Partie II Extraction de motifs séquentiels - Application à la santé et l'environnement

Maguelonne Teisseire

Ecole thématique BDA, Les Houches, 16-21 mai 2010

Cemagref
Sciences, eaux & territoires

TETIS

Avant Propos

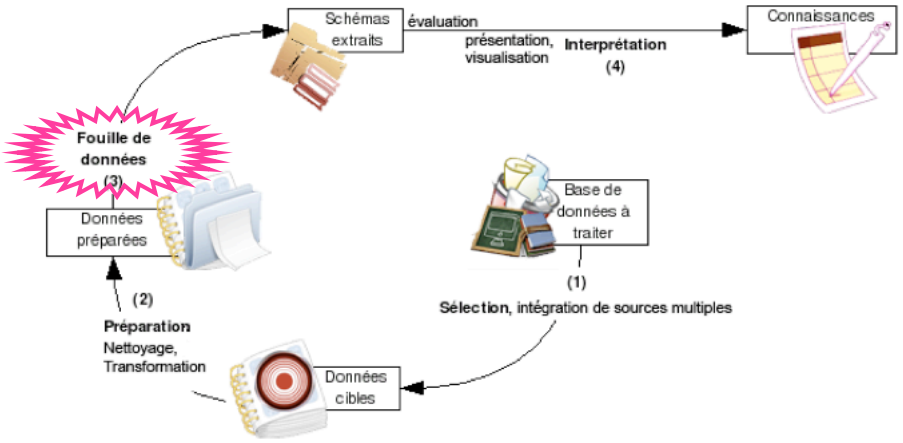
- Extraction de Connaissances

  Identifier dans les données des schémas valides, nouveaux, potentiellement utiles et compréhensibles

  [Fayyad et al., 1996]

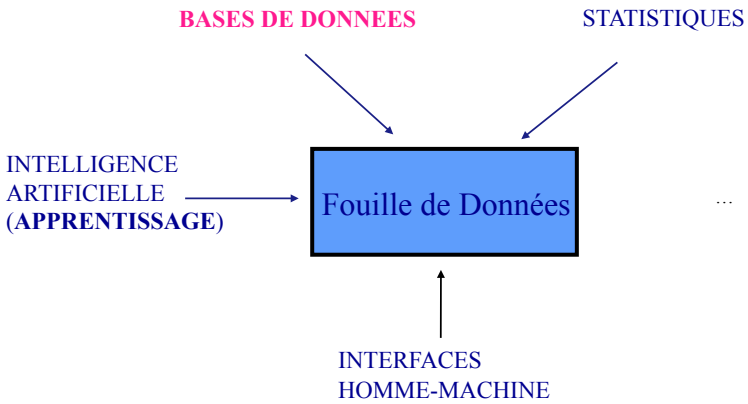
2

Le Processus d'Extraction de Connaissances



3

Fouille de données ?



4

Recherche d'associations et motifs séquentiels

- **Recherche d'associations**
  - recherche de corrélations entre attributs (items)
- **Recherche de motifs séquentiels**
  - recherche de corrélations entre attributs (items) mais en prenant en compte le temps entre items => comportement

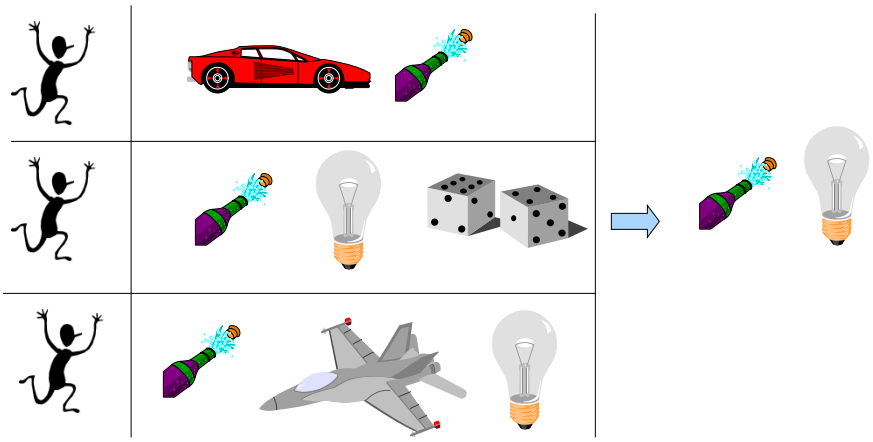
Plan

1. Règles d'association
2. Motifs séquentiels
3. Application 1 : Identification de marqueurs de la maladie d'Alzheimer
4. Application 2 : Détection d'anomalies dans les données ferroviaires
5. Conclusion

Plan

1. **Règles d'association**
2. Motifs séquentiels
3. Application 1 : Identification de marqueurs de la maladie d'Alzheimer
4. Application 2 : Détection d'anomalies dans les données ferroviaires

Recherche de règles d'association



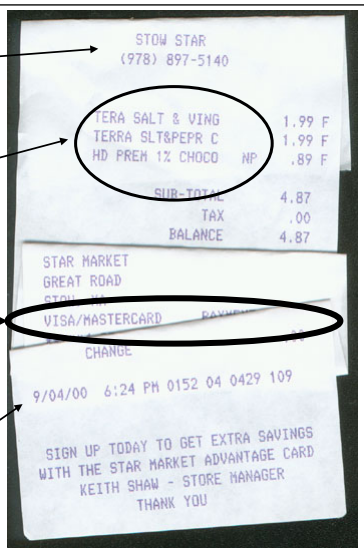
Panier de la ménagère

Localisation

Produits achetés

Identification

Date, heure



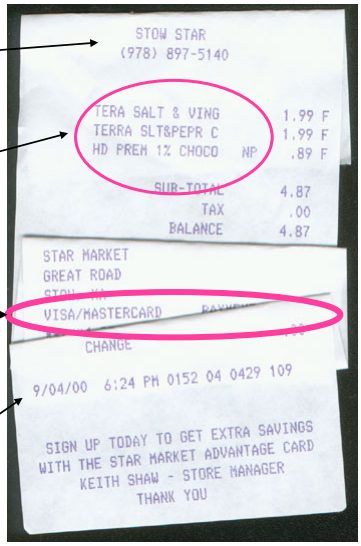
Panier de la ménagère

Localisation

Produits achetés

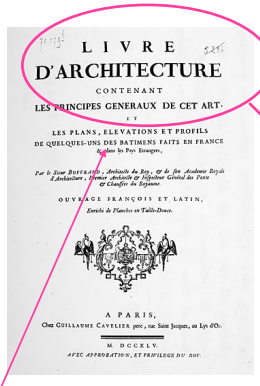
Identification

Date, heure



Panier de la ménagère

Localisation



Premier paragraphe

« Livre d'architecture contenant les principes généraux … »

Identification

Position # Date

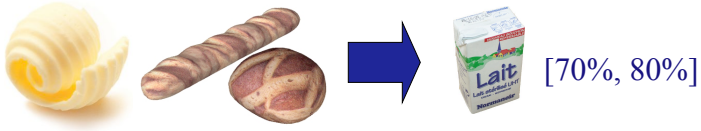
Mots # Produits

Recherche de règles d'association

- **Règles de la forme**

ANTECEDENT → CONSEQUENT [Support, Confiance]

(support et confiance sont des mesures d'intérêt définies par l'utilisateur)



[70%, 80%]

Interprétation

- R : X →Y (A%, B%)
  - ***Support*** *: portée de la règle*
    Proportion de paniers contenant tous les attributs
    A% des clients ont acheté les 2 articles X et Y
  - ***Confiance*** *:*
    Proportion de paniers contenant le conséquent parmi ceux qui contiennent l'antécédent
    B% des clients qui ont acheté X ont aussi acheté Y

[70%, 80%]

Utilisation des règles d'association

…. [70%, 80%]

- **Couches** comme conséquent
  Déterminer ce qu'il faut faire pour augmenter les ventes
- **Bière** comme antécédent
  Quel(s) produit(s) serai(en)t affecté(s) si on n'arrête de vendre de la bière ?
- **Bière** comme antécédent et **Couche** comme conséquent
  Quels produits devraient être vendus avec la Bière pour promouvoir la vente de couches ?

