Introduction
OLAP et ECD
Motifs séquentiels
Contexte Multidimensionnel
Hiérarchies
Contributions
Modèle de données
Définitions
Algorithmes
Expérimentations
Conclusion et perspectives

Prise en compte des hiérarchies dans l'extraction de motifs séquentiels multidimensionnels

MARC PLANTEVIT, ANNE LAURENT, MAGUELONNE TEISSEIRE

LIRMM, UNIVERSITÉ MONTPELLIER II, FRANCE

EDA06, Versailles, 19 juin 2006

Plan

HYPE

Introduction
OLAP et ECD
Motifs séquentiels
Contexte Multidimensionnel
Hiérarchies
Contributions
Modèle de données
Définitions
Algorithmes
Expérimentations
Conclusion et perspectives

1 Introduction

- OLAP et Fouille de données
- Motifs Séquentiels
- Contexte Multidimensionnel
- Hiérarchies

2 Contributions

- Modèle de données
- Définitions
- Algorithmes
- Expérimentations

3 Conclusion et perspectives

Plan

HYPE

Introduction
OLAP et ECD
Motifs séquentiels
Contexte Multidimensionnel
Hiérarchies
Contributions
Modèle de données
Définitions
Algorithmes
Expérimentations
Conclusion et perspectives

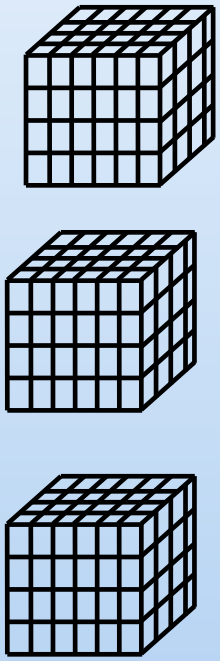
1. Introduction
   - OLAP et Fouille de données
   - Motifs Séquentiels
   - Contexte Multidimensionnel
   - Hiérarchies
2. Contributions
   - Modèle de données
   - Définitions
   - Algorithmes
   - Expérimentations
3. Conclusion et perspectives

La fouille de données pour l'aide à la décision

HYPE

Introduction
OLAP et ECD
Motifs séquentiels
Contexte Multidimensionnel
Hiérarchies
Contributions
Modèle de données
Définitions
Algorithmes
Expérimentations
Conclusion et perspectives

Le décideur confronté à de gros volumes de données . . .



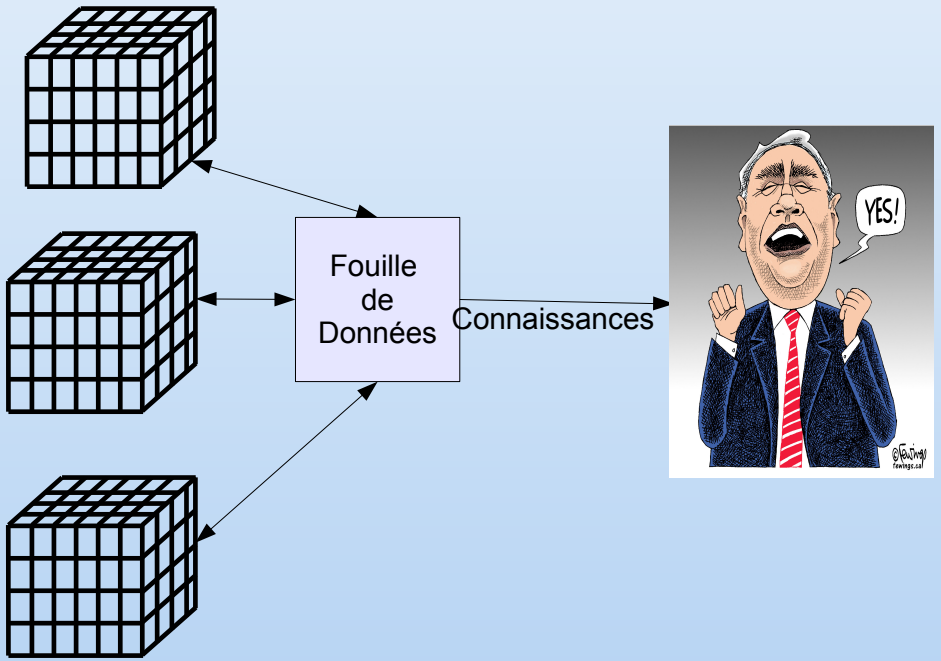
La fouille de données pour l'aide à la décision

HYPE

Introduction
OLAP et ECD
Motifs séquentiels
Contexte Multidimensionnel
Hiérarchies
Contributions
Modèle de données
Définitions
Algorithmes
Expérimentations
Conclusion et perspectives

La fouille de données comme une aide à la décision

. . .



