Fouille de Données Multidimensionnelles :
Différentes Stratégies pour Prendre en Compte la Mesure

Marc Plantevit, Anne Laurent et Maguelonne Teisseire

Université Montpellier 2

Laboratoire d'Informatique de Robotique et de
Micro-électronique de Montpellier

5 Juin 2008

Plan

Les enjeux de la fouille de données

De plus en plus de données

- Avènement des TIC
- Volumes de données brassés devenus énormes

Décideurs

- Croulent sous l'information
- Difficile d'avoir une bonne vision des données
- Des marchés de plus en plus concurrentiels

La fouille de données

- Extraire des connaissances dans des grands volumes de données
- Une des technologies émergentes qui « changeront le monde » au $XXI^{ème}$ siècle (MIT)

Les motifs séquentiels multidimensionnels

Extraction de corrélations dans des données

- multidimensionnelles
- organisées suivant une relation d'ordre (e.g. temps)
- etc.

Plan

Partition des Dimensions

Soit $\mathscr{D}$, l'ensemble des dimensions d'une table DB, on considère la partition de $\mathscr{D}$:

- D_A : l'ensemble des dimensions d'*analyse*
- D_R : l'ensemble des dimensions de *référence*
- D_T : l'ensemble des dimensions *temporelles*
- D_I : l'ensemble des dimensions *ignorées*

Division en bloc

Etant donnée une base de données DB, on appelle *bloc* l'ensemble des n-uplets ayant la même valeur r sur D_R. L'ensemble des blocs de DB est noté B_{DB,D_R}. Ainsi chaque bloc B_r de B_{DB,D_R} est décrit pas le n-uplet r qui le définit.

$D_I = \emptyset, D_R = \{\text{CID}\}, D_T = \{Date\}$ et $D_A = \{City, Cust\text{-}Grp, A\text{-}Grp, Product\}$

CID	Date	City	Customer Informations		Product
			Cust-Grp	*Cust-Age*	
C_1	1	*NY*	*Educ.*	*Middle*	*A*
C_1	1	*NY*	*Educ.*	*Middle*	*B*
C_1	2	*LA*	*Educ*	*Middle*	*C*
C_2	1	*SF*	*Prof.*	*Middle*	*A*
C_2	2	*SF*	*Prof.*	*Middle*	*C*
C_3	1	*DC*	*Business*	*Retired*	*A*
C_3	1	*LA*	*Business*	*Retired*	*B*

$D_I = \emptyset, D_R = \{CID\}, D_T = \{Date\}$ et $D_A = \{City, Cust\text{-}Grp, A\text{-}Grp, Product\}$

CID	Date	City	Customer Informations		Product
			$Cust\text{-}Grp$	$Cust\text{-}Age$	
C_1	1	NY	$Educ.$	$Middle$	A
C_1	1	NY	$Educ.$	$Middle$	B
C_1	2	LA	$Educ$	$Middle$	C
C_2	1	SF	$Prof.$	$Middle$	A
C_2	2	SF	$Prof.$	$Middle$	C
C_3	1	DC	$Business$	$Retired$	A
C_3	1	LA	$Business$	$Retired$	B

Item, itemset et séquence

Item multidimensionnel

$e = (d_1, d_2, \ldots, d_m)$ t.q
$d_i \in Dom(D_i) \cup \{*\}, \forall D_i \in D_A$ et où $*$ joue le rôle de valeur *joker*.

- $(NY, Educ., Middle, A)$
- $(*, *, Young, A)$

Itemset multidimensionnel

$i = \{e_1, \ldots, e_k\}$ est un ensemble non vide d'items multidimensionnels.

$\{(NY, *, M, B), (*, Educ, Y, C)\}$

Séquence multidimensionnelle

$s = \langle i_1, \ldots, i_l \rangle$ est une liste ordonnée d'itemsets multidimensionnels

$\langle \{(NY, *, M, A), (*, Educ, Y, B)\}$
$\{(*, *, M, C)\} \rangle$

Support

Séquence et Bloc

Un bloc de données B_r supporte une séquence $\varsigma = \langle i_1, \ldots, i_l \rangle$ si :

- $\forall j = 1 \ldots l$, $\exists d_j \in Dom(D_t)$, $\forall e = (a_{i_1}, \ldots, a_{i_m}) \in i_j$, $\exists c = (f, r, (x_{i_1}, \ldots, x_{i_m}), d_j) \in B_r$ avec $a_{i_k} = x_{i_k}$ ou $a_{i_k} = *$
- $d_1 < d_2 < \cdots < d_l$.