Recherche des règles

- **Items :** A, B, C, D, E, F

Trans ID	Items
1	A, D
2	A, C
3	A, B, C
4	A, B, E, F

Recherche des règles

- **Items :** A, B, C, D, E, F
- 4 **transactions** (sous ensemble d'items)
  - T1 : {A,D}

Trans ID	Items
1	A, D
2	A, C
3	A, B, C
4	A, B, E, F

Recherche des règles

- **Items :** A, B, C, D, E, F
- 4 **transactions** (sous ensemble d'items)
  - T1 : {A,D}
- **Support** pour un itemset
  - Supp ({A,D})=1/4

Trans ID	Items
1	A, D
2	A, C
3	A, B, C
4	A, B, E, F

Recherche des règles

- **Items :** A, B, C, D, E, F
- 4 **transactions** (sous ensemble d'items)
  - T1 : {A,D}
- **Support** pour un itemset
  - Supp ({A,D})=1/4
  - Supp ({A,C})=2/4

Trans ID	Items
1	A, D
2	A, C
3	A, B, C
4	A, B, E, F

Recherche des règles

- **Items :** A, B, C, D, E, F
- 4 **transactions** (sous ensemble d'items)
  - T1 : {A,D}
- **Support** pour un itemset
  - Supp ({A,D})=1/4
  - Supp ({A,C})=2/4
- **Itemsets fréquents** (minSupp=50%)
  - {A,C} est un itemset fréquent

Trans ID	Items
1	A, D
2	A, C
3	A, B, C
4	A, B, E, F

Recherche des règles

- **Items :** A, B, C, D, E, F
- 4 **transactions** (sous ensemble d'items)
  - T1 : {A,D}
- **Support** pour un itemset
  - Supp ({A,D})=1/4
  - Supp ({A,C})=2/4
- **Itemsets fréquents** (minSupp=50%)
  - {A,C} est un itemset fréquent
- **Règles :** (minSupp et minConf = 50%)
  - A → C [50%, 50%]

Trans ID	Items
1	A, D
2	A, C
3	A, B, C
4	A, B, E, F

Recherche des règles

- **Items :** A, B, C, D, E, F
- 4 **transactions** (sous ensemble d'items)
  - T1 : {A,D}
- **Support** pour un itemset
  - Supp ({A,D})=1/4
  - Supp ({A,C})=2/4
- **Itemsets fréquents** (minSupp=50%)
  - {A,C} est un itemset fréquent
- **Règles :** (minSupp et minConf = 50%)
  - A → C [50%, 50%]
  - C → A [50%, 100%]

Trans ID	Items
1	A, D
2	A, C
3	A, B, C
4	A, B, E, F

Schéma algorithmique de base

- I = {A, B,C}

- L 'algorithme naïf procède en **2 phases** :
  - 1) **Génération de tous les ensembles fréquents**

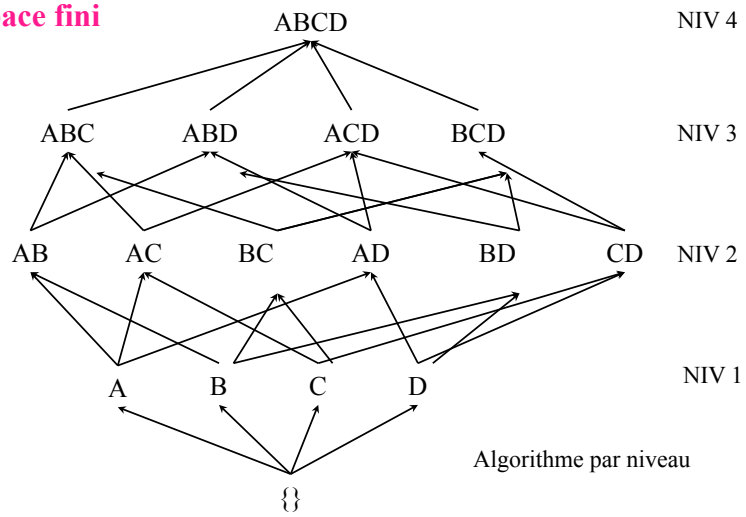
    {∅},{A}, {B}, {C},
    {A,B}, {A,C}, {B,C}
    {A,B,C}
  - 2) **Comptage du support et confiance**

Génération des ensembles fréquents

- Le nombre d'ensembles fréquents potentiels est égal à la taille du produit cartésien de tous les items ….
- ….qui **croit exponentiellement** en fonction du nombre d'items considérés.
- **Exemple :** 1000 items => 2^{1000} ensembles à considérer

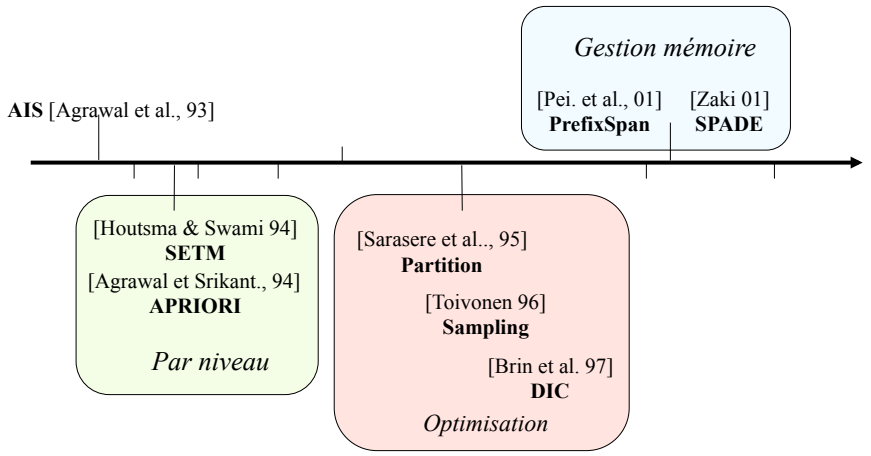
Espace de recherche

Espace fini



Algorithme par niveau

De nombreux algorithmes



Plan

1. Règles d'association

3. **Motifs séquentiels**

5. Application 1 : Identification de marqueurs de la maladie d'Alzheimer
6. Application 2 : Détection d'anomalies dans les données ferroviaires
7. Application 3 : Analyse des usages dans les logs web

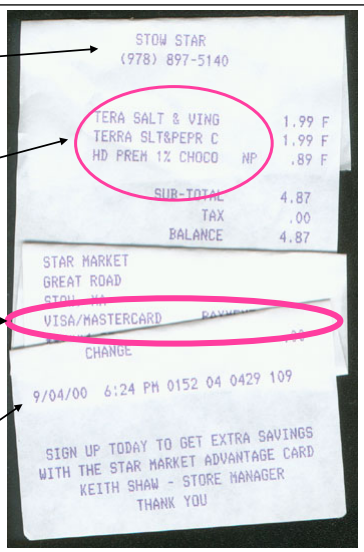
Panier de la ménagère

Localisation

Produits achetés

Identification

Date, heure



Règles d'association vs Motifs Séquentiels

- **Corrélation entre les produits (RA)**

  Les personnes qui achètent des couches achètent de la bière

- **Comportement des clients au cours du temps (MS)**

  *Les personnes qui achètent des couches achètent **trois jours après** de la bière*

28

Extraction de Motifs Séquentiels

- Chaque transaction *T* est composée de :
  - Un **identifiant** du client noté C_{id}
  - Une **estampille temporelle**, notée *time*
  - Un **ensemble d'items** (itemset) intervenants dans la transaction, noté i_t

I = {a,b,c,d} :

C_1	t_1	a,b,c,d
	t_2	a,b
C_2	t_3	a,b
C_3	t_1	b,c,d
	t_4	a,b

29

Extraction de Motifs Séquentiels

- Chaque transaction *T* est composée de :
  - Un **identifiant** du client noté C_{id}
  - Une **estampille temporelle**, notée *time*
  - Un **ensemble d'items** (itemset) intervenants dans la transaction, noté i_t

Séquence de données du client C_3

S = <(bcd)(ab)>

- Signifie que le client C_3 a acheté b,c et d en même temps puis a et b en même temps

I = {a,b,c,d} :

C_1	t_1	a,b,c,d
	t_2	a,b
C_2	t_3	a,b
C_3	t_1	b,c,d
	t_4	a,b

30

Extraction de Motifs Séquentiels

- Une **séquence *S*** : liste ordonnée d'itemsets
  - i1 = (bc), i2 = (ab) 2 itemsets
  - S = <(bc)(ab)> 4-séquence

I = {a,b,c,d} :

C_1	t_1	a,b,c,d
	t_2	a,b
C_2	t_3	a,b
C_3	t_1	b,c,d
	t_4	a,b

31

Extraction de Motifs Séquentiels

- Une **séquence *S*** : liste ordonnée d'itemsets
  - i1 = (bc), i2 = (ab) 2 itemsets
  - S = <(bc)(ab)> 4-séquence
- Inclusion de séquences
  - S est incluse dans la séquence de données du client C_3
  - <(bc)(ab)> ≤ <(bcd)(ab)>
  - <(a)(b)> <> <(bcd)(ab)>

I = {a,b,c,d} :

C_1	t_1	a,b,c,d
	t_2	a,b
C_2	t_3	a,b
C_3	t_1	b,c,d
	t_4	a,b

32

Extraction de Motifs Séquentiels

- Support

$$Support(S,D) = \frac{|\{C \in D \mid S \leq C_{trans}\}|}{|\{C \in D\}|}$$

- Attention! l'occurrence n'est prise en compte qu'une fois dans la séquence

  Support (20) dans <(10) (20 30) (40) (20)>=1

33

Problématique

- Extraction de Motifs Séquentiels

  Extraire toutes les séquences fréquentes S, i.e. vérifiant :

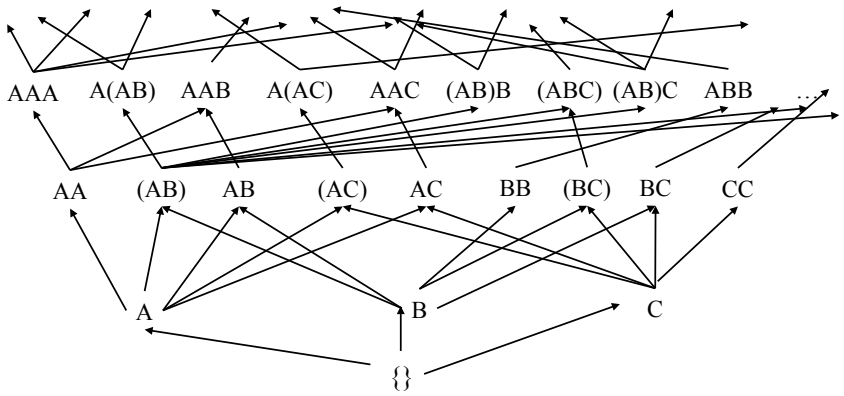
  Support (S,D) ≥ *minsup* avec 0 ≤ minsup ≤ *1*

