Les Motifs séquentiels multidimensionnels étoilés

M^2SP : une nouvelle approche

Marc Plantevit[*,1] Anne Laurent[*,1] Maguelonne Teisseire[*,1] Dominique Laurent[**,1] Y.W Choong[**,***]

[*] LIRMM, Montpellier
[**] LICP, Cergy Pontoise
[***] HELP, Kuala Lumpur

20 mai 2005

[1] la virgule est un *opérateur commutatif n-aire*

Plan

Plan

Introduction

Motivations

- Vers une extraction d'information plus riche
- Développement d'OLAP
- limites des motifs séquentiels "classiques"

exemple

- Motif séquentiel : $\langle\{parapluie, écharpe\}, \{crème\ solaire\}\rangle$
- Motif séquentiel multidimensionnel : Ce que l'on veut extraire
  $\langle\{(Paris, Professeur, parapluie), (Paris, Professeur, écharpe)\}$
  $\{(Palavas, Professeur, crème\ solaire)\}\rangle$

Plan

Les approches existantes abordant le problème des motifs séquentiels multidimensionnels

J Han, H Pinto & co

- l'approche fondatrice
- recherche de motifs séquentiels *intra pattern*

Yu & Chen

- extraction d'information sur des données spécifiques (weblog) organisées en différents niveaux d'agrégation

de Amo, Furtado & co

- approche basée sur la logique temporelle

J Han, H Pinto & co

Motif séquentiel multidimensionnel β :

$$\beta = \underbrace{\textbf{PATTERN}}_{\text{déf. sur 1 ou pls. dim.}} + \underbrace{\textbf{séquence d'items}}_{\text{déf. sur une \textbf{seule} dim.}}$$

Exemple : Problème du "panier de la ménagère"

- Pattern :
  - informations relatives aux clients
  - Age, ville, etc
- Séquence d'items : les produits achetés
- (*Paris*, *Sportif*, ⟨tee-shirt JO-2012, drapeau JO-2012⟩)

Les Patterns

- Pattern = n-uplet défini sur n dimensions $\mathscr{D}_i$

$$t = (d_1, \cdots, d_n) \quad d_i \in Dom(D_i) \cup \{*\}$$

- * : un joker

Principe

- Essai d'instancier le plus possible de dimensions afin d'obtenir des pattern maximalement spécifiques (moins * possible)

2 méthodes équivalentes d'extraction

patterns fréquents puis séquences

1. Pattern p $\overset{identifie}{\longrightarrow}$ un ensemble de transactions T
2. Dans T on va extraire des motifs séquentiels classiques s
3. (p, s) est un motif séquentiel multidimensionnel

méthode réciproque : séquences fréquentes puis patterns

1. Sur la dimension *Produits* extraire les motifs séquentiels classiques.
2. Matching des séquences fréquentes avec les patterns qui les identifient

exemples

exemples de motifs séquentiels multidimensionnels

$minsupp = 2$

- $\langle(business, middle), a\rangle$
- $\langle(chicago), bf\rangle$

cid	Cust-Grp	City	Age-grp	product-sequence
10	**business**	Boston	**middle**	$\langle$(b*d*)*cb***a**$\rangle$
20	professional	**Chicago**	young	$\langle$**bf**(*ce*)(*fg*)$\rangle$
30	**business**	**Chicago**	**middle**	$\langle$(*ah*)**abf**$\rangle$
40	education	New York	retired	$\langle$(*be*)(*ce*)$\rangle$

Critique

☺

- approche fondatrice
- permet une analyse plus fine

☹

- pas de prise en compte de valeurs de mesure
- que des information *intra pattern* (pas de corrélations possibles entre patterns différents)
  On ne pas extraire des informations du type :

$$(N.Y., business, age\_middle, \langle Clou \rangle), (N.Y., Retired, age\_old, \langle pneu \rangle)$$

Plan

Base de données multidimensionnelles (ROLAP)

Cellule $= \langle (d_1, \ldots, d_n), \mu \rangle$

- $\forall i \in [1 \cdots n],\ d_i \in Dom(D_i)$
- $\mu \in Dom(M)$, μ est la mesure de la cellule.