- Le *support absolu* d'une séquence multidimensionnelle s dans une base de données DB correspond au nombre de blocs de B_{DB,D_R} qui contiennent s.
- Le *support relatif* correspond au pourcentage de blocs de B_{DB,D_R} qui contiennent s ($\frac{absolute\_support(S)}{|B_{DB,D_R}|}$).

$$support(\langle\{(*,*,Middle,A)\}\{((*,*,Middle,C)\}\rangle) = 2 \text{ ou } \frac{2}{3}$$

CID	Date	City	Customer Informations		Product
			Cust-Grp	*Cust-Age*	
C_1	1	*NY*	*Educ.*	*Middle*	*A*
C_1	1	*NY*	*Educ.*	*Middle*	*B*
C_1	2	*LA*	*Educ*	*Middle*	*C*
C_2	1	*SF*	*Prof.*	*Middle*	*A*
C_2	2	*SF*	*Prof.*	*Middle*	*C*
C_3	1	*DC*	*Business*	*Retired*	*A*
C_3	1	*LA*	*Business*	*Retired*	*B*

Définition du problème

Séquence fréquente

Etant donné un seuil de support fixé a priori par l'utilisateur, noté σ $(0 < \sigma \leq 1)$, une séquence s est *fréquente* sur la base de données DB si $relative\_support(S) \geq \sigma$.

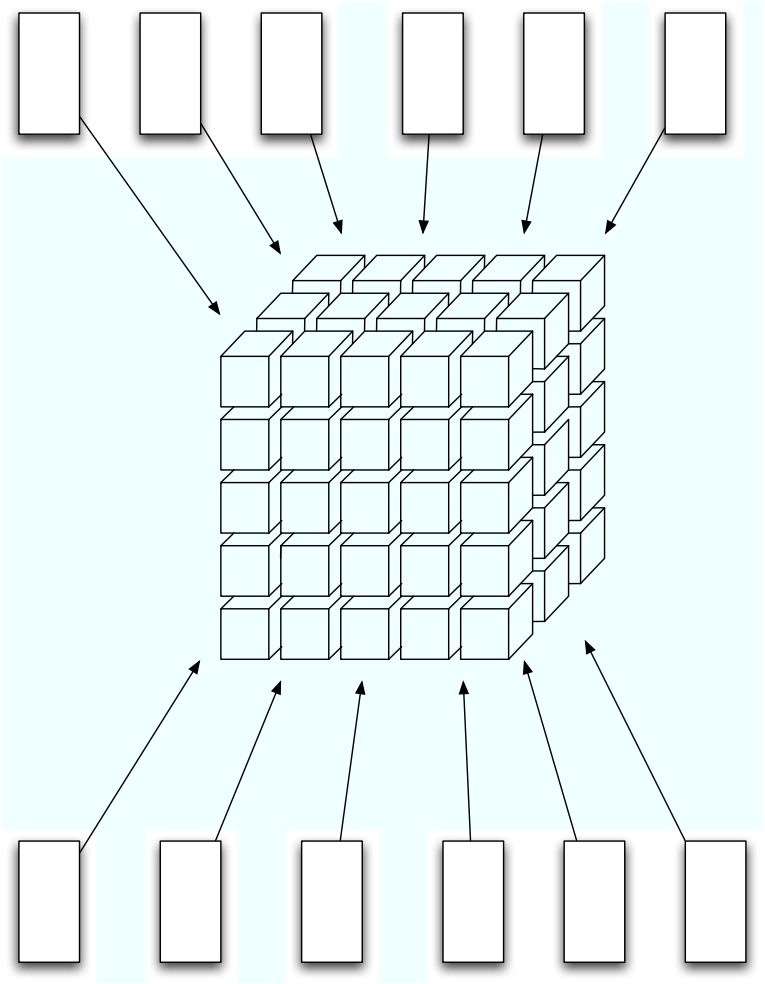
Problématique de l'extraction de motifs séquentiels multidimensionnels

Le but de l'extraction des motifs séquentiels multidimensionnels est de découvrir l'ensemble complet des séquences fréquentes, étant donnés une base de données DB et un seuil de support minimum σ.