- 50% des personnes qui achètent du vin et du fromage achètent aussi plus tard du pain

<( , ) ()>

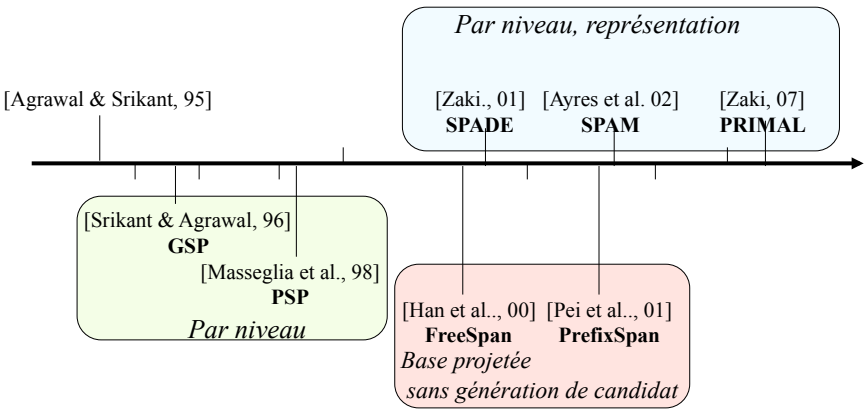
Motifs Séquentiels : l'espace de recherche

Espace infini mais borné par la longueur de la plus grande séquence



35

Des algorithmes de plus en plus efficaces



Extraction rapide de motifs séquentiels

36

D'autres challenges…

- Utiliser les motifs pour répondre à des problèmes réels
  - Quid des données complexes ? Autres que booléennes ? Numérique, texte… multidimensionnelles
  - Quid de la présentation des motifs aux utilisateurs ? Plutôt que tous les générer, ne peut on pas les contraindre ?
  - Quid de l'évolution des données dans les bases ? Comment extraire de nouvelles connaissances sans tout recalculer ? Flots ?
  - ….

Le Flou et les Motifs Séquentiels

" We did not consider quantitites or values of the items bought in a transaction, which are important for some applications. Finding such rules need further work" [Aggrawal et al 96]

- Les personnes qui achètent **des** couches achètent trois jours après **de la** bière
- Les personnes qui achètent **trois paquets** de couches achètent trois jours après **six paquets** de bière
- Les personnes qui achètent **un peu** de couches achètent trois jours après **beaucoup** de bière
- Les personnes qui achètent **un peu** de couches achètent **quelques jours** après **beaucoup** de bière

38

Les Motifs Séquentiels Flous

- Comment convertir des attributs numériques en items flous ?
- Quid de la fréquence en fonction des degrés des items flous ?

- Item flou = [item, sous-ensemble flou]
- *SpeedyFuzzy* : Comptage Binaire
- *MiniFuzzy* : Comptage Seuillé
- *TotallyFuzzy* : Σ- Comptage Seuillé

39

Les Motifs Séquentiels Flous

- Comptage binaire (degré d'appartenance > 0)

		Maux Tête		Fièvre			Démangeaisons	
Pat.	*H.*	*oui*	*non*	*Peu*	*Moy.*	*Bcp.*	*oui*	*non*
X	2	0	1	0.8	0.2	0	1	0
X	52	0	1	0	0.6	0.4	1	0
X	53	0	1	0	0.1	0.9	1	0
X	54	1	0	0	0	1	1	0

Support (X,<([Fièvre, Bcp])>) = 1

$Support_{SpeedyFuzzy}(c,(X,A))=1$ si $\forall$ [x,a] $\in$(X,A), $\mu_a(t[x])> 0$ et 0 sinon

40

Les Motifs Séquentiels Flous

- Comptage seuillé (degré d'appartenance $> \omega$)

  si $\omega = 0.5$

		Maux Tête		Fièvre			Démangeaisons	
Pat.	*H.*	*oui*	*non*	*Peu*	*Moy.*	*Bcp.*	*oui*	*non*
X	2	0	1	0.8	0.2	0	1	0
X	52	0	1	0	0.6	0.4	1	0
X	53	0	1	0	0.1	0.9	1	0
X	54	1	0	0	0	1	1	0

Support (X,<([Fièvre, Bcp])>) =1

$\text{Support}_{\text{MiniFuzzy}}(c,(X,A))=1$ si $\forall$ [x,a] $\in(X,A)$, $\mu_a(t[x])> \omega$ et 0 sinon

41

Les Motifs Séquentiels Flous

- Σ- Comptage seuillé(degré $> \omega$ et pondération)

		Maux Tête		Fièvre			Démangeaisons	
Pat.	*H.*	*oui*	*non*	*Peu*	*Moy.*	*Bcp.*	*oui*	*non*
X	2	0	1	0.8	0.2	0	1	0
X	52	0	1	0	0.6	0.4	1	0
X	53	0	1	0	0.1	0.9	1	0
X	54	1	0	0	0	1	1	0

Support (X,<([Fièvre, Bcp])>)

$$\alpha_a(t[x]) = \begin{cases} \mu_a(t[x]) \text{ si } \mu_a(t[x])> \omega \\ 0 \text{ sinon} \end{cases}$$

$\text{Support}_{\text{TotallyFuzzy}}(c,(X,A))=\perp_{j=1}^{\theta c} \mp$ [x,a] $\in(X,A)$ $[\alpha_a(t_j[x])]$ où

$\mp$ et $\perp$ sont les opérateurs de t-norme et t-conorme généralisés 42

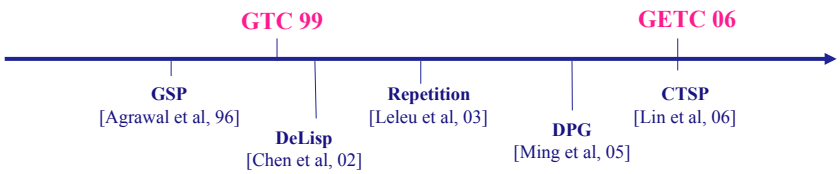
Le Flou et les Motifs Séquentiels

- Les personnes qui achètent **des** couches achètent trois jours après **de la** bière
- Les personnes qui achètent **trois paquets** de couches achètent trois jours après **six paquets** de bière
- Les personnes qui achètent **un peu** de couches achètent trois jours après **beaucoup** de bière
- Les personnes qui achètent **un peu** de couches achètent **quelques jours** après **beaucoup** de bière

43

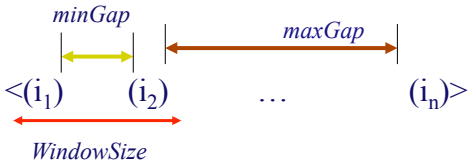
Contraintes Temporelles

- Comment prendre en compte des contraintes temporelles à la volée ? (*windowSize*, *mingap*, *maxgap*)
- Est il possible d'introduire plus de souplesse pour des contraintes trop strictes ?
- GTC (*Graph for Time Constraints*)
  Traiter les contraintes à la volée
- GETC (*Graph Extended for Time Constraints)*
  Prise en compte du flou



44

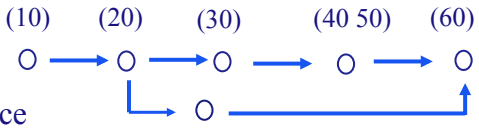
GTC : pré-traitement des contraintes de temps



Client	Tps	Items
C_1	1	10
	7	20
	13	30
	17	40
	18	50
	24	60

minGap=1
windowSize=5

<(10) (20) (30) (50) (60)>
<(10) (20) (30) (40) (60)>
<(10) (20) (30) (40 50) (60)>
<(10) (20) (30 40) (60)
<(10) (20) (30 40 50) (60)>



Graphe de séquence

45

GETC : un exemple

- Extension de windowSize

  S=<(10 20 30 40)>

ws=1

Date	1	2	3	4	5	6	7
Client 1	10	20	30 40	-	50 60	80	90
Client 2	10	20 30 40	50 60	70 80	-	-	-

Avec des contraintes de temps strictes, $s \not\subseteq C_1$, $s \subseteq C_2$

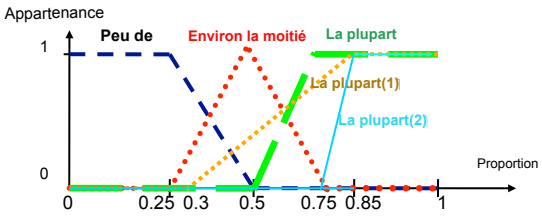
Avec windowSize étendue : ρ_{ws}=0.5, M = 6, $ws_\rho = 2$

Dans ce cas, $s \subseteq C_1$, $s \subseteq C_2$

46

Résultats obtenus

- ***Revues***
  IEEE Transactions of Fuzzy Sets, IJWET
- ***Conférences***
  FUZZ-IEEE, FLINS, TIME, EUSFLAT – BDA, LFA, EGC
- ***Projets***
  BPSolar (2003), Airtist (2006), KEOSIA (2007)
- ***Thèse*** : C. Fiot (2007)
  (avec A. Laurent, MCF UMII)
- ***DEA/Master*** :
  R. Baraër

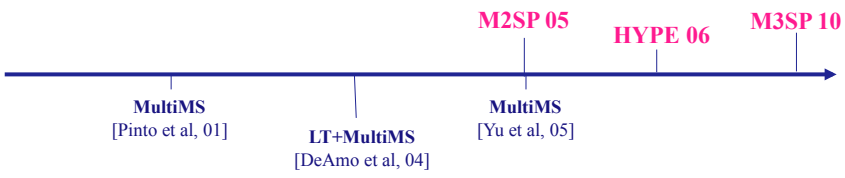


47

Motifs Multidimensionnels

- Comment offrir des connaissances plus riches à l'utilisateur ?
- Pourquoi se contenter d'une seule dimension ?

