Bases de Données Inductives
Des langages de requêtes au service de l'extraction de connaissances

Jean-François Boulicaut
LISI/LIRIS - INSA Lyon
http://lisi.insa-lyon.fr/~jfboulic

Tutoriel EGC 2003 (Lyon) - 22 janvier 2003

Un point de vue sur l'ECBD

- Knowledge Discovery in Databases (KDD)

  … process of identifying valid, novel, potentially useful and ultimely understandable information in data …

  données → propriétés → connaissances

  expression de gènes → associations → groupes de synexpressions

- Au carrefour de multiples disciplines

  … dont l 'apprentissage automatique et les bases de données

Exemples de tâches descriptives

- Données ventes → associations → stratégies de ventes

  beer, sausage ⇒ mustard (0.04,0.72)

- Alarmes → motifs → schéma de corrélation

  alarm=12 < alarm=13 ⇒ < alarm=43 (40) (227,0.6)

- Base de données → dépendances → contraintes

  Puhelinluettelo[<nimi,huone>] ⊆ Etudiants[<id,chambre>]

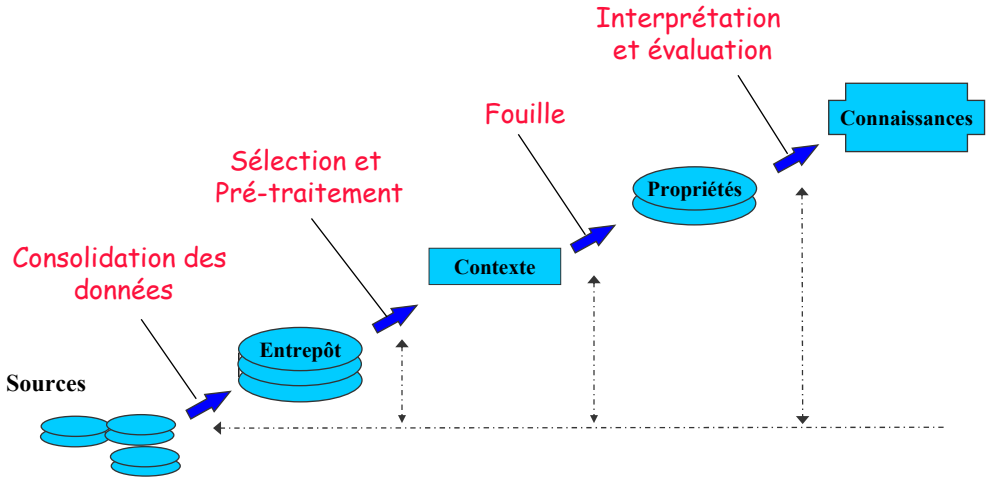
- Tâches descriptives vs. prédictives

Notre point de vue

- Une perspective « Bases de données » sur l 'ECBD
  - ✓ Assister l'exploration, l'analyse et la visualisation de données par une meilleure interaction entre les utilisateurs et les Systèmes de Gestion de Bases de Données
- Des langages de requête

  données → motifs
  modèles → connaissances

Processus d'extraction de connaissances



Une perspective BD à court-terme

- Algorithmes de fouille de données
  - ✓ Passage à l 'échelle
- Architectures logicielles (SGBD)

  Systèmes « SQL-aware »

  « Data Warehousing » et « Data Mining »

  SQL → OLAP → DM

Une perspective BD à long-terme (1)

- Pourquoi le modèle relationnel est-il un succès ?

  ✓Un langage de requête général avec de bonnes propriétés

  - fondements théoriques
  - sémantique déclarative
  - principe de clôture

  Il faut rechercher la même chose pour l 'ECBD

Une perspective BD à long-terme (2)

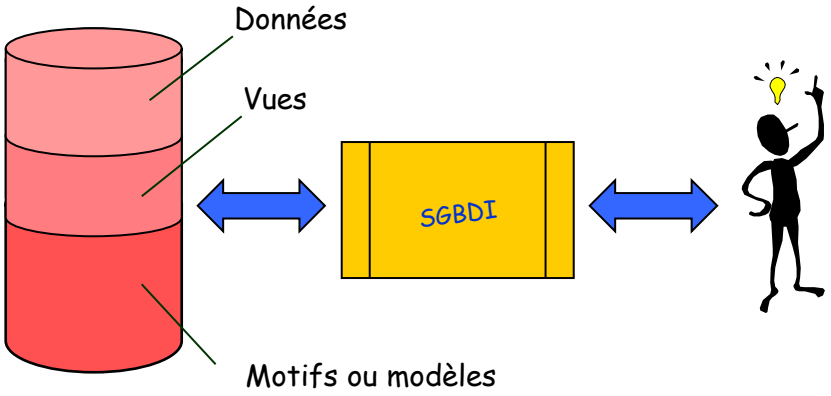
- Une vision

  « There is no such thing as real discovery, just a matter of the expressive power of the query languages »