Plan

Agrégation des données

- A des fins d'analyse
- Apparition de *mesure(s)*



CID	Date	City	Customer Informations		Product
			Cust-Grp	*Cust-Age*	
C_1	1	*NY*	*Educ.*	*Middle*	*A*
C_1	1	*NY*	*Educ.*	*Middle*	*B*
C_1	2	*LA*	*Educ*	*Middle*	*C*
C_2	1	*SF*	*Prof.*	*Middle*	*A*
C_2	2	*SF*	*Prof.*	*Middle*	*C*
C_3	1	*DC*	*Business*	*Retired*	*A*
C_3	1	*LA*	*Business*	*Retired*	*B*

Date	City	Customer Informations		Product	Mesure
1	*NY*	*Educ.*	*Middle*	*A*	123
1	*NY*	*Educ.*	*Middle*	*B*	234
2	*LA*	*Educ*	*Middle*	*C*	120
1	*SF*	*Prof.*	*Middle*	*A*	125
2	*SF*	*Prof.*	*Middle*	*C*	115
1	*DC*	*Business*	*Retired*	*A*	1
1	*LA*	*Business*	*Retired*	*B*	24

Une gestion symbolique qui entraîne des incohérences

Date	City	Customer Informations		Product	Mesure
1	*NY*	**Educ.**	*Middle*	*A*	123
1	*NY*	**Educ.**	*Middle*	*B*	234
2	*LA*	**Educ.**	*Middle*	*C*	120
1	*SF*	**Prof.**	*Middle*	*A*	125
2	*SF*	**Prof.**	*Middle*	*C*	115
1	*DC*	**Business**	*Retired*	*A*	1
1	*LA*	**Business**	*Retired*	*B*	24

- $support(\langle\{(*, M, A, 125)\}\rangle) = 1$. 123 et 125 sont considérés comme totalement différents.
- $support(\langle\{(*, *, A, *)\}\rangle) = 3$. $(*, *, A, 125)$ et $(*, *, A, 1)$ ont le *même impact* dans le calcul du support de la séquence $\langle\{(*, *, A, *)\}\rangle$.

Plan

Etat de l'art

La présence de valeurs numériques pour des « approches symboliques » est un problème relativement étudié.

- A. Laurent : une architecture basée sur les bases de données multidimensionnelles floues pour générer des résumés flous.
- D. Dubois et al. : règles d'association sur des attributs numériques.
- R. Ben Messaoud et al. : la mesure pour calculer différentes mesures d'intérêt
- C. Fiot et al. : théorie des sous ensembles flous pour l'extraction de motifs séquentiels

Contraintes d'agrégation sur la mesure

- Ne considérer que les cellules dont la mesure associée respecte la *condition d'iceberg*
- Ex : Cellules dont la mesure est supérieure à 50

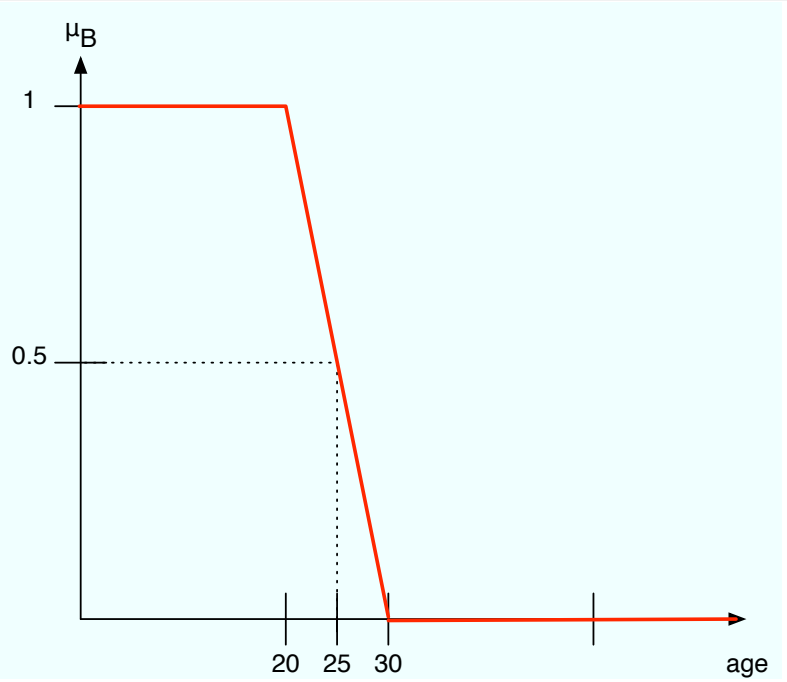
Date	City	Customer Informations		Product	Mesure
1	*NY*	**Educ.**	*Middle*	*A*	123
1	*NY*	**Educ.**	*Middle*	*B*	234
2	*LA*	**Educ.**	*Middle*	*C*	120
1	*SF*	**Prof.**	*Middle*	*A*	125
2	*SF*	**Prof.**	*Middle*	*C*	115
1	*DC*	**Business**	*Retired*	*A*	1
1	*LA*	**Business**	*Retired*	*B*	24