- M2SP, HYPE, M3SP
- Extraction de motifs sur plusieurs dimensions



48

Motifs Multidimensionnels

- Item multidimensionnel
  (France, c, 100)
- Séquence multidimensionnelle
  <{((France, c, 100), (Allemagne, d, 54)} {(, b, 2)}> au lieu de <(c, d),b>*
- Qu'est-ce qu'un client ?
  Une base peut être partitionnée en blocs selon différentes dimensions
- Partition des dimensions

$$D = D_T \oplus D_A \oplus D_R \oplus D_F$$

- Tuple c = (d1, d2, … dn) = (f, r, a, t)

49

Motifs Multidimensionnels

- D_R = {Cust-Grp, City}
- D_A = {Age, Product, Quantity}
- blocs définis sur D_R

Cust-Grp	City
Educ	NY
Educ	LA
Reti.	SF

Date	Cust-Grp	City	Age	Product	Qty
1	Educ	NY	Y	O	50
1	Educ	NY	M	P	2
2	Educ	NY	M	R	10
2	Educ	NY	M	P	3
1	Educ	LA	Y	C	30
2	Educ	LA	Y	C	20
3	Educ	LA	M	R	15
1	Reti.	SF	O	C	20
1	Reti.	SF	O	m	2

Partition de la base avec Cust-Grp et City

50

Motifs Multidimensionnels

- s=<{(Y,C,50)(M,P,2)},{(M,R,10)}>

Date	Cust-Grp	City	Age	Product	Qty
1	Educ	NY	Y	C	50
1	Educ	NY	M	P	2
2	Educ	NY	M	P	3
2	Educ	NY	M	R	10

Bloc(Educ,NY)

2 dates différentes : le bloc (Educ, NY) supporte la séquence s

Date	Cust-Grp	City	Age	Product	Qty
1	Educ	LA	Y	C	30
2	Educ	LA	Y	C	20
3	Educ	LA	M	R	15

Bloc(Educ,LA)

?

Ne contient pas (M,P,2) : le bloc (Educ, LA) ne supporte pas s

51

Motifs Multidimensionnels

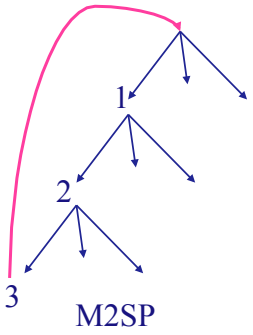
- Les items (Y,C), (M,C), (O,C) sont non fréquents mais ((Y ∨ M ∨ O), C) est fréquent
  - Utilisation d'un caractère joker * pour réduire les contraintes sur les dimensions d'analyse : (*,c) est un item α-joker
  - Recherche des plus spécifiques puis utilisation de *

→ (*,C) est fréquent

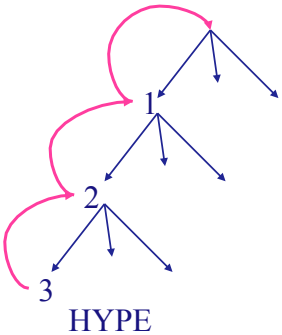
52

Résultats (1/2)

- De nouveaux types de motifs
  <{(surf, NY), (Housse, NY)}, {(combi, SF)}>
- M2SP : gestion « binaire » de la valeur de *
- HYPE : prise en compte des hiérarchies



M2SP



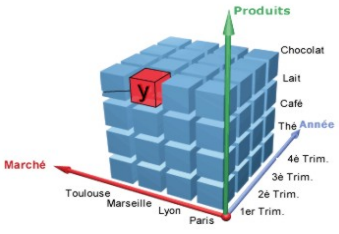
HYPE

53

Résultats obtenus

- ***Revues***
  ISI, ACM TKDD
- ***Conférences***
  PKDD, DOLAP – EGC,
  EDA, BDA
- ***Projet***
  EDF R&D (2007-2008)

- ***Thèse*** : M. Plantevit (2008)
  (avec A. Laurent, MCF UMII)
- ***DEA/Master*** :
  D. Jouve



54

Domaines d'application

- Web Mining, Text Mining, Schema Mining, Tree Mining, Stream Mining …
- Santé :
  - 3 partenariats (Inserm Bordeaux, Montpellier I et II)
  - 2 Programmes Exploratoires Pluridisciplinaires 2008
    - PEPS ST2I « GeneMining »
    - PEPS STI-SHS « Langage, Mémoire et Alzheimer »
- Environnement :
  - 2 partenariats (BP Solar, Fatronik)

55

Plan

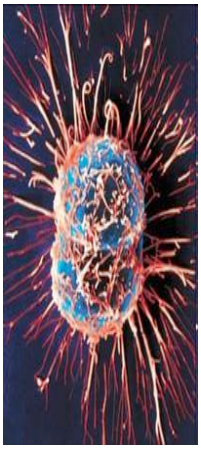
1. Règles d'association
2. Motifs séquentiels

3. Application 1 : Identification de marqueurs de la maladie d'Alzheimer

4. Application 2 : Détection d'anomalies dans les données ferroviaires

Maladie d'Alzheimer : problème majeur de la société moderne

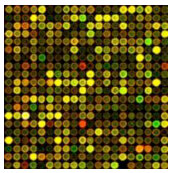
- **Maladie d'Alzheimer (AD) :** la forme la plus commune de démence
  - 26.6 millions de personnes atteintes de la maladie d'Alzheimer's (2006)
- **Augmentation du nombre de patients** (*4 en 2050) ➔ Intérêt de la communauté biomédicale pour la découverte des gènes impliqués dans le développement la maladie



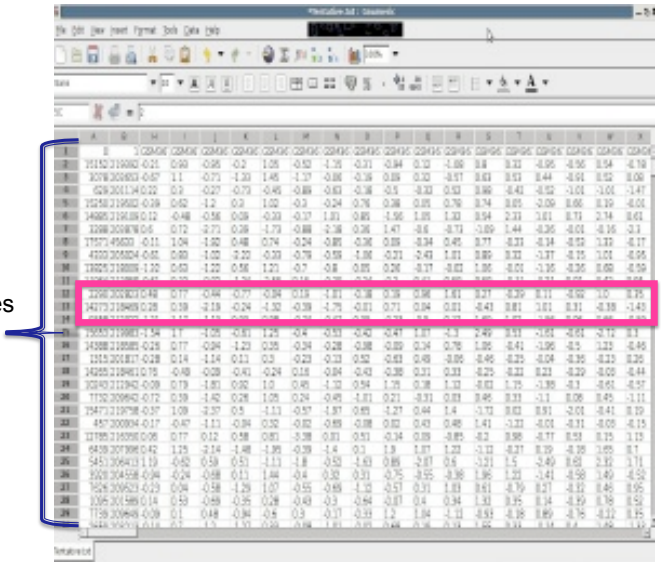
57

Puces à ADN

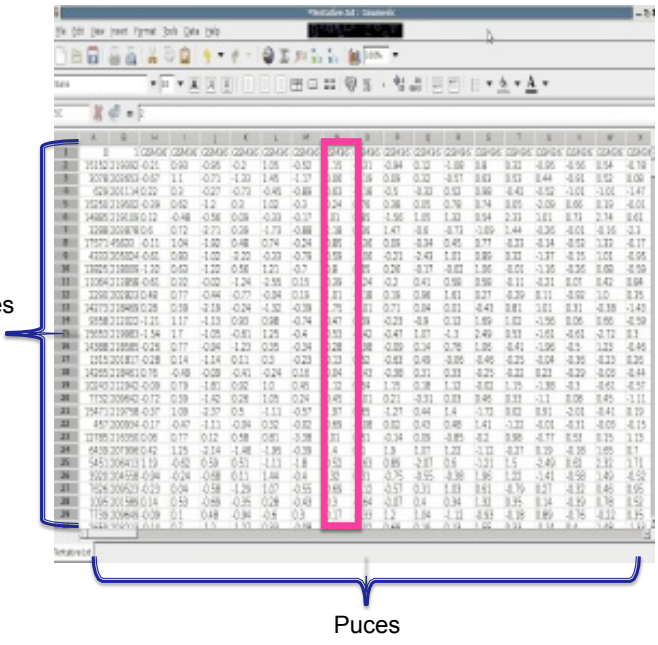
- Puces : pour comparer l'expression de millier de gènes dans différents tissus, cellules ou conditions
- Exploiter les analyses des puces à ADN pour obtenir un sens biomédical : challenge - à cause des gros volumes de données
- Importance des techniques de fouille : pour découvrir des connaissances jusqu'alors inconnues dans les gros volumes de données
  - ▶ Adaptations



