ARQAT: une plate-forme d'analyse exploratoire pour la qualité des règles d'association

Xuan-Hiep Huynh, Fabrice Guillet, et Henri Briand

LINA - Equipe COD

Ecole polytechnique de l'université de Nantes

Objectifs

- Présenter une nouvelle approche dans le domaine de qualité des connaissances, avec une forte orientation vers les techniques d'exploratoire de données
- Etudier le comportement des mesures d'intérêt sur un jeu de règles d'association de l'utilisateur
- Développer une plate-forme expérimentale ARQAT (Association Rule Quality Analysis Tool)
- Aider graphiquement l'utilisateur à repérer dans ses données les meilleures mesures et au final les meilleures règles

Etat de l'art et travaux connexes

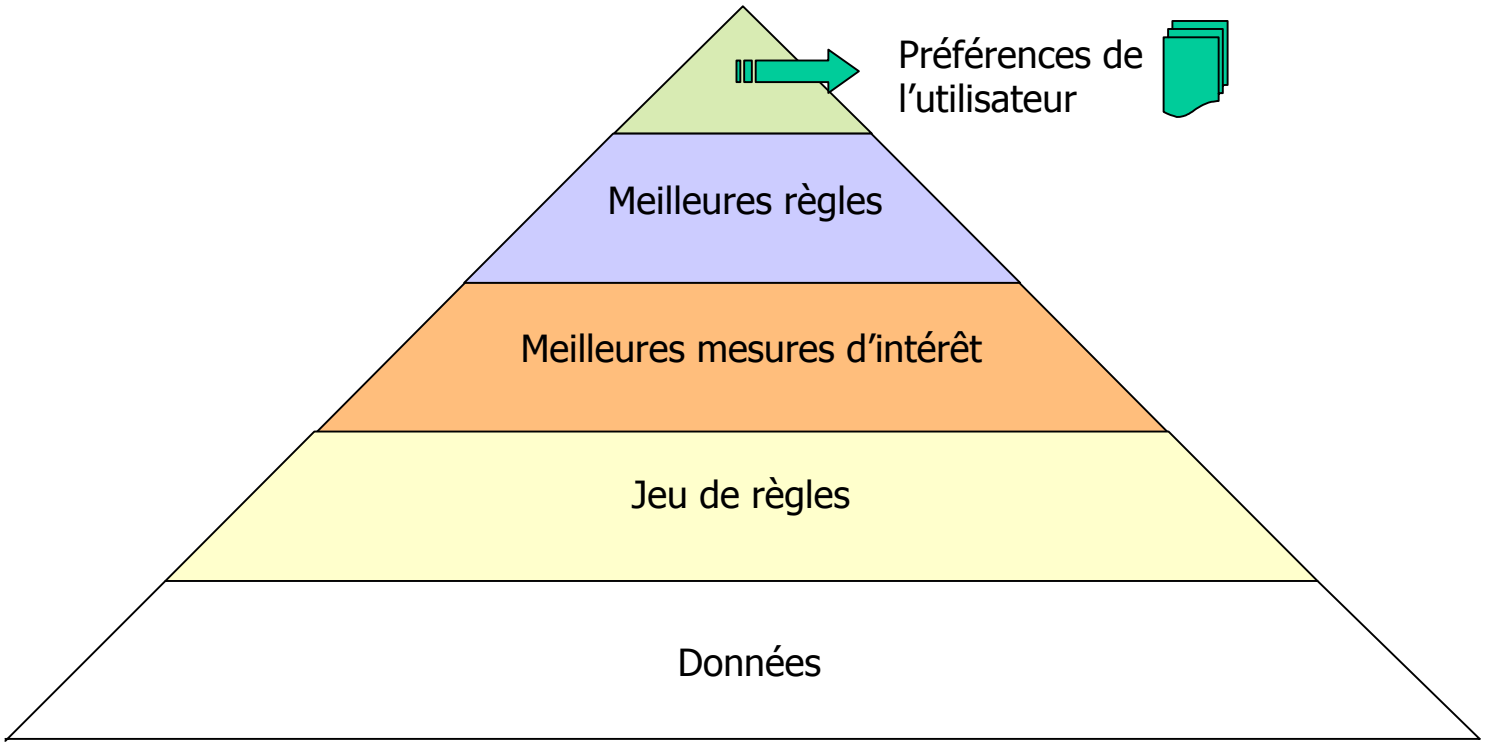
- Deux types de mesure (Freitas 1999): subjective et objective

- Piatetsky-Shapiro 1991: 3 propriétés et 1 mesures
- Hilderman et Hamilton 2001: 5 propriétés et 16 mesures
- Freitas 1999: 1 propriété, 1 mesure
- Tan et al. 2002, Tan et al. 2004: 5 propriétés et 20 mesures

- Lenca et al. 2004: 8 propriétés et 20 mesures
- Guillet 2004: synthèse sur les propriétés et les mesures

