

Fouille de données, analyse d'évènements

Christophe Cérin

cerin@laria.u-picardie.fr

x placement, ordonnancement ;



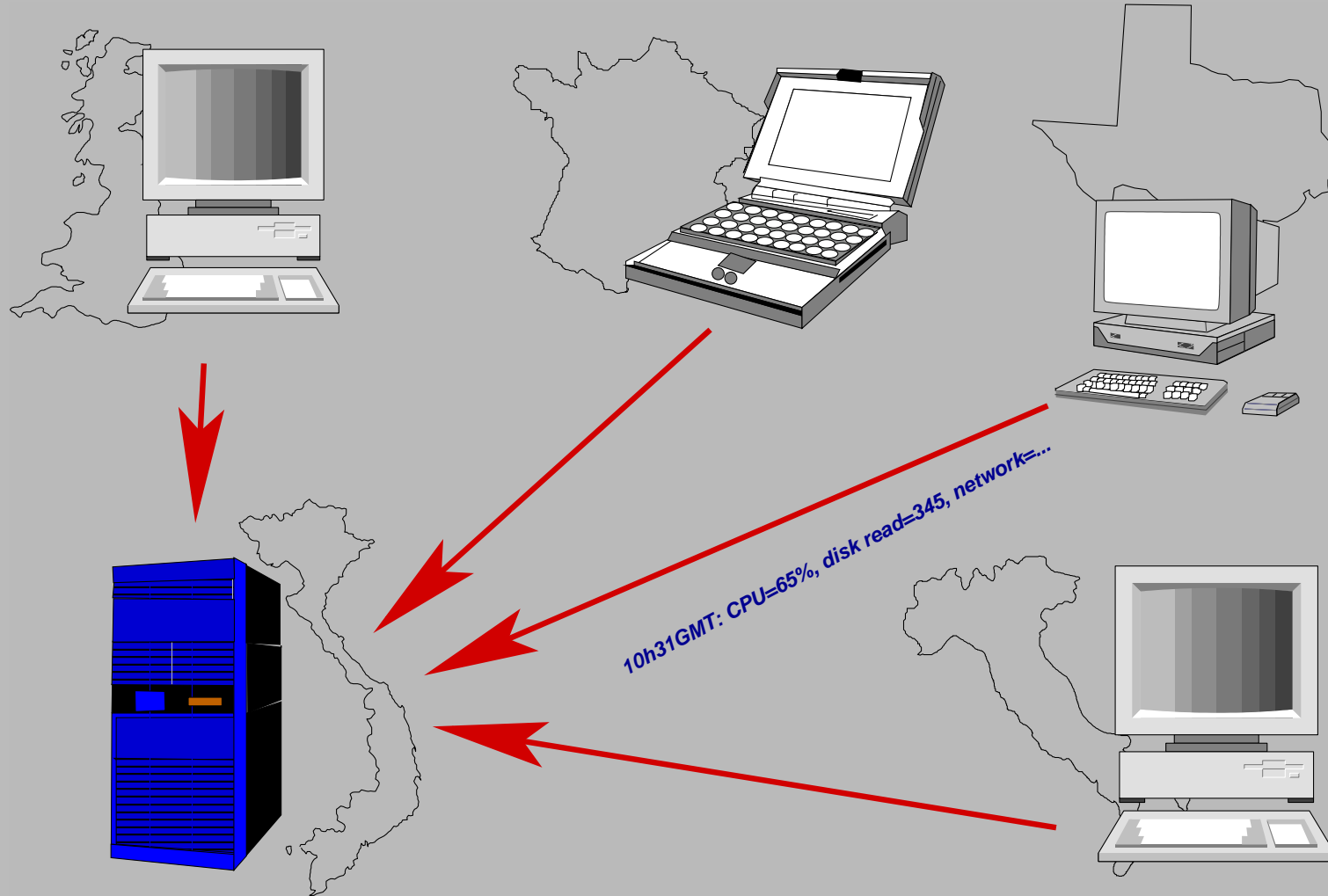
-
- X Placement, ordonnancement ;
 - X Représentation information ;