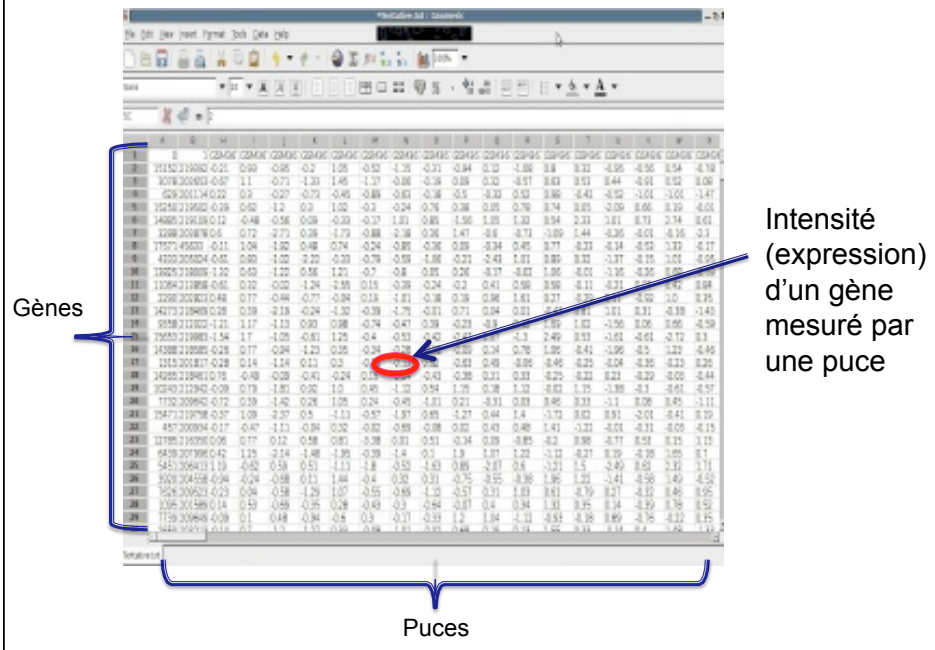
58



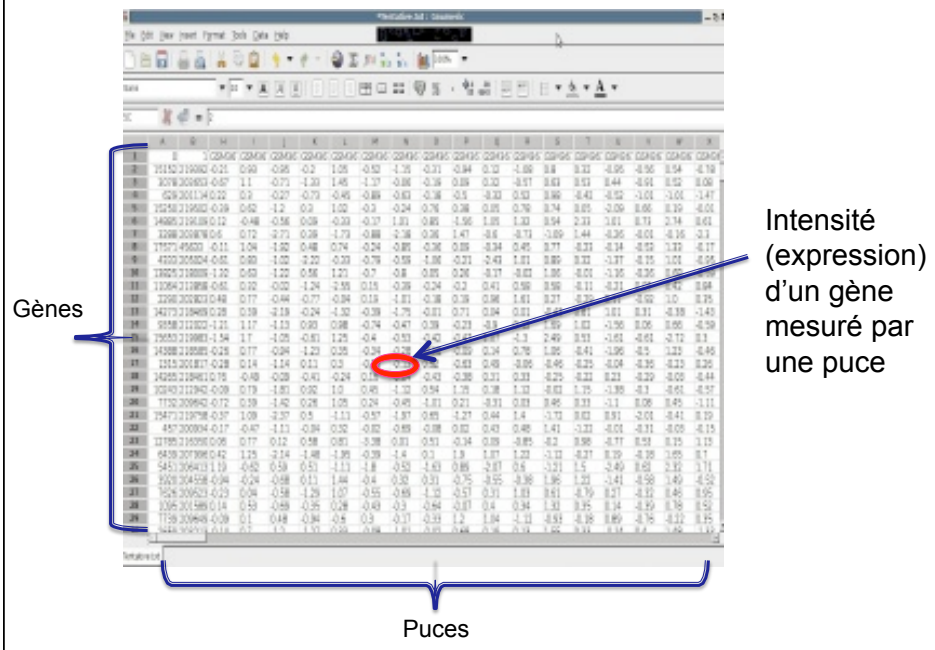
59



60

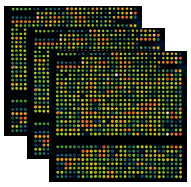


61

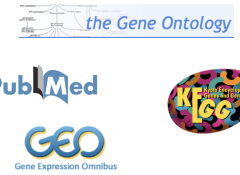


Très grande densité : Affymetrix U-133 plus 2.0 Array 54,675 probesets

Données biologiques massives



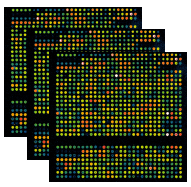
Technologies puces à ADN



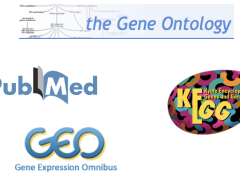
Bases de connaissances et bases bibliographiques disponibles en ligne

Nouvelles connaissances

Données biologiques massives



Technologies puces à ADN



Bases de connaissances et bases bibliographiques disponibles en ligne

Nouvelles connaissances

☹ Challenge : exploiter toutes ces données en terme de signification biologique

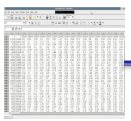
Projet GeneMining

- Collaboration

- **Objectif :** proposer un processus qui permette aux experts d'interpréter les données transcriptomiques
- **Application :** Déchiffrer les mécanismes du vieillissement et des pathologies associées (Alzheimer)
- **Données :**
  - Transcriptome du cortex temporel du Microcebus murinus
  - Puces Affymetrix

65

Processus général



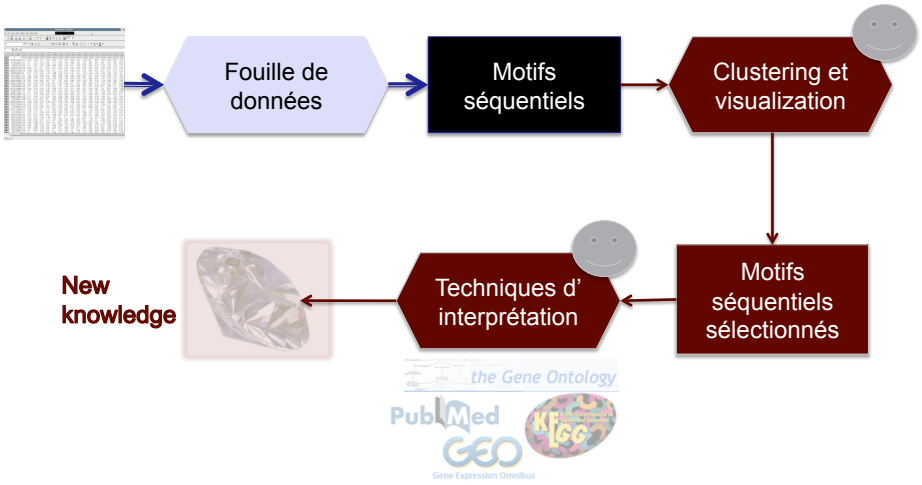
Fouille de données → Motifs séquentiels → Clustering et visualisation → Sélection de motifs séquentiels → Techniques d' interprétation → Nouvelles connaissances biologiques

the Gene Ontology

PubMed GEO Gene Expression Omnibus

66

Processus général



67

Recherche de motifs séquentiels (Salle, AIME 2009)

Puces	Séquences
M1	<(G2)(G1 G5)(G3)(G4)>
M2	<(G2)(G1 G5)(G4)(G3)>
M3	<(G2)(G4)(G1 G5)(G3) >
M4	<(G2)(G3)(G1 G5)(G4)>

<(G2)(G1 G5)(G3)>

- Le gène G2 a une expression plus petite que les gènes G1 et G5 qui ont une expression similaire et plus petite que le gène G3

68

Recherche de motifs séquentiels (Salle, AIME 2009)

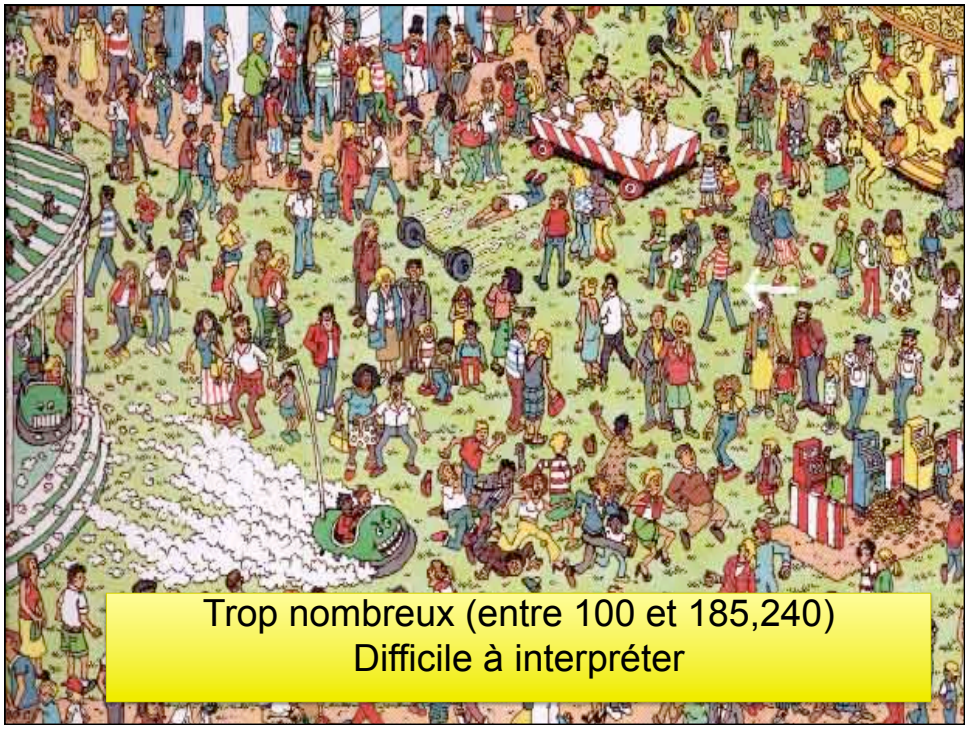
Puces	Séquences
M1	<(G2)(G1 G5)(G3)(G4)>
M2	<(G2)(G1 G5)(G4)(G3)>
M3	<(G2)(G4)(G1 G5)(G3) >
M4	<(G2)(G3)(G1 G5)(G4)>

