Aix∗Marseille université

Approximation de bordures de motifs fréquents par le calcul de traverses minimales approchées d'hypergraphes

Nicolas Durand & Mohamed Quafafou

LSIS UMR 7296 - Aix-Marseille Université

Conférence Francophone sur l'Apprentissage Automatique (CAP)
4 juillet 2013

Plan

Introduction

- Découverte de motifs fréquents, règles d'association [AIS93]
- Contexte de fouille de données : $\mathcal{D} = (\mathcal{T}, \mathcal{I}, \mathcal{R})$ où $\mathcal{R} \subseteq \mathcal{T} \times \mathcal{I}$
  Trouver l'ensemble S des motifs fréquents :
  $\{X \subseteq \mathcal{I} \text{ tq. } |\{t \in \mathcal{T} \text{ tq. } \forall i \in X, (t,i) \in \mathcal{R}\}| \geq \mathit{minsup}\}$

Exemple

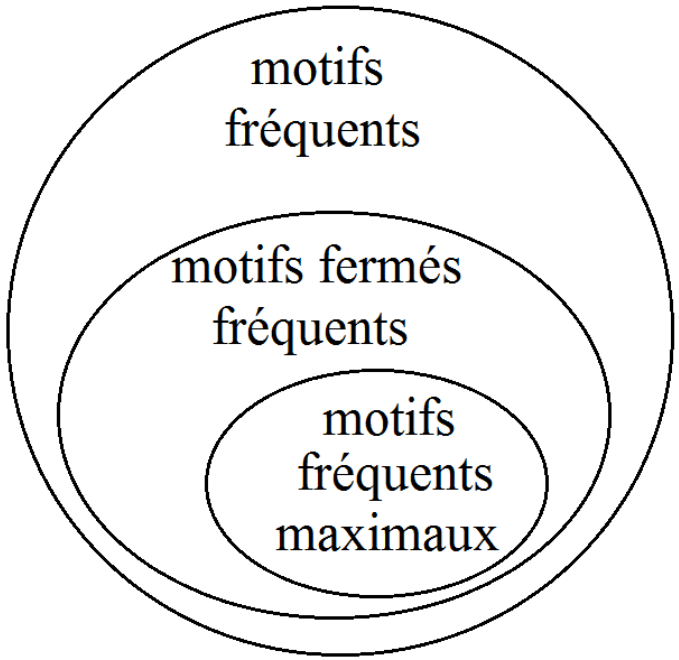
Id	Items
t_1	A C E G
t_2	B C E G
t_3	A C E H
t_4	A D F H
t_5	B C F H
t_6	B C E F H

minsup = 3
A fréquent (support=3)
BC fréquent (support=3)
EG infréquent (support=2)

motifs 3-fréquents = $\{A, B, C, E, F, H, BC, CE, CH, FH\}$

Introduction

- Problèmes rencontrés :
  - Vaste espace de recherche
  - Nombre élevé de motifs produits
- Notre contribution : réduction du nombre de motifs fréquents maximaux via un calcul approché (de bordures de motifs fréquents)



Exemple

motifs fermés 3-fréquents $= \{A, C, H, BC, CE, CH, FH\}$

Plan

Bordures de motifs fréquents

Définition (Bordure positive et bordure négative (MT97))

La bordure positive (resp. négative) de S, notée $Bd^+(S)$ (resp. $Bd^-(S)$), est constituée par les motifs fréquents maximaux (resp. infréquents minimaux) (au sens de l'inclusion) de $\mathcal{D}$.

$Bd^+(S) = \{X \in S \mid \forall Y \ tq \ X \subset Y, \ Y \notin S\}$

$Bd^-(S) = \{X \in 2^{\mathcal{I}} \setminus S \mid \forall Y \ tq \ Y \subset X, \ Y \in S\}$

Exemple

Si *minsup*=3

$Bd^+(S) = \{A, BC, CE, CH, FH\}$

$Bd^-(S) = \{D, G, AB, AC, AE, AF, AH, BE, BF, BH, CF, EF, EH\}$

Hypergraphes et traverses minimales

Définition (Hypergraphe (Berge89))

Hypergraphe $\mathcal{H} = (V, E)$, ensemble V de sommets et ensemble E d'hyperarêtes ($\forall e \in E$, $e \subseteq V$).

