Clustering de motifs séquentiels

Application à la détection d'intrusions

SANEIFAR Hassan

– Université **M**ontpellier **II –**

Laboratoire d'**I**nformatique, de **R**obotique et de **M**icroélectronique de **M**ontpellier (LIRMM)

Équipe TATOO

INRIA Sophia Antipolis

Sous la direction :
Sandra Bringay et Anne Laurent

ARC SéSur INRIA

10 septembre 2008

Données

Logs des connexions et des activités de réseaux $\Longrightarrow$ Données séquentielles.

Système de Détection d'Intrusions (IDS)

- Intrusion : Une série d'actions qui compromettent l'intégrité, la confidentialité ou la disponibilité d'une ressource [SSM06].
- Détection d'intrusions : Suivi et analyse des événements survenus dans un système pour trouver les signes d'intrusions.

Extraction de Connaissances dans les grandes bases de Données.

Processus d'extraction de connaissances nouvelles, potentiellement utiles et ayant un degré de plausibilité, dans de grands volumes de données [Fio07]

Objectif

Utilisation des techniques d'ECD pour analyser les événements survenus afin de détecter des intrusions.

Système de détection d'intrusions (IDS)
Approches générales d'IDS

Anomaly Detection

- **On sait ce qui est bon.**
- Construire des profils représentant les comportements généraux d'un utilisateur, d'un groupe ou d'un domaine de réseau,
- Une déviation par rapport à un comportement normal $\Longrightarrow$ Intrusion.

Misuse Detection

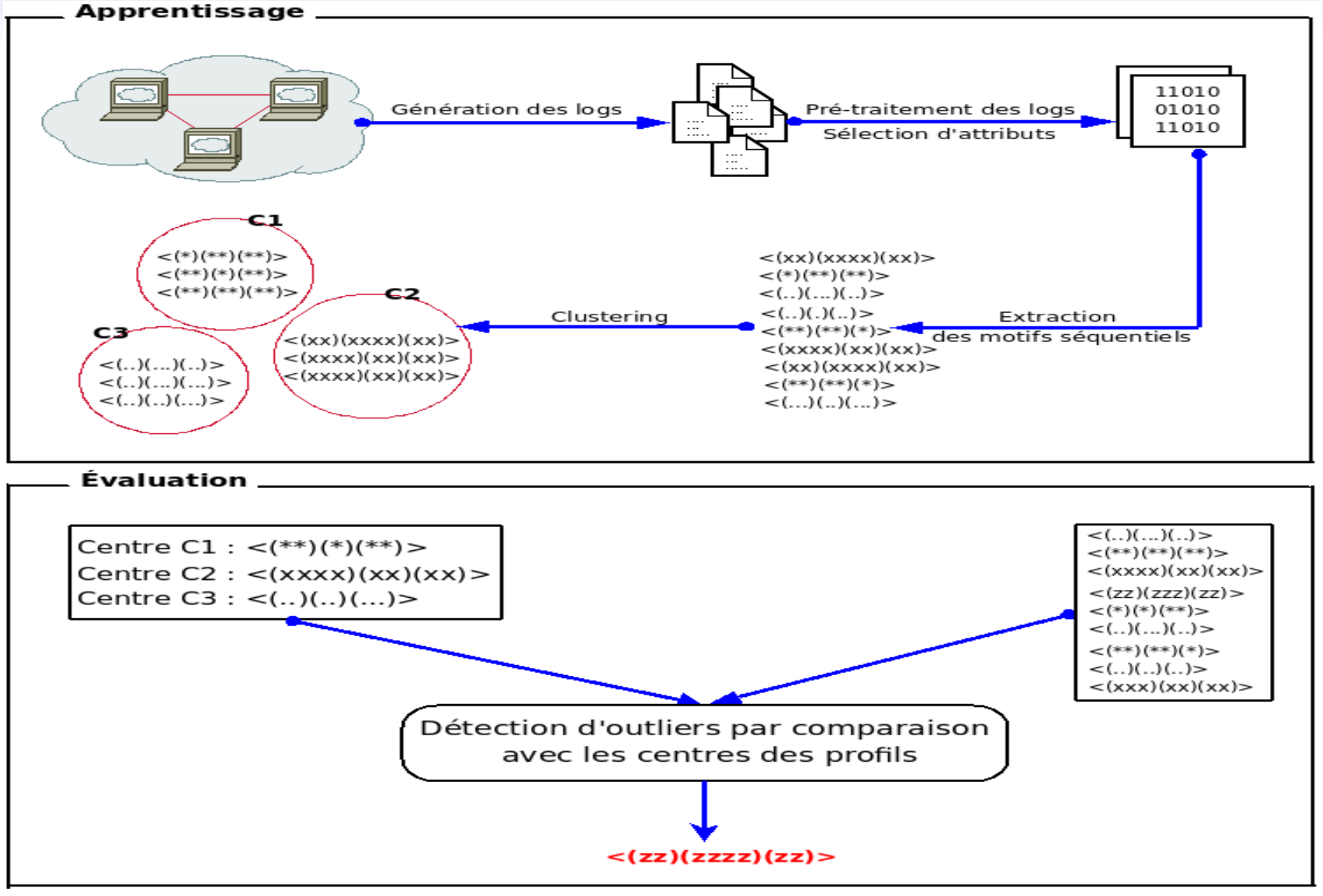
- **On sait ce qui est mauvais.**
- Modélisation des intrusions connues sous forme de signatures d'attaques,
- Identification directe des attaques déjà connues en comparant des activités avec ces signatures.

Plan

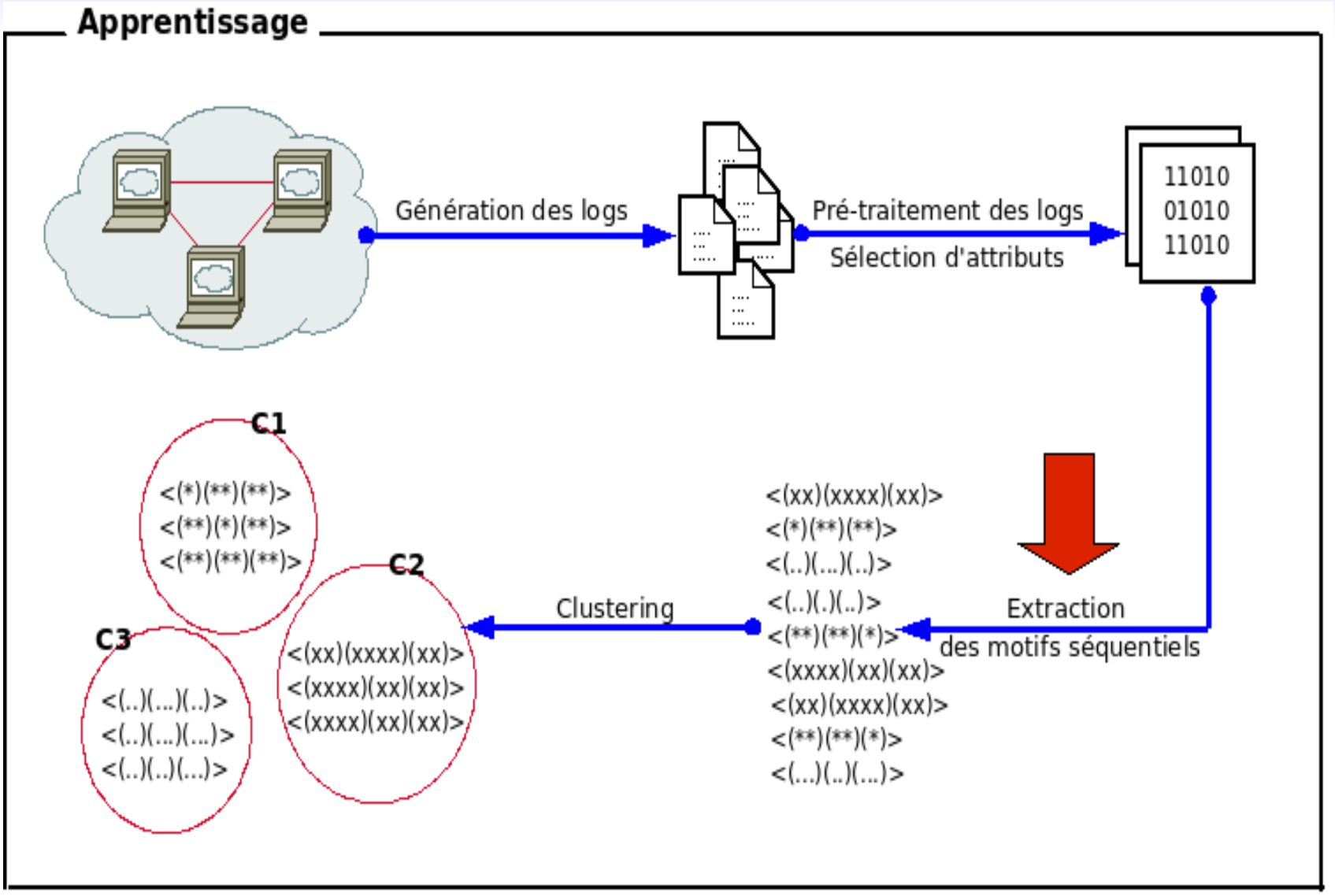
- **Présentation de l'approche proposée pour la détection d'anomalies**
- **Mesure de similarité pour les motifs séquentiels**
  - Ce qui existe...
  - Proposition d'une mesure de similarité pour les motifs séquentiels
  - Discussions
- **Expérimentations**
- **Conclusions et perspectives**

Plan

Notre approche de la détection d'anomalies
Schéma Global



Notre approche de la détection d'anomalies
Modélisation de comportements par motifs séquentiels