- Vaillant et al. 2003: outil HERBS

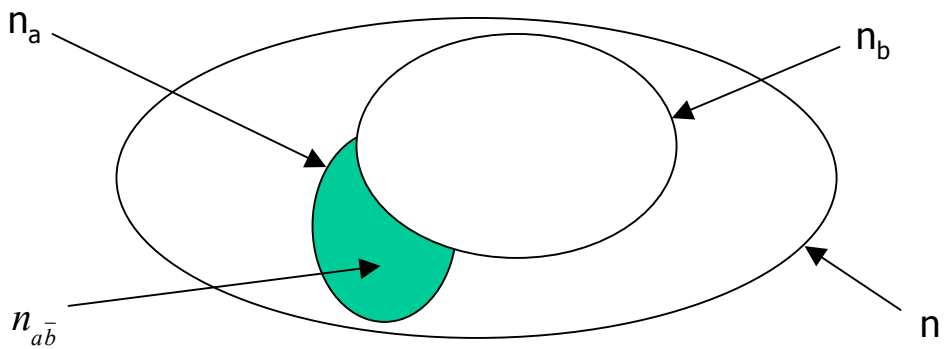
Qualité des connaissances



Définition des règles d'association

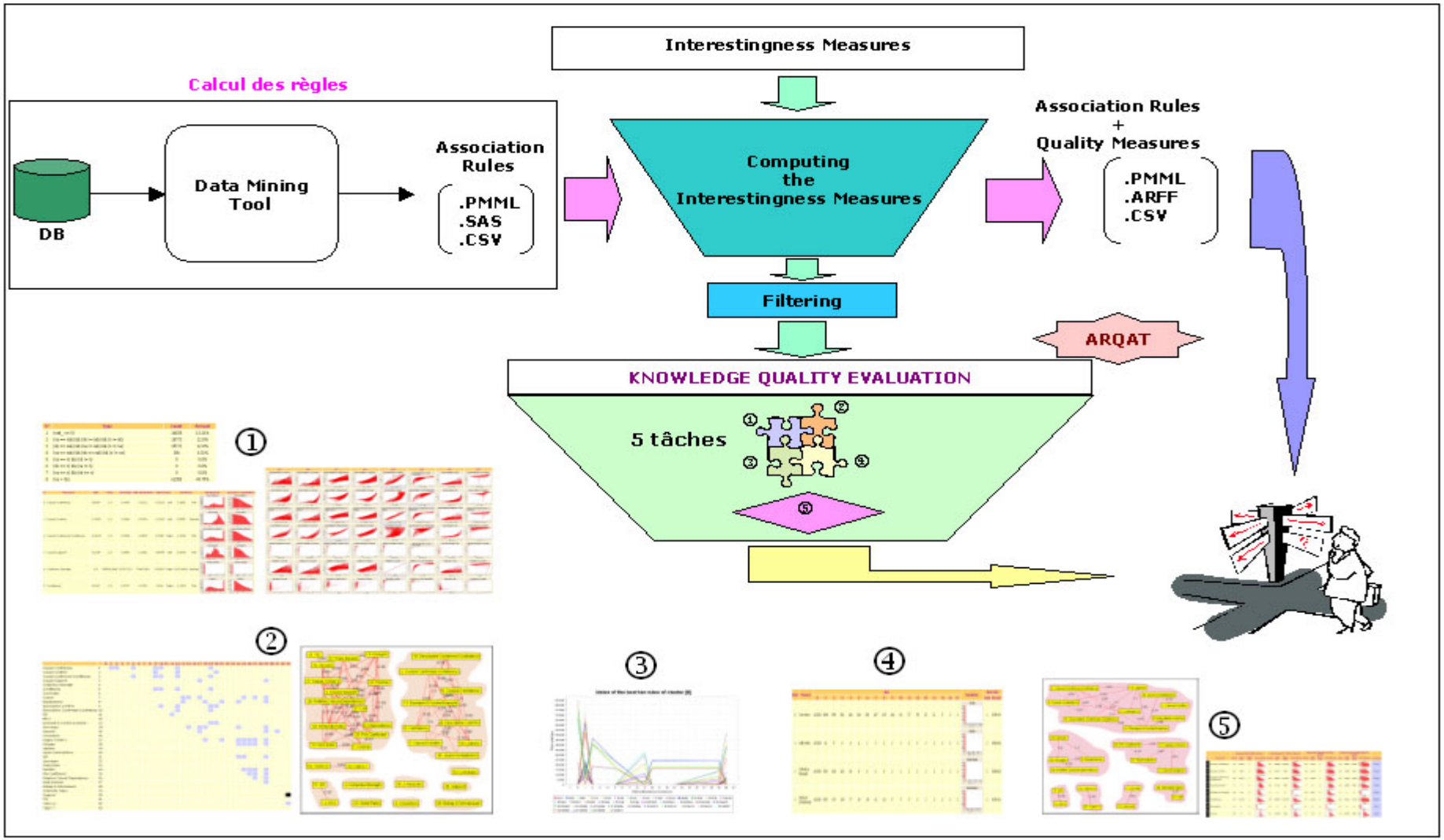
- Une règle d'association est une implication $a \Rightarrow b$ (*a* et *b* sont deux disjoints itemsets)

- Une règle d'association est déterminée par 4 cardinalités: $(n, n_a, n_b, n_{a\bar{b}})$