-
- X Placement, ordonnancement ;
 - X Représentation information ;
 - X Algorithmes de fouilles ;

-
- X Placement, ordonnancement ;
 - X Représentation information ;
 - X Algorithmes de fouilles ;
 - X Quelques résultats ;



Contexte architectural



X Network Time Protocol (NTP) synchronise le temps d'un client ;

X Organisation hiérarchique ;

```
[root@m24 etc]# crontab -e  
5 * * * * /usr/local/bin/ntpdate ntp.obspm.fr
```

Serveurs secondaire de temps de l'Observatoire de Paris.

X NTP assure une précision de 10ms ;

X <http://www.ntp.org/>

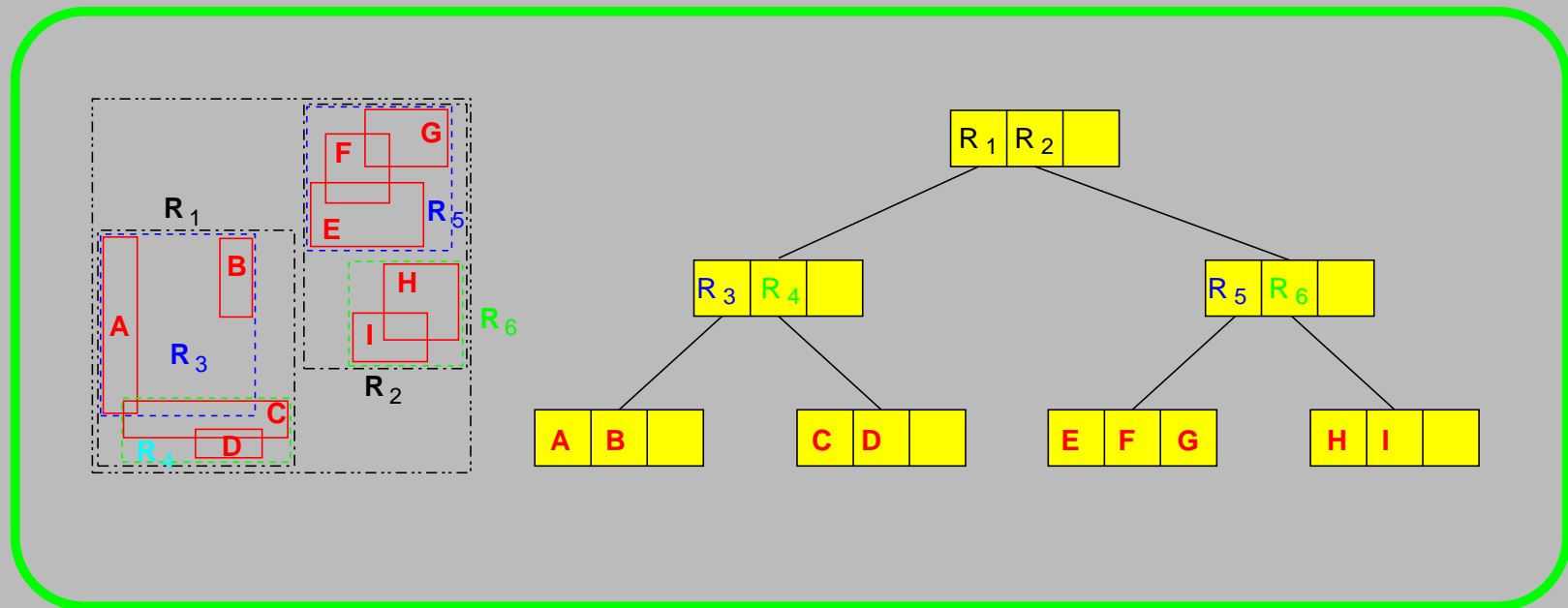
✘ Trace utilisée : 15 évènements, prise toutes les 15min, 100 machines ;

✘ 78,2Mo décompressé ;

⇒ 11Go pour 1000 machines échantillonnées toutes les minutes (pessimiste) ;



Doit-on stocker à plat les données ?