Définition (Traverse et traverse minimale (Berge89))

$\tau \subseteq V$ est une traverse de $\mathcal{H}$ ssi $\forall e \in E, \tau \cap e \neq \emptyset$.
Traverse τ de $\mathcal{H}$ est minimale si $\nexists \tau' \subset \tau$ tq. τ' est une traverse de $\mathcal{H}$.
$MinTr(\mathcal{H})$: ensemble des traverses minimales de $\mathcal{H}$.

Exemple

Exemple précédent $= \mathcal{H}$. BC n'est pas une traverse. ABC est une traverse mais pas minimale car AC est une traverse (minimale).
$MinTr(\mathcal{H}) = \{AB, AC, CD, CF, CH, EF, EH, GH, AFG, BDE\}$

Bordures et traverses minimales : *dualisation*

Propriété (De la bordure positive à la bordure négative (MT97))

$Bd^-(S) = MinTr(\overline{Bd^+(S)})$
où $\overline{Bd^+(S)}$ *représente l'hypergraphe dont les sommets sont les items de* $\mathcal{I}$ *et les hyperarêtes sont les complémentaires des motifs de la bordure positive de* S.

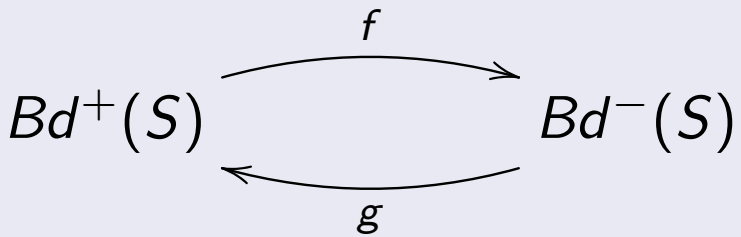
Propriété (De la bordure négative à la bordure positive (DMP03))

$Bd^+(S) = \overline{MinTr(Bd^-(S))}$
où $Bd^-(S)$ *représente l'hypergraphe dont les sommets sont les items de* $\mathcal{I}$ *et les hyperarêtes sont les motifs de la bordure négative de* S.

Plan

Approche proposée d'approximation de bordures

- Exploitation des dualisations entre la bordure positive et la bordure négative
- Notons f et g les fonctions qui permettent de passer respectivement de $Bd^+(S)$ à $Bd^-(S)$ et de $Bd^-(S)$ à $Bd^+(S)$

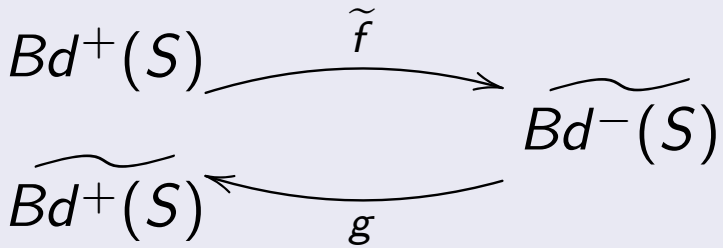


- Nouvelle fonction $\widetilde{f}$ qui utilise un calcul de traverses minimales approchées noté $\widetilde{MinTr}$

Approche proposée d'approximation de bordures

$$\widetilde{f}(Bd^+(S)) = \widetilde{MinTr}(\overline{Bd^+(S)}) = \widetilde{Bd^-(S)}$$

$$g(\widetilde{Bd^-(S)}) = \overline{MinTr(\widetilde{Bd^-(S)})} = \widetilde{Bd^+(S)}$$



- Résultat : bordure négative approchée $\widetilde{Bd^-(S)}$ et la bordure positive approchée correspondante $\widetilde{Bd^+(S)}$

Approche proposée d'approximation de bordures

Exemple

$\widetilde{Bd^-(S)} = \widetilde{f}(Bd^+(S)) = \widetilde{MinTr}(\overline{Bd^+(S)})$
$= \widetilde{MinTr}(\{\overline{A}, \overline{BC}, \overline{CE}, \overline{CH}, \overline{FH}\}$
$= \widetilde{MinTr}(\{BCDEFGH, ADEFGH, ABDFGH, ABDEFG, ABCDEG\})$
Supposons que le calcul de traverses minimales approchées nous donne :
$\widetilde{Bd^-(S)} = \{D, E, G, AF, AH, BF, BH\}$.
Dualisons pour obtenir la bordure positive approchée :
$\widetilde{Bd^+(S)} = g(\widetilde{Bd^-(S)}) = MinTr(\widetilde{Bd^-(S)})$
$= \{\overline{ABDEG}, \overline{DEFGH}\} = \{ABC, CFH\}$.