<(G2)(G1 G5)(G3)>

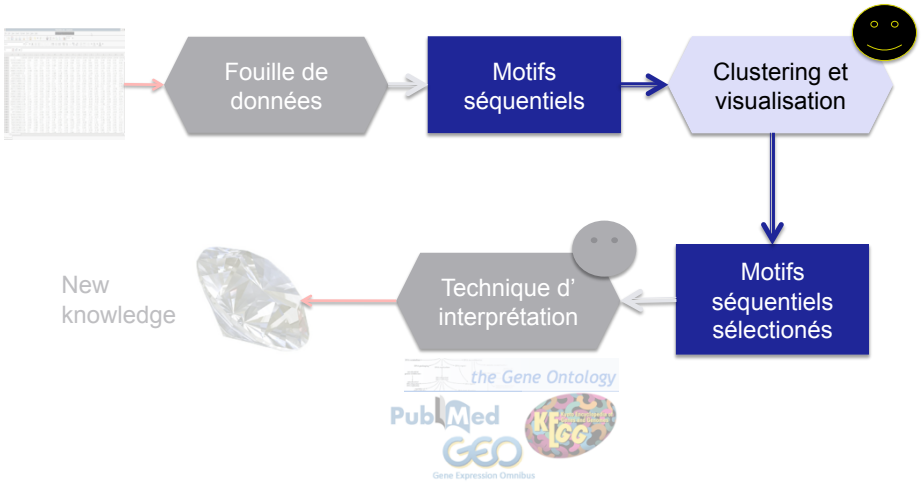
Support: 3/4

- Motifs séquentiels discriminants
  - Fréquents dans une classe (malades)
  - Non fréquents dans la classe complémentaire (sains)

69



Processus général



71

Comment comparer les motifs ?

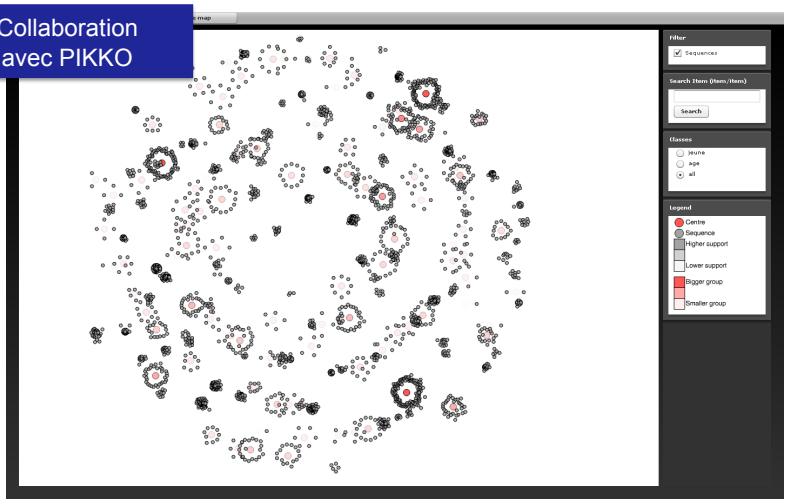
S75%=<(G1)(G2 G3)>
S'75%=< (G2 G3) (G1)>

- Mesure de similarité [Saneifar et al., AusDM'08]
  - Gènes communs et non communs
  - Ordre des gènes
  - Support

72

Clustering simple (k-means)

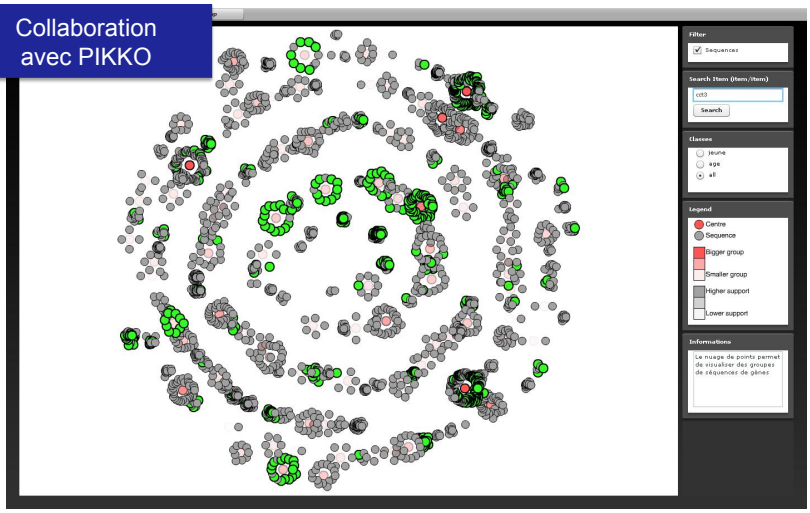
Collaboration avec PIKKO



73

Clustering simple (k-means)

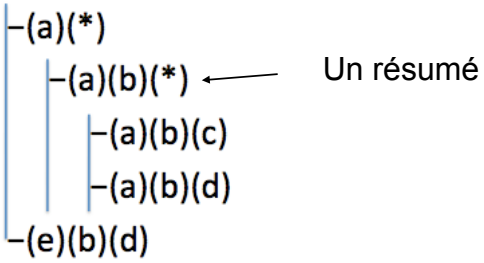
Collaboration avec PIKKO



74

Clustering hiérarchique

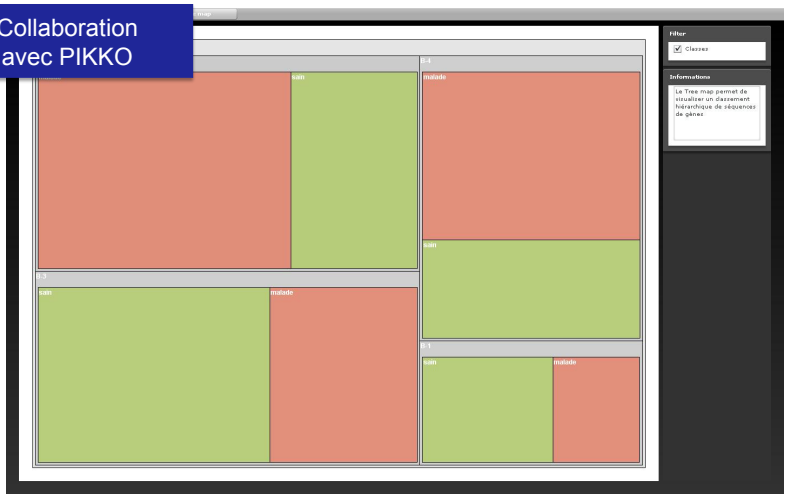
- Méthode de clustering hierarchique [Nin Guerero et al., AusDM'08]
- Exemple: (a)(b)(c), (a)(b)(d), (e)(b)(d)