Partitionnement du domaine de la mesure

$M \in D_A$

- Discrétiser cette dimension :
  - equi-depth
  - equi-width
- partition stricte ou floue

Person	Age	degree of youth
Johan	10	1.00
Edwin	21	0.90
Parthiban	25	0.50
Arosha	26	0.40
Chin Wei	28	0.20
Rajkumar	83	0.00



Date	City	Customer Informations		Product	Mesure		
					little	*medium*	*lot*
1	*NY*	**Educ.**	*Middle*	*A*	0	1	0
1	*NY*	**Educ.**	*Middle*	*B*	0	0	1
2	*LA*	**Educ.**	*Middle*	*C*	0	1	0
1	*SF*	**Prof.**	*Middle*	*A*	0	1	0
2	*SF*	**Prof.**	*Middle*	*C*	0	1	0
1	*DC*	**Business**	*Retired*	*A*	1	0	0
1	*LA*	**Business**	*Retired*	*B*	1	0	0

Date	City	Customer Informations		Product	Mesure		
					little	*medium*	*lot*
1	*NY*	**Educ.**	*Middle*	*A*	0	1	0
1	*NY*	**Educ.**	*Middle*	*B*	0	0.2	0.8
2	*LA*	**Educ.**	*Middle*	*C*	0	1	0
1	*SF*	**Prof.**	*Middle*	*A*	0	1	0
2	*SF*	**Prof.**	*Middle*	*C*	0.1	0.0	0
1	*DC*	**Business**	*Retired*	*A*	1	0	0
1	*LA*	**Business**	*Retired*	*B*	0.9	0.1	0

Utiliser la mesure pour calculer directement le support

- Agrégation : un pré-calcul du support ?
- Eviter les pré-traitements dus aux partitionnements

Deux philosophies :

1. « Tous les blocs sont égaux » (vote démocratique)
2. L'impact du bloc dans le calcul du support est fonction de son importance (IdF vs Lozère)

Gérer l'antimonotonie du support

t-norme $\otimes$ pour garantir l'antimotonie

Généralisation de la conjonction logique

- Opérateur $[0,1] \times [0,1] \rightarrow [0,1]$ qui est associatif, commutatif, monotone.
- Satisfait les conditions s $\alpha \otimes 0 = 0$ et $\alpha \otimes 1 = \alpha$.
- **minimum** $(\alpha, \beta) \mapsto \min(\alpha, \beta)$, le produit $(\alpha, \beta) \mapsto \alpha\beta$ et la t-norme de Lukasiewicz $(\alpha, \beta) \mapsto \max(\alpha + \beta - 1, 0)$.

t-conorme $\oplus$ pour exhiber la meilleure solution

- Opérateur $[0,1] \times [0,1] \rightarrow [0,1]$ qui est associatif, commutatif, monotone.
- Satisfait les conditions $\alpha \oplus 1 = 1$ et $\alpha \oplus 0 = \alpha$.
- **maximum** $(\alpha, \beta) \mapsto \max(\alpha, \beta)$, la somme probabiliste $(\alpha, \beta) \mapsto \alpha + \beta - \alpha\beta$, la somme bornée $(\alpha, \beta) \mapsto \min(\alpha + \beta, 1)$, etc.