- ABC n'est pas fréquent ni existant, mais A, B, C et BC sont fréquents. CFH n'est pas fréquent mais il l'est presque.
- Notons que les hyperarêtes de $\overline{Bd^+(S)}$ ont de fortes intersections.

Plan

Traverses minimales approchées

- La méthode proposée est constituée de 2 phases :
  1. Réduction de l'hypergraphe $\mathcal{H}$
     avec algorithme spécialement conçu pour le calcul de traverses minimales
     (se base sur les intersections des hyperarêtes et le nombre d'occurrences de chaque sommet)
  2. Calcul des traverses minimales de l'hypergraphe réduit $\mathcal{H}_R$

$$\widetilde{MinTr}(\mathcal{H}) = MinTr(\mathcal{H}_R)$$

Traverses minimales approchées

- Réduction d'un hypergraphe $\mathcal{H} = (V, E)$ en 3 étapes :
  1. Calcul du nbre d'occurrences de chaque sommet dans les hyperarêtes de $\mathcal{H}$,
  2. Calcul des intersections de chaque paire d'hyperarêtes et construction d'un graphe valué $G = (V', E')$ où un sommet v'_i représente une hyperarête e_i de $\mathcal{H}$, une arête entre v'_i et v'_j traduit une intersection non vide entre les deux hyperarêtes e_i, e_j. Poids d'une arête (v'_i, v'_j) : $w_{(v'_i,v'_j)} = \sum_{v \in e_i \cap e_j} occur[v]$
  3. Sélection d'arêtes de G et génération de l'hypergraphe réduit $\mathcal{H}_R$=$(V_R$,$E_R)$: algorithme glouton sélectionnant l'arête ayant le plus fort poids tant qu'il reste des arêtes à selectionner. Chaque arête sélectionnée est transformée en une hyperarête pour $\mathcal{H}_R$ (elle contient les sommets de $\mathcal{H}$ correspondant à l'intersection).
- Algorithme en $O(m^2)$ où $m = |E|$

Traverses minimales approchées : exemple

Exemple

Considérons notre exemple comme un hypergraphe $\mathcal{H}$.

Calcul des nombres d'occurrences :

$occur[A] = 3$, $occur[B] = 3$, $occur[C] = 5$, $occur[D] = 1$, $occur[E] = 4$, $occur[F] = 3$, $occur[G] = 2$ et $occur[H] = 4$.

Calcul des intersections des hyperarêtes :

par exemple, $e_5 \cap e_6 = \{B, C, F, H\}$.

$w_{(v'_5, v'_6)} = occur[B] + occur[C] + occur[F] + occur[H] = 15$.

Matrice d'adjacence du graphe valué généré :

$$\begin{pmatrix} 0 & 11 & 12 & 3 & 5 & 9 \\ 11 & 0 & 9 & 0 & 8 & 12 \\ 12 & 9 & 0 & 7 & 9 & 13 \\ 3 & 0 & 7 & 0 & 7 & 7 \\ 5 & 8 & 9 & 7 & 0 & 15 \\ 9 & 12 & 13 & 7 & 15 & 0 \end{pmatrix}$$

Traverses minimales approchées : exemple

Sélection d'arête pour générer l'hypergraphe réduit :
(v'_5, v'_6) est sélectionnée car poids le plus élevé. Les arêtes où figurent v'_5 ou v'_6 sont supprimées. On a $V_R = \{B, C, F, H\}$ et $E_R = \{\{B, C, F, H\}\}$.
Le matrice d'adjacence du graphe restant est :

$$\begin{pmatrix} 0 & 11 & 12 & 3 & 0 & 0 \\ 11 & 0 & 9 & 0 & 0 & 0 \\ 12 & 9 & 0 & 7 & 0 & 0 \\ 3 & 0 & 7 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

(v'_1, v'_3) est sélectionnée. Après suppression, plus d'arêtes donc fin.
Donc $\mathcal{H}_R = (V_R, E_R)$ où $V_R = \{A, B, C, E, F, H\}$ et $E_R = \{\{A, C, E\}, \{B, C, F, H\}\})$.