(Hyper-)Cube

Un hypercube C de données est un ensemble de cellules qui sont définies sur les mêmes dimensions.

$$C\ :\ Dom(D_1) \times Dom(D_2) \times \cdots \times Dom(D_n) \longrightarrow M$$

exemple

Représentation d'un hypercube de données sous forme bidimensionnelle

Date	Cust-Grp	City	Age	Product	Measure
1	Educ	Chigago	A	clou	50
1	Educ	Chigago	B	pneu	2
1	Educ	Los Angeles	A	clou	30
1	Reti.	Miami	C	clou	20
1	Reti.	Miami	C	marteau	2
2	Educ	Chigago	B	rustine	10
2	Educ	Chigago	B	pneu	3
2	Educ	Los Angeles	A	clou	20
3	Educ	Los Angeles	B	rustine	15

Sous-cube de données

- un cube de données peut être partitionné en différents **sous-cubes** en fonction de certaines dimensions

Partition par rapport à *Cust-Grp et City*

Date	Cust-Grp	City	Age	Product	Measure
1	Educ	Chigago	A	clou	50
1	Educ	Chigago	B	pneu	2
2	Educ	Chigago	B	rustine	10
2	Educ	Chigago	B	pneu	3
1	Educ	Los Angeles	A	clou	30
2	Educ	Los Angeles	A	clou	20
3	Educ	Los Angeles	B	rustine	15
1	Reti.	Miami	C	clou	20
1	Reti.	Miami	C	marteau	2

Données manipulées

Partition des dimensions

$$D = \mathscr{D}_{\mathscr{F}} \oplus \mathscr{D}_{\mathscr{R}} \oplus \mathscr{D}_{\mathscr{A}} \oplus \{D_t\}$$

- D_t pour la dimension temporelle
- $\mathscr{D}_{\mathscr{A}}$ pour les dimensions dites d'**analyse**
- $\mathscr{D}_{\mathscr{R}}$ pour les dimensions dites de **référence**
- $\mathscr{D}_{\mathscr{F}}$ pour les dimensions **ignorées**.

Notation

Chaque cellule $cell = \langle (d_1, \cdots, d_n), \mu \rangle$ d'un cube peut être notée :

$$cell = \langle (f, r, a, t), \mu \rangle$$

avec :

- f la restriction sur $\mathscr{D}_{\mathscr{F}}$ de *cell*
- r la restriction sur $\mathscr{D}_{\mathscr{R}}$ de *cell*
- a la restriction sur $\mathscr{D}_{\mathscr{A}}$ de *cell*
- t la restriction sur D_t de *cell*

M^2SP : item,itemset,séquence multidimensionnels

Définitions

- item multidimensionnel :

$$e = \langle a, \mu \rangle$$

- itemset multidimensionnel :

$$i = \{e_1, \cdots, e_p\}$$

- séquence multidimensionnelle :

$$s = \langle i_1, \cdots, i_l \rangle$$

Support d'une séquence

Un cube de données $\mathscr{C}_c$ supporte une séquence $s = \langle i_1, \cdots, i_l \rangle$ si et seulement si :

$$\forall j = 1 \cdots l, \exists d_j \in D_t, \quad \forall e = \langle a, \mu_e \rangle \in i_j,$$
$$\exists cell = \langle (f, a, r, d_j), \mu \rangle \in \mathscr{C}_c \quad avec\ d_1 \prec_t d_2 \prec_t \cdots \prec_t d_l$$
$$et\ \mu_e = \mu$$

- **Support d'une séquence** *Soient s un séquence, un cube DC composé d'un ensemble de sous-cubes $\mathscr{C}_c$ construits à partir de $\mathscr{D}_{\mathscr{R}}$ est :*

$$support(s) = \frac{|\{\mathscr{C}_c\ supportant\ s\}|}{|\{\mathscr{C}_c\}|}$$

- **séquence fréquente**
- **k-fréquent**

Exemple récapitulatif

- Soient $\mathscr{D}_{\mathscr{R}} = \{Cust - Grp, City\}$ $\mathscr{D}_{\mathscr{A}} = \{Age, Product\}$ $suppmin = 0.3$
- Sous-cubes définis sur $\mathscr{D}_{\mathscr{R}}$:

Cust-Grp	City
Educ	Chigago
Educ	Los Angeles
Reti.	Miami

- Soit la séquence
  $s = \langle \{(A, clou, 50)(B, pneu, 2)\}, \{(B, rustine, 10)\} \rangle$
- cherchons le support

Sous-cube 1

Date	Cust-Grp	City	Age	Product	Measure
1	Educ	Chigago	**A**	**clou**	**50**
1	Educ	Chigago	**B**	**pneu**	**2**
2	Educ	Chigago	B	pneu	3
2	Educ	Chigago	**B**	**rustine**	**10**

- 2 dates différentes

$\rightarrow$ le sous-cube 1 supporte la séquence s

exemple (suite)

sous-cube 2 :

Date	Cust-Grp	City	Age	Product	Measure
1	Educ	Los Angeles	**A**	**clou**	30
2	Educ	Los Angeles	A	clou	20
3	Educ	Los Angeles	B	rustine	15

pas possible : trop peu de nuplets par date pour pouvoir supporter l'itemset de la séquence

sous-cube 3 :

Date	Cust-Grp	City	Age	Product	Measure
1	Reti.	Miami	C	clou	20
1	Reti.	Miami	C	marteau	2

pas assez de date pour pouvoir supporter la séquence

Support

$$Support(s) = \frac{1}{3} \longrightarrow \textbf{fréquente}$$

vers de nouveaux horizons...

