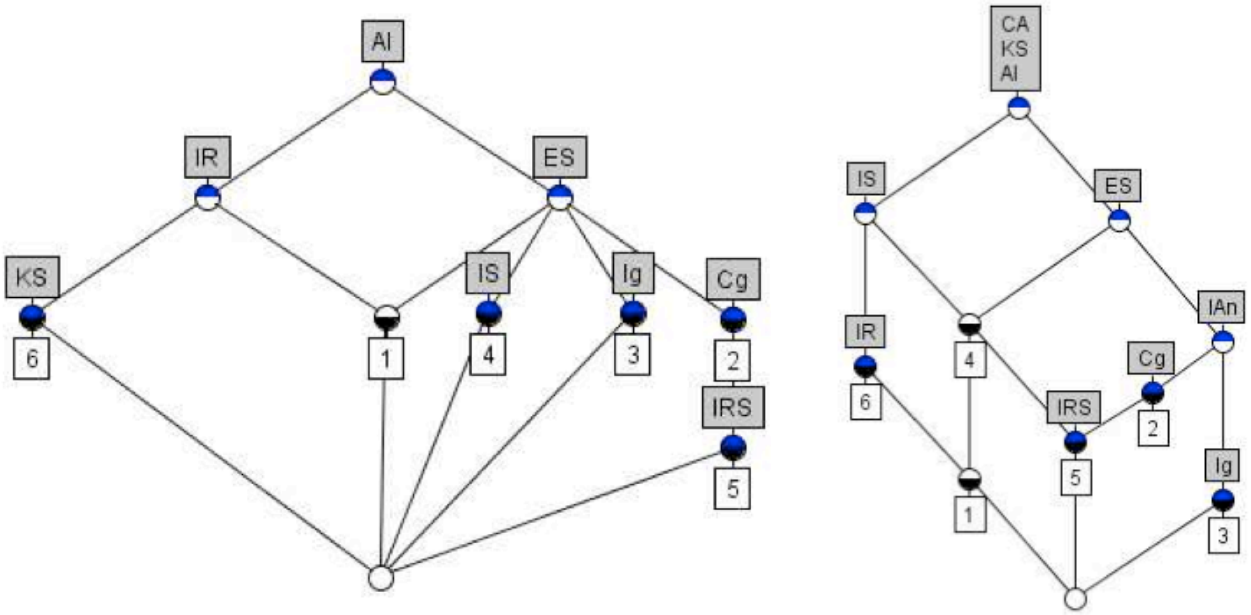
Enrichissement du contexte avec la taxonomie



	Artificial Intelligence (AI)	Expert Systems (ES)	Information Retrieval (IR)	Cataloguing (Cg)	Indexing (Ig)	Information Science (IS)	Information Retrieval Systems (IRS)	Knowledge-based Systems (KS)	Computer-Applications (CA)	Information-Analysis (IAn)
1	x	x	x			x		x	x	
2	x	x		x				x	x	x
3	x	x			x			x	x	x
4	x	x				x		x	x	
5	x	x		x		x	x	x	x	x
6	x		x			x		x	x	



Enrichissement du contexte avec la taxonomie



1 (resp. 5) est maintenant dans une sous-catégorie de la catégorie de 4 (car IR sous-catégorie de IS, resp. IRS sous-catégorie de IS)

2, 3, 5 sont regroupés dans la catégorie IAn

Matériel utilisé

- *Formal Concept Analysis*, Ganter et Wille, Springer, 1999
- *Analyse de concepts formels guidée par des connaissances du domaine : Application à la découverte de ressources génomiques sur le Web*, Nizar Messai, thèse de doctorat, Université H. Poincaré-Nancy 1, mars 2009