$MinTr(\mathcal{H}) = MinTr(\mathcal{H}_R) = \{C, AB, AF, AH, BE, EF, EH\}$
(rappel : $MinTr(\mathcal{H}) = \{AB, AC, CD, CF, CH, EF, EH, GH, AFG, BDE\}$)

Plan

Expérimentations : données

- 4 jeux de données classiques : Mushroom, Chess, Connect et Kosarak
- Couvrent les différents types selon 2 classifications : [GZ01] [FMP10]

Données	Nb transactions	Nb items	Taille moy transaction
Mushroom	8124	119	23
Chess	3196	75	37
Connect	67557	129	43
Kosarak	990002	41270	8,1

Expérimentations : protocole

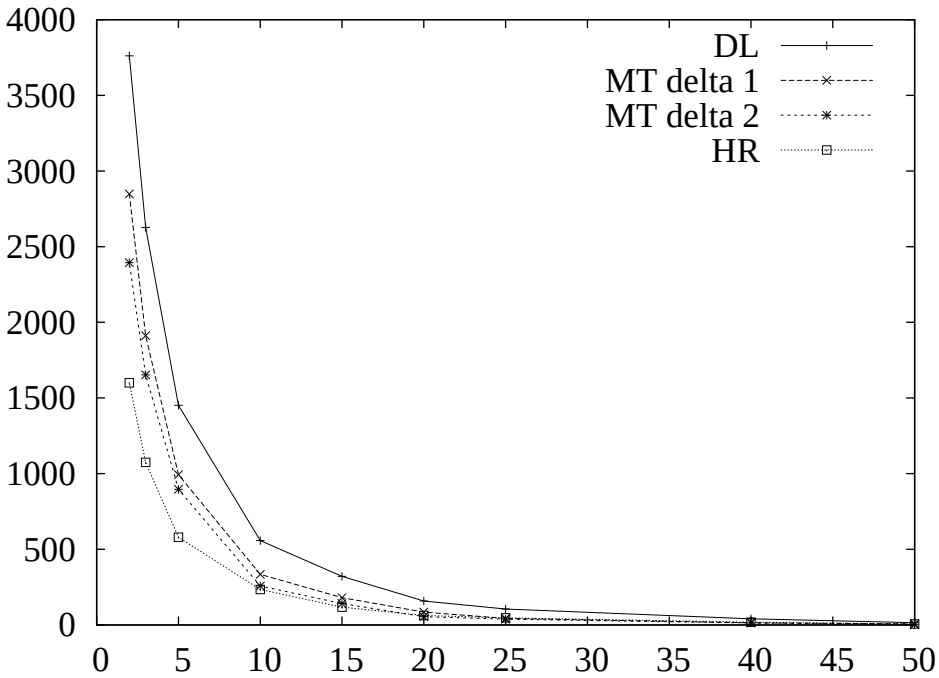
- Pour chaque jeu de données et pour différentes valeurs de seuil minimum de support,
  1. Calcul de $Bd^+(S)$ en utilisant *IBE* [SU03],
  2. Calcul de $Bd^-(S)$ avec *DL* [DL05] (la référence),
     de $\widetilde{Bd^-(S)}$ avec δ-*MTminer* [RZC10,HBC07] (pour δ=1 et δ=2)
     et de $\widetilde{Bd^-(S)}$ avec notre méthode (appelée *HR*),
  3. Dualisation vers 1 $Bd^+(S)$ et 3 $\widetilde{Bd^+(S)}$ avec *DL*.
- Pour chaque bordure calculée : nombre de motifs, taille moyenne d'un motif, distance avec la bordure exacte.

$$D(\mathcal{X},\mathcal{Y}) = \max\{\, h(\mathcal{X},\mathcal{Y}) \,,\, h(\mathcal{Y},\mathcal{X}) \,\}$$