75

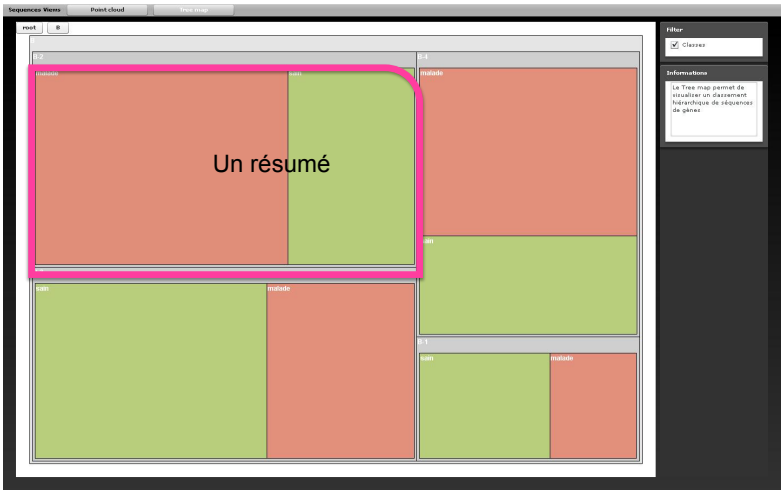
Clustering hiérarchique

Collaboration avec PIKKO



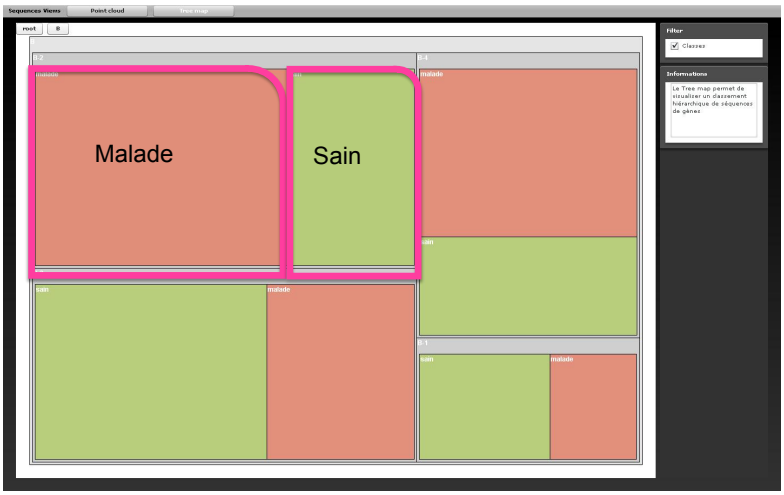
76

Clustering hiérarchique

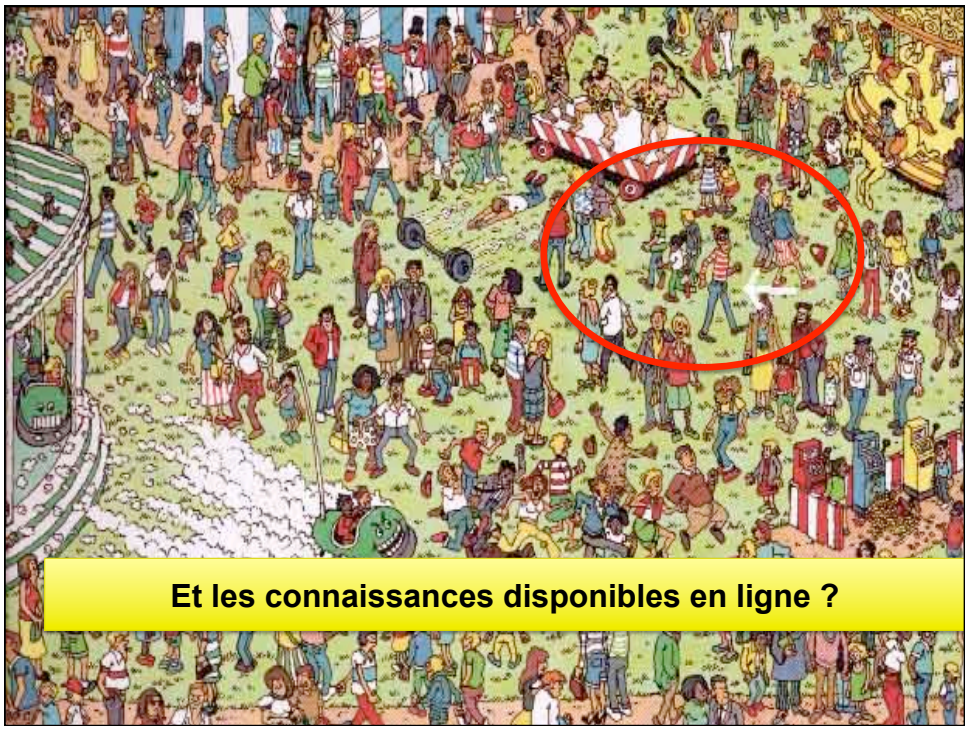


77

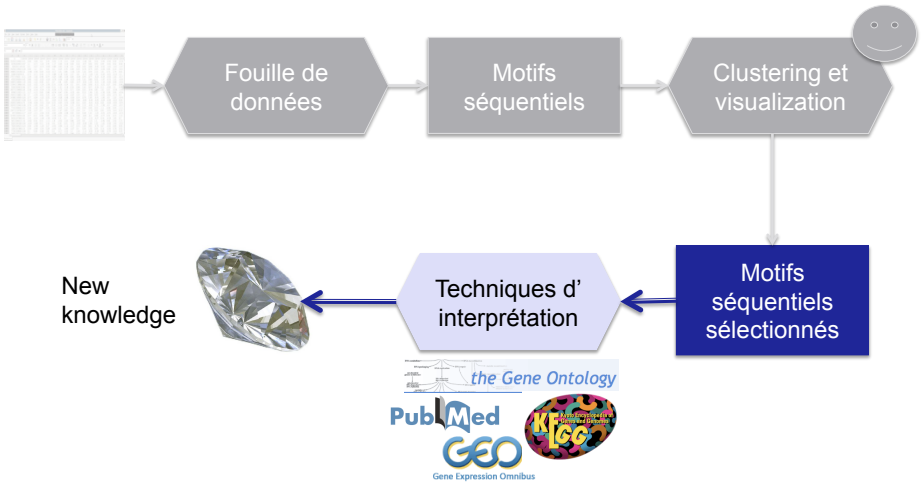
Clustering hiérarchique



78



Processus général



80

Interprétation des motifs grâce à documents

S75%,25%=<(G1)(G2 G3)> → Texts PubMed

- Recherche de documents associés avec les gènes des motifs

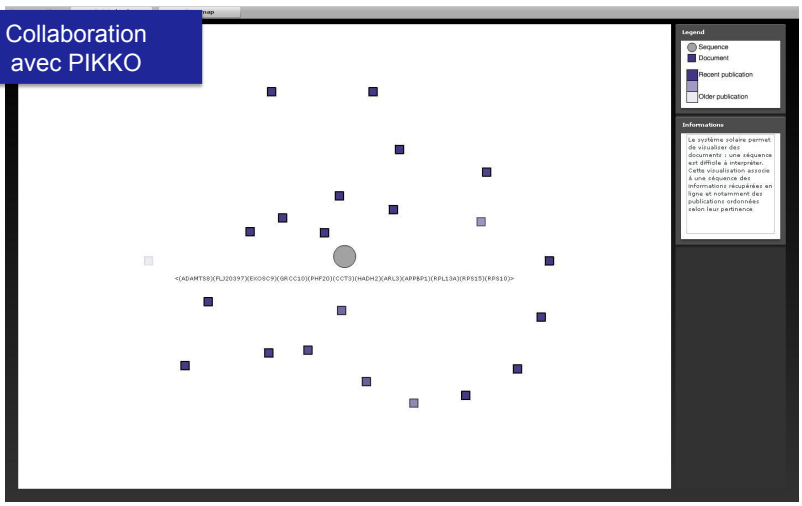
- **Objectifs:** validation + recherche de nouveautés

Séquences populaires and innovantes

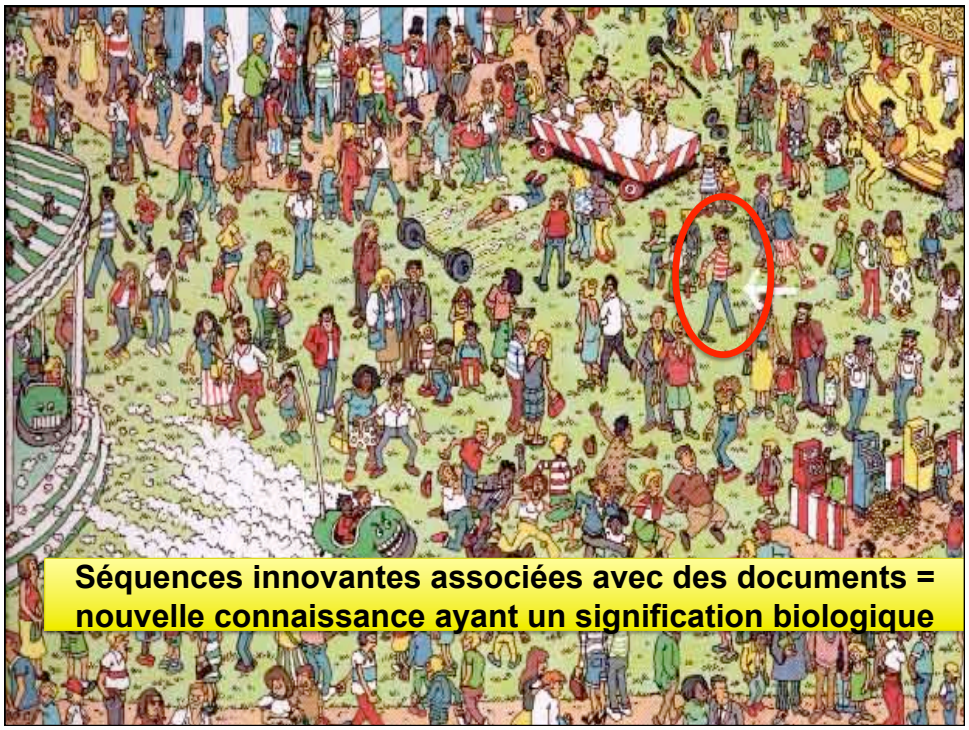
81

Visualisation de documents

Collaboration avec PIKKO



82



Un motif pertinent

S_{75}=<(MRVI1)(PGAP1)(PLA2R1)(A2M)(GSK3B)>

- Protéines impliquées dans les mécanismes de signalisation et du métabolisme
- Certaines interfèrent avec les événements cellulaires de la maladie d'Alzheimer

84

Conclusion

- Processus en 3 étapes
- Découverte de nouvelles connaissances à partir de données transcriptomiques ayant une signification biologique
- Avancés en terme de fouille et de biologie

85

Perspectives

- **Améliorer chaque étape du processus**
  - D'autres types de motifs (Fuzzy patterns)
  - D'autres méthodes de clustering
- **Généraliser** ce processus à d'autres types de données massives
- Utilisation des motifs séquentiels pour la **prédiction**
- http://www.lirmm.fr/tatoo/spip.php?page=prototypes

86

Plan

1. Règles d'association
2. Motifs séquentiels
3. Application 1 : Identification de marqueurs de la maladie d'Alzheimer

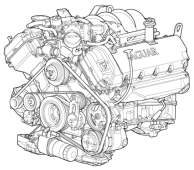
4. Application 2 : Détection d'anomalies dans les données ferroviaires

Maintenance ferroviaire

- Mieux comprendre le comportement de systèmes complexes (trains, machines outils, réseaux routiers, etc)
- Des problèmes de maintenance

fatronik francetecnalia

- Quelles solutions ?