Définition d'un motif séquentiel

Motif séquentiel

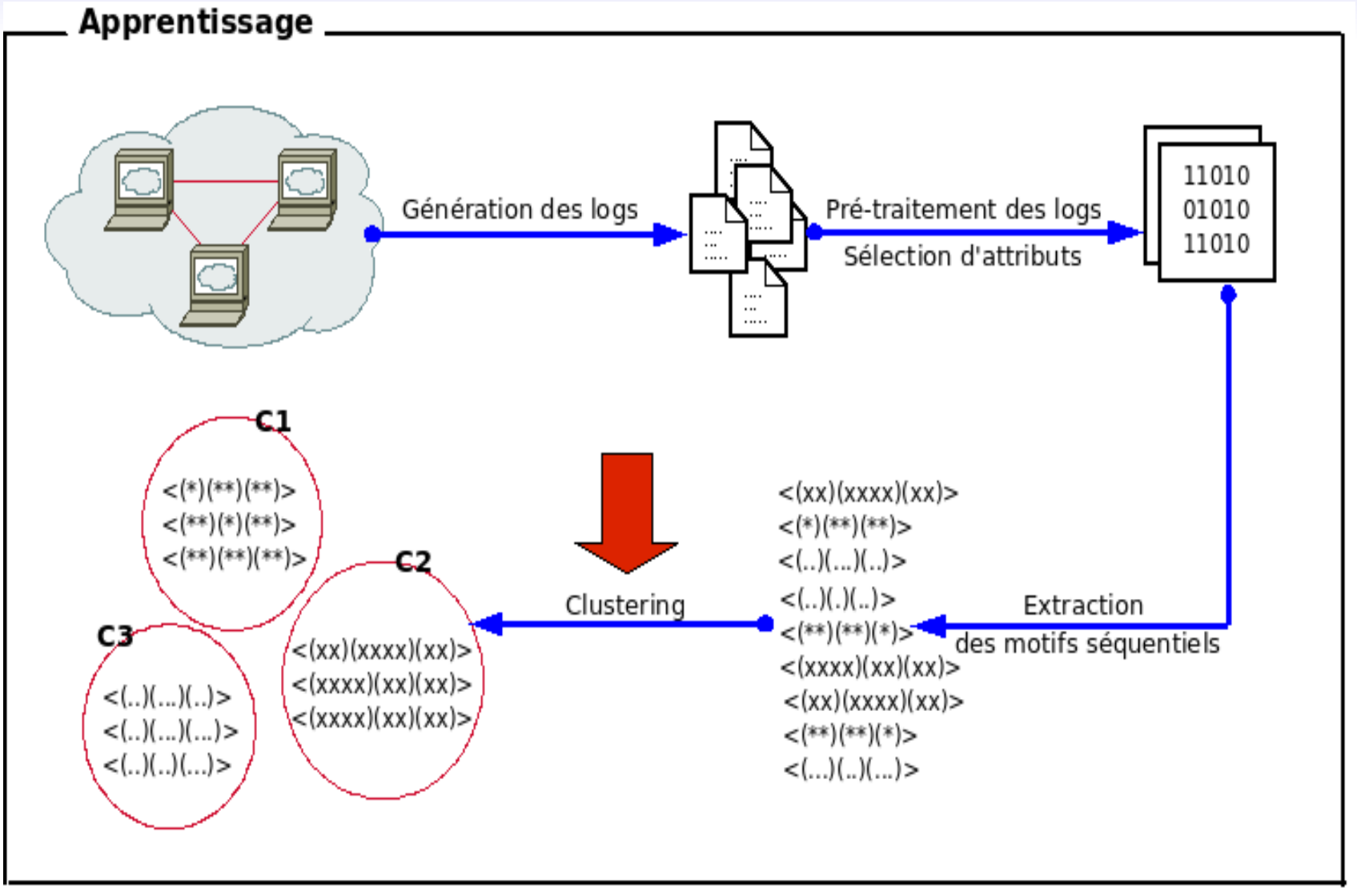
Une liste **ordonnée** non vide d'**itemsets** où **itemset** est un ensemble **non vide** et **non-ordonné** d'items

Exemple

$\langle (Chocolat, Soda)(gâteau, chips)(minceur) \rangle$

Fréquemment les clients achètent d'abord du chocolat et du soda, puis dans une prochain achat, ils achètent des gâteaux et des chips et ensuite ils reviennent plus tard pour acheter des produits minceurs.

Notre approche de la détection d'anomalies
Création des profils de comportements : clustering de motifs séquentiels



Création des profils de comportements
Pourquoi un clustering de motifs séquentiels ?

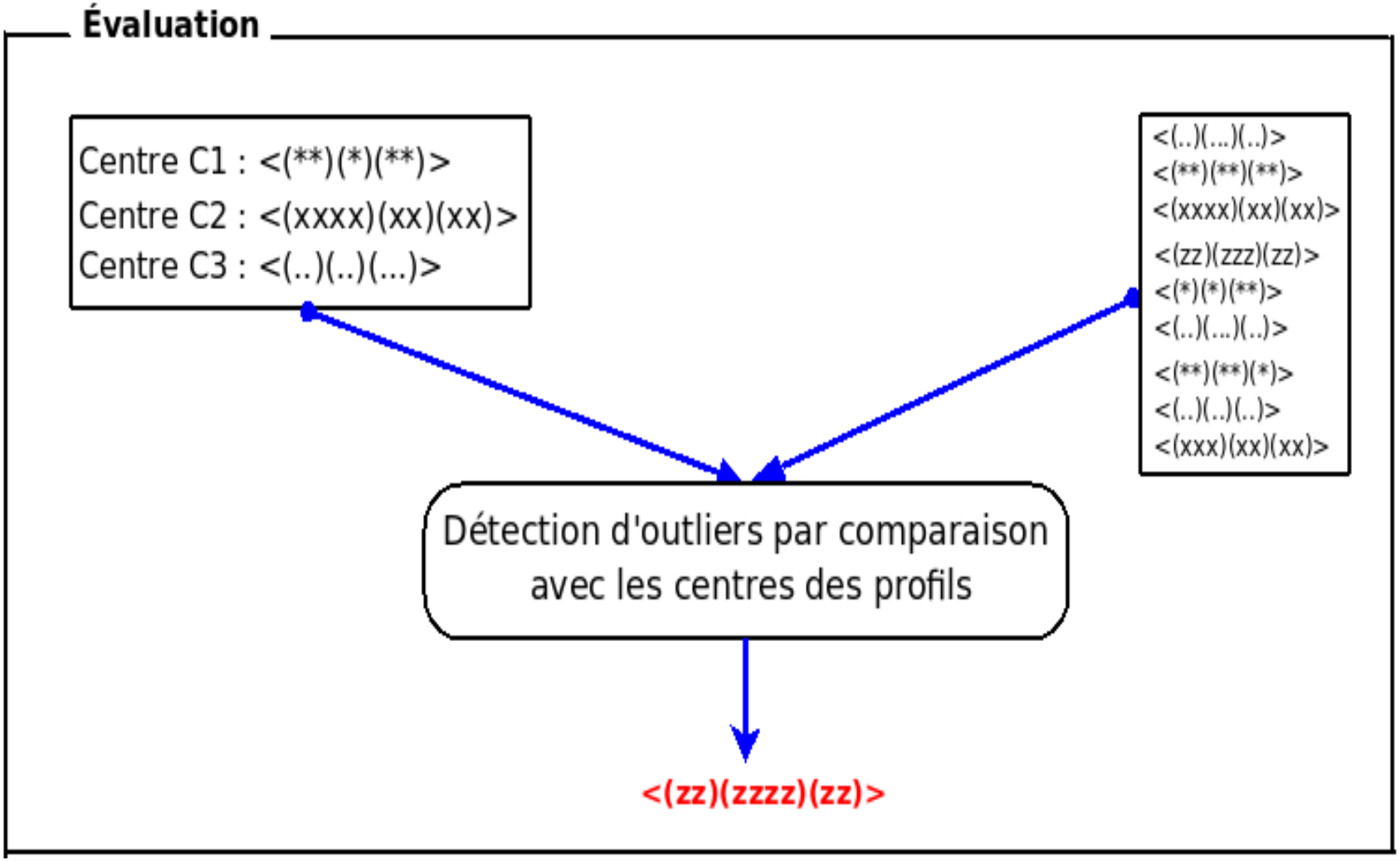
- Des motifs extraits **volumineux**
- Besoin de **regrouper les comportements similaires** (Profils de comportements)
- Choisir d'un motif comme le **représentant de chaque groupe** (profil) qui décrit **mieux** le comportement du groupe que les autres motifs
- **Facilité** pour trouver des déviations uniquement par comparaison avec les représentants de chaque groupe

Création des profils de comportements
Pourquoi un clustering de motifs séquentiels ?

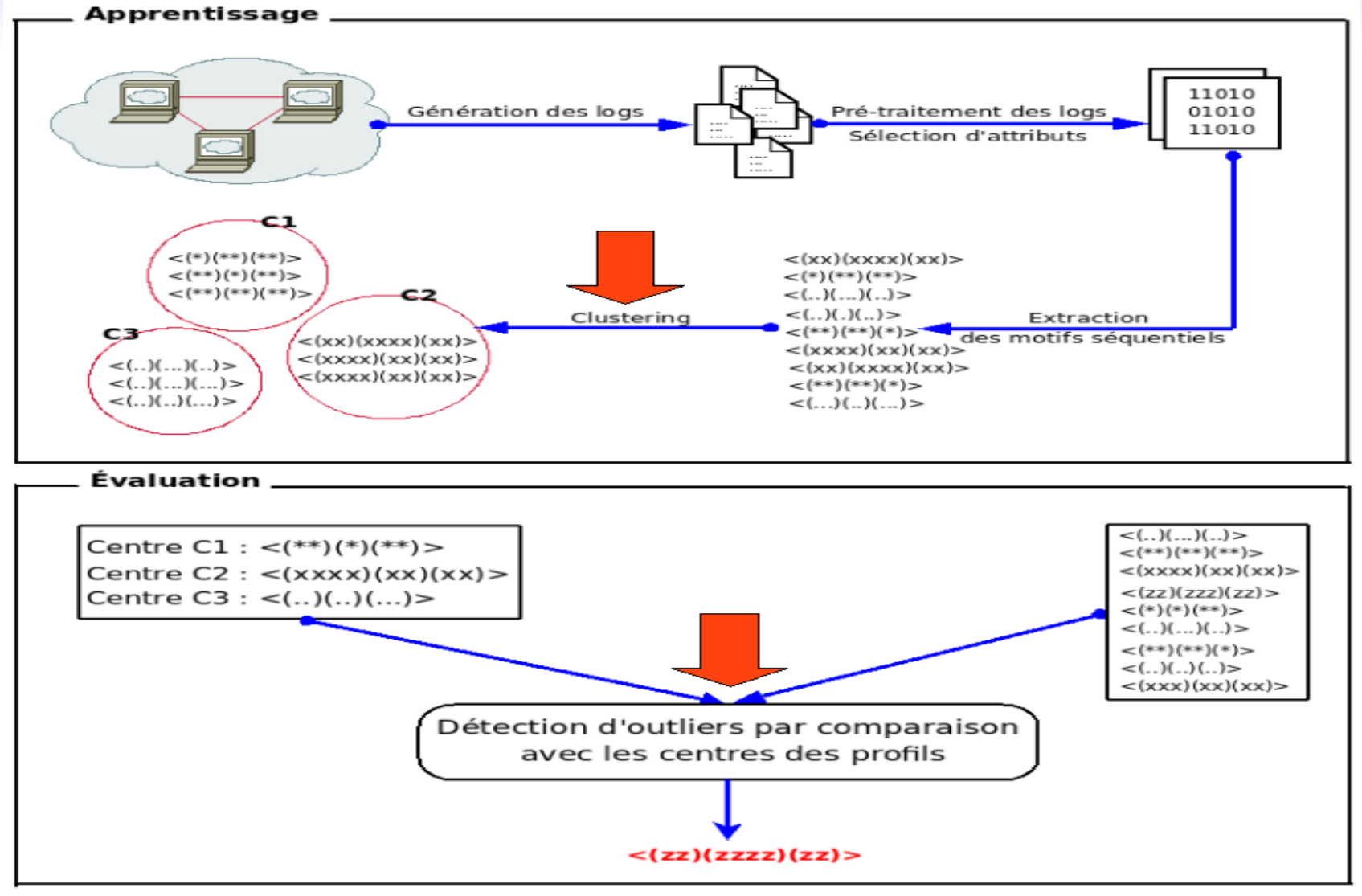
- Des motifs extraits **volumineux**
- Besoin de **regrouper les comportements similaires** (Profils de comportements)
- Choisir d'un motif comme le **représentant de chaque groupe** (profil) qui décrit **mieux** le comportement du groupe que les autres motifs
- **Facilité** pour trouver des déviations uniquement par comparaison avec les représentants de chaque groupe

Clustering de motifs séquentiels

Notre approche de la détection d'anomalies
Phase évaluation : détection d'intrusions



Notre approche de la détection d'anomalies



Clustering de motifs séquentiels

Les problèmes

- **Clustering** de motifs séquentiels
- Comparer **la ressemblance de deux motifs séquentiels** dans la phase d'évaluation
- Besoin d'une **mesure de similarité** pour les motifs séquentiels
- **Peu de travaux** sur la comparaison des séquences d'itemsets (motifs séquentiels)

Plan

Mesure de similarité pour les motifs séquentiels
Mesures existant / Edit Distance

Edit Distance :
Le nombre minimal de modifications pour passer d'une séquence à une autre.