Problème :

- $\langle clou, Chicago, \mu \rangle, \langle clou, Miami, \mu \rangle, \langle clou, L.A., \mu \rangle$ non fréquents
- $\langle clou, (Chicago \vee Miami \vee L.A.), \mu \rangle$ fréquent
- Que faire ?

Solution :

- Réduire ponctuellement les axes de recherche : * un méta-symbole

M^2SP-α : Item multidimensionnel α-étoilé (*)

Un item multidimensionnel α-étoilé est de la forme :

$$\boxed{e = \langle a, \mu \rangle \text{ où } \forall d_i \in a,\ d_i \in Dom(D_i) \cup \{*\}}$$

1. $\forall e = \langle a, \mu \rangle, \exists d_{i_j} \in a \text{ tq } d_{i_j} \neq *$
   (au moins une dimension fixée)
2. $\forall d_{i_j} = *,\ \nexists x \in Dom(D_{i_j}) \text{ tq } e' = \langle a_{[d_{i_j}/x]}, \mu \rangle,\ support(e') \geq suppmin$
   (rôle de seconde chance)

* et support

Un cube de données $\mathscr{C}_c$ supporte une séquence $s = \langle i_{s1}, \cdots, i_{sl} \rangle$ si et seulement si :

$\forall j = 1 \cdots l, \quad \exists d_j \in D_t, \quad \forall e = \langle (a_{i_1}, \cdots, a_{i_m}), \mu_e \rangle \in i_j$

$\exists cell = \langle (f, (x_{i_1}, \cdots, x_{i_m}), r, d_j), \mu \rangle \in \mathscr{C}_c,$

$$avec \boxed{a_i = x_i \text{ ou } a_i = *} \text{ et } d_1 \prec_t d_2 \prec_t \cdots \prec_t d_l$$

M^2SP-μ : Item multidimensionnel « μ-étoilé »($\circledast$)

définition

Un item multidimensionnel μ-étoilé est de la forme :

$$\boxed{e = \langle a, \mu \rangle}$$

où

- $\mu \in Dom(M) \cup \circledast$

$\circledast$ et support

Un cube de données $\mathscr{C}_c$ supporte une séquence $s = \langle i_{s1}, \cdots, i_{sl} \rangle$ si et seulement si :

$$\forall j = 1 \cdots l, \quad \exists d_j \in D_t, \quad \forall e = \langle a, \mu_e \rangle \in i_j, \quad \exists cell = \langle (f, a, \mu \rangle \in \mathscr{C}_c,$$