Motifs Séquentiels

HYPE

Introduction
OLAP et ECD
Motifs séquentiels
Contexte Multidimensionnel
Hiérarchies
Contributions
Modèle de données
Définitions
Algorithmes
Expérimentations
Conclusion et perspectives

- Découverte de corrélations entre évènements
- Données qui supportent une relation d'ordre (*e.g.* temporelle)
- Un vaste champ d'application : marketing, prise de décision, bioinformatique, sécurité informatique . . .

Motifs Séquentiels

HYPE

Introduction
OLAP et ECD
Motifs séquentiels
Contexte Multidimensionnel
Hiérarchies
Contributions
Modèle de données
Définitions
Algorithmes
Expérimentations
Conclusion et perspectives

- Découverte de corrélations entre évènements
- Données qui supportent une relation d'ordre (*e.g.* temporelle)
- Un vaste champ d'application : marketing, prise de décision, bioinformatique, sécurité informatique . . .

Motifs Séquentiels

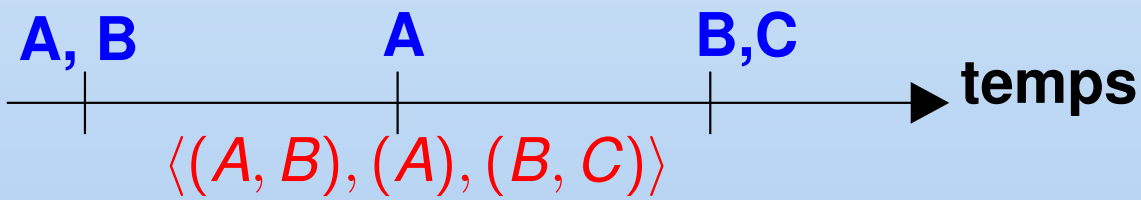
HYPE

Introduction
OLAP et ECD
Motifs séquentiels
Contexte Multidimensionnel
Hiérarchies
Contributions
Modèle de données
Définitions
Algorithmes
Expérimentations
Conclusion et perspectives

- Découverte de corrélations entre évènements
- Données qui supportent une relation d'ordre (*e.g.* temporelle)
- Un vaste champ d'application : marketing, prise de décision, bioinformatique, sécurité informatique . . .

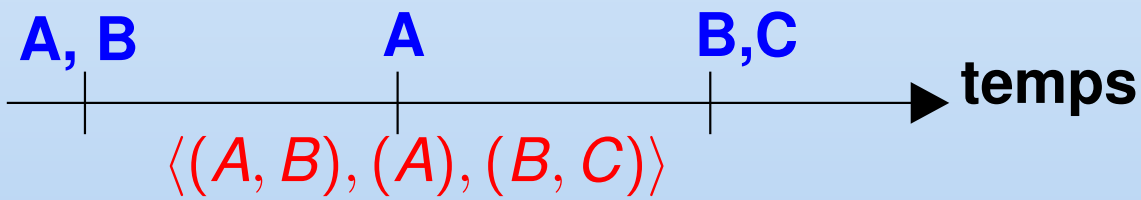
Motifs Séquentiels

- Découverte de corrélations entre évènements
- Données qui supportent une relation d'ordre (*e.g.* temporelle)
- Un vaste champ d'application : marketing, prise de décision, bioinformatique, sécurité informatique . . .



Motifs Séquentiels

- Découverte de corrélations entre évènements
- Données qui supportent une relation d'ordre (*e.g.* temporelle)
- Un vaste champ d'application : marketing, prise de décision, bioinformatique, sécurité informatique . . .