Les **opérateurs de modification** (*fréquemment utilisés*) :
insertion, suppression et substitution.

Exemple

$S_1 = \{A, B, A, C, B, D\}, S_2 = \{A, B, C, C, A, D\}$.
$O = \{Supp(A_3), Ins(C_4), Supp(B_5), Ins(A_5)\}$

Mesure de similarité pour les motifs séquentiels
Mesures existant / Edit Distance

Edit distance non pertinente pour comparer les motifs séquentiels ?

Itemset≡Symbole (selon la définition dans [CMB02]).

Mesure de similarité pour les motifs séquentiels
Mesures existant / Edit Distance

Edit distance non pertinente pour comparer les motifs séquentiels ?

Itemset≡Symbole (selon la définition dans [CMB02]).

- $M_1 = \{(ab)(c)\}, M_2 = \{(a)(c)\}$:

Mesure de similarité pour les motifs séquentiels
Mesures existant / Edit Distance

Edit distance non pertinente pour comparer les motifs séquentiels ?

Itemset≡Symbole (selon la définition dans [CMB02]).

- $M_1 = \{(ab)(c)\}, M_2 = \{(a)(c)\}$:

$M_1\{(ab)(c)\} \Rightarrow$ X $\rightarrow$ Y | X=ab, Y=c (X,Y = symboles)

$M_2\{(a)(c)\} \Rightarrow$ Z$\rightarrow$ Y | Z=a, Y=c (Y,Z = symboles)

Mesure de similarité pour les motifs séquentiels
Mesures existant / Edit Distance

Edit distance non pertinente pour comparer les motifs séquentiels ?

Itemset≡Symbole (selon la définition dans [CMB02]).

- $M_1 = \{(ab)(c)\}, M_2 = \{(a)(c)\}$:

$M_1\{(ab)(c)\} \Rightarrow$ X $\rightarrow$ Y | X=ab, Y=c (X,Y = symboles)

$M_2\{(a)(c)\} \Rightarrow$ Z$\rightarrow$ Y | Z=a, Y=c (Y,Z = symboles)

- $O_{Edit} = \{Substitution(X, Z, 1)\}$

Mesure de similarité pour les motifs séquentiels
Mesures existant / Edit Distance

Edit distance non pertinente pour comparer les motifs séquentiels ?

Itemset≡Symbole (selon la définition dans [CMB02]).

- $M_1 = \{(ab)(c)\}, M_2 = \{(a)(c)\}$:

$M_1\{(ab)(c)\} \Rightarrow$ X $\rightarrow$ Y | X=ab, Y=c (X,Y = symboles)

$M_2\{(a)(c)\} \Rightarrow$ Z$\rightarrow$ Y | Z=a, Y=c (Y,Z = symboles)

- $O_{Edit} = \{Substitution(X, Z, 1)\}$

(ab) et (a) sont vus comme deux éléments complétement différents.

Mesure de similarité pour les motifs séquentiels
Mesures existant / Edit Distance

Edit distance non pertinente pour comparer les motifs séquentiels ?

Itemset≡Symbole (selon la définition dans [CMB02]).

- $M_1 = \{(ab)(c)\}, M_2 = \{(a)(c)\}$:

$M_1\{(ab)(c)\} \Rightarrow$ X $\rightarrow$ Y | X=ab, Y=c (X,Y = symboles)

$M_2\{(a)(c)\} \Rightarrow$ Z$\rightarrow$ Y | Z=a, Y=c (Y,Z = symboles)

- $O_{Edit} = \{Substitution(X, Z, 1)\}$

(ab) et (a) sont vus comme deux éléments complétement différents.

(ab) et (a) sont deux comportements ressemblants.

Edit Distance et Séquences d'itemsets

Origine du problème : les motifs séquentiels vus comme des séquences d'événements

- Dans un système d'alarme automatique :
- $Alarm_{\mathbf{A}} = (cap_1 = 0, cap_2 = 1, cap_3 = 0)$
- $Alarme_{\mathbf{B}} = (cap_1 = 0, cap_2 = 1, cap_3 = 1)$

Selon contexte $Alarme_{\mathbf{B}}$ représente une situation complétement différente de l'$Alarme_{\mathbf{A}}$ (Or seul Cap_3 diffère).

Edit Distance et Séquences d'itemsets

Origine du problème : les motifs séquentiels vus comme des séquences d'événements

- Dans un système d'alarme automatique :
- $Alarm_{\mathbf{A}} = (cap_1 = 0, cap_2 = 1, cap_3 = 0)$
- $Alarme_{\mathbf{B}} = (cap_1 = 0, cap_2 = 1, cap_3 = 1)$

Selon contexte $Alarme_{\mathbf{B}}$ représente une situation complétement différente de l'$Alarme_{\mathbf{A}}$ (Or seul Cap_3 diffère).

- $M = \{(chips, soda, pains)(pizza)(chips, soda, chocolat)(farine)\}$
- Dans $(chips, soda, pains)$ et $(chips, soda, chocolat)$: Un seul item diffère

$(chips, soda, pains)$ représente-t'il un comportement complétement différent de $(chips, soda, chocolat)$?

Edit Distance et Séquences d'itemsets

Alors...

- Interprétation de l'itemset comme un événement n'est toujours pas pertinente
- Prend en compte le contexte et le sens des itemsets dans le contexte

Mesure de similarité pour les motifs séquentiels
Mesures existant / LCS

LCS :

La longueur de la sous-séquence la plus longue commune entre deux séquences

$S_1 = \{A, \mathbf{C}, F, \mathbf{G}\}$ et $S_2 = \{\mathbf{C}, \mathbf{G}, F, A\}$

La sous-séquence la plus longue : $\{\mathbf{C}, \mathbf{G}\}$
$\mathrm{LCS}(S_1, S_2) = 2$

Mesure de similarité pour les motifs séquentiels

LCS non pertinente pour comparer les motifs séquentiels ?

$$M_1 = \{(ab)(c)(xy)(d)\}$$
$$N_1 = \{(bce)(ad)(abf)\}, N_2 = \{(ab)(rt)(c)(mn)(lk)(ad)\}$$

Mesure de similarité pour les motifs séquentiels

LCS non pertinente pour comparer les motifs séquentiels ?

$M_1 = \{(ab)(c)(xy)(d)\}$
$N_1 = \{(bce)(ad)(abf)\}, N_2 = \{(ab)(rt)(c)(mn)(lk)(ad)\}$
La sous-séquence commune : $\{(c)(d)\}$

Mesure de similarité pour les motifs séquentiels

LCS non pertinente pour comparer les motifs séquentiels ?

$M_1 = \{(ab)(c)(xy)(d)\}$
$N_1 = \{(bce)(ad)(abf)\}, N_2 = \{(ab)(rt)(c)(mn)(lk)(ad)\}$
La sous-séquence commune : $\{(c)(d)\}$

$$LCS(M_1, N_1) = 2 \qquad LCS(M_1, N_2) = 2$$

Mesure de similarité pour les motifs séquentiels

LCS non pertinente pour comparer les motifs séquentiels ?

$M_1 = \{(ab)(c)(xy)(d)\}$
$N_1 = \{(bce)(ad)(abf)\}, N_2 = \{(ab)(rt)(c)(mn)(lk)(ad)\}$

La sous-séquence commune : $\{(c)(d)\}$

$$LCS(M_1, N_1) = 2 \qquad LCS(M_1, N_2) = 2$$

Mesure de similarité pour les motifs séquentiels

LCS non pertinente pour comparer les motifs séquentiels ?

$$M_1 = \{(ab)(c)(xy)(d)\}$$
$$N_1 = \{(bce)(ad)(abf)\}, N_3 = \{(ab)(rt)(yu)(ze)(c)(xy)(d)\}$$

Mesure de similarité pour les motifs séquentiels

LCS non pertinente pour comparer les motifs séquentiels ?

$M_1 = \{(ab)(c)(xy)(d)\}$
$N_1 = \{(bce)(ad)(abf)\}, N_3 = \{(ab)(rt)(yu)(ze)(c)(xy)(d)\}$
La sous-séquence commune : $\{(c)(d)\}$

Mesure de similarité pour les motifs séquentiels

LCS non pertinente pour comparer les motifs séquentiels ?

$M_1 = \{(ab)(c)(xy)(d)\}$
$N_1 = \{(bce)(ad)(abf)\}, N_3 = \{(ab)(rt)(yu)(ze)(c)(xy)(d)\}$

La sous-séquence commune : $\{(c)(d)\}$

$$LCS(M_1, N_1) = 2 \qquad LCS(M_1, N_3) = 2$$

Mesure de similarité pour les motifs séquentiels

LCS non pertinente pour comparer les motifs séquentiels ?

$M_1 = \{(ab)(c)(xy)(d)\}$
$N_1 = \{(bce)(ad)(abf)\}, N_3 = \{(ab)(rt)(yu)(ze)(c)(xy)(d)\}$

La sous-séquence commune : $\{(c)(d)\}$

$$LCS(M_1, N_1) = 2 \qquad LCS(M_1, N_3) = 2$$

Mesure de similarité pour les motifs séquentiels

LCS non pertinente pour comparer les motifs séquentiels ?

$$M_1 = \{(ab)(c)(xy)(d)\}$$
$$N_1 = \{(bce)(ad)(abf)\}, N_4 = \{(c)(be)(ad)(abf)\}$$

Mesure de similarité pour les motifs séquentiels

LCS non pertinente pour comparer les motifs séquentiels ?

$M_1 = \{(ab)(c)(xy)(d)\}$
$N_1 = \{(bce)(ad)(abf)\}, N_4 = \{(c)(be)(ad)(abf)\}$
La sous-séquence commune : $\{(c)(d)\}$

Mesure de similarité pour les motifs séquentiels

LCS non pertinente pour comparer les motifs séquentiels ?

$M_1 = \{(ab)(c)(xy)(d)\}$
$N_1 = \{(bce)(ad)(abf)\}, N_4 = \{(c)(be)(ad)(abf)\}$

La sous-séquence commune : $\{(c)(d)\}$

$$LCS(M_1, N_1) = 2 \qquad LCS(M_1, N_4) = 2$$

Mesure de similarité pour les motifs séquentiels

LCS non pertinente pour comparer les motifs séquentiels ?

$M_1 = \{(ab)(c)(xy)(d)\}$
$N_1 = \{(bce)(ad)(abf)\}, N_4 = \{(c)(be)(ad)(abf)\}$
La sous-séquence commune : $\{(c)(d)\}$

$$LCS(M_1, N_1) = 2 \qquad LCS(M_1, N_4) = 2$$

Une mesure de similarité pour les motifs séquentiel devrait ...