$$avec \ \boxed{\mu_e = \circledast \ ou \ \mu_e = \mu}$$

Algorithmes

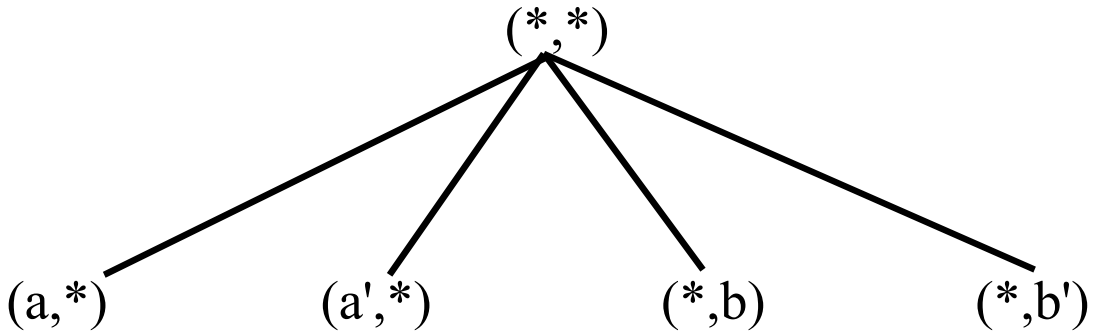
Génération des items fréquents

- obtention des **items maximalement spécifiques**
- génération par niveau

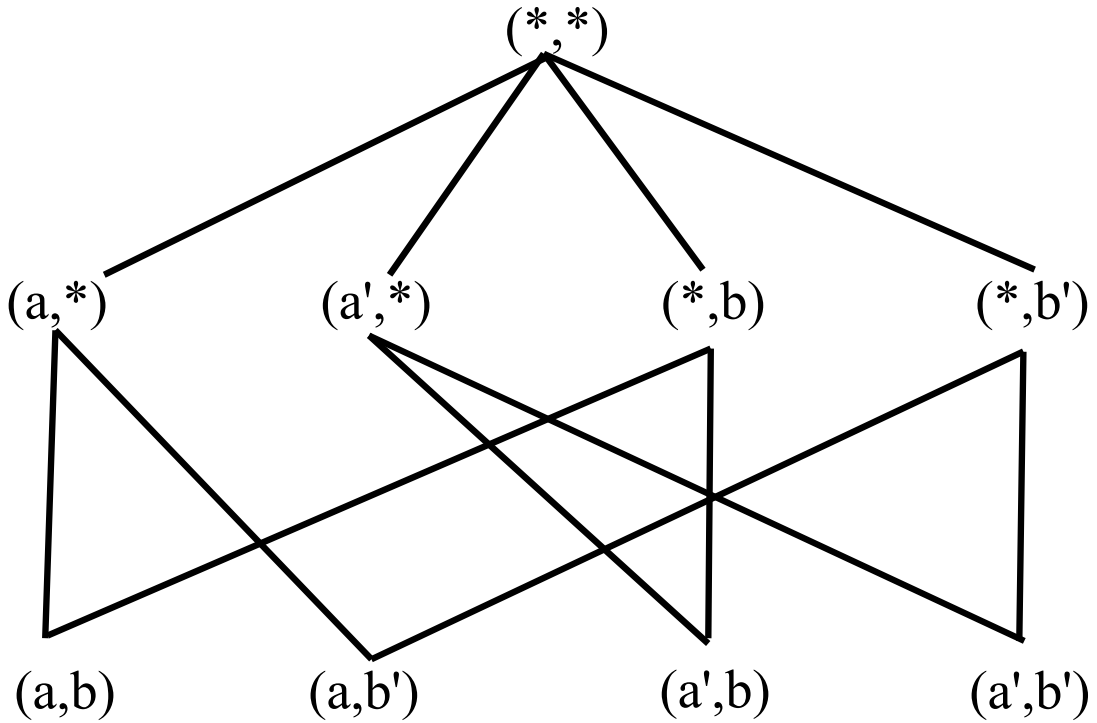
Génération des séquences fréquentes

- méthode générer/élaguer
- utilisation de l'algorithme psp

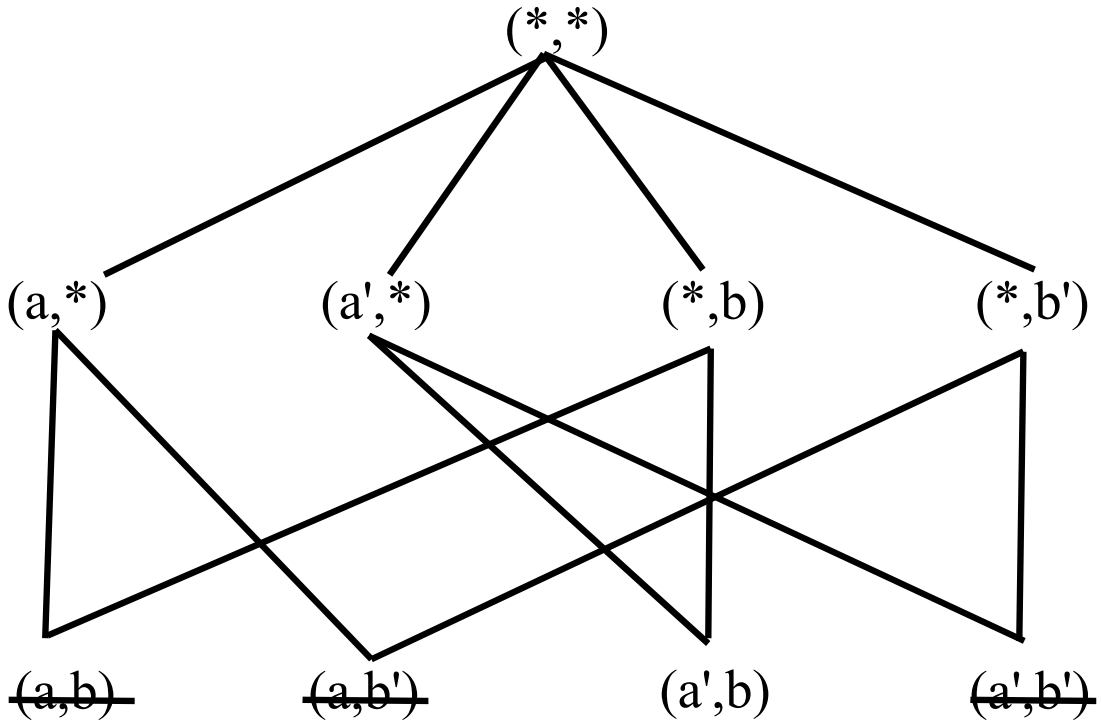
Exemple : génération des items fréquents