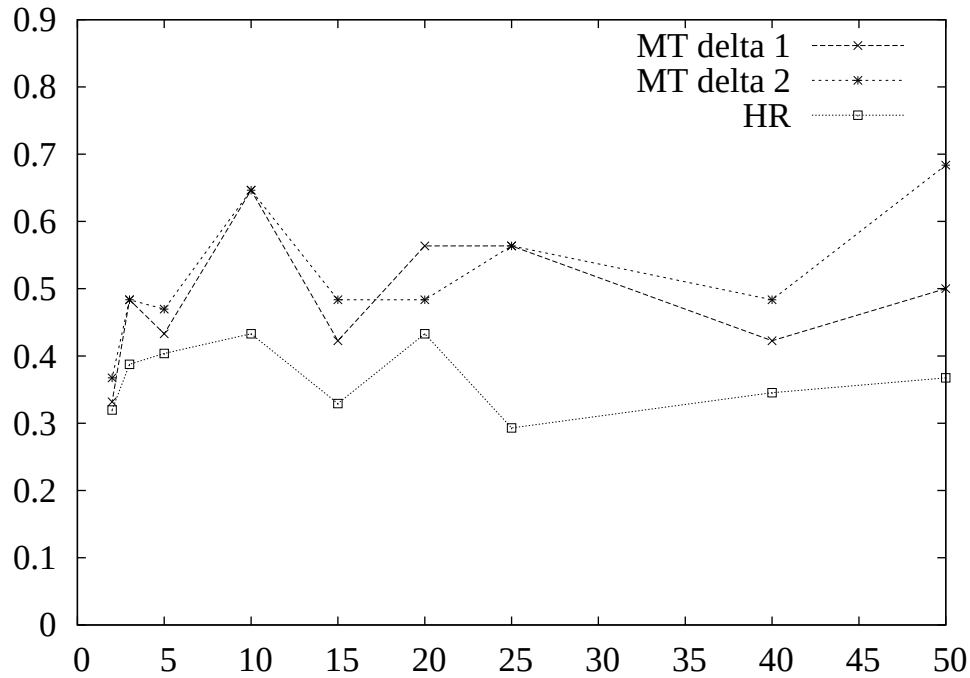
$$\text{avec } h(\mathcal{X},\mathcal{Y}) = \max_{X\in\mathcal{X}}\{\min_{Y\in\mathcal{Y}} d(X,Y)\} \text{ et } d(X,Y) = 1 - \frac{|X\cap Y|}{\sqrt{|X|\times|Y|}}$$

Expérimentations : résultats

Mushroom : $\widetilde{Bd^+(S)}$



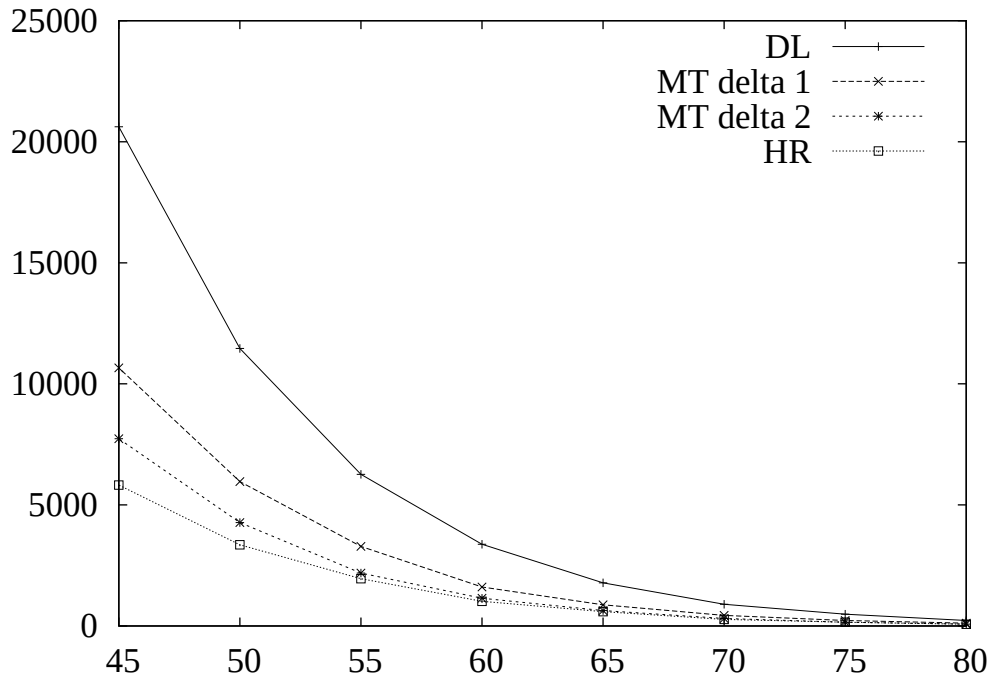
Nombre de motifs
en fonction de *minsup* (%)