- Traiter les motifs séquentiels comme des **séquences ordonnées d'itemsets** et **non d'items**
- **Comparer** les séquences **au niveau des itemsets** et également **au niveau des items dans des itemsets**
- Prendre en compte les **positions** (distance dans l'ordre) **des itemsets** lors du calcul de la similarité
- Prendre en compte **le nombre d'items non communs** au niveau de **la séquence** et aussi au niveau des **itemsets correspondant**

Proposition d'une mesure de similarité pour les motifs séquentiels
Description générale

La mesure de similarité proposée est composée de deux scores :

1. Le **score de mapping** mesurant la ressemblance des deux motifs en fonction des liens qu'il est possible d'établir entre les itemsets
2. Le **score d'ordre** qui mesure la ressemblance des deux séquences vis-vis de l'ordre des itemsets

Proposition d'une mesure de similarité pour les motifs séquentiels

Phase 1 - Mapping entre les itemsets :

- Mapper chaque itemset $Seq_1(i)$ de la séquence 1 avec l'itemset le plus ressemblant $Seq_2(j)$ dans la séquence 2
- La comparaison de similarité entre deux itemsets :

$$Poids(i)(j) = \frac{\mid Seq_1(i) \cap Seq_2(j) \mid}{(\mid Seq_1(i) \mid + \mid Seq_2(j) \mid)/2}$$

Score de mapping = la moyenne des poids du Mapping des itemsets

Proposition d'une mesure de similarité pour les motifs séquentiels

Exemple de mapping

Phase 1 - Mapping des itemsets :

<(abc)(ab)(cd)>

<(ab)(ca)(a)>

Proposition d'une mesure de similarité pour les motifs séquentiels

Exemple de mapping

Phase 1 - Mapping des itemsets :

<(abc)(ab)(cd)>

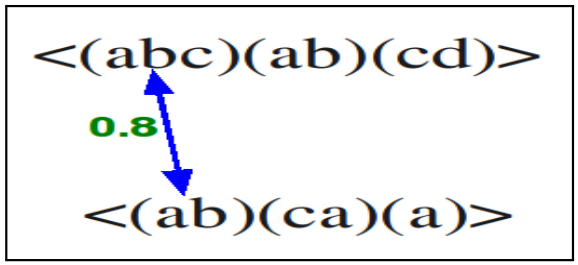
<(ab)(ca)(a)>

- Itemset $Seq_1(1)$:
  - $Poids\ ((abc), (ab)) = 0.8$
  - $Poids\ ((abc), (ca)) = 0.8$
  - $Poids\ ((abc), (a)) = 0.5$

Proposition d'une mesure de similarité pour les motifs séquentiels

Exemple de mapping

Phase 1 - Mapping des itemsets :



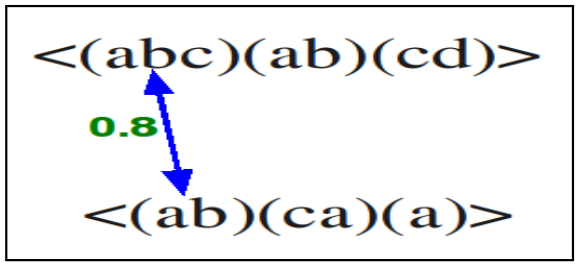
- Itemset $Seq_1(1)$:
  - $Poids\ ((abc), (ab)) = 0.8$
  - $Poids\ ((abc), (ca)) = 0.8$
  - $Poids\ ((abc), (a)) = 0.5$

  $MappedItemSets.put\ ((abc), (ab))$

Proposition d'une mesure de similarité pour les motifs séquentiels

Exemple de mapping

Phase 1 - Mapping des itemsets :



- Itemset $Seq_1(1)$:
  – $Poids\ ((abc), (ab)) = 0.8$
  – $Poids\ ((abc), (ca)) = 0.8$
  – $Poids\ ((abc), (a)) = 0.5$

  $MappedItemSets.put\ ((abc), (ab))$

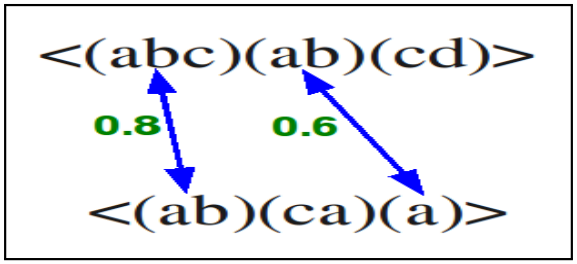
- Itemset $Seq_1(2)$:
  – $Poids((ab), (ab)) = 1$
  – $Poids((ab), (ca)) = 0.5$
  – $Poids((ab), (a)) = 0.6$

  Résoudre le conflit entre (ab) et (abc) pour mapCandidat=(ab) :

Proposition d'une mesure de similarité pour les motifs séquentiels

Exemple de mapping

Phase 1 - Mapping des itemsets :



- Itemset $Seq_1(1)$:
  – $Poids\ ((abc), (ab)) = 0.8$
  – $Poids\ ((abc), (ca)) = 0.8$
  – $Poids\ ((abc), (a)) = 0.5$

  $MappedItemSets.put\ ((abc), (ab))$

- Itemset $Seq_1(2)$:
  – $Poids((ab), (ab)) = 1$
  – $Poids((ab), (ca)) = 0.5$
  – $Poids((ab), (a)) = 0.6$

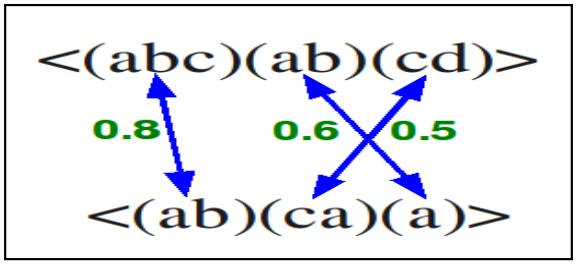
  Résoudre le conflit entre (ab) et (abc) pour mapCandidat=(ab) :

  $MappedItemSets.put((ab), (a))$

Proposition d'une mesure de similarité pour les motifs séquentiels

Exemple de mapping

Phase 1 - Mapping des itemsets :



- Itemset $Seq_1(1)$:
  – $Poids\ ((abc),(ab)) = 0.8$
  – $Poids\ ((abc),(ca)) = 0.8$
  – $Poids\ ((abc),(a)) = 0.5$

  $MappedItemSets.put\ ((abc),(ab))$

- Itemset $Seq_1(2)$:
  – $Poids((ab),(ab)) = 1$
  – $Poids((ab),(ca)) = 0.5$
  – $Poids((ab),(a)) = 0.6$

  Résoudre le conflit entre (ab) et (abc) pour mapCandidat=(ab) :

  $MappedItemSets.put((ab),(a))$

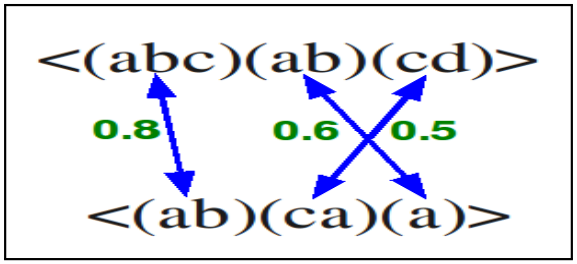
- Itemset $Seq_1(3)$:

  $MappedItemSets.put((cd),(ca))$

Proposition d'une mesure de similarité pour les motifs séquentiels

Exemple de mapping

Phase 1 - Mapping des itemsets :



- Itemset $Seq_1(1)$:
  – $Poids\ ((abc),(ab)) = 0.8$
  – $Poids\ ((abc),(ca)) = 0.8$
  – $Poids\ ((abc),(a)) = 0.5$

  $MappedItemSets.put\ ((abc),(ab))$

- Itemset $Seq_1(2)$:
  – $Poids((ab),(ab)) = 1$
  – $Poids((ab),(ca)) = 0.5$
  – $Poids((ab),(a)) = 0.6$

  Résoudre le conflit entre (ab) et (abc) pour mapCandidat=(ab) :

  $MappedItemSets.put((ab),(a))$

- Itemset $Seq_1(3)$:

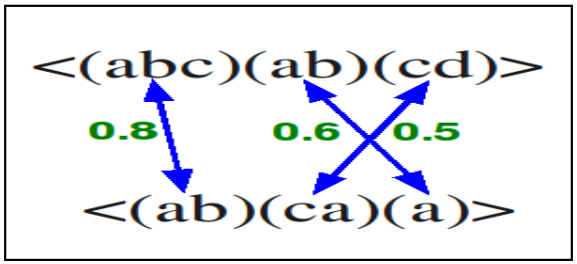
  $MappedItemSets.put((cd),(ca))$

- $AveWeightScore = \frac{0.8+0.6+0.5}{3} = 0.63$

Proposition d'une mesure de similarité pour les motifs séquentiels

Phase 2 - Calcul du score de l'ordre

But : Trouver les **mappings non croisés** et déterminer en même temps **la distance (dans l'ordre)** entre les couples mappés



Comment :

Chercher **des itemsets mappés de séquence 2** placés dans **la même ordre** que les **itemsets de la séquence 1** :

- Réordonner les itemsets de la séquence 2 dans l'ordre du mapping avec les itemsets de la séquence 1 (***Création du mapOrder***)

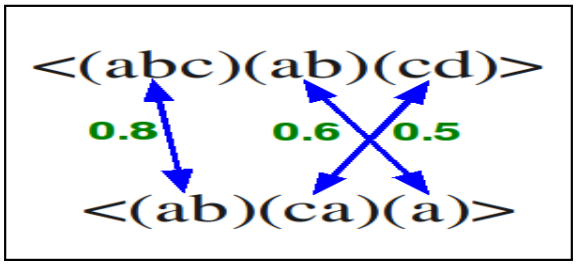
$$mapOrder = \{t_1, t_2, \ldots, t_i, \ldots t_n\}$$

t_i = le **timeStamp** de l'itemset de la Seq_2 associés à **i'eme** itemset de la Seq_1.

Proposition d'une mesure de similarité pour les motifs séquentiels

Phase 2 - Calcul du score de l'ordre

But : Trouver les **mappings non croisés** et déterminer en même temps **la distance (dans l'ordre)** entre les couples mappés



Comment :

Chercher **des itemsets mappés de séquence 2** placés dans **la même ordre** que les **itemsets de la séquence 1** :

- Réordonner les itemsets de la séquence 2 dans l'ordre du mapping avec les itemsets de la séquence 1 (***Création du mapOrder***)

$$mapOrder = \{t_1, t_2, \ldots, t_i, \ldots t_n\}$$

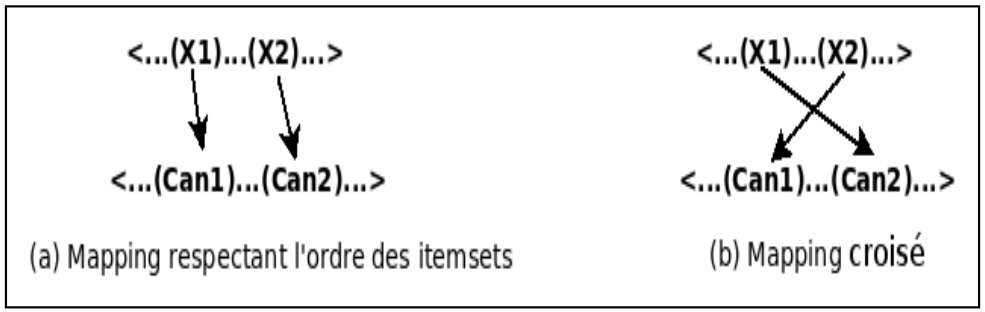
t_i = le **timeStamp** de l'itemset de la Seq_2 associés à **i'eme** itemset de la Seq_1.

$$mapOrder = \{1, 3, 2\}$$

Proposition d'une mesure de similarité pour les motifs séquentiels

Phase 2 - Calcul du score de l'ordre

- Calculer pour toutes les sous-séquences croissantes et maximales du *mapOrder* :