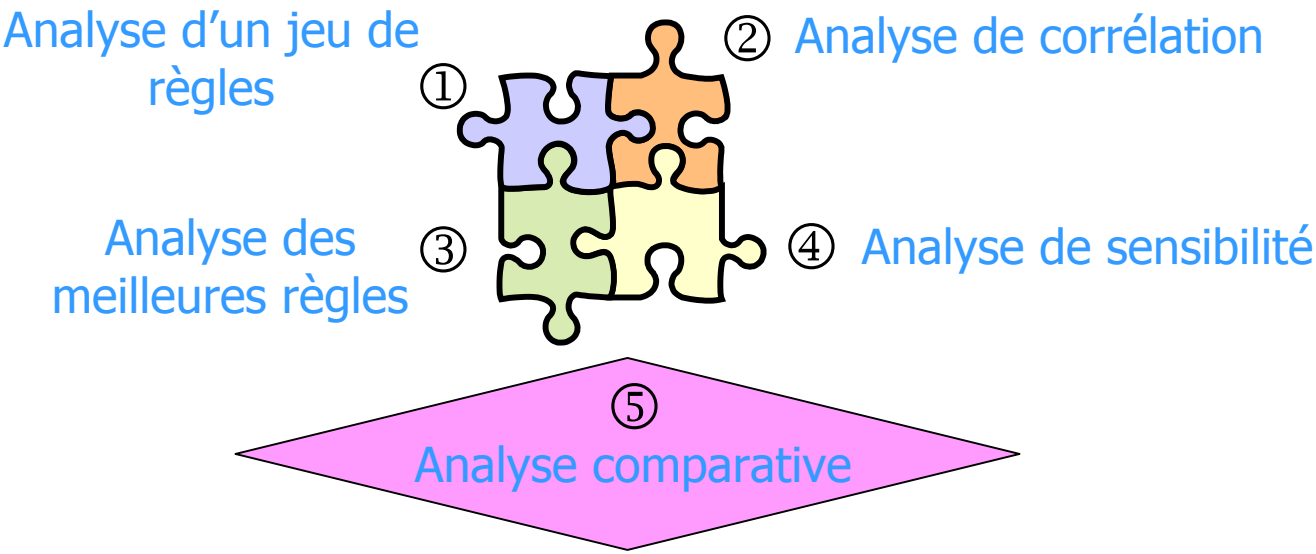
- Mesure de qualité: la qualité des règles d'association est calculée par une fonction de 4 paramètres: $f(n, n_a, n_b, n_{a\bar{b}})$

Plate-forme ARQAT



Principales tâches

- 5 tâches d'analyse avec 14 vues graphiques complémentaires
  (présenterons les trois premières tâches)



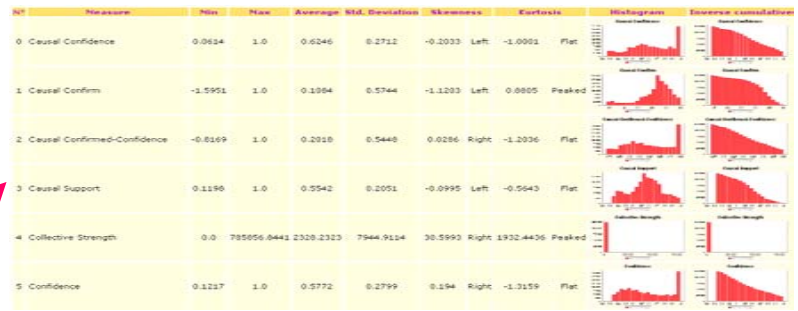
- 34 mesures d'intérêt
- Etude sur la base mushroom (120000 règles)

Tâche ①: analyse d'un jeu de règles

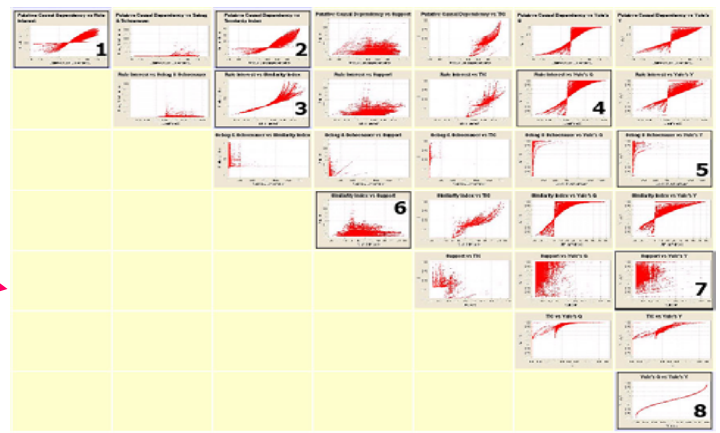
Distribution des contingences

N°	Type	Count	Percent
1	(nab_ == 0)	16158	13.11%
2	(na == nab) && (nb != nab) && (n != nb)	15772	12.8%
3	(nb == nab) && (na != nab) && (n != na)	15772	12.8%
4	(na == nab) && (nb == nab) && (n != na)	386	0.31%
5	(na == n) && (nb != n)	0	0.0%
6	(nb == n) && (na != n)	0	0.0%
7	(na == n) && (nb == n)	0	0.0%
8	(na > nb)	61355	49.79%

Distribution des mesures



Distributions croisées



Tâche ①: analyse d'un jeu de règles

Distribution des contingences: facilite la détection des cas limites

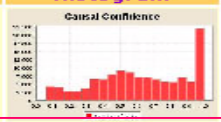
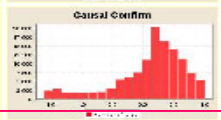
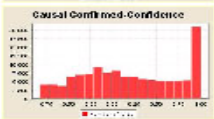
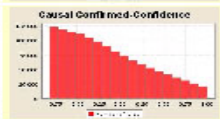
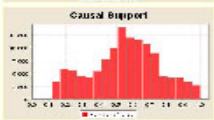
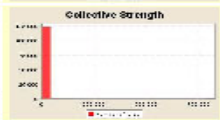
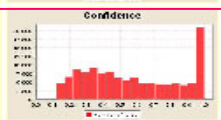
$$(n_{a\bar{b}} = 0, n_a > n_b, \ldots)$$

Règles « logiques »