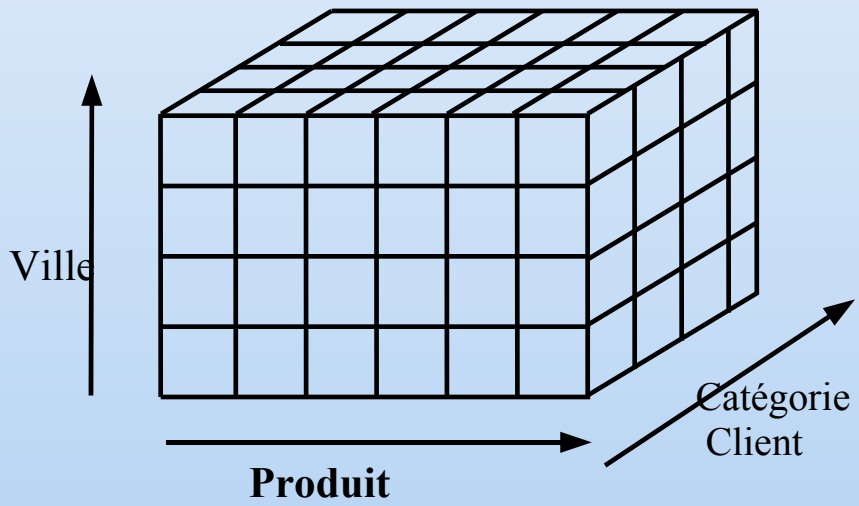
- ☹ : Motifs séquentiels pauvres par rapport aux données qu'ils décrivent (une seule dimension explorée)

cube de données

HYPE

Introduction
OLAP et ECD
Motifs séquentiels
Contexte Multidimensionnel
Hiérarchies
Contributions
Modèle de données
Définitions
Algorithmes
Expérimentations
Conclusion et perspectives

- Les connaissances extraites seulement au sein de la dimension *produit*.
- Quid des autres dimensions ?



Bases de Données Relationnelles : Exemple

HYPE

Introduction
OLAP et ECD
Motifs séquentiels
Contexte Multidimensionnel
Hiérarchies
Contributions
Modèle de données
Définitions
Algorithmes
Expérimentations
Conclusion et perspectives

Date	Distributor	Place	Product	Quantity
1	Auchan	France	c	100
2	Carrefour	Chine	e	60
3	walmart	New York	b	80
...	...	...	...	...

- Base de données relationnelles peut être vue comme une table de faits dans une **base de données multidimensionnelles**.
- **dimension** $\approx$ attribut

Combiner plusieurs dimensions d'analyse

- un item d'une séquence n'est plus défini sur une seule dimension mais plusieurs

Combiner plusieurs dimensions d'analyse

- un item d'une séquence n'est plus défini sur une seule dimension mais plusieurs
- item classique : c

Combiner plusieurs dimensions d'analyse

- un item d'une séquence n'est plus défini sur une seule dimension mais plusieurs
- item classique : c
- item multidimensionnel : $(France, c, 100), (France, c, *)$

Combiner plusieurs dimensions d'analyse

- un item d'une séquence n'est plus défini sur une seule dimension mais plusieurs
- item classique : c
- item multidimensionnel : $(France, c, 100), (France, c, *)$
- séquence multidimensionnelle :

$\langle\{(France, c, 100), (Allemagne, d, 54)\}\{(*, b, 2)\}\rangle$
au lieu de $\langle(c, d), b\rangle$

La prise en compte des Hiérarchies

HYPE

Introduction
OLAP et ECD
Motifs séquentiels
Contexte Multidimensionnel
Hiérarchies
Contributions
Modèle de données
Définitions
Algorithmes
Expérimentations
Conclusion et perspectives

dilemme Support/Nombre de connaissances extraites

- Support trop élevé : très peu de connaissances fréquentes.
- Support trop faible : un nombre trop important de connaissances extraites, inutilisables pour le décideur.

Difficulté d'extraire des connaissances de qualité en fonction du support.

La prise en compte des Hiérarchies

HYPE

Introduction
OLAP et ECD
Motifs séquentiels
Contexte Multidimensionnel
Hiérarchies
Contributions
Modèle de données
Définitions
Algorithmes
Expérimentations
Conclusion et perspectives

dilemme Support/Nombre de connaissances extraites

- Support trop élevé : très peu de connaissances fréquentes.
- Support trop faible : un nombre trop important de connaissances extraites, inutilisables pour le décideur.

Difficulté d'extraire des connaissances de qualité en fonction du support.

La prise en compte des hiérarchies pour résoudre ce dilemme

- Extraire des règles sur plusieurs niveaux de hiérarchies
- faculté de subsomption

Approches existantes

Agrawal & Srikant (1996) :

- Prise en compte des hiérarchies dans l'extraction de séquences.
- Réécriture de la base avec les items et **leurs ancêtres**.
- Approche inenvisageable dans contexte multidimensionnel.

J. Han (2001) :

- Extraction de connaissances par niveaux (contexte classique)
- Des connaissances extraites sur un unique niveau de hiérarchie.

Yu & Chen (2005) :

- extraction de connaissance à partir de données spécifiques (weblogs) organisées en différents niveaux de hiérarchie
- Hiérarchie utilisée pour représenter le temps (jours, sessions, pages visitées).