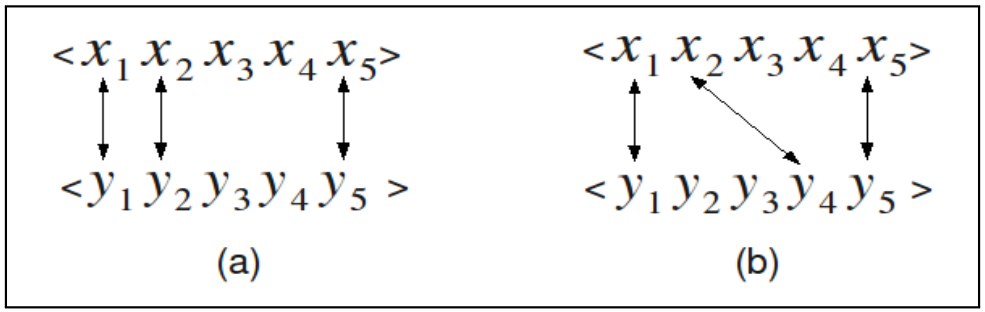


$$totalOrder = \frac{nbOrderedItemSets}{aveNbItemSets}$$

Proposition d'une mesure de similarité pour les motifs séquentiels

Phase 2 - Calcul du score de l'ordre

- Calculer pour toutes les sous-séquences croissantes et maximales du *mapOrder* :



$$positionOrder = \sum_{i=1}^{|sub|} \frac{|sub(i)-sub(i-1)|-|mapOrder^{-1}(sub(i))-mapOrder^{-1}(sub(i-1))|}{aveNbItemSets}$$

Proposition d'une mesure de similarité pour les motifs séquentiels

Phase 2 - Calcul du score de l'ordre

- Calculer pour toutes les sous-séquences croissantes et maximales du *mapOrder* :

$$orderScore = \max\{totalOrder(sub) \times (1 - positionOrder(sub))\}$$
$$sub \in \{sous\_seqs\ croissantes\ et\ maximales\ du\ mapOrder\}$$

Phase 3 - Calcul de la similarité :

$$\textbf{SimDegree} = \frac{(orderScore \times Co_1) + (AveWeightScore \times Co_2)}{Co_1 + Co_2}$$

Mesure de similarité proposée pour les motifs séquentiels

Exemple de calcul de la mesure de similarité

Phase 2 - Calcul de l'ordre :

<(abc)(ab)(cd)>

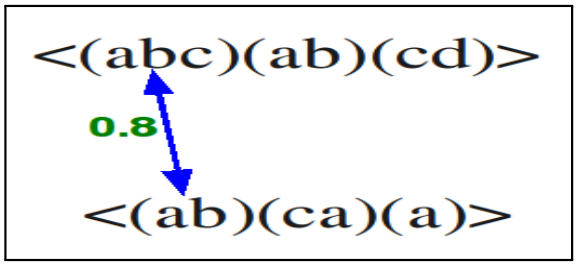
<(ab)(ca)(a)>

- $mapOrder =$

Mesure de similarité proposée pour les motifs séquentiels

Exemple de calcul de la mesure de similarité

Phase 2 - Calcul de l'ordre :

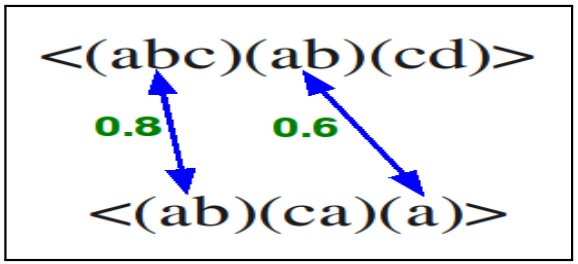


- $mapOrder = \{1,$

Mesure de similarité proposée pour les motifs séquentiels

Exemple de calcul de la mesure de similarité

Phase 2 - Calcul de l'ordre :

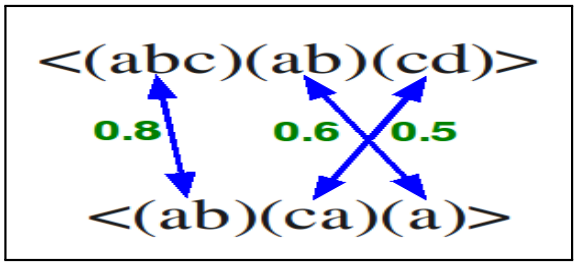


- $mapOrder = \{1, 3,$

Mesure de similarité proposée pour les motifs séquentiels

Exemple de calcul de la mesure de similarité

Phase 2 - Calcul de l'ordre :

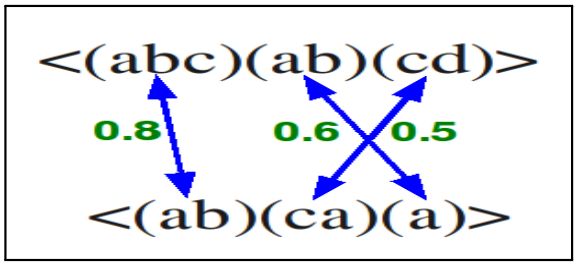


- $mapOrder = \{1, 3, 2\}$

Mesure de similarité proposée pour les motifs séquentiels

Exemple de calcul de la mesure de similarité

Phase 2 - Calcul de l'ordre :

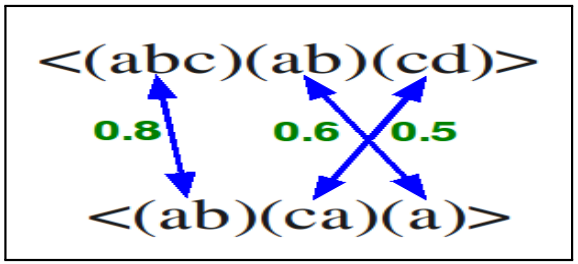


- $mapOrder = \{1, 3, 2\}$
- **Les sous-séq croissantes maximales :**
  $\{1, 2\}, \{1, 3\}$

Mesure de similarité proposée pour les motifs séquentiels

Exemple de calcul de la mesure de similarité

Phase 2 - Calcul de l'ordre :



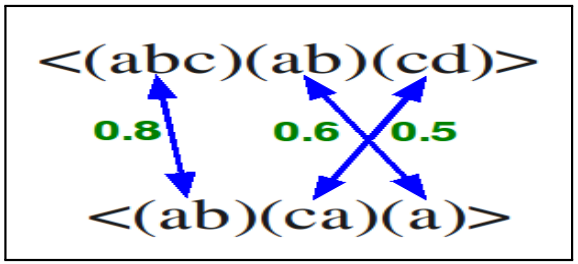
- $mapOrder = \{1, 3, 2\}$
- **Les sous-séq croissantes maximales :**
  $\{1, 2\}, \{1, 3\}$

$$orderScore = \max\{totalOrder(sub) \times (1 - positionOrder(sub))\}$$

Mesure de similarité proposée pour les motifs séquentiels

Exemple de calcul de la mesure de similarité

Phase 2 - Calcul de l'ordre :



- $mapOrder = \{1, 3, 2\}$
- **Les sous-séq croissantes maximales :** $\{1, 2\}, \{1, 3\}$

$$orderScore = \max\{totalOrder(sub) \times (1 - positionOrder(sub))\}$$

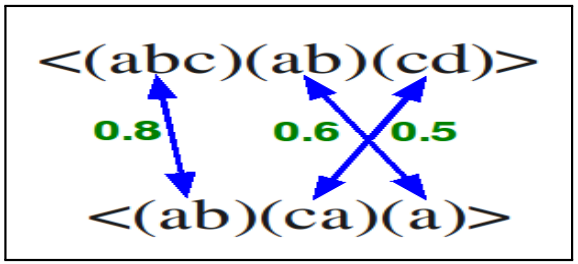
Le score de l'ordre pour chacun est :

- $orderScore(\{1, 2\}) = \frac{2}{3} \times (1 - \frac{|1-2|}{3}) = 0.44$
- $orderScore(\{1, 3\}) = \frac{2}{3} \times (1 - \frac{|2-1|}{3}) = 0.44$

Mesure de similarité proposée pour les motifs séquentiels

Exemple de calcul de la mesure de similarité

Phase 2 - Calcul de l'ordre :



- $mapOrder = \{1, 3, 2\}$
- **Les sous-séq croissantes maximales :** $\{1, 2\}, \{1, 3\}$

$$orderScore = \max\{totalOrder(sub) \times (1 - positionOrder(sub))\}$$

Le score de l'ordre pour chacun est :

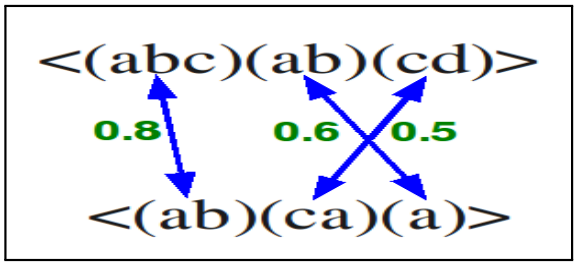
- $orderScore(\{1, 2\}) = \frac{2}{3} \times (1 - \frac{|1-2|}{3}) = 0.44$
- $orderScore(\{1, 3\}) = \frac{2}{3} \times (1 - \frac{|2-1|}{3}) = 0.44$

Phase 3 - Calcul de la similarité :

Mesure de similarité proposée pour les motifs séquentiels

Exemple de calcul de la mesure de similarité

Phase 2 - Calcul de l'ordre :



- $mapOrder = \{1, 3, 2\}$
- **Les sous-séq croissantes maximales :** $\{1, 2\}, \{1, 3\}$

$$orderScore = \max\{totalOrder(sub) \times (1 - positionOrder(sub))\}$$

Le score de l'ordre pour chacun est :

- $orderScore(\{1, 2\}) = \frac{2}{3} \times (1 - \frac{|1-2|}{3}) = 0.44$
- $orderScore(\{1, 3\}) = \frac{2}{3} \times (1 - \frac{|2-1|}{3}) = 0.44$

Phase 3 - Calcul de la similarité :

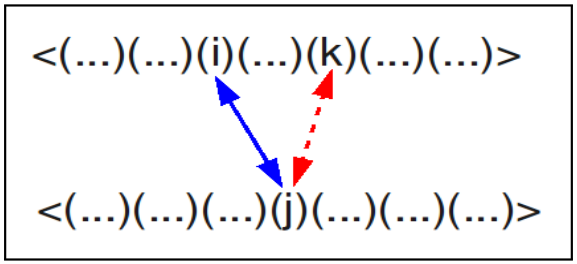
En multipliant par *AveWeightScore* :

$$SimDegree = \frac{0.44 + 0.63}{2} = 53\%$$

Mesure de similarité proposée pour les motifs séquentiels

Conflits lors du mapping 1/2

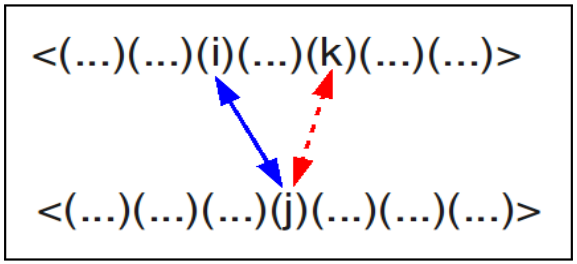
La fonction *SolveConflict* **:**



Mesure de similarité proposée pour les motifs séquentiels

Conflits lors du mapping 1/2

La fonction *SolveConflict* :

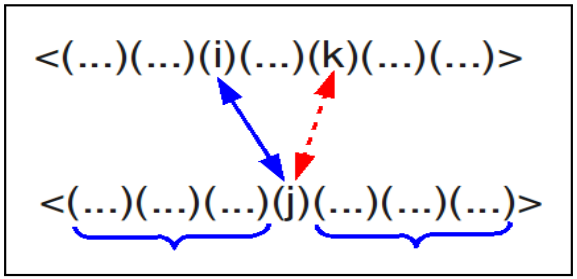


Chercher deux autres candidats pour les i et k :