N°	Type	Count	Percent
1	(nab_ == 0)	16158	13.11%
2	(na == nab) && (nb != nab) && (n != nb)	15772	12.8%
3	(nb == nab) && (na != nab) && (n != na)	15772	12.8%
4	(na == nab) && (nb == nab) && (n != na)	386	0.31%
5	(na == n) && (nb != n)	0	0.0%
6	(nb == n) && (na != n)	0	0.0%
7	(na == n) && (nb == n)	0	0.0%
8	(na > nb)	61355	49.79%

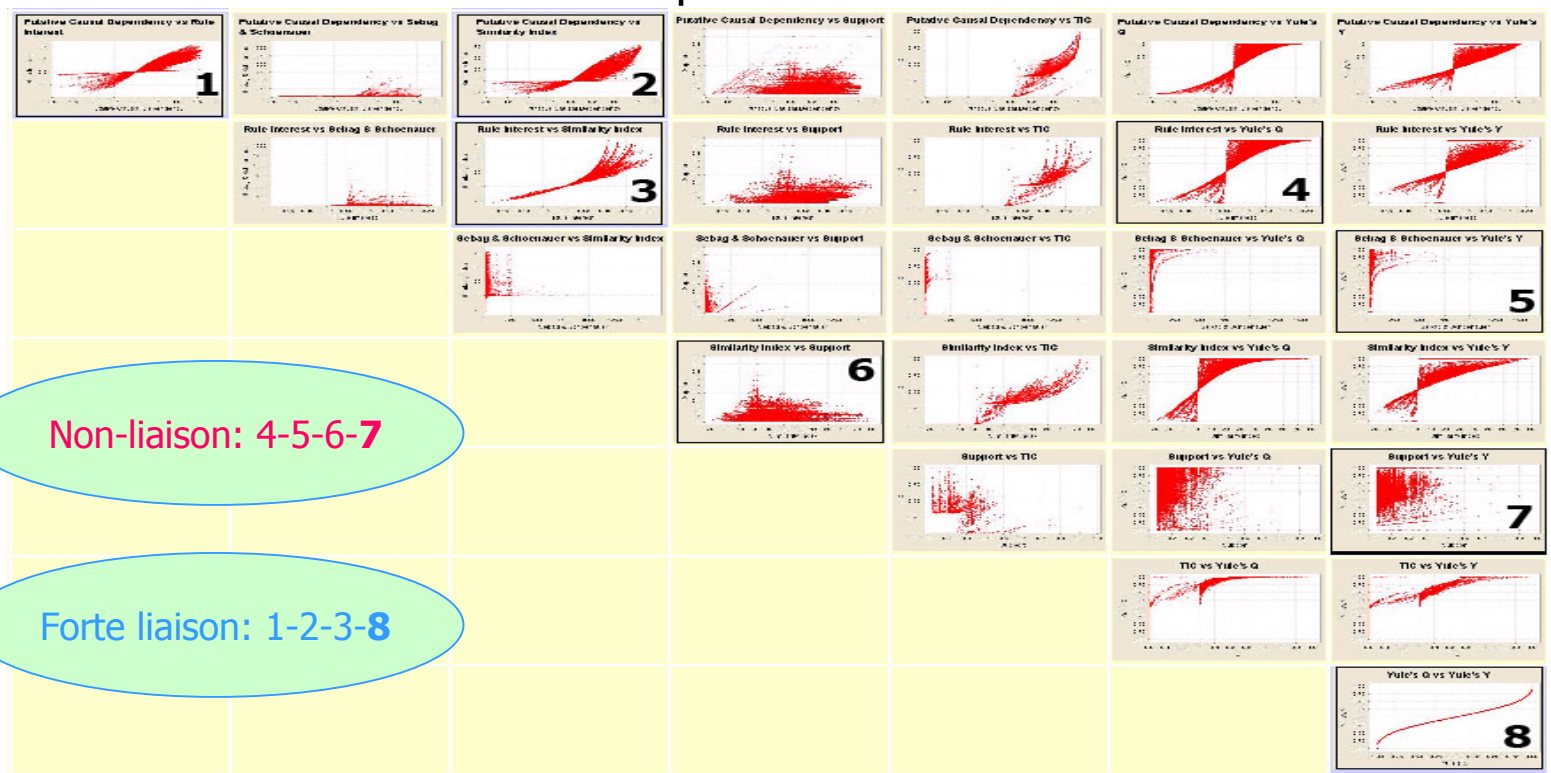
Tâche ①: analyse d'un jeu de règles

Distribution des mesures: sortie des histogrammes (fréquence, inverse-cumulative) de chaque mesure, en complétant de divers indicateurs comme minimum, maximum, écart-type,...

N°	Measure	Min	Max	Average	Std. Deviation	Skewness		Kurtosis		Histogram	Inverse cumulative
0	Causal Confidence	0.0614	1.0	0.6246	0.2712	-0.2033	Left	-1.0001	Flat		
1	Causal Confirm	-1.5951	1.0	0.1084	0.5744	-1.1203	Left	0.8805	Peaked		
2	Causal Confirmed-Confidence	-0.8169	1.0	0.2018	0.5448	0.0286	Right	-1.2036	Flat		
3	Causal Support	0.1198	1.0	0.5542	0.2051	-0.0995	Left	-0.5643	Flat		
4	Collective Strength	0.0	785856.8441	2328.2323	7944.9114	30.5993	Right	1932.4436	Peaked		
5	Confidence	0.1217	1.0	0.5772	0.2799	0.194	Right	-1.3159	Flat		

Tâche ①: analyse d'un jeu de règles

Distributions croisées: concerne des couples de mesures, représente graphiquement une matrice très utile pour visualiser la forme de la liaison existant entre les couples de mesures