Définitions

Micro Count

Soit une g-k-séquence $s = \langle s_1, s_2, \ldots, s_g \rangle$, le support relatif de s dans un cube de données DB avec la technique micro count est égal à :

$$\text{Relative support}(s) = \sum_{B_r \in B_{DB,D_R}} \bigoplus \bigotimes_{s_i \in s} \bigotimes_{s_{i_j} \in s_i} \frac{(m[B_r, s_{i_j}])}{m[(*,*,\ldots,*)]}$$

Macro Count

Soit une g-k-séquence $s = \langle s_1, s_2, \ldots, s_g \rangle$, le support relatif de s dans un cube de données DB avec la technique macro count est égale à :

$$\text{Relative support}(s) = \frac{1}{|B_{DB,D_R}|} \times \sum_{B_r \in B_{DB,D_R}} \bigoplus \bigotimes_{s_i \in s} \bigotimes_{s_{i_j} \in s_i} \frac{(m[B_r, s_{i_j}])}{m[(r,*,\ldots,*)]}$$

D_R	D_A			M
Educ.	*	*	*	477
Prof.	*	*	*	240
Business	*	*	*	25
*.	*	*	*	742

« Démocratie vs Poids »

$B_{Educ.}$, $B_{Prof.}$ et $B_{Business}$ ont des influences respectives d'environ 64%, 32% et 4% dans le calcul MicroCount du support d'une séquence alors qu'avec le comptage MacroCount, ils ont tous des influences égales.

1	*NY*	*Middle*	*A*	123
1	*NY*	*Middle*	*B*	234
2	*LA*	*Middle*	*C*	120

1	*SF*	*Middle*	*A*	125
2	*SF*	*Middle*	*C*	115

1	*DC*	*Retired*	*A*	1
1	*LA*	*Retired*	*B*	24

Séquences	MicroCount	MacroCount
$\langle\{(*,*,A)\}\rangle$	0.34	0,27
$\langle\{(*,Middle,*)\}\rangle$	0,65	0,42
$\langle\{(*,*,A)\},\{(*,Middle,*)\}\rangle$	0.32	0,24
$\langle\{(*,*,A),(*,*,B)\}\rangle$	0,17	0,10

$\langle\{(*,*,A)\}\rangle$

1	*NY*	*Middle*	*A*	123
1	*NY*	*Middle*	*B*	234
2	*LA*	*Middle*	*C*	120

1	*SF*	*Middle*	*A*	125
2	*SF*	*Middle*	*C*	115

1	*DC*	*Retired*	*A*	1
1	*LA*	*Retired*	*B*	24

- Pour MicroCount, on divise la mesure des cellules par la mesure totale : $\frac{125+123+1}{742} = 0,34$.
- Pour MacroCount, on divise la mesure de chaque cellule par la mesure associée au bloc contenant la cellule, puis on calcule la moyenne : $\frac{\frac{123}{477}+\frac{125}{240}+\frac{1}{25}}{3} = 0,27$.

$\langle\{(*, Middle, *)\}\rangle$

1	*NY*	*Middle*	*A*	123
1	*NY*	*Middle*	*B*	234
2	*LA*	*Middle*	*C*	120

1	*SF*	*Middle*	*A*	125
2	*SF*	*Middle*	*C*	115

1	*DC*	*Retired*	*A*	1
1	*LA*	*Retired*	*B*	24

- La séquence apparaît plusieurs fois dans deux blocs.
- Il faut retenir la meilleure solution : t-conorme
- Pour MicroCount, le support de la séquence $\langle\{(*, Middle, *)\}\rangle$ est égal à : $\frac{\max(357,120)+\max(125,115)}{742} = 0,65$.