Mesure de similarité proposée pour les motifs séquentiels

Conflits lors du mapping 1/2

La fonction *SolveConflict* :



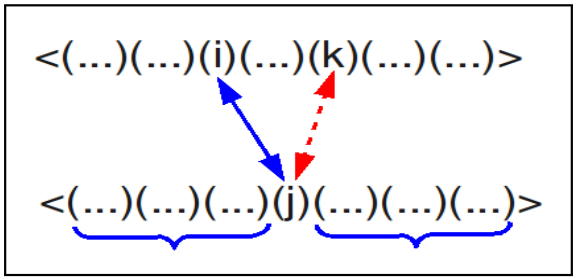
Chercher deux autres candidats pour les *i* et *k* :

- 1er candidat = l'itemset **se situant avant j** ayant également **le poids maximum** : nextMaxBefore
- 2ème candidat = sélectionné de la même manière mais étant recherché **après j** : nextMaxAfter

Mesure de similarité proposée pour les motifs séquentiels

Conflits lors du mapping 1/2

La fonction *SolveConflict* :



Chercher deux autres candidats pour les i et k :

- 1er candidat = l'itemset **se situant avant j** ayant également **le poids maximum** : nextMaxBefore
- 2ème candidat = sélectionné de la même manière mais étant recherché **après j** : nextMaxAfter

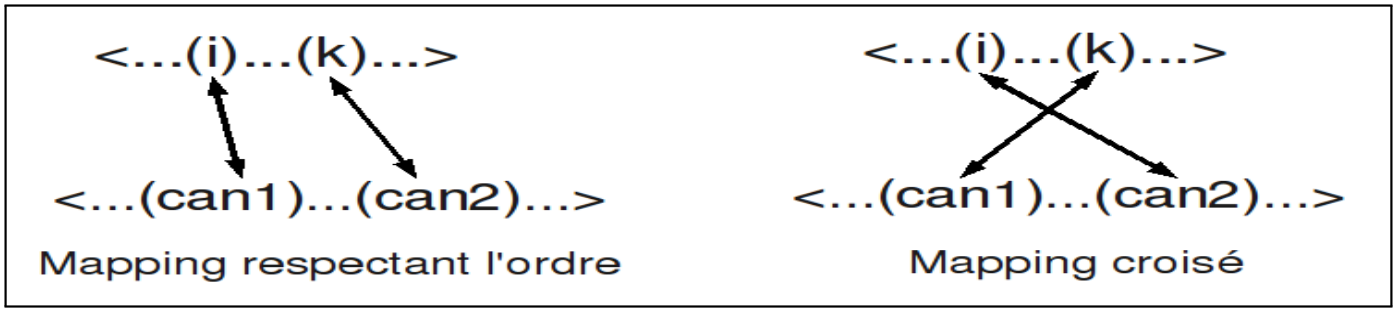
Tous les couples possibles de mappings :

- $(i,j),(k,nextMaxBefore_k)$
- $(i,j),(k,nextMaxAfter_k)$
- $(k,j),(i,nextMaxBefore_i)$
- $(k,j),(i,nextMaxAftre_i)$

• Calculer la pertinence de chaque cas pour les proposer comme nouveaux candidats (*calcul localSim*).

Mesure de similarité proposé pour les motifs séquentiels

Conflits lors du mapping 2/2



- **Mapping respectant l'ordre :**

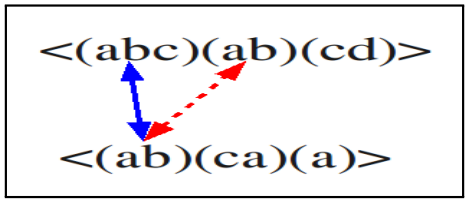
  $\textbf{localSim}(i, Can_1)(k, Can_2) = \frac{Poids(i,Can_1)+Poids(k,Can_2)}{2}$

- **Mapping croisé** (*l'ordre étant à moitié respecté*) **:**

  $\textbf{localSim}(k, Can_1)(i, Can_2) = \frac{1}{2} \times \frac{Poids(k,Can_1)+Poids(i,Can_2)}{2}$

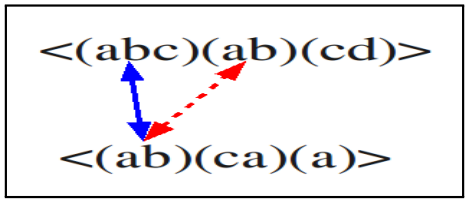
Mesure de similarité proposé pour les motifs séquentiels

Exemple de résolution du conflit



Mesure de similarité proposé pour les motifs séquentiels

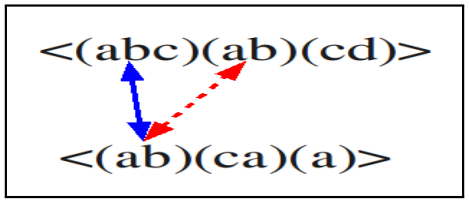
Exemple de résolution du conflit



- **Pour l'itemSet (abc) :**
  $nextMaxBefore_1 = \emptyset$
  $nextMaxAfter_1 = (ca)$

Mesure de similarité proposé pour les motifs séquentiels

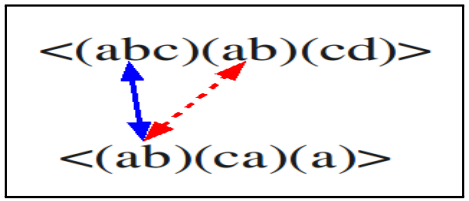
Exemple de résolution du conflit



- **Pour l'itemSet (abc) :**
  $nextMaxBefore_1 = \emptyset$
  $nextMaxAfter_1 = (ca)$
- **Pour l'itemSet (ab) :**
  $nextMaxBefore_2 = \emptyset$
  $nextMaxAfter_2 = (a)$

Mesure de similarité proposé pour les motifs séquentiels

Exemple de résolution du conflit

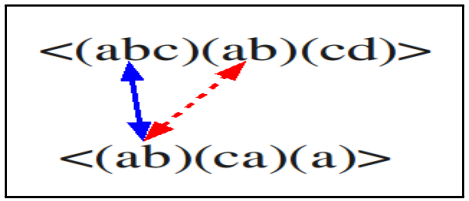


- **Pour l'itemSet (abc) :**
  $nextMaxBefore_1 = \emptyset$
  $nextMaxAfter_1 = (ca)$
- **Pour l'itemSet (ab) :**
  $nextMaxBefore_2 = \emptyset$
  $nextMaxAfter_2 = (a)$

- **Les couples Mappings possibles :**
  $(\langle (abc), (ca) \rangle , \langle (ab), (ab) \rangle)$
  $(\langle (abc), (ab) \rangle , \langle (ab), (a) \rangle)$
- $localSim(\langle (abc), (ca) \rangle , \langle (ab), (ab) \rangle) = 0.45$
- $localSim(\langle (abc), (ab) \rangle , \langle (ab), (a) \rangle) = 0.7$

Mesure de similarité proposé pour les motifs séquentiels

Exemple de résolution du conflit



- **Pour l'itemSet (abc) :**
  $nextMaxBefore_1 = \emptyset$
  $nextMaxAfter_1 = (ca)$
- **Pour l'itemSet (ab) :**
  $nextMaxBefore_2 = \emptyset$
  $nextMaxAfter_2 = (a)$

- **Les couples Mappings possibles :**
  $(\langle(abc),(ca)\rangle, \langle(ab),(ab)\rangle)$
  $(\langle(abc),(ab)\rangle, \langle(ab),(a)\rangle)$
- $localSim(\langle(abc),(ca)\rangle, \langle(ab),(ab)\rangle) = 0.45$
- $localSim(\langle(abc),(ab)\rangle, \langle(ab),(a)\rangle) = 0.7$

$\Longrightarrow$ Le couple choisi : $(\langle(abc),(ab)\rangle, \langle(ab),(a)\rangle)$

Discussions

Caractéristiques de notre mesure de similarité

- **Définir une mesure de similarité** qui prend en compte les **caractéristiques** des motifs séquentiels
- Combinaison de deux scores : (1) **score de mapping des itemsets** (2) **score d'ordre** des itemsets dans les deux séquences
- Comparaison à la fois **au niveau des itemsets** et de **leurs positions** dans la séquence et aussi **au niveau des items** dans les **itemsets ressemblants**
- **Surmonter** les problèmes liés aux mesures **LCS** et **Edit distance** par ces deux scores
- Une mesure de similarité a priori **Asymétrique**
- **Possibilité** de le rendre **symétrique**

Discussions

Domaine d'application de notre mesure de similarité

- Comme une mesure **Asymétrique** : Comparaison **directionnelle**
  - L'extraction de motifs séquentiels sous contrainte de similarité
  - La visualisation des motifs proches d'un motif sélectionné
  - Lorsqu'il y a des comparaisons de motifs séquentiels avec un motif de référence
- Comme une mesure **Symétrique** :
  - Le clustering de motifs séquentiels
  - La compression des motifs séquentiels