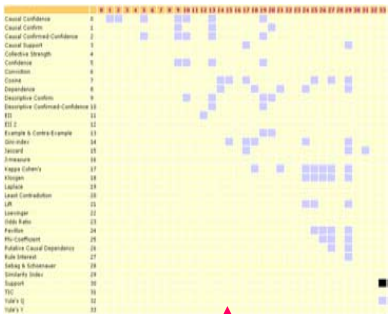


Non-liaison: 4-5-6-**7**

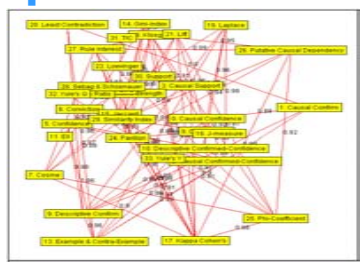
Forte liaison: 1-2-3-**8**

Tâche ②: analyse de corrélation

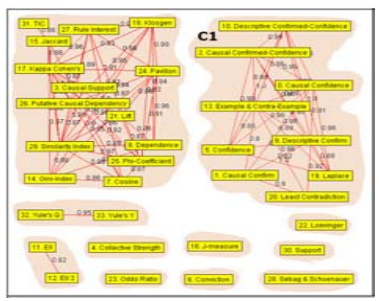
Matrice de corrélation



Graphes de corrélation



Clustering



Corrélation d'un cluster

N°	Versus measures	Value	Image
0	32. Yule's Q <-----> 33. Yule's Y	0.9482	

N°	Measure	Min	Max	Average	Std. Deviation	Skewness	Kurtosis	Histogram	Invers cumulative
32	Yule's Q	-1.0	1.0	0.5628	0.4693	-1.0356	Left 0.3553 Peaked		
33	Yule's Y	-1.0	1.0	0.4561	0.4217	-0.1827	Left -0.7804 Flat		

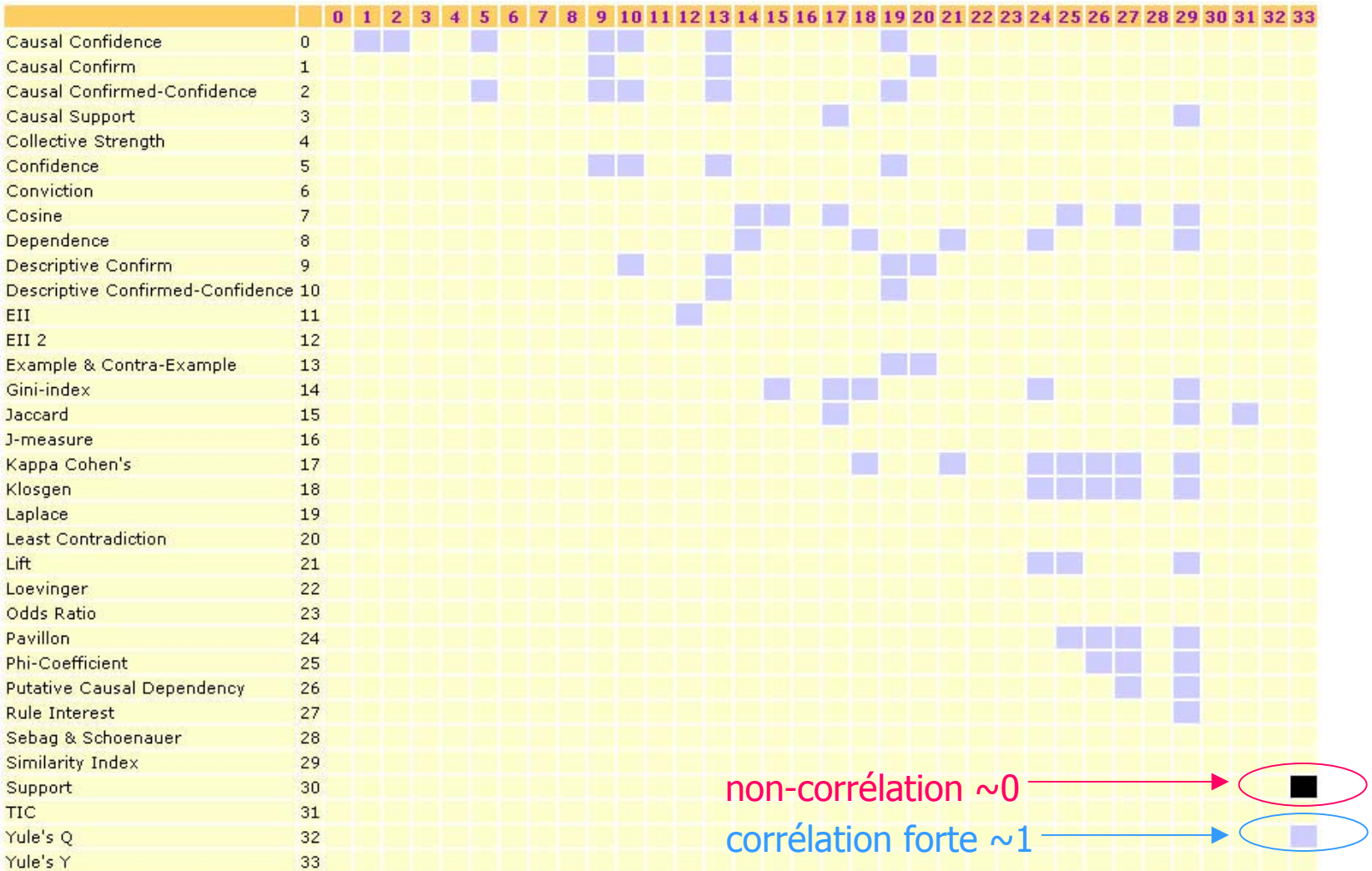
Tâche ②: analyse de corrélation

Matrice

- Utilise le coefficient de corrélation linéaire de Pearson et stocker dans une matrice de corrélation ($I \times I$)