$\langle\{(*,*,A)\},\{(*,Middle,*)\}\rangle$

1	NY	Middle	A	123
1	NY	Middle	B	234
2	LA	Middle	C	120

1	SF	Middle	A	125
2	SF	Middle	C	115

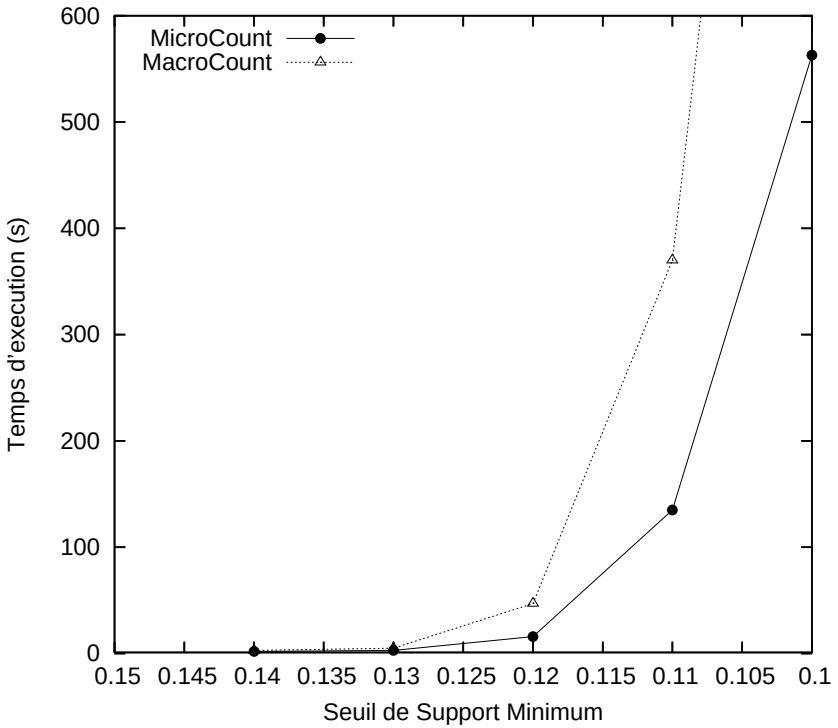
1	DC	Retired	A	1
1	LA	Retired	B	24

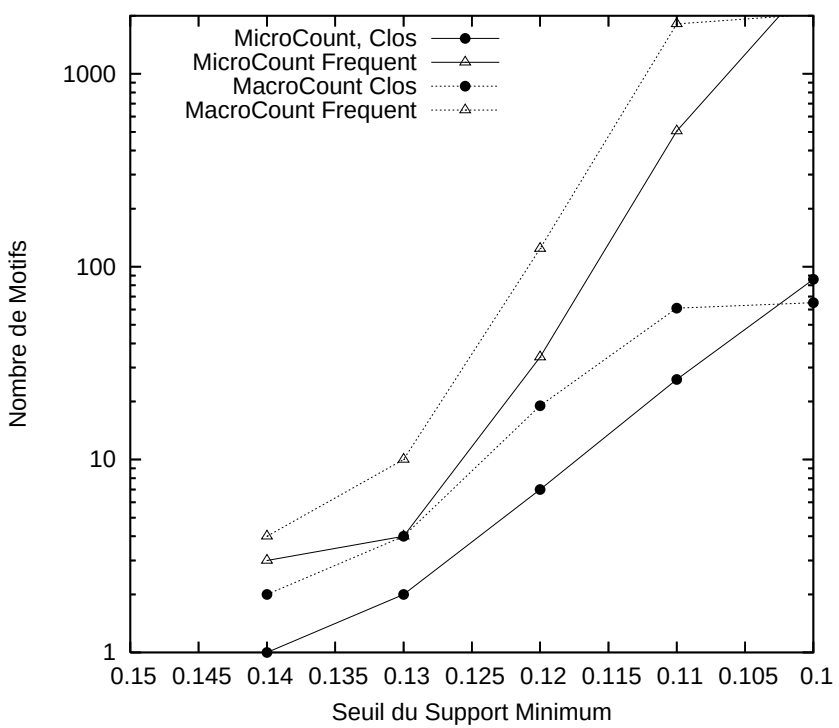
- Une seule combinaison est possible pour chacun de ces deux blocs. .
- La t-norme (min) va nous permettre de garantir l'antimonotonie du support.
- Support de la séquence $\langle\{(*,*,A)\},\{(*,Middle,*)\}\rangle$ doit être inférieur ou égal aux supports des séquences $\langle\{(*,*,A)\}\rangle$ et $\langle\{(*,Middle,*)\}\rangle$.
- Pour MicroCount, le support de la séquence est égal à : $\frac{\min(123,120)+\min(125,115)}{742} = 0,32$