Indexation : Contient, Est Contenu, Est Egal.

× Placement : pb difficile

× Fouille : produire des règles qui décrivent les relations entre différentes séquences ;



Ex: si (BF) arrive 4 fois tandis que - (ABF) arrive 3 fois on produit la règle "si BF arrive alors il y a 75% de chance de voir arriver A"

⇨ Algo génération règles :

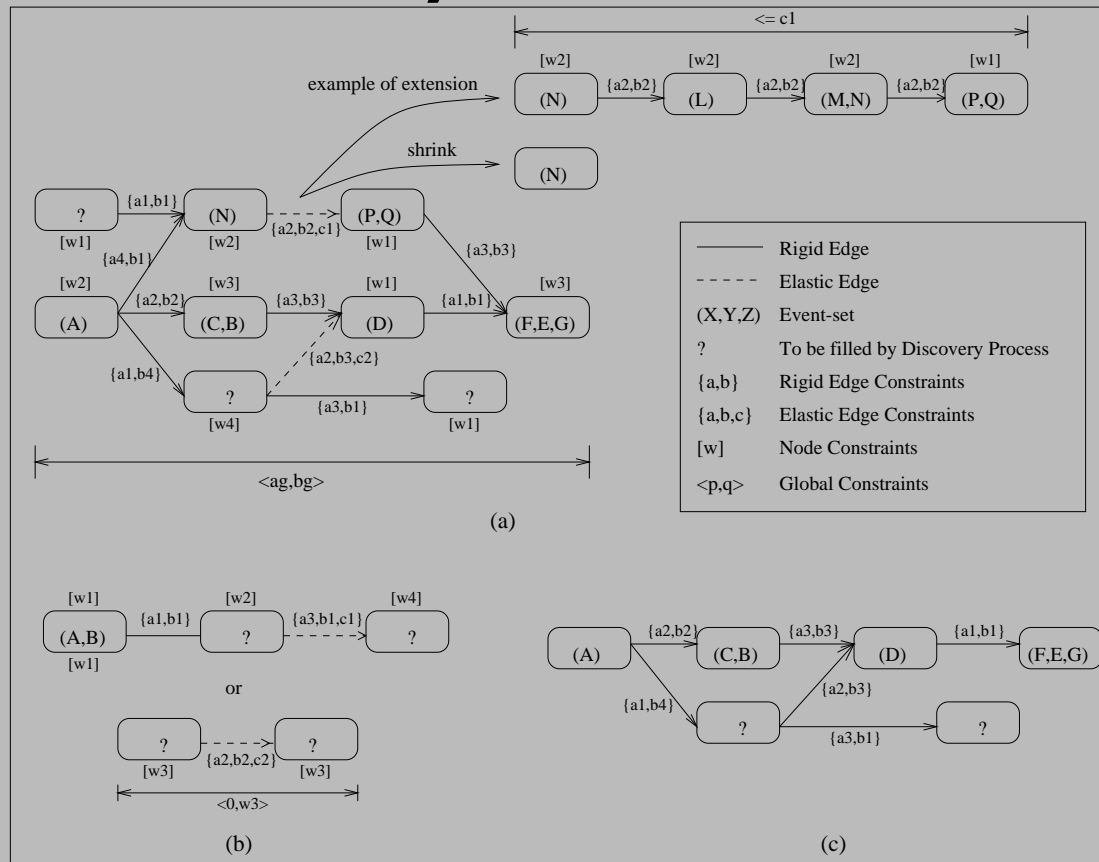
```
for all frequent sequences  $\beta$  do
  for all subsequences  $\alpha < \beta$  do
    conf = fr( $\alpha$ ) / fr( $\beta$ )
    if conf > min_conf then
      output  $\alpha \Rightarrow \beta$ 
      output conf
```

⇨ Le problème est celui de la découverte des **épisodes fréquents**.