- La matrice de niveau de gris: est une présentation importante, chaque valeur de significativité est visualisée avec un niveau de gris

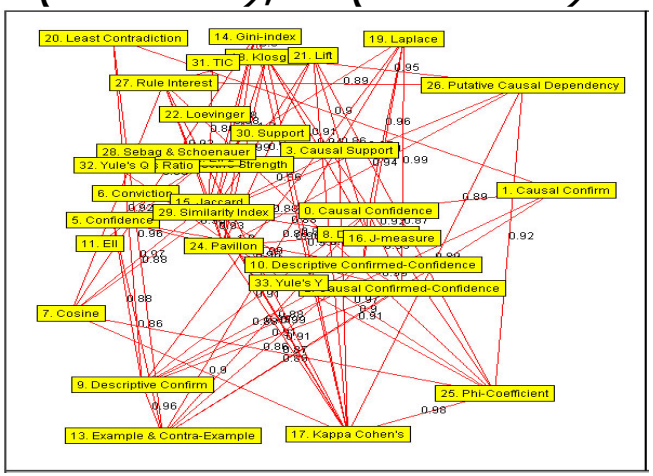
Matrice de niveau de gris



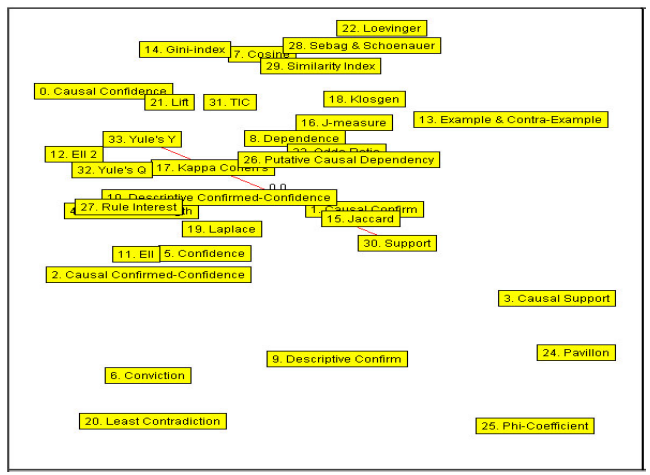
Tâche ②: analyse de corrélation

Graphes

- Matrice de niveau de gris sous formes des graphes de corrélation: non-orienté et valué:
  - Sommet: mesure
  - Arête: valeur de corrélation entre deux sommets/mesures
- Deux sous-graphes partiels **CG+**, **CG0** obtenus à partir les deux seuils: τ *(minimale)*, θ *(maximale)*