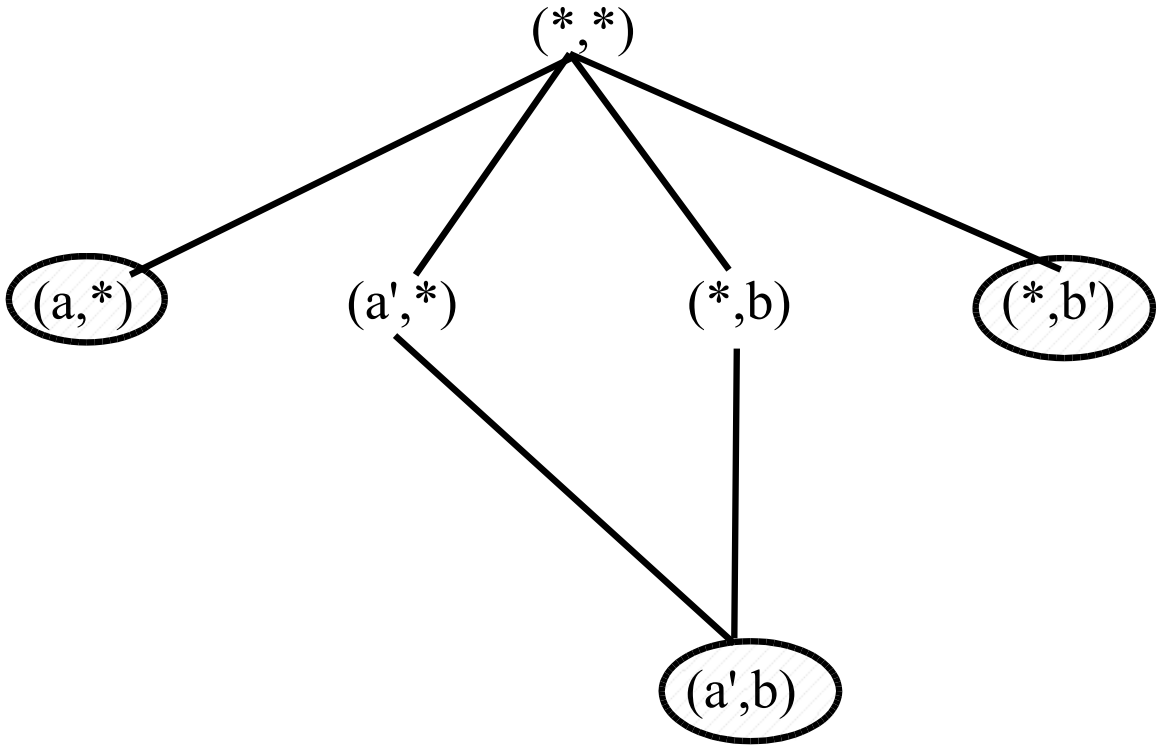
Exemple : génération des items fréquents



Exemple : génération des items fréquents



Exemple : la bordure



Calcul du support

- prérequis : prétraitement des données (cube by *date*, $D_1, \cdots, D_n$)
- calcul du support d'une séquence
  - pour chaque sous-cube : **supportCount(**s, DC, $\mathscr{D}_{\mathscr{R}}$, *comptage***)**
  - dans un sous-cube donné : **supportCube(**s, C, *comptage***)**
- ancrage ($\sigma_{condition}(C) \longmapsto C'$ avec $C' \subseteq C$)

complexité

- n_C est le nombre de cellules du cube C
- $m = |\mathscr{D}_{\mathscr{A}}|$ est le nombre de dimensions des items multidimensionnels.
- supportCount :$O(n_C \times m \times \log n_C)$.
- supportCube :$O(l \times n_{max} \times m \times \log n_{max})$

Plan

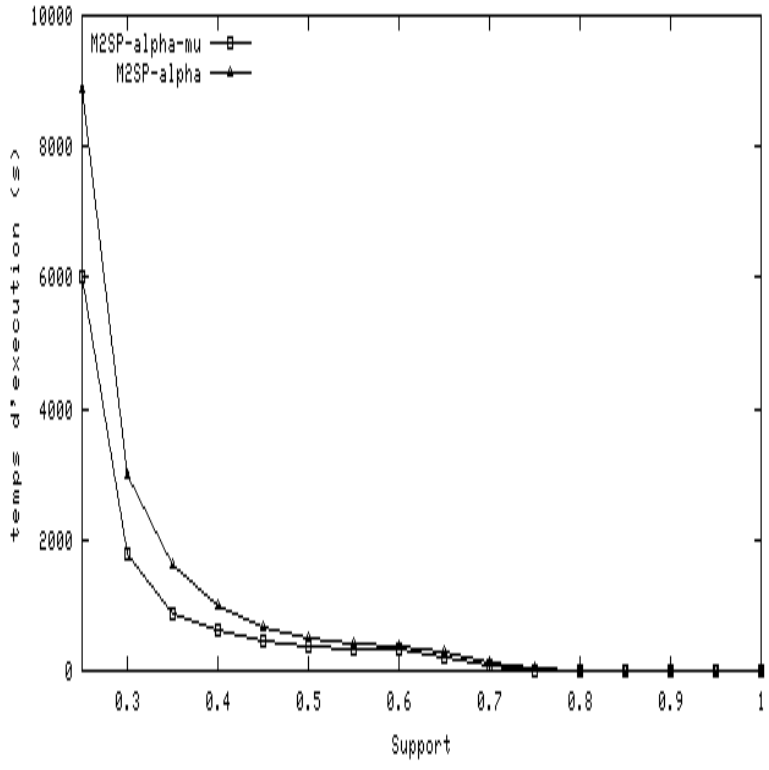
généralisation des motifs séquentiels "classiques"