Aucune approche ne permet d'extraire des séquences *multidimensionnelles* sur plusieurs *niveaux de hiérarchie*

Introduction
OLAP et ECD
Motifs séquentiels
Contexte Multidimensionnel
Hiérarchies
Contributions
Modèle de données
Définitions
Algorithmes
Expérimentations
Conclusion et perspectives

Plan

1 Introduction
- OLAP et Fouille de données
- Motifs Séquentiels
- Contexte Multidimensionnel
- Hiérarchies

2 Contributions
- Modèle de données
- Définitions
- Algorithmes
- Expérimentations

3 Conclusion et perspectives

Base de données & Blocs

HYPE

Introduction
OLAP et ECD
Motifs séquentiels
Contexte Multidimensionnel
Hiérarchies
Contributions
Modèle de données
Définitions
Algorithmes
Expérimentations
Conclusion et perspectives

BLOC :

- Une base de données peut être partionnée en différents **blocs** selon certaines dimensions.

BlocID	Date	Place	Product
1	1	Allemagne	2
1	1	Allemagne	5
1	2	Allemagne	8
1	3	Allemagne	1
1	4	Allemagne	7
2	1	France	6
2	2	France	2
2	2	France	4
2	3	France	8
3	1	UK	2
3	1	UK	3
3	2	UK	8
4	1	LA	1
4	2	LA	7
4	3	NY	3
4	4	NY	6

Partition des Dimensions

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

- D_t : dimensions temporelles
- D_A : dimensions d'analyse
- D_R : dimensions de référence

nuplet $c = (d_1, \cdots, d_n) = (r, a, t)$ où :

- r : la restriction de c sur $\mathscr{D}_{\mathscr{R}}$
- a : la restriction de c sur $\mathscr{D}_{\mathscr{A}}$
- t : la restriction de c sur $\mathscr{D}_t$

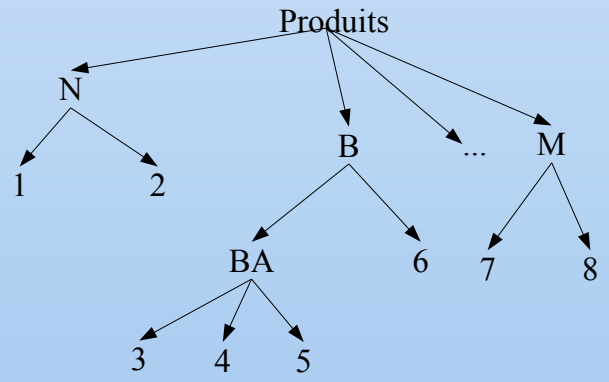
Taxonomies

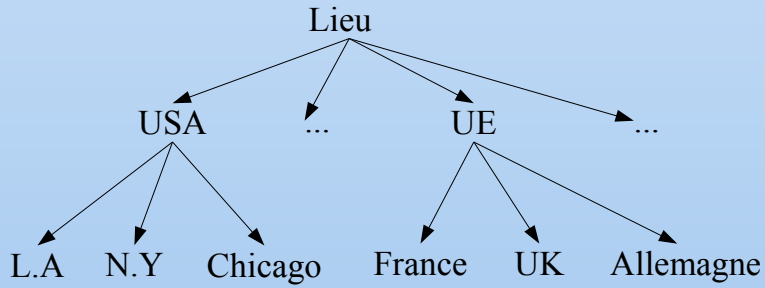
HYPE

Introduction
OLAP et ECD
Motifs séquentiels
Contexte Multidimensionnel
Hiérarchies
Contributions
Modèle de données
Définitions
Algorithmes
Expérimentations
Conclusion et perspectives

- Des relations hiérarchiques entre éléments peuvent être établies.
- Ces relations sont matérialisées par des **taxonomies**.
- Seules des *feuilles* de la taxonomie peuvent apparaître dans la base

Taxonomies représentant les dimensions PLACE et PRODUIT :





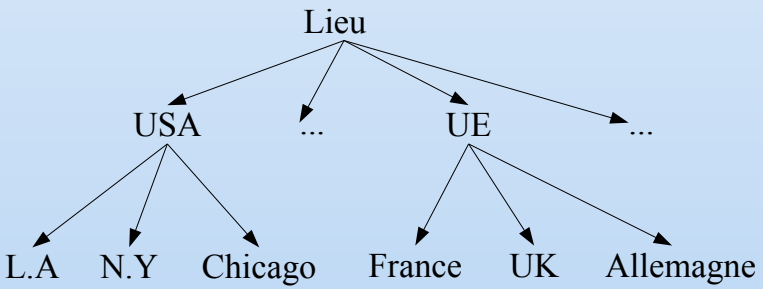
Relation entre élément de la taxonomie

ancêtre : $\hat{x}$ un ancêtre de x dans la taxonomie.