CG+ : corrélation forte ~1

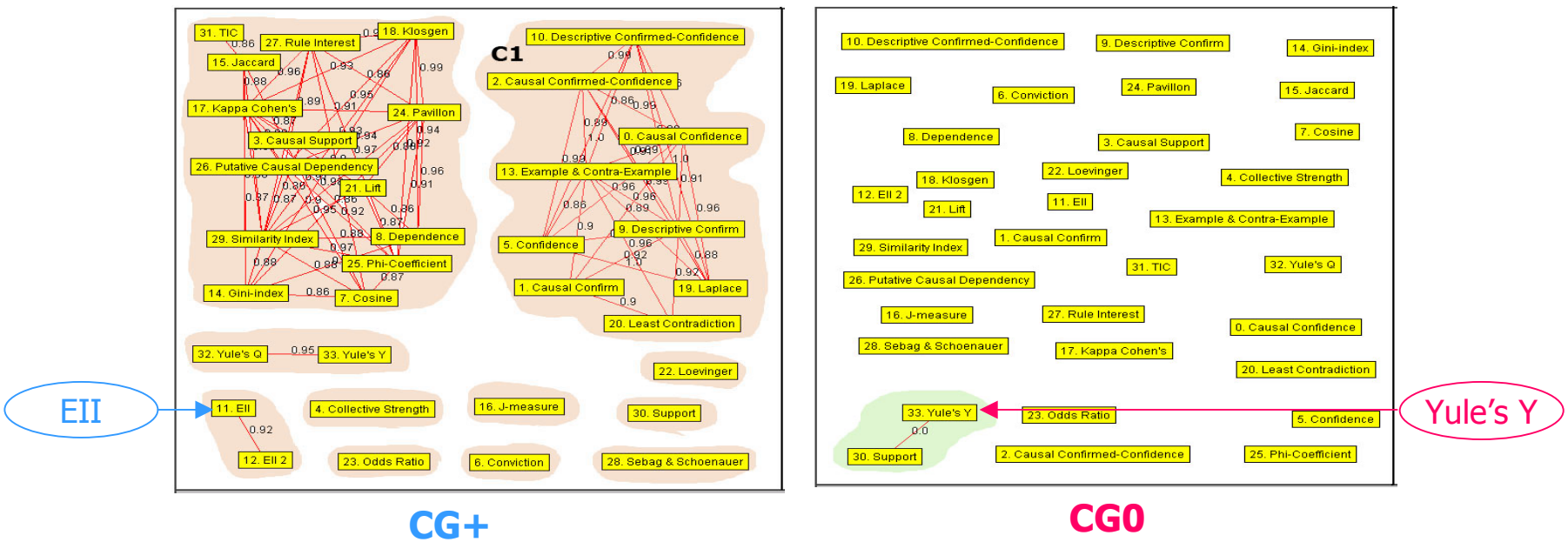


CG0 : non-corrélation ~0

Tâche ②: analyse de corrélation

Clustering: visualisation rapide

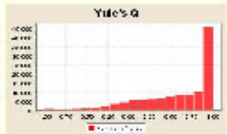
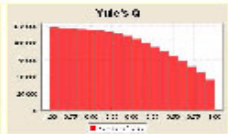
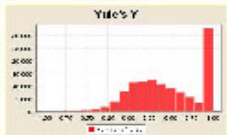
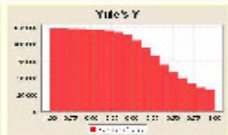
- Extrait des parties connexes des mesures, chaque partie est définie comme un maximale connexe sous-graphe
- Sélection des mesures les plus représentatives

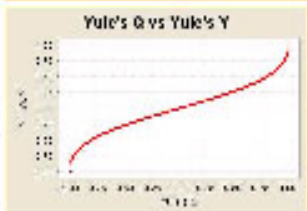


Tâche ②: analyse de corrélation d'un cluster

Vues sur les mesures d'un cluster

- Les informations concernent: les distributions des mesures, les distributions croisée et le sous-graphe

N°	Measure	Min	Max	Average	Std. Deviation	Skewness		Kurtosis		Histogram	Invers cumulative
32	Yule's Q	-1.0	1.0	0.5628	0.4693	-1.0356	Left	0.3553	Peaked		
33	Yule's Y	-1.0	1.0	0.4561	0.4217	-0.1827	Left	-0.7804	Flat		

N°	Versus measures	Value	Image
0	32. Yule's Q <-----> 33. Yule's Y	0.9482	

Tâche ③: Analyse des meilleures règles

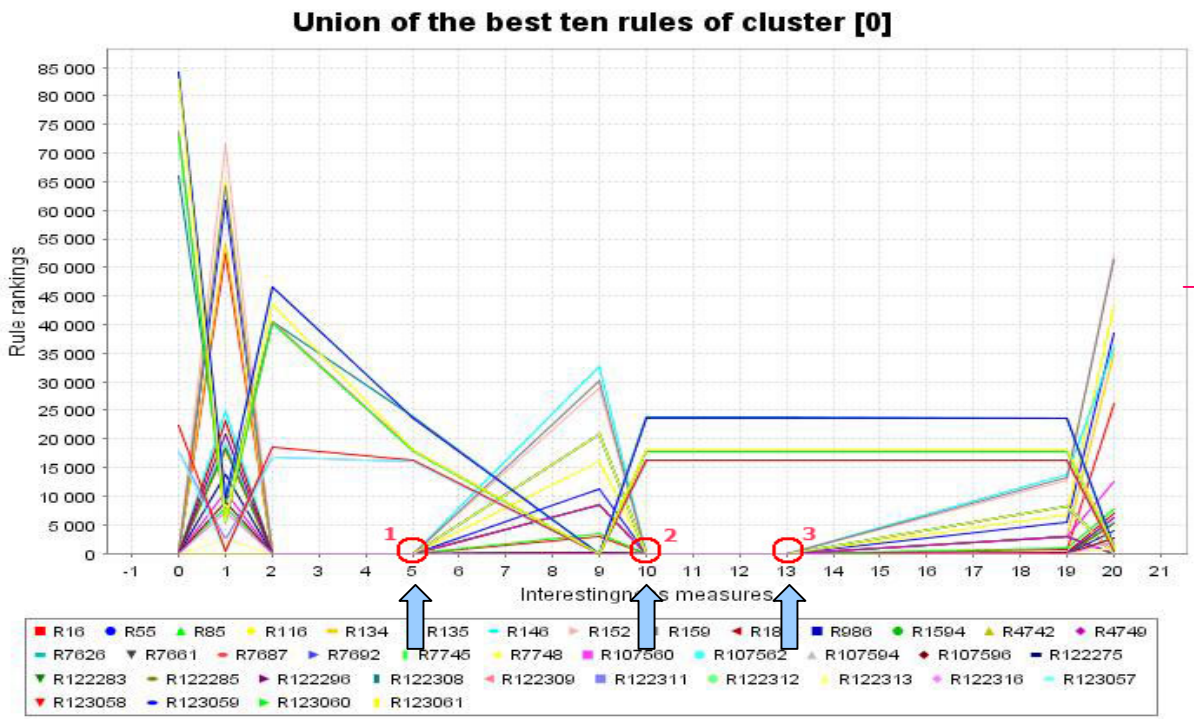
D'un cluster:

- Collecte un ensemble des n meilleures règles de chaque mesure d'intérêt d'une classe (partie connexe d'un graphe)

- Représentation en coordonnées parallèles d'une règle: une interprétation rapide des mesures de chaque règle et de leur variation