-
- ⇒ Discovery and Monitoring Event Description (DAMED) Working Group.
 - ⇒ “The aim of this WG is to develop standard representations of the most widely used measurement values (the “top N”.)”
 - ⇒ RIEN DE FAIT SUR LE THEME “DISCOVERY” !

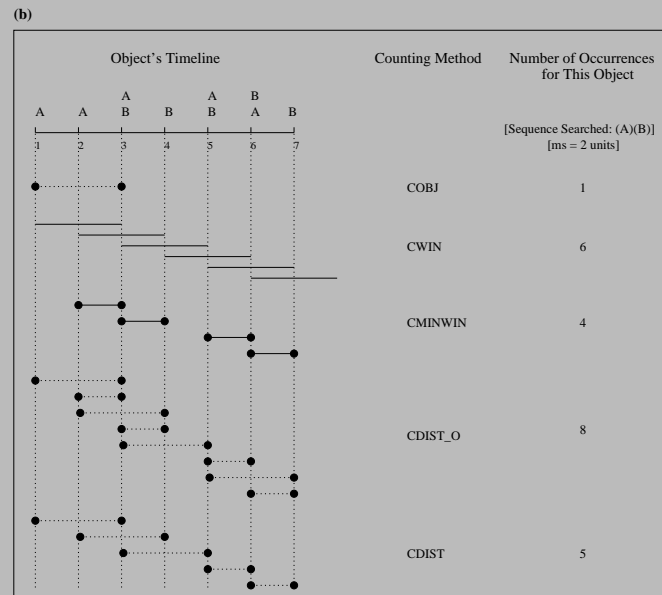
➡ on recherche les *relations séquentielles* entre les événements ;



- Une séquence est intéressante si elle apparaît suffisamment de fois ;
- Différentes méthodes de comptage (>8) ;

(a)

	Count All	Count Minimal
Count Windows	CWIN	CMINWIN
Count Occurrences	CDIST_O	CDIST





-
- ⇨ Déterminer Algo(R-arbre, séquence, seuil)
 - ⇨ Explosion combinatoire : le nombre maximum de sequences est $\mathcal{O}(m^k \cdot 2^{k-1})$.
Ex: m=k=15 : 717445350000000000000000
séquences !
 - ⇨ Utiliser des techniques d'élagage ;
 - ⇨ Deux techniques d'élagage ;

- ⇒ Règle d'antimonotonie : pour qu'une séquence soit fréquente il faut que toutes les sous-séquences le soient
- ⇒ on génère les séquences fréquentes de longueur i à partir de séquences candidates de longueur $k-1$;
- ⇒ Utilisation du seuil ;
- ⇒ WINEPI & Co

- ⇒ Exprimer des contraintes au niveau des motifs ;
- ⇒ SPIRIT [VLDB 1999] : introduit un langage souple de spécification des contraintes ;
- ⇒ cSPADE [Zaki, 2001] : reconnu comme le plus performant...mais problème avec le format d'entrée ;



- nrgrep-1.1.1/ (Gonzalo Navarro at the University of Chile)
- version améliorée de grep ; "approximate string matching" : insert, del, substitution.



Candidat potentiel pour la fouille.

- bit-parallélisme : structure de donnée permettant un gain > 2 vis à vis de KMP
- **Bug trouvé !!!**

↗ Charge CPU :

00%	-	09%	45932.0	observations
10%	-	19%	5994.0	observations
20%	-	29%	1801.0	observations
30%	-	39%	838.0	observations
40%	-	49%	1350.0	observations
50%	-	59%	789.0	observations
60%	-	69%	635.0	observations
70%	-	79%	850.0	observations
80%	-	90%	327.0	observations
90%	-	99%	591.0	observations

- ✗ Étude statistique disponible ;
- ✗ Les choix des outils de fouille et de représentation des données se précisent ;
- ✗ Défi : fouille multi-dimensionnelle ;
- ✗ Problématiques sous-étudiées dans le Grid-Forum ;