descendant : noté $\check{x}$.

$$UE = \widehat{France}$$

$$Lieu = \widehat{Allemagne}$$



Item, itemset et séquence h-généralisés

HYPE

Introduction
OLAP et ECD
Motifs séquentiels
Contexte Multidimensionnel
Hiérarchies
Contributions
Modèle de données
Définitions
Algorithmes
Expérimentations
Conclusion et perspectives

Item multidimensionnel h-généralisé :

un m-uplet $e = (d_1, \ldots, d_m)$ défini sur les dimensions d'analyse D_A telles que $d_i \in \{label(T_i)\}$.

Exemples : $(France, 2)$, $(Allemagne, B)$

Item, itemset et séquence h-généralisés

Item multidimensionnel h-généralisé :

un m-uplet $e = (d_1, \ldots, d_m)$ défini sur les dimensions d'analyse D_A telles que $d_i \in \{label(T_i)\}$.

Exemples : $(France, 2)$, $(Allemagne, B)$

Inclusion hiérarchique

Soient $e = (d_1, \ldots, d_m)$ et $e' = (d'_1, \ldots, d'_m)$, alors :

- e est **plus général** que e' ($e >_h e'$) si $\forall d_i, d_i = \hat{d}'_i$ *ou* $d_i = d'_i$
- e est **plus spécifique** que e' ($e <_h e'$) si $\forall d_i, d_i = \check{d}'_i$ *ou* $d_i = d'_i$
- e et e' sont **incomparables** s'il n'existe pas de relation entre eux ($e \not>_h e'$ *et* $e' \not>_h e$)

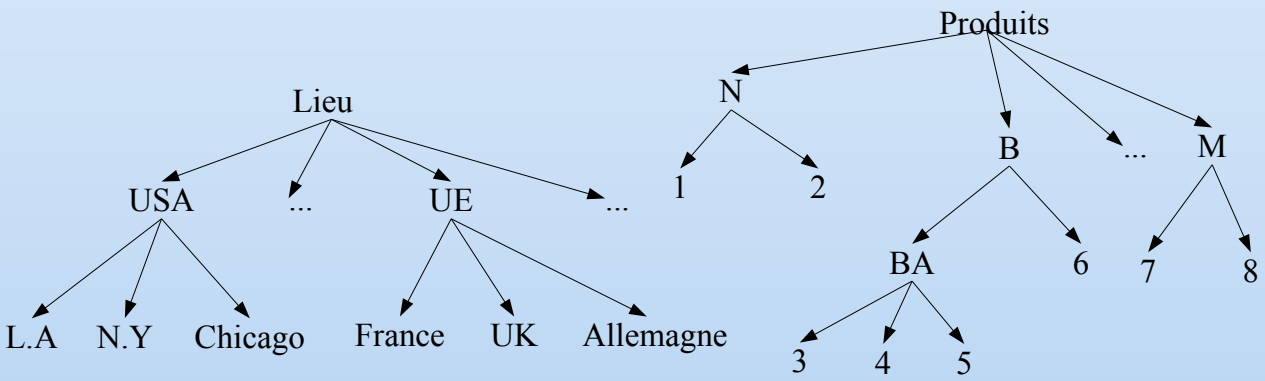
Support

HYPE

Introduction
OLAP et ECD
Motifs séquentiels
Contexte Multidimensionnel
Hiérarchies
Contributions
Modèle de données
Définitions
Algorithmes
Expérimentations
Conclusion et perspectives

Exemple de relation entre items :

- $(USA, B) >_h (USA, 6)$.
- $(France, 4) <_h (UE, BA)$.
- $(France, 4)$ et $(USA, 6)$ sont incomparables.



Itemset et séquence h-généralisés

Itemset h-généralisé :

$i = \{e_1, \ldots, e_k\}$ où tous les items sont incomparables entre eux.

$\{(France, 4), (USA, 6)\}$ OUI
$\{(France, 4), (U.E, BA)\}$ NON car
$(France, 4) <_h (U.E, BA)$

Séquence Multidimensionnelle h-généralisée

$s = \langle i_1, \ldots, i_j \rangle$ est une liste ordonnée non vide d'itemsets multidimensionnels h-généralisés.