Coordonnées parallèles d'une règle

Measure Order		0	1	2	5	9	10	13	19	20	Rule's presentation
21	R107560	1	19121	1	1	41	1	1	8	5388	BROAD FREE ONE ==>veil_color=WHITE
22	R107562	1	18997	1	1	41	1	1	8	5361	BROAD ONE veil_color=WHITE ==>FREE
23	R107594	1	8972	1	1	18	1	1	3	2574	CLOSE FREE ONE ==>veil_color=WHITE
24	R107596	1	8914	1	1	18	1	1	3	2564	CLOSE ONE veil_color=WHITE ==>FREE
25	R122275	1	13800	1	1	32	1	1	5	3977	BROAD FREE ==>veil_color=WHITE
26	R122283	1	18299	1	1	38	1	1	6	5145	FREE stalk_surf_above=SMOOTH ==>veil_color=WHITE
27	R122285	1	18179	1	1	38	1	1	6	5134	stalk_surf_above=SMOOTH veil_color=WHITE ==>FREE
28	R122296	1	20903	1	1	55	1	1	10	6193	FREE stalk_surf_below=SMOOTH ==>veil_color=WHITE
29	R122308	65969	8772	40612	23743	10	23743	23743	23714	1013	FREE ==>ONE veil_color=WHITE



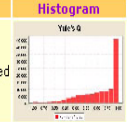
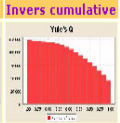
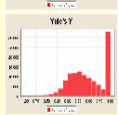
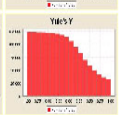
Classé par rangs

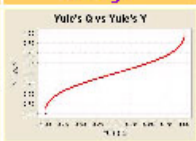
n = 10

Complémentarité des vues sur un cluster

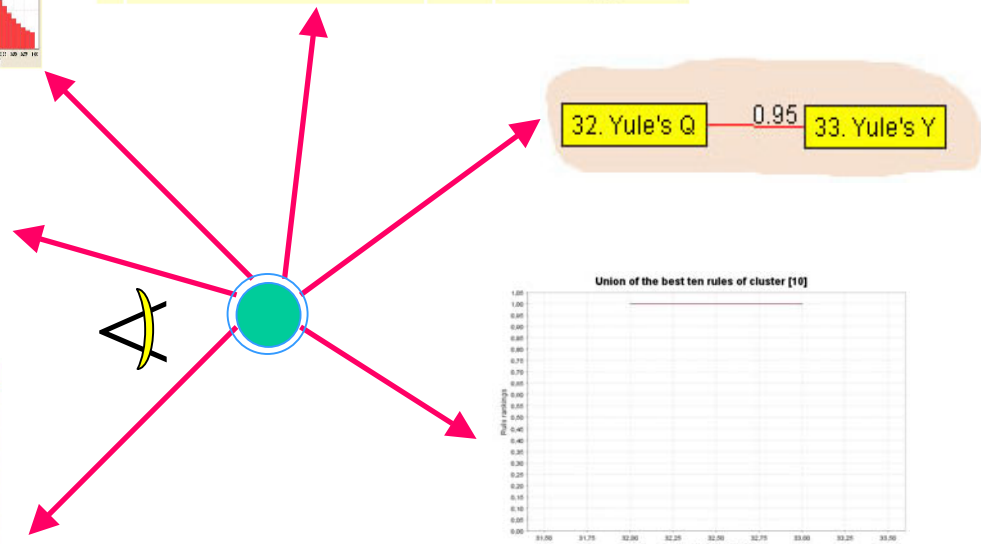
Plusieurs vues complémentaires

- Pour la compréhension plus fine/précise des résultats

N°	Measure	Min	Max	Average	Std. Deviation	Skewness	Kurtosis	Histogram	Invers cumulative
32	Yule's Q	-1.0	1.0	0.5628	0.4693	-1.0356 Left	0.3553 Peaked		
33	Yule's Y	-1.0	1.0	0.4561	0.4217	-0.1827 Left	-0.7804 Flat		

N°	Versus measures	Value	Image
0	32. Yule's Q <-----> 33. Yule's Y	0.9482	

Rank	32. Yule's Q		33. Yule's Y	
~1	R4	1.0	R4	1.0
~2	R12	1.0	R12	1.0
~3	R16	1.0	R16	1.0
~4	R32	1.0	R32	1.0
~5	R33	1.0	R33	1.0
~6	R41	1.0	R41	1.0
~7	R42	1.0	R42	1.0
~8	R55	1.0	R55	1.0
~9	R84	1.0	R84	1.0
~10	R85	1.0	R85	1.0



Measure Order		32	33	Rule's presentation
1	R4	1	1	veil_color=WHITE ==>WOODS
2	R12	1	1	FREE ==>EVANESCENT
3	R16	1	1	stalk_surf_below=SILKY ==>veil_color=WHITE
4	R32	1	1	FREE ==>NARROW
5	R33	1	1	ONE ==>NARROW
6	R41	1	1	FREE ==>BULBOUS
7	R42	1	1	veil_color=WHITE ==>BULBOUS
8	R55	1	1	SOLITARY ==>veil_color=WHITE
9	R84	1	1	ONE ==>FOUL
10	R85	1	1	FOUL ==>POISONOUS

Plus informations…

- Ecrit en Java, sortie sous fichiers HTML
- Formats supportés pour importer/exporter les jeux de règles: PMML, CSV, ARFF
- Téléchargeable à www.polytech.univ-nantes.fr/arqat

Conclusion

Points forts

- Nouvelle approche exploratoire (graphique + interactif): rapidité d'interprétation des résultats
- Richesse des mesures d'intérêt: 34
- Divers points de vue (vues complémentaires) et mesures implémentées permettent à un expert de mieux comprendre les corrélations existant entre les mesures d'intérêt sur son jeu de règles
- Richesse des tâches/vues: permettent une étude fine/précise des règles d'association

Conclusion

Perspectives

- Amélioration l'analyse de corrélation avec un coefficient plus performant
- Amélioration la classification des mesures pour la sélection des meilleures mesures d'intérêt, dirigé par les préférences de l'utilisateur

Merci de votre attention.