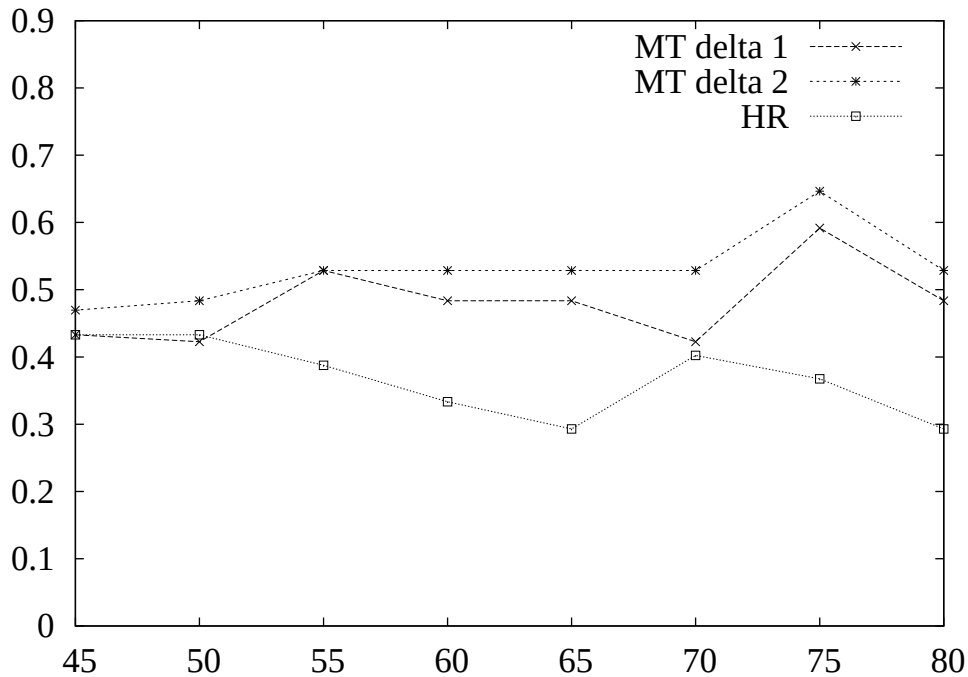
Distance avec $Bd^+(S)$
en fonction de *minsup* (%)

Expérimentations : résultats

Chess : $\widetilde{Bd^+(S)}$



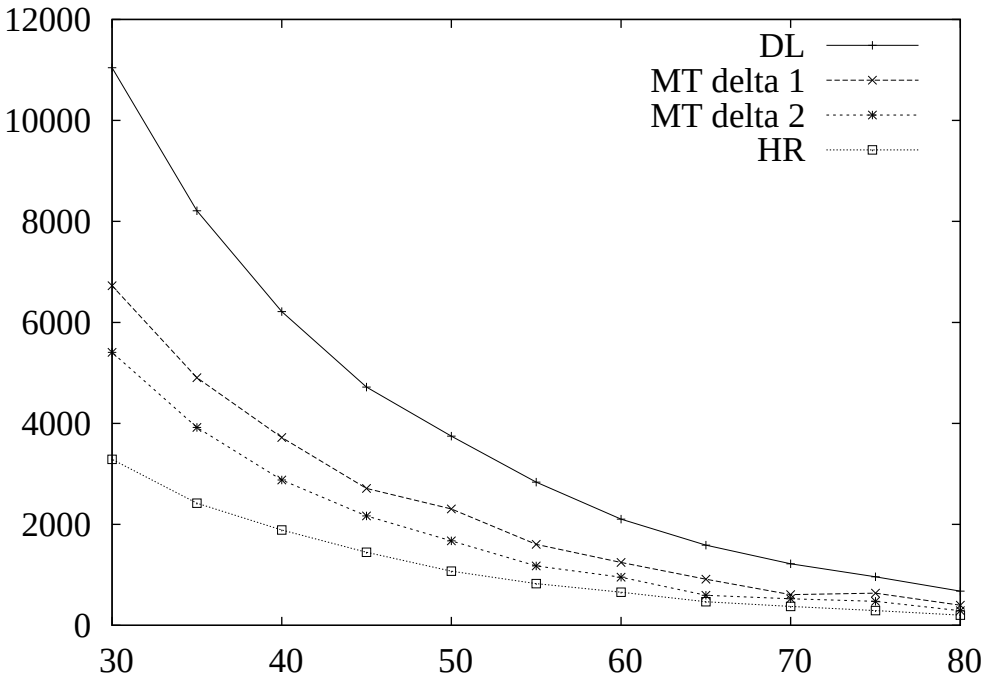
Nombre de motifs
en fonction de *minsup* (%)