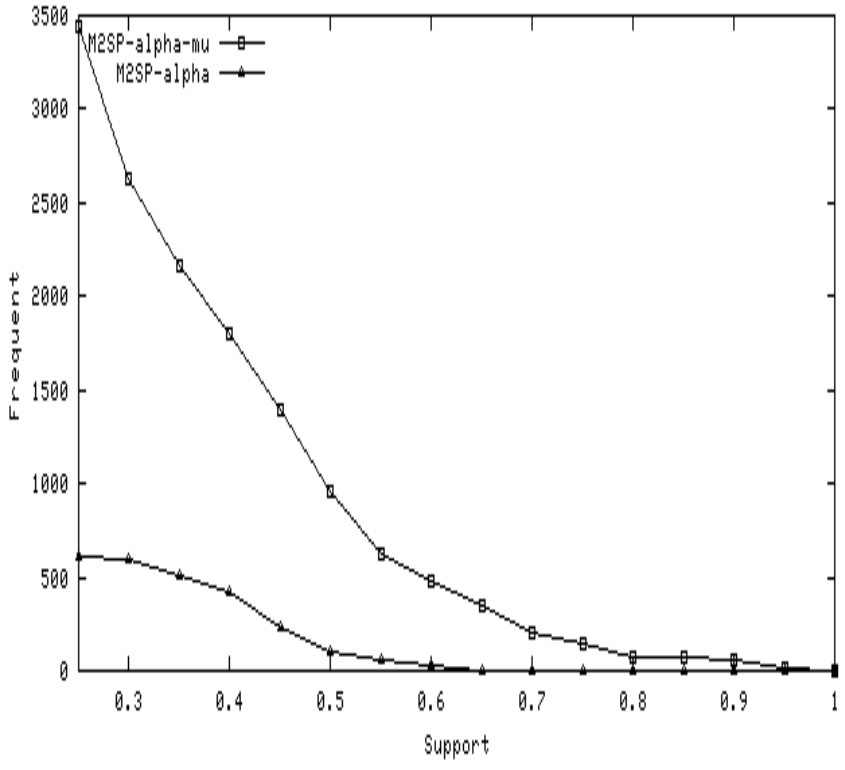
- $\mathscr{D}_{\mathscr{R}}$ = dimension relative aux clients
- $\mathscr{D}_{\mathscr{A}}$ = dimension relative aux produits
- D_t = la date
- $\mathscr{D}_{\mathscr{F}}$ =le reste

généralisation

- tous les fréquents extraits par J.Han, peuvent être extraits par M²SPα-μ
- on en trouve plus (*inter pattern*)

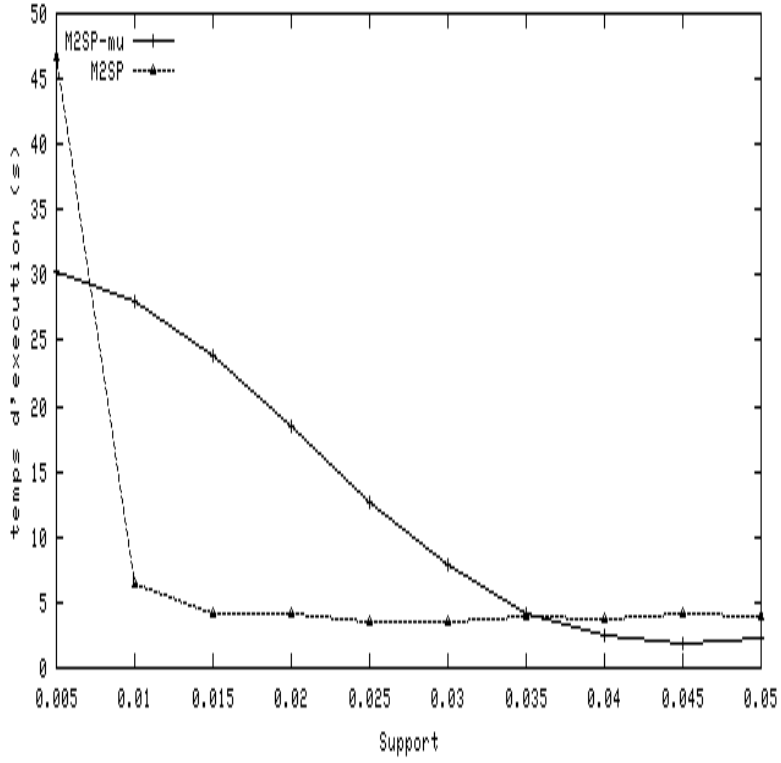
Support M^2SP-α(-μ)

