M1 MIAGE Option RIM
Cours 6 : Mining de patterns complexes

Alexandre Termier

2010-2011 S2

Limites des itemsets

- Itemsets ne prennent pas en compte les relations temporelles
- Relations temporelles peuvent exprimer :
  - *Causalité*

```
T0    : {marteau, clous}
T0+2h : {pansements, alcool, coton}
```

  - *Rythmes*

```
T0     : {viande, légumes, fruits, pain}
T0+3j  : {pain, fruits}
T0+7j  : {viande, légumes, fruits, pain}
T0+10j : {pain, fruits}
...
```

Données séquentielles : exemple

- Données clients
  - Historique des achats d'un client
  - Transaction = liste des produits achetés au temps t
- Données Web
  - Navigation d'un visiteur de site web identifé par IP/cookie
  - Transaction = liste des fichiers envoyés après un clic
- Données d'événements
  - Evénements générés par un capteur / fichiers de log
  - Transaction = événements du capteur/log au temps t
- ADN
  - Séquence de base pour une espèce particulière
  - Transaction = élément de la séquence

Exemple : historique navigateur

```
16:47:51        weka sequence mining - Recherche Google
16:47:53        Modified WEKA
16:48:12        Old Nabble - WEKA - Event Sequence Mining
16:48:31        Weka 3 - Data Mining with Open Source Machine Learning Software in Java
16:49:03        Weka 3 - Data Mining with Open Source Machine Learning Software in Java
16:49:43        Machine Learning Group Publications
16:50:01        Weka 3 - Data Mining with Open Source Machine Learning Software in Java
16:50:08        weka - Frequently Asked Questions
16:50:34        Data Mining: Practical Machine Learning Tools and Techniques
16:54:21        Overview | MOA Massive Online Analysis
16:54:36        People | MOA Massive Online Analysis
16:54:39        http://www.cs.waikato.ac.nz/~abifet
16:54:39        Albert BIFET
16:55:04        Adaptive Stream Mining
16:55:32        MOA Massive Online Analysis
16:56:20        Machine Learning Project at the University of Waikato in New Zealand
16:56:30        weka - Recherche Google
16:56:40        Weka---Machine Learning Software in Java | Download Weka---Machine Learning Software in Ja
16:56:43        http://www.cs.waikato.ac.nz/~ml
16:56:43        Machine Learning Project at the University of Waikato in New Zealand
16:57:15        http://www.google.com/
16:57:18        knime - Recherche Google
16:57:20        KNIME | Konstanz Information Miner
16:57:36        KNIME | Features
16:59:00        KNIME | Example Workflows
16:59:14        KNIME | KNIME Desktop
16:59:19        KNIME | KNIME
```

Exemple : /var/log/messages

```
Mar  5 16:38:03 Osaka kernel: [   28.723047] vboxdrv: fAsync=0 offMin=0x1b9 offMax=0x14a4
Mar  5 16:38:03 Osaka kernel: [   28.723078] vboxdrv: TSC mode is 'synchronous',
                                             kernel timer mode is 'normal'.
Mar  5 16:38:03 Osaka kernel: [   28.925477] vboxnetflt: no symbol version for SUPDrvLinuxIDC
Mar  5 16:38:03 Osaka kernel: [   28.925479] vboxnetflt: Unknown symbol SUPDrvLinuxIDC
Mar  5 16:38:03 Osaka kernel: [   28.955950] ppdev: user-space parallel port driver
Mar  5 16:38:13 Osaka pulseaudio[2221]: alsa-mixer.c: Your kernel driver is broken:
                                it reports a volume range from 0,00 dB to 0,00 dB which makes no sense.
Mar  5 16:46:03 Osaka kernel: [  508.030060] __ratelimit: 36 callbacks suppressed
Mar  5 16:46:03 Osaka kernel: [  508.030066] operapluginwrap[2715]: segfault at 0 ip (null)
                        sp 00007fff9a370278 error 14 in operapluginwrapper-native[400000+3b000]
Mar  5 16:46:03 Osaka kernel: [  508.076292] operapluginwrap[2747]: segfault at 0 ip (null)
                        sp 00007fffb8a321c8 error 14 in operapluginwrapper-native[400000+3b000]
Mar  5 16:46:03 Osaka kernel: [  508.093336] operapluginwrap[2763]: segfault at 0 ip (null)
                        sp 00007fffc7a24f18 error 14 in operapluginwrapper-native[400000+3b000]
Mar  5 16:46:03 Osaka kernel: [  508.110190] operapluginwrap[2779]: segfault at 0 ip (null)
                        sp 00007fff523bcef8 error 14 in operapluginwrapper-native[400000+3b000]
Mar  5 16:46:03 Osaka kernel: [  508.149040] operapluginwrap[2811]: segfault at 0 ip (null)
                        sp 00007fffcf4994e8 error 14 in operapluginwrapper-native[400000+3b000]
Mar  5 16:46:03 Osaka kernel: [  508.165336] operapluginwrap[2827]: segfault at 0 ip (null)
                        sp 00007ffff6b3ff58 error 14 in operapluginwrapper-native[400000+3b000]
Mar  5 16:46:03 Osaka kernel: [  508.181488] operapluginwrap[2843]: segfault at 0 ip (null)
                        sp 00007fff5eb47698 error 14 in operapluginwrapper-native[400000+3b000]
Mar  5 16:46:03 Osaka kernel: [  508.197743] operapluginwrap[2859]: segfault at 0 ip (null)
                        sp 00007fff3296fe18 error 14 in operapluginwrapper-native[400000+3b000]
```