Plan

Expérimentations de mesure de similarité

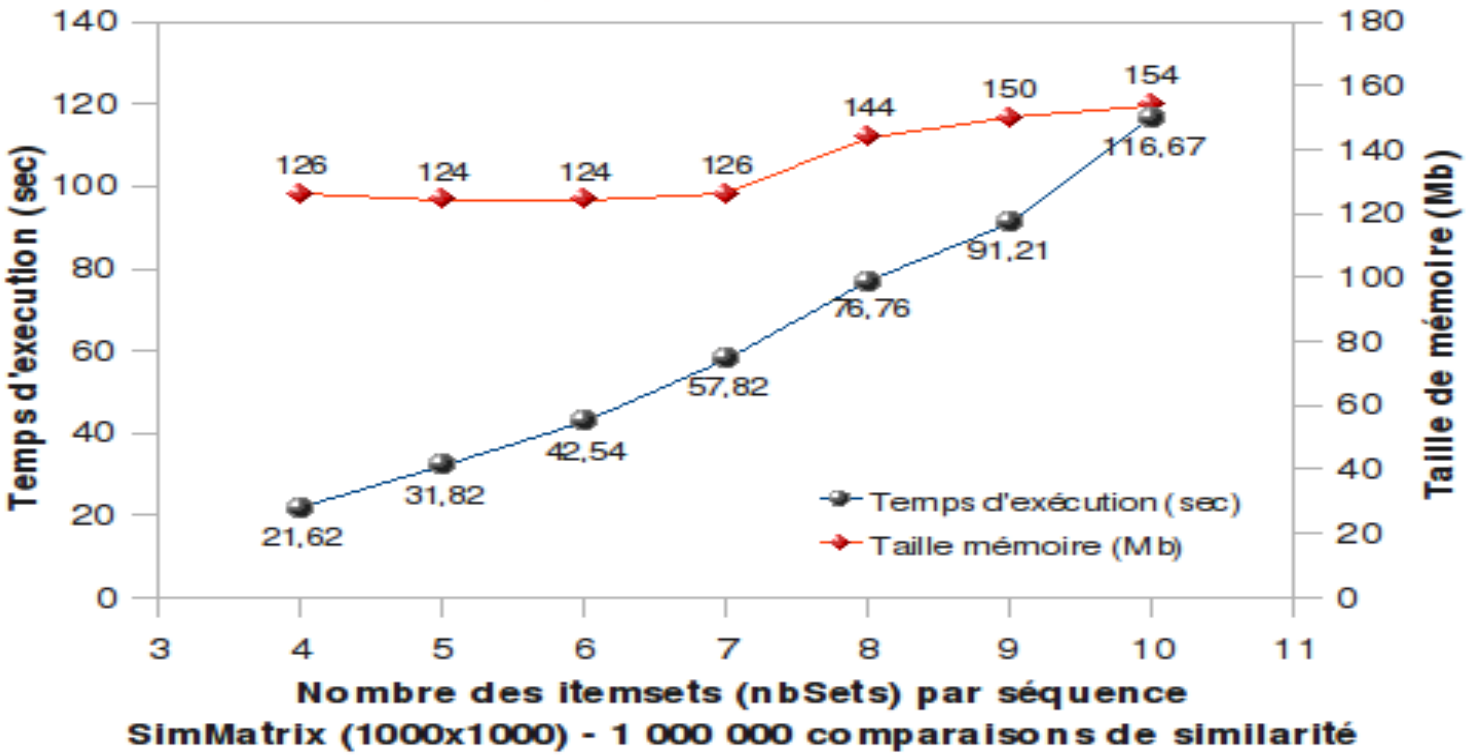


FIG.: Temps de calcul et taille mémoire en fonction du nombre d'itemsets

Expérimentations de mesure de similarité

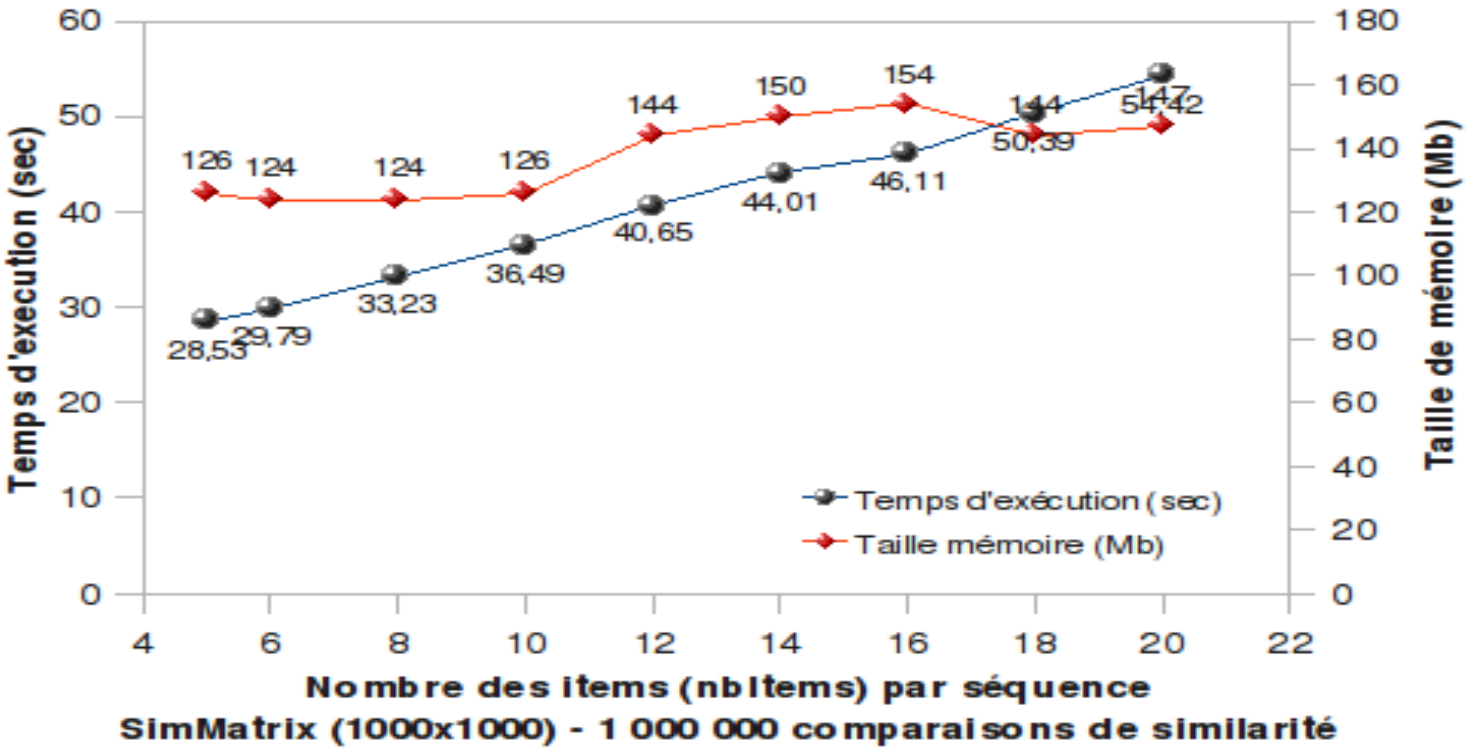


FIG.: Temps de calcul et taille mémoire en fonction du nombre d'items par séquence

Expérimentations de mesure de similarité

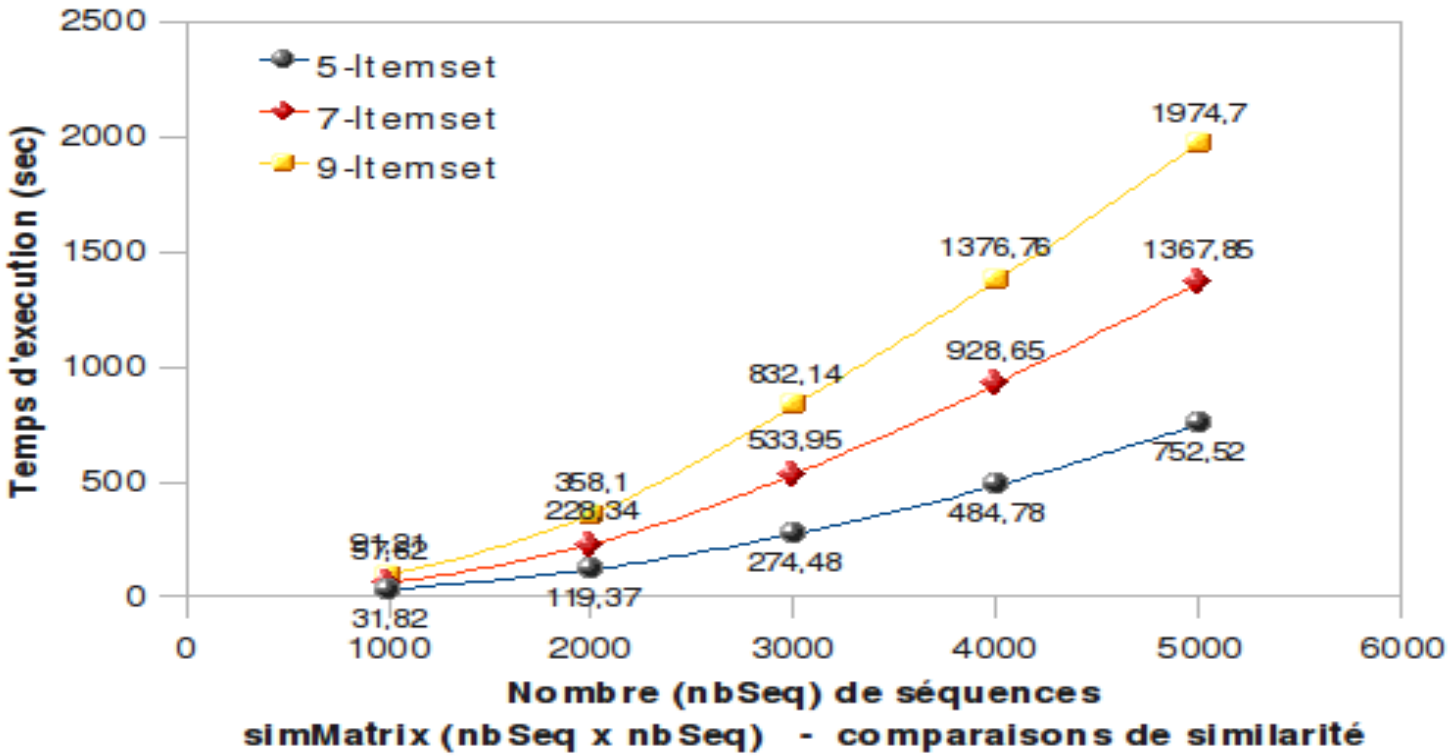


FIG.: Temps de calcul de la matrice de similarité en fonction du nombre de séquences

Expérimentations de mesure de similarité

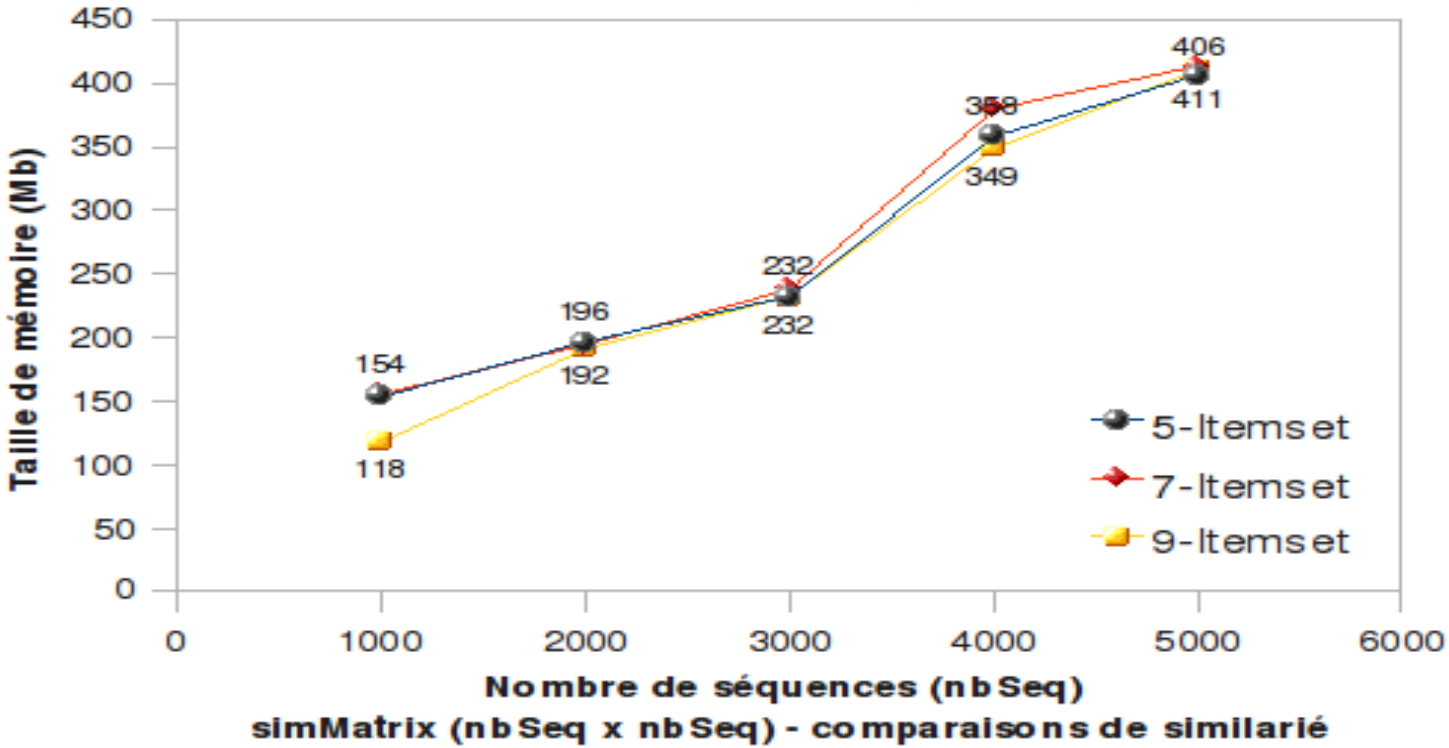


FIG.: Taille mémoire utilisée lors du calcul de la matrice de similarité en fonction du nombre de séquences