$\langle \{(France, 4), (USA, 6)\}\{(Allemagne, 5)\} \rangle$

Support

HYPE

Introduction
OLAP et ECD
Motifs séquentiels
Contexte Multidimensionnel
Hiérarchies
Contributions
Modèle de données
Définitions
Algorithmes
Expérimentations
Conclusion et perspectives

Item supporté par une transaction

Une transaction b supporte un item e si $\Pi_{D_A}(b) <_h e$.

La transaction $(1, 1, \mathit{France}, 4)$ supporte l'item $(\mathit{UE}, \mathit{BA})$.

Séquence supportée par un bloc

Un bloc supporte une séquence $\langle i_1, \ldots, i_l \rangle$ si $\forall j = 1 \ldots l, \ \exists d_j \in \mathit{Dom}(D_j)$, pour chaque item e de i_j, $\exists t = (r, e, d_j)$ *ou* $t = (r, \check{e}, d_j) \in T$ avec $d_1 < d_2 < \ldots < d_l$.

Support d'une séquence

Étant donnés :

- D_R l'ensemble des dimensions de référence
- DB l'ensemble des transactions partitionné en un ensemble de blocs B_{T,D_R}
- une séquence ς

Support de ς

$$support(\varsigma) = \frac{|\{B \in B_{DB,D_R} \ t.q. \ B \ supporte \ \varsigma\}|}{|B_{DB,D_R}|}$$

Exemple

$D_R = \{B_{id}\}$, $D_A = \{Lieu,\ Produit\}$ et $D_T = \{Date\}$, $minsupp = 2$

Cherchons le support de la séquence

$\varsigma = \langle\{(UE, 2), (UE, BA)\}\ \{(UE, 8)\}\rangle$

BlocID	Date	Place	Product
1	1	Allemagne	2
1	1	Allemagne	5
1	2	Allemagne	8
1	3	Allemagne	1
1	4	Allemagne	7
2	1	France	6
2	2	France	2
2	2	France	4
2	3	France	8
3	1	UK	2
3	1	UK	3
3	2	UK	8
4	1	LA	1
4	2	LA	7
4	3	NY	3
4	4	NY	6

$\varsigma = \langle \{(UE, 2), (UE, BA)\} \{(UE, 8)\} \rangle$

Bloc 1

1	1	**Allemagne**	**2**
1	1	**Allemagne**	**5**
1	2	**Allemagne**	**8**
1	3	Allemagne	1
1	4	Allemagne	7

Bloc 1 supporte ς : $support(\varsigma)$ ++

Bloc 2

2	1	France	6
2	2	**France**	**2**
2	2	**France**	**4**
2	3	**France**	**8**

Bloc 2 supporte ς : $support(\varsigma)$ ++

$$\varsigma = \langle \{(UE, 2), (UE, BA)\} \{(UE, 8)\} \rangle$$

HYPE

Introduction
OLAP et ECD
Motifs séquentiels
Contexte Multidimensionnel
Hiérarchies
Contributions
Modèle de données
Définitions
Algorithmes
Expérimentations
Conclusion et perspectives

Bloc 3

3	1	**UK**	**2**
3	1	**UK**	**3**
3	2	**UK**	**8**

Bloc 3 supporte ς : $support(\varsigma)$ ++

Bloc 4

4	1	LA	1
4	2	LA	7
4	3	NY	3
4	4	NY	6

Bloc 4 ne supporte pas ς

$\varsigma = \langle \{(UE, 2), (UE, BA)\} \{(UE, 8)\} \rangle$

HYPE

Introduction
OLAP et ECD
Motifs séquentiels
Contexte Multidimensionnel
Hiérarchies
Contributions
Modèle de données
Définitions
Algorithmes
Expérimentations
Conclusion et perspectives

Bloc 3

3	1	**UK**	**2**
3	1	**UK**	**3**
3	2	**UK**	**8**

Bloc 3 supporte ς : $support(\varsigma)$ ++

Bloc 4

4	1	LA	1
4	2	LA	7
4	3	NY	3
4	4	NY	6

Bloc 4 ne supporte pas ς

$support(\varsigma) = 3 \geq minsupp$

- ς est fréquente

Fonctionnement Général

HYPE

Introduction
OLAP et ECD
Motifs séquentiels
Contexte Multidimensionnel
Hiérarchies
Contributions
Modèle de données
Définitions
Algorithmes
Expérimentations
Conclusion et perspectives