Definitions

Definition (Séquence)

Liste ordonnée de transactions.

$s = <t_1, t_2, \ldots, t_n>$

Definition (Transaction)

Collection d'événements/items *(+ timestamp)*

t_i = *(timestamp :)* $\{e_1, \ldots, e_k\}$

Example (achats de la semaine)

```
< Mardi   : {pain, chocolat}
  Jeudi   : {bananes, bière, chips}
  Samedi  : {surgelés, viande, yaourts} >
```

Définitions, suite

Definition (Sous-séquence)

Une séquence $<a_1,\ldots,a_n>$ est contenue dans une séquence $<b_1,\ldots,b_m>$ $(m \geq n)$ s'il existe $i_1 < i_2 < \ldots < i_n$ tels que $a_1 \subseteq b_{i_1},\ldots,a_n \subseteq b_{i_n}$

Example (sous-séquences)

Séquence	Sous-séquence	Inclusion ?
<{2,4} {3,5,6} {8}>	<{2} {3,5}>	oui
<{1,2} {3,4}>	<{1} {2}>	non
<{2,4} {2,4} {2,5}>	<{2} {4}>	oui

Definition (Support, fréquence)

Support d'une sous-séquence : nombre de séquences des données là contenant. Une sous-séquence s est fréquente si $support(s) \geq minsup$.

Contraintes temporelles

- Nécessité de rajouter des contraintes sur les sous-séquences découvertes
  - Signification des résultats
  - Temps de calcul
- Contraintes
  - max-gap : temps maximum autorisé entre 2 transactions
  - min-gap : temps minimum autorisé entre 2 transactions
  - max-span : temps maximal de la sous-séquence dans la séquence

Exemple

max-gap = 2
min-gap = 0
max-span = 3
Sous-séquence = <{1} {2} {5}>

- Dans <1:{1,2} 2:{3} 3:{4} 4:{2} 5:{5}>

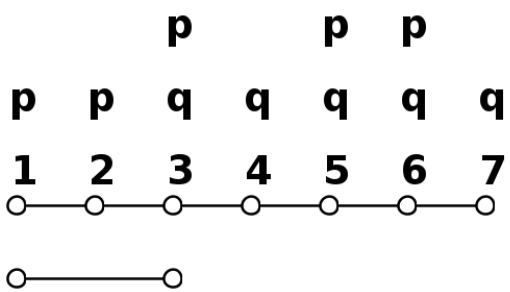
  Non, car ici gap entre {1} et {2} = 4-1 = 3 > max-gap
- Dans <1:{1,2} 2:{3} 3:{2} 4:{6} 5:{5}>

  Ici gaps sont OK.
  Mais span = 5-1 = 4 > max-span

Comptage de sous-séquences

Sous séquence $ss = < \{p\}\{q\} >$, séquence S

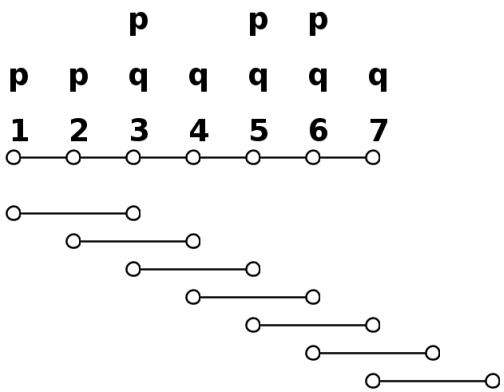
- Simple
  - 1 si ss apparaît dans S, 0 sinon



  - ici : **compte = 1**

Comptage de sous-séquences

- Fenêtre glissantes
  - Comptage des apparitions dans une fenêtre glissant par pas d'une unité, de longueur max-span



  - ici : compte = 4

Ressources

- Algorithmes : PrefixSpan, CloSpan
- Sequential Pattern Mining Framework

  `http://www.philippe-fournier-vigier.com/spmf/index.php`