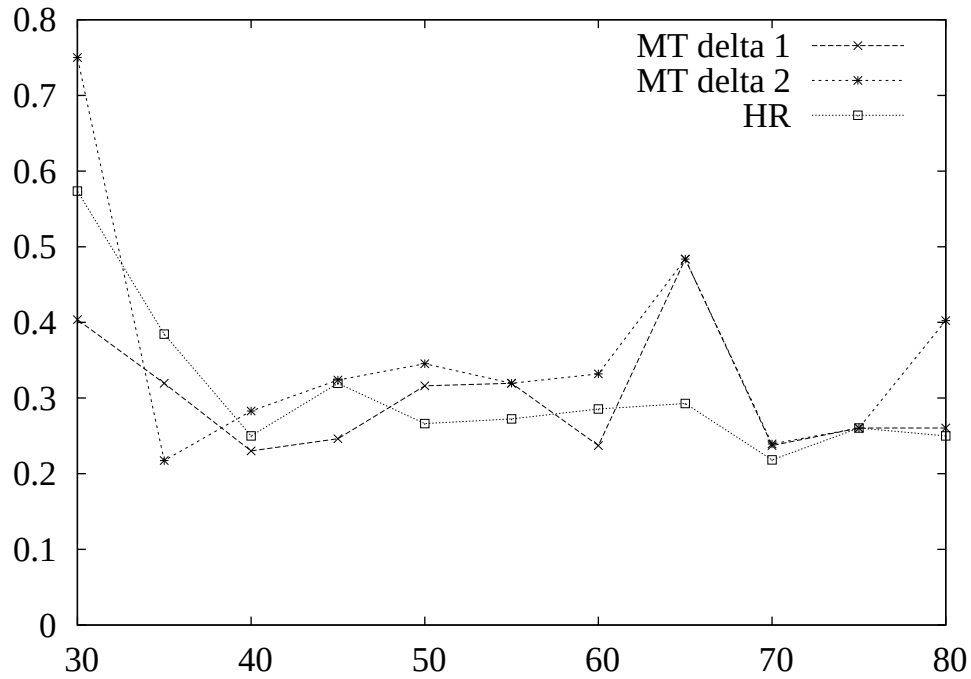
Distance avec $Bd^+(S)$
en fonction de *minsup* (%)

Expérimentations : résultats

Connect : $\widetilde{Bd^+(S)}$



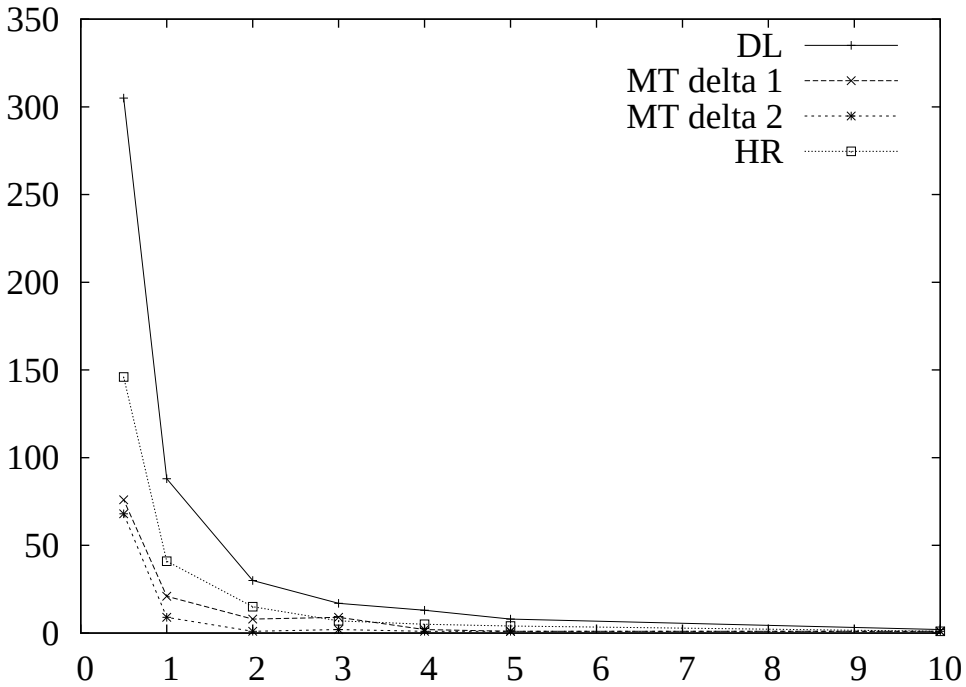
Nombre de motifs
en fonction de *minsup* (%)