Génération des items candidats

- extraire tous les **items maximalement spécifiques**
- génération par niveau

Génération de séquences candidates

- propriété d'anti-monotonie du support
- approche Apriori (générer - élaguer)
- utilisation d'un arbre préfixé pour stocker les séquences candidates (PSP)

Calcul du support :

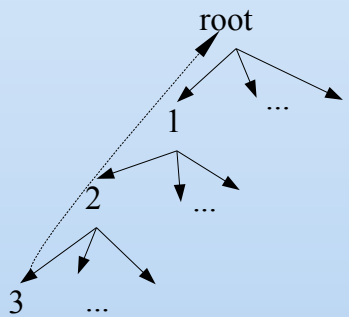
- prérequis : prétraitement des données (group by $date, D_1, \cdots, D_n$)
- calcul du support d'une séquence : **compterSupport(**$s, DB, \mathscr{D}_{\mathscr{R}}$**)**
  - Pour chaque bloc : **supportbloc(**s, B**)**
- ancrage ($\sigma_{condition}(B) \longmapsto C'$ with $B' \subseteq B$)

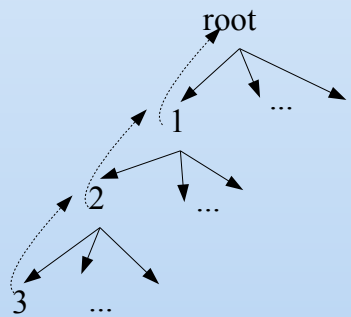
complexité

- n_B : nombre de nuplets dans B
- $m = |D_A|$: nombre de dimensions
- P_{max} la profondeur maximale des taxonomies
- supportBloc :$O(P_{max} \times n_B \times m \times \log n_B)$
- compterSupport :$O(I \times P_{max} \times n_{max} \times m \times \log n_{max})$

Une gestion plus fine des valeurs jokers

- Avant (M^2SP), une gestion "binaire" des valeurs jokers.
- Maintenant, une gestion plus fine grâce à la prise en compte des hiérarchies permettant ainsi l'extraction de connaissances plus précises.



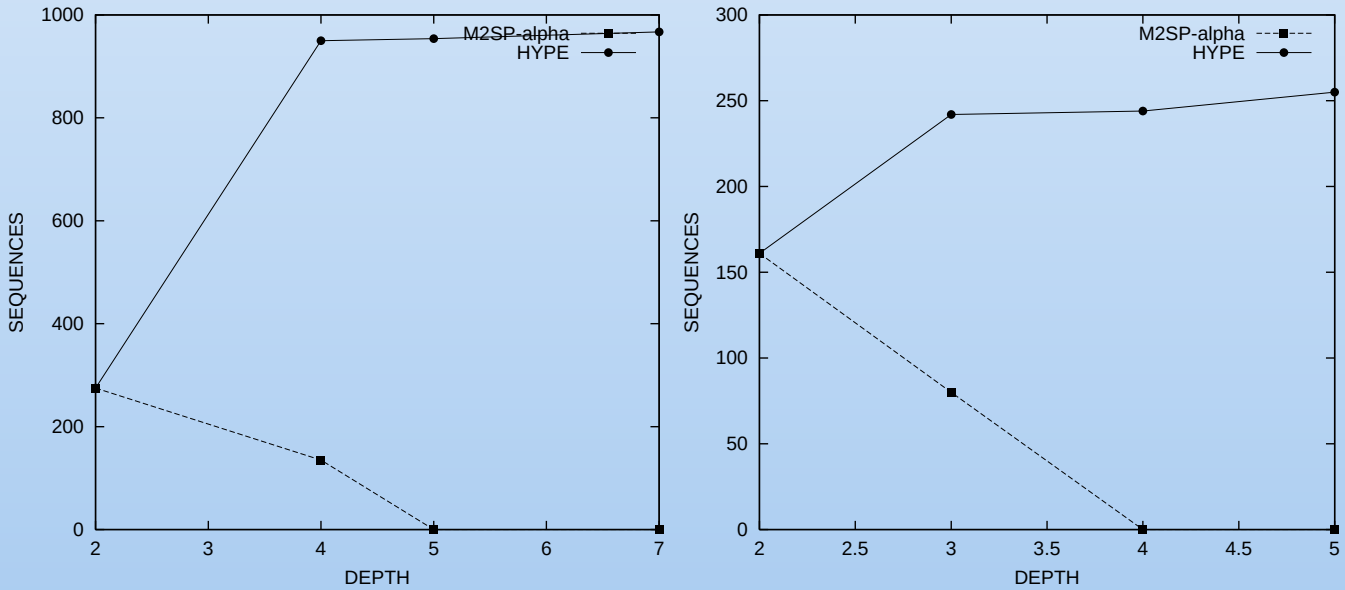


Protocole expérimental

- jeu de données synthétiques
- 5000 n-uplets
- $|D_A| = 5$
- Étude du nombre de fréquents extraits en fonction de la profondeur des taxonomies (degré de spécialisation) et du seuil de support considéré.
- Comparaison avec M^{2}SP(-α) afin d'étudier la qualité des connaissances extraites.