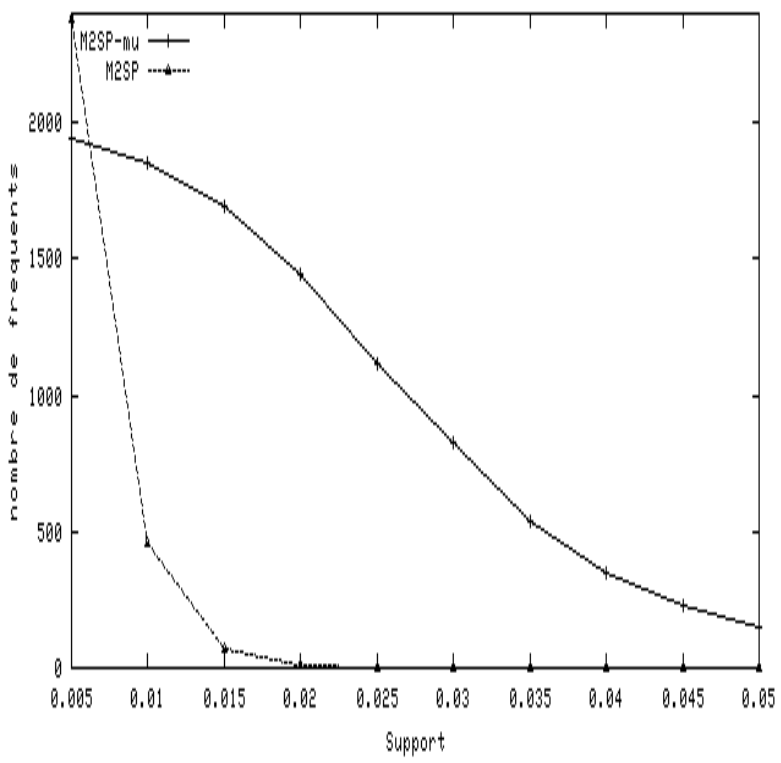




- La mesure : une contrainte

Support $M^2SP(-\mu)$

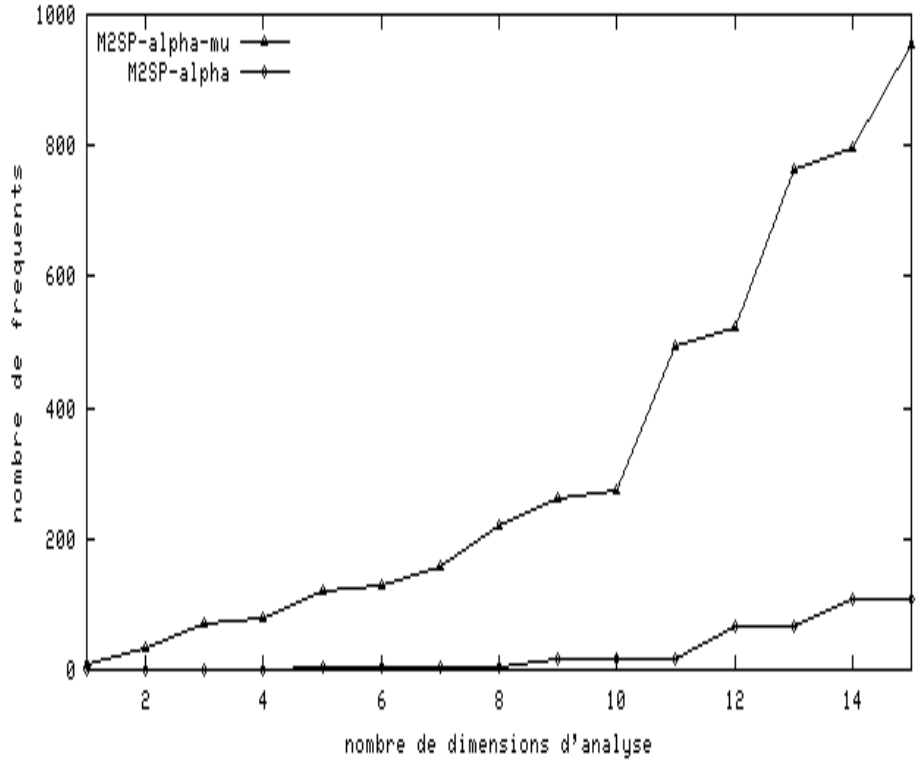




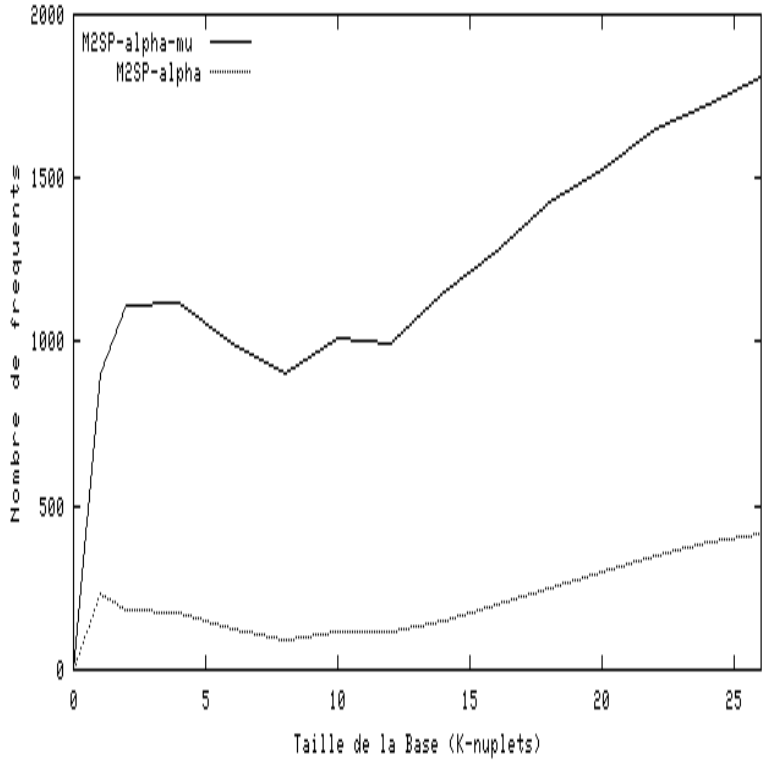
- Une contrainte supplémentaire : pour des extractions sur des données spécifiques

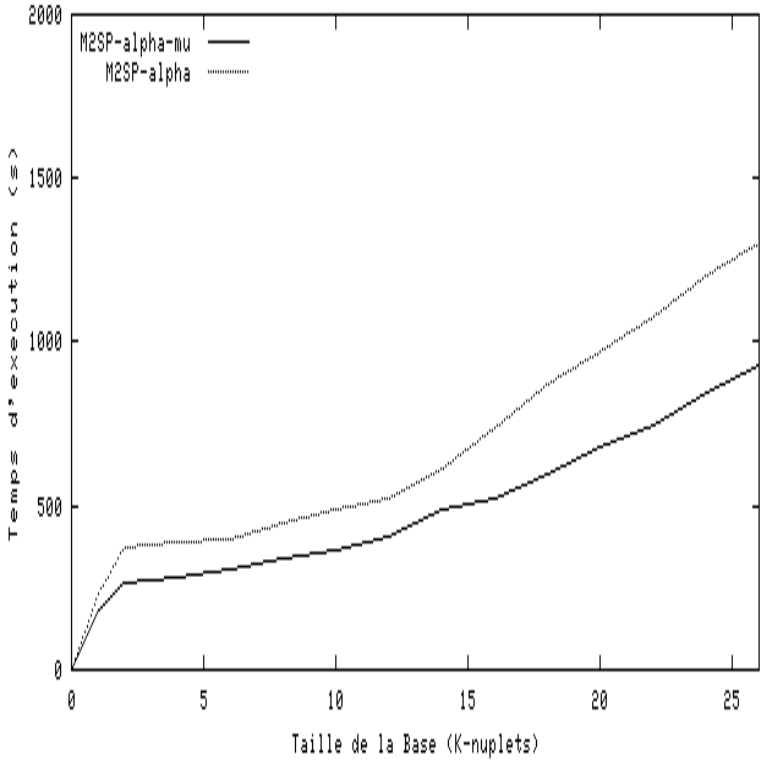
Propriété

- le nombre de fréquents croît en fonction du nombre de dimensions



Passage à l'échelle





Plan

Conclusion et perspectives

une approche plus multidimensionnelle et plus générale

- *inter cube*
- prise en compte des quantités possibles ou non(⊛)
- plusieurs types de motifs séquentiels
- une plus grande liberté dans le choix des axes

perspectives de travail

- utilisation de la théorie des sous-ensembles flous
- choix automatique des axes de références et d'analyse
- relation d'ordre (spatiale, spatio-temporelle, . . .)
- classifieur multidimensionnel
- motifs séquentiels multidimensionnels généralisés (contraintes de temps,. . .)