Expérimentations de mesure de similarité

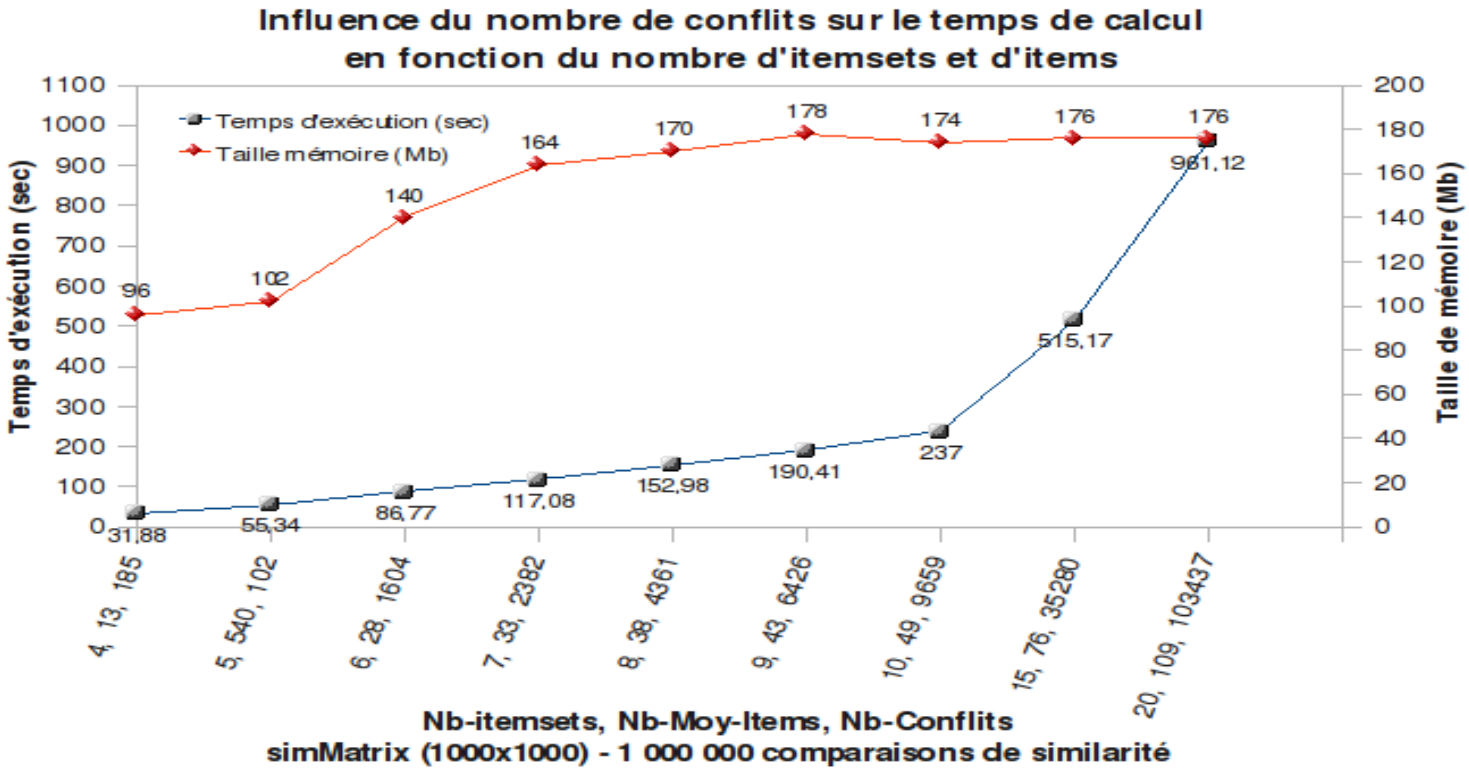


FIG.: Influence des conflits sur le temps de calcul de la mesure de similarité

Plan

Conclusions

- Définition d'une approche de détection d'anomalies fondée sur le clustering de motifs séquentiels fréquents
  - Apprentissage :
    - Utilisation de **motifs séquentiels** pour **modéliser les comportements** généraux
    - Réalisation un **clustering de motifs séquentiels** pour regrouper les comportements similaires
  - Évaluation
    - Détecter les intrusions en **identifiant des déviation** par rapport aux comportements généraux
    - En comparant les nouveaux motifs séquentiels extraits dans la phase d'évaluation avec les représentant des profils de comportements

Conclusions

- Définition d'une mesure de similarité adaptée aux motifs séquentiels pour le clustering et la comparaison des motifs séquentiels
  - Elle se calcule très rapidement même lorsqu'il y a beaucoup de séquences ayant plusieurs itemsets
  - Utilisable pour **d'autres applications que le clustering** de motifs séquentiels
    - L'extraction des motifs séquentiels sous contrainte de similarité
    - La compression des motifs séquentiels
    - La visualisation des motifs similaires etc

Perspectives

À court terme :

- Expérimenter notre approche de la détection d'anomalies sur des **données réelles**
- Implémenter des méthodes d'extraction d'attributs à partir des logs

- Considérer une **version symétrique** de notre mesure
- Comparer les résultats de la mesure symétrique avec celui de la mesure asymétrique
- Comparer avec **Edit distance**

Perspectives

À plus long terme :

- Évaluer les **différentes manières d'extraire des motifs séquentiels** en fonction du format de logs
- Adapter d'autres méthodes de clustering notamment le **clustering dynamique**
- Implémenter des méthodes de **post-traitement pour améliorer** les résultats de clustering

- Trouver d'autres applications dans lesquelles notre mesure de similarité peut être utilisée
  - Le domaine de **fouille de données biologiques**
  - Les motifs séquentiels extraits volumineux
  - Fournir une visualisation des motifs séquentiels pour pouvoir observer les motifs séquentiels souhaités
  - Choisir un motif séquentiel comme référence et présenter d'autres motifs séquentiels ressemblants parmi tous les motifs extraits

Fin

Merci de votre attention

Rakesh Agrawal, Tomasz Imielinski, and Arun N. Swami.
Mining association rules between sets of items in large databases.
In Peter Buneman and Sushil Jajodia, editors, *Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data*, pages 207–216, Washington, D.C., 26–28 1993.

Rakesh Agrawal and Ramakrishnan Srikant.
Fast algorithms for mining association rules.
In Jorge B. Bocca, Matthias Jarke, and Carlo Zaniolo, editors, *Proc. 20th Int. Conf. Very Large Data Bases, VLDB*, pages 487–499. Morgan Kaufmann, 12–15 1994.

Rakesh Agrawal and Ramakrishnan Srikant.
Mining sequential patterns.
In Philip S. Yu and Arbee S. P. Chen, editors, *Eleventh International Conference on Data Engineering*, pages 3–14, Taipei, Taiwan, 1995. IEEE Computer Society Press.

Eric Bloedorn, Alan D. Christiansen, Willian Hill, Clement Skorupka, Lisa M. Talbot, and Jonathan Tivel.
Data mining for network intrusion detection : How to get started, August 2001.

A. Chittur.
Model generation for an intrusion detection system using genetic algorithms.
PhD thesis, Ossining High School. In cooperation with Columbia Univ., 2001.

Matthieu Capelle, Cyrille Masson, and Jean-François Boulicaut.
Mining frequent sequential patterns under a similarity constraint.
In *IDEAL*, pages 1–6, 2002.

J. E. Dickerson and J. A. Dickerson.
Fuzzy network profiling for intrusion detection.
In *In Proc. of NAFIPS 19th International Conference of the North American Fuzzy Information Processing Society, Atlanta*, page 301–306. North American Fuzzy Information Processing Society (NAFIPS), July 2000.

Paul Dokas, Levent Ertoz, Vipin Kumar, Aleksandar Lazarevic, Jaideep Srivastava, and Pang-Ning Tan.
Data mining for network intrusion detection.
University of Minnesota, Minneapolis, MN 55455, USA, 2002.

Dorothy E. Denning.
An intrusion-detection model.

J. Hartigans.
clustering algorithms.
John Wiley and Sons, Inc., 1975.

Jiawei Han and Micheline Kamber.
Data Mining : Concepts and Techniques.
Morgan Kaufmann Publishers, 2000.

S. Kumar.
Classification and Detection of Computer Intrusions.
PhD thesis, Purdue Univ., West Lafayette, IN, 1995.

Hye-Chung(Monica) Kum.
Approximate Mining of Consensus Sequential Patterns.
PhD thesis, University of North Carolina, August 2004.

Carbone P. L.
Data mining or knowledge discovery in databases : An overview. In Data Management Handbook.
New York : Auerbach Publications, 1997.

T. Lane.

Wenke Lee, Salvatore J. Stolfo, and K. Mok.
Algorithms for mining system audit data.
In T. Y. Lin and N. Cercone, editors, *Data Retrieval and Data Mining*. Kluwer Academic Publishers.

Wenke Lee, Salvatore J. Stolfo, and Kui W. Mok.
Mining audit data to build intrusion detection models.
In *Knowledge Discovery and Data Mining*, pages 66–72, 1998.

Wenke Lee, Salvatore J. Stolfo, and Kui W. Mok.
A data mining framework for building intrusion detection models.
In *IEEE Symposium on Security and Privacy*, pages 120–132, 1999.

Wenke Lee, Salvatore J. Stolfo, and Kui W. Mok.
Adaptive intrusion detection, a data mining approach.
Artificial Intelligence Review, 14(6) :533–567, 2000.

J. Lue.
Integrating fuzzy logic with data mining methods for intrusion detection.
Master's thesis, Mississippi State Univ, 1999.

Teresa F. Lunt.

Filippo Neri.
Comparing local search with respect to genetic evolution to detect intrusion in computer networks.
In *Proceedings of the 2000 Congress on Evolutionary Computation CEC00*, pages 238–243, La Jolla Marriott Hotel La Jolla, California, USA, 6-9 2000. IEEE Press.

L. Portnoy, E. Eskin, and S. Stolfo.
Intrusion detection with unlabeled data using clustering, 2001.

Marc Plantevit, Anne Laurent, and Maguelonne Teisseire.
Extraction d'outliers dans des cubes de données : une aide à la navigation.
In *3èmes journées francophones sur les Entrepôts de Données et l'Analyse en ligne (EDA 2007), Poitiers*, volume B-3 of *RNTI*. Cépaduès, Address = Toulouse, Pages = 113-130, Juin 2007.

Julien Rabatel, Yuan Lin, Yoann Pitarch, Hassan Saneifar, Claire Serp, Mathieu Roche, and Anne Laurent.
Visualisation des motifs séquentiels extraits à partir dun corpus en ancien français.
In *EGC*, pages 237–238, 2008.

Abhinav Srivastava, Shamik Sural, and Arun K. Majumdar.
Weighted intra-transactional rule mining for database intrusion detection.
In *PAKDD*, pages 611–620, 2006.

Karlton Sequeira and Mohammed Javeed Zaki.
Admit : anomaly-based data mining for intrusions.
In *KDD*, pages 386–395, 2002.

Doru Tanasa.
Web Usage Mining : Contributions to Intersites Logs Preprocessing and Sequential Pattern Extraction with Low Support.
PhD thesis, Université de Nice Sophia Antipolis, Juin 2005.

L.A. Zadeh.
Fuzzy sets. information and control.
pages 8 (3) 338–353, 1965.