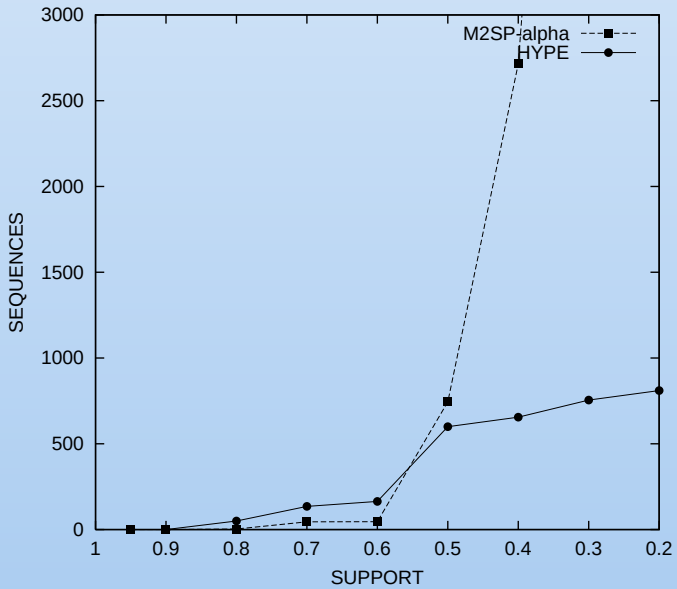
Nombre de séquences fréquentes par rapport à la profondeur de la taxonomie :

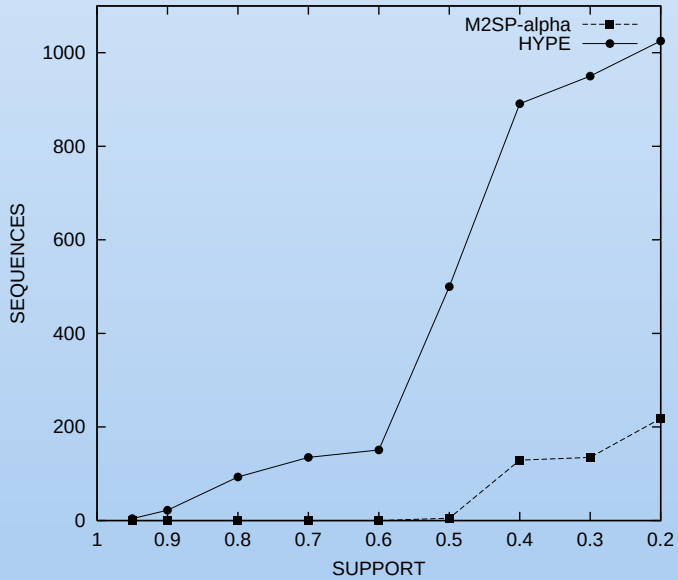
- minsup=0.3, deg = 3
- minsup=0.4, deg = 4



Nombre de séquences fréquentes par rapport au support :

- Données denses (faible degré)
- Données non denses (degré plus important)





Plan

HYPE

Introduction
OLAP et ECD
Motifs séquentiels
Contexte Multidimensionnel
Hiérarchies
Contributions
Modèle de données
Définitions
Algorithmes
Expérimentations
Conclusion et perspectives

1 Introduction
- OLAP et Fouille de données
- Motifs Séquentiels
- Contexte Multidimensionnel
- Hiérarchies

2 Contributions
- Modèle de données
- Définitions
- Algorithmes
- Expérimentations

3 Conclusion et perspectives

Prise en compte des hiérarchies dans l'extraction de motifs séquentiels multidimensionnels

- Définition des motifs séquentiels h-généralisés.
- Des connaissances extraites plus "*fines*".
- Une approche plus robuste face au dilemme "support/Connaissances extraites" grâce à faculté de subsomption.

Perspectives

- Utilisation de *représentations condensées* (clos, libres).
- Gestion modulaire des hiérarchies pour permettre à l'utilisateur l'extraction de connaissances *ciblées*
- . . .