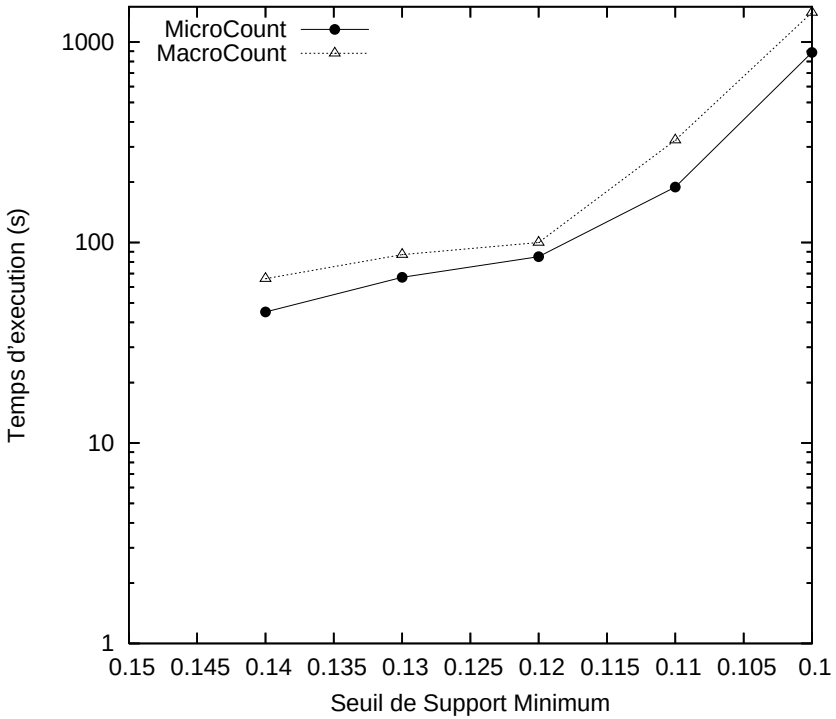
Mise en œuvre

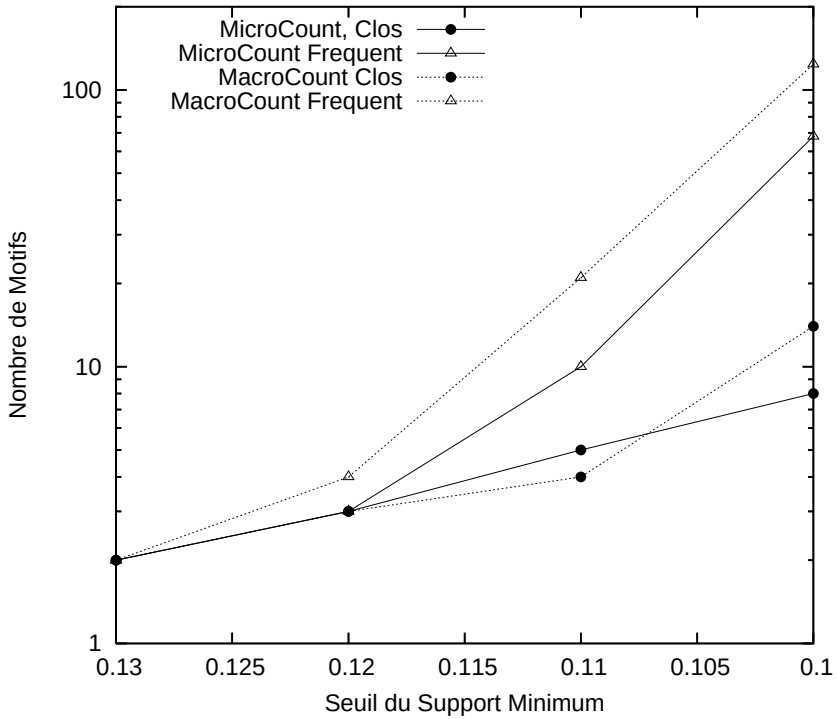
Adaptation de l'algorithme *CMSP*

- Extraction de motifs séquentiels multidimensionnels fermés
- Recherche de la meilleure solution pour chaque séquence de données


MicroCount
MacroCount
Temps d'execution (s)
600
500
400
300
200
100
0
0.15 0.145 0.14 0.135 0.13 0.125 0.12 0.115 0.11 0.105 0.1
Seuil de Support Minimum


MicroCount, Clos
MicroCount Frequent
MacroCount Clos
MacroCount Frequent
Nombre de Motifs
1000
100
10
1
0.15 0.145 0.14 0.135 0.13 0.125 0.12 0.115 0.11 0.105 0.1
Seuil du Support Minimum


MicroCount
MacroCount
1000
100
10
1
Temps d'execution (s)
0.15 0.145 0.14 0.135 0.13 0.125 0.12 0.115 0.11 0.105 0.1
Seuil de Support Minimum


MicroCount, Clos
MicroCount Frequent
MacroCount Clos
MacroCount Frequent
100
10
1
Nombre de Motifs
0.13 0.125 0.12 0.115 0.11 0.105 0.1
Seuil du Support Minimum

Plan

Conclusion

Bilan

- Utilisation de la mesure pour le calcul des motifs séquentiels multidimensionnels
- Deux méthodes complémentaires

Perspectives

- Opérateurs autres que comptage et somme
- Dimensions non additives
- Connaissances inattendues

MERCI