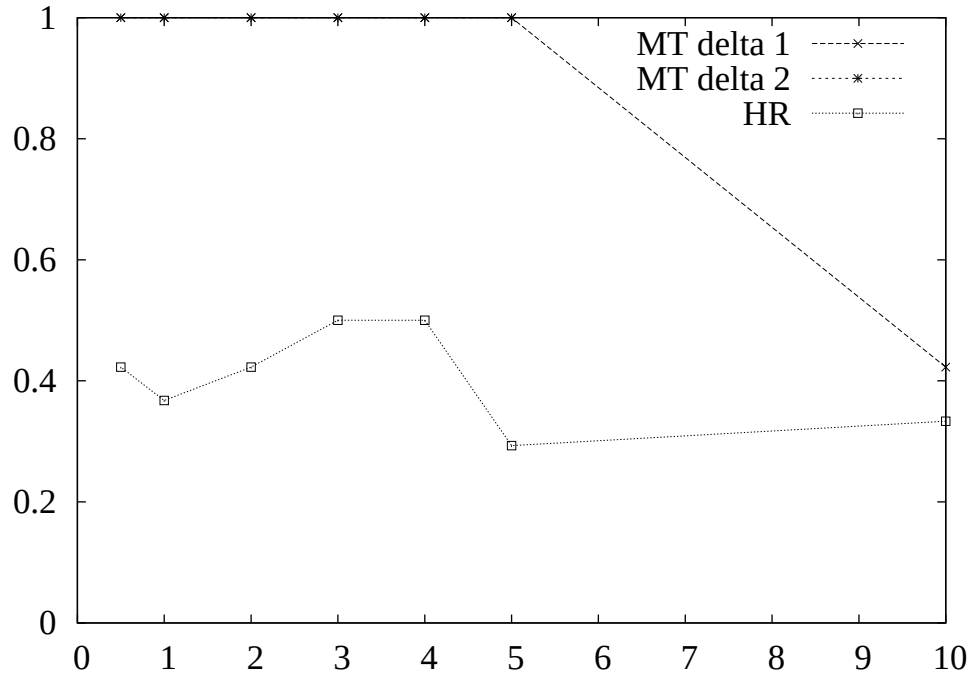
Distance avec $Bd^+(S)$
en fonction de *minsup* (%)

Expérimentations : résultats

Kosarak : $\widetilde{Bd^+(S)}$



Nombre de motifs
en fonction de *minsup* (%)



Distance avec $Bd^+(S)$
en fonction de *minsup* (%)

Plan

Conclusion

- Nouvelle approche d'approximation de bordures de motifs fréquents via le calcul de traverses minimales approchées d'hypergraphes.
- Nouvelle méthode pour calculer les traverses minimales approchées, basée sur la réduction d'hypergraphes.
- Pas de nouveaux paramètres à fixer.
- Expérimentations ont montré que
  - Notre proposition produit une bordure positive approximative plus petite que la bordure positive exacte, tout en gardant une distance raisonnable avec elle,
  - Semble robuste par rapport aux différents types de jeux de données que nous pouvons rencontrés,

Perspectives

- Utilisation de notre méthode pour introduire une part d'approximation dans des algorithmes qui se basent sur la dualisation (comme ABS - Adaptive Borders Search [FMP04])
- Développement de systèmes de recommandation utilisant les motifs de bordures positives approchées (dans le domaine de la découverte de services Web, la recherche de documents, ...).

Merci de votre attention