Maintenance curative

- Maintenance après une panne
  - Très coûteuse
    - Quelques heures d'arrêt d'une chaîne automobile peuvent coûter très cher (millions de dollars de pertes)
  - Problème de sécurité
    - 5% des accidents des véhicules à moteurs sont causés par un mauvais fonctionnement des équipements ou un manque de maintenance
  - Perte d'énergie

Maintenance systématique

- Programmée à date fixe
  - Réalisée régulièrement pour anticiper les problèmes
  - Trop coûteuse

Maintenance préventive

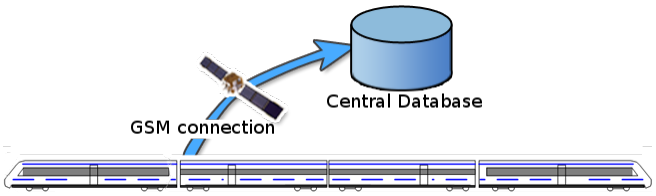
- Pour rendre la maintenance plus rapide et efficace
  - Contrôler les comportements (monitoring)
  - Détecter les comportements anormaux
  - Prévenir les pannes

- Problème général : comment détecter automatiquement les comportements anormaux et guider l'expert dans sa tâche de maintenance

Du monitoring à la maintenance…

- Collecte des données : difficiles à exploiter
  - Volumineuses
  - Bruitées
  - Avec des erreurs de transmission
  - Hétérogènes
- Maintenance préventive
  - Manque de connaissances sur les comportements normaux
  - Comment définir une anomalie ?

Données ferroviaires



- Capteurs : A, B, C…
- Mesures : le capteur A mesure la valeur 82.5 au temps 06:41:39

TIME	A	B	C	...
2008/03/27 06: 36: 39	0	16	16	...
2008/03/27 06: 41: 39	82.5	16	16	...
2008/03/27 06: 46: 38	135.6	19	21	...
2008/03/27 06: 51: 38	105	22	25	...

Caractérisation des comportements

- Qu'est-ce qu'un comportement normal ?
  - Étiqueté par un expert
  - Comme ne contenant pas d'anomalie
- Comment les caractériser ?
  - En décrivant les comportement fréquents dans des données historisées et étiquetées comme normales par des experts

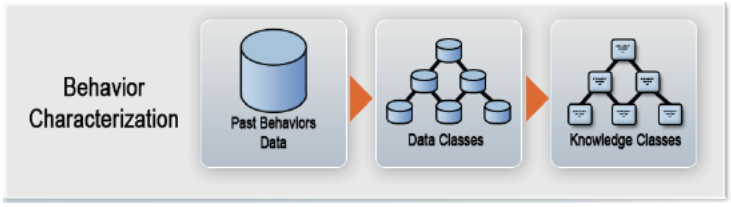
… Avec des motifs séquentiels

Définitions

- Item : un capteur et sa valeur discrétisée à un moment donné
  - A_{low} : le capteur A mesure une valeur basse
- Itemset : un ensemble non ordonné d'items
  - (A_{low}, B_{avg}) : A un certain moment, A mesure une valeur basse et B mesure une valeur moyenne
- Séquence : une liste ordonnée d'itemsets
  - $<(A_{low}, C_{low})\ (B_{avg})\ (A_{high}, C_{avg})>$: A et C mesurent une valeur basse PUIS B mesure une valeur moyenne, PUIS A…
- Séquence fréquente (minsup = 60%) :
  - Si plus de 60% des trajets contiennent la séquence précédente

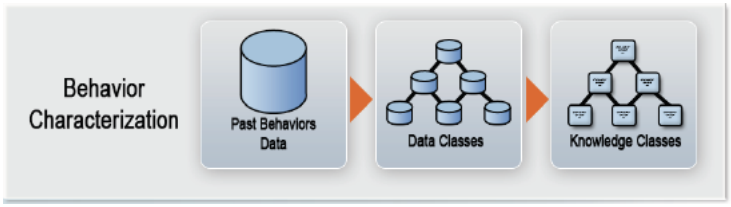
Prise en compte du contexte (Rabatel, ICDM 2009)

- Météo, température extérieure…
- Données indexées en fonction de classes de contexte
  - (Tlow,Hhigh), (Tlow,*)
  - Notion de généralisation => un treillis



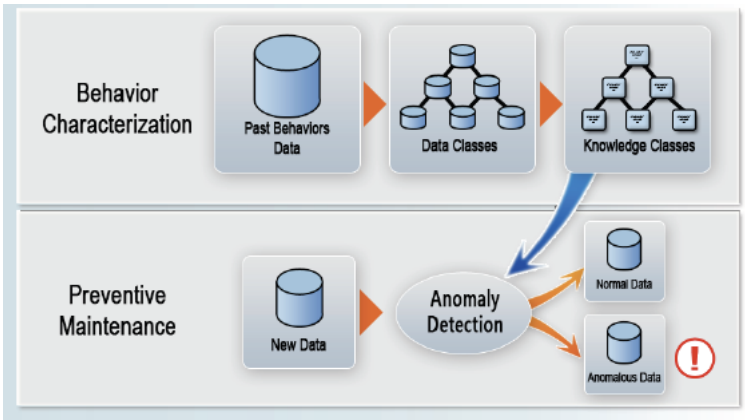
Prise en compte du contexte (Rabatel, ICDM 2009)

- Les motifs sont extraits dans chaque classe de données
- Chaque classe décrite par un ensemble de motifs
- Questions ?
  - Quels sont les comportements qui apparaissent uniquement quand la température est haute ?
  - Quels sont les comportements indépendants du contexte ?



Détection d'anomalies (Rabatel, ICDM 2009)

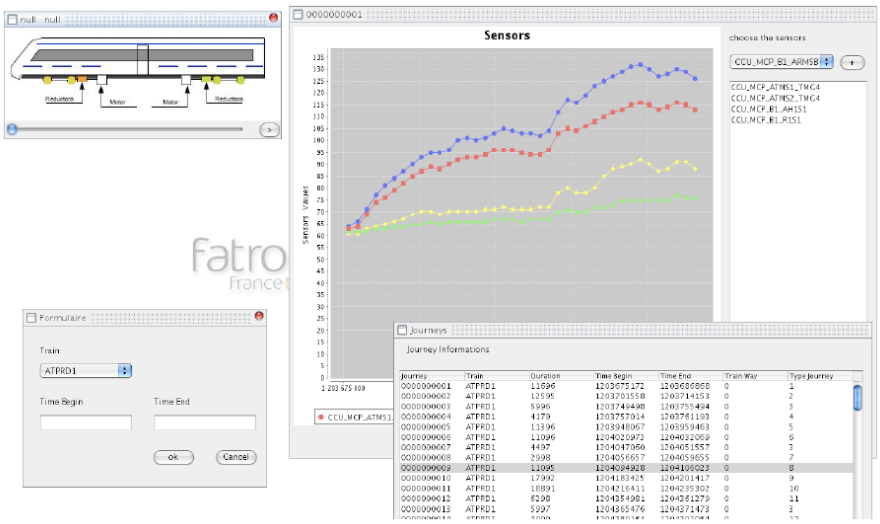
- Arrivée d'un nouveau trajet
- Le comparer avec la classe associée (la plus spécialisée)



Score de conformité (Rabatel, ICDM 2009)

- Pour chaque capteur, pour chaque mesure…
- Score entre -1 et 1 :
  - -1 comportement anormal
  - 0 comportement incertain
  - 1 comportement normal
- Ce score prend en compte :
  - le nombre de séquences dans la classe
  - la taille des séquences
  - la position de la classe dans le treillis

Outil de visualisation



Perspectives

- Détecter des tendances dans l'évolution des données
  - Mettre à jour les données dans la base
    - Prendre en compte le vieillissement du matériel
- Optimiser le problème de comparaisons entre la base de connaissances et les nouvelles données d'un train
  - Clustering de séquences
  - …
- Prédiction en temps réel

Plan

1. Règles d'association
2. Motifs séquentiels
3. Application 1 : Identification de marqueurs de la maladie d'Alzheimer
4. Application 2 : Détection d'anomalies dans les données ferroviaires

5. Conclusion

Travaux en cours

Des motifs fréquents ? Oui, mais …

- Dans un contexte de données difficiles (données médicales), comment proposer des connaissances
  - Données patients + résultats d'analyse
  - Questionnaire + Diagnostic
  - Règles graduelles (items graduels, motifs graduels, cubes graduels)

Thèse de Lisa Di-Jorio (Co-encadrement A. Laurent UM2)

103

Travaux en cours

Des motifs fréquents ? Oui, mais …

- Dans un contexte d'aide à la décision il est important de connaître les comportements atypiques
  - Capteurs défectueux, comportements différents, données incomplètes, …
  - Par rapport à l'ensemble de la base (*outliers*)
  - Par rapport à une base de croyances (*motifs inattendus*)

Thèse de Cécile Low-Kam (Co-encadrement A. Mas UM2)

104

Conclusions

- Gros volume de données
- Données estampillées (notion d'ordre) : Motifs séquentiels
- Données complexes ou agrégées : Bases de Données multi dimensionnelles (cubes de données)
- Résultats (données) approximatifs : Logique floue

105