  Imielinski & Mannila 96 (cacm)
  Mannila 97 (ilps)
  Boulicaut & al. 98 (pkdd) 99 (dawak)
  Giannotti & Manco 99 (pkdd)
  De Raedt 00 (ilp)

Le cadre des bases de données inductives

cInQ IST-2000-26469 (05/2001 - 05/2004)



Project funded by the Future and Emerging Technologies arm of the IST Programme FET-Open scheme

Plan du tutoriel

1. Qu 'est-ce qu'une base de données inductive ?
2. Deux exemples
   - Analyse de fragments moléculaires
   - Extraction de motifs fréquents
3. Quels langages de requêtes ?
   - … et quelles techniques d'évaluation ?
4. Perspectives

1. Qu 'est-ce qu'une base de données inductive ?

- Une base de données et une base de motifs ou de modèles
  - ✓ Processus ECBD = Processus d'interrogation
- Requêtes
  - Pré-traitement
  - Extraction
  - Post-traitement

Requêtes (1)

- Pré-traitement
  - ✓ Sélections des sources, agrégations, échantillonnage, discrétisations, etc
- Extraction (« Data Mining »)
  - ✓ « Sucre syntaxique » sur un algorithme d 'extraction
  - ✓ Vers des requêtes inductives

Requêtes (2)

- Post-traitement
  - ✓ Sélections des motifs ou modèles, élimination de la redondance, « crossing over »
  - ✓ Interrogation de collections matérialisées ?
- Exploitation
  - ✓ Voir e.g., Imielinski & Virmani 99 (dmkd)
  - ✓ PMML

Requêtes (3)

- Itération ... Aide à l 'interactivité
  - ✓ Propriété de clôture

$$Th(L,r,q) = \{\phi \in L \mid q(r,\varphi) \text{ est vrai}\}$$

Schéma (R,S_L)

Requêtes $(r, s) \Rightarrow (r', s')$

Des exemples « simples » ont été étudiés ... essentiellement pour la découverte de motifs

2. Deux exemples

Analyse de fragments moléculaires

Extraction de motifs fréquents

Ensembles (règles d 'association)

Motifs séquentiels

Abstraction cInQ

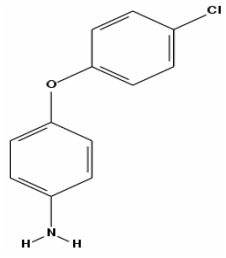
Molfea - Molecular Feature mining

- Trouver des fragments intéressants (sous-structures) dans des ensembles de molécules

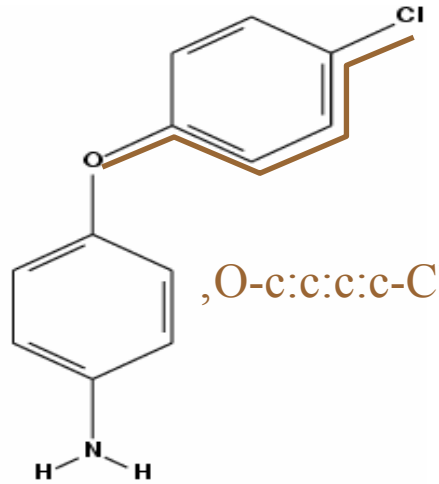
  Kramer et al. 01 (icml, sigkdd)

  De Raedt & Kramer 01 (ijcai)

- Pourquoi ?
  - ✓ Découverte de connaissances
  - ✓ Utilisation prédictive
    - SAR (Structure Activity Relationship)



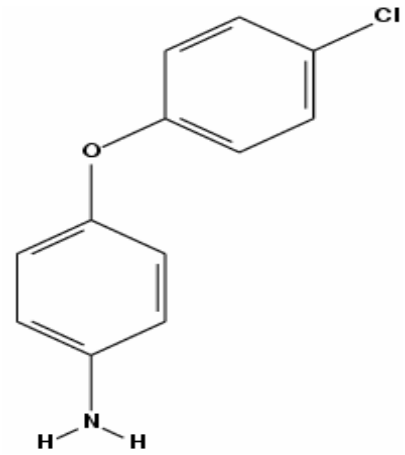
Molécules et fragments



‚O-c:c:c:c-Cl‘

- Structure 2D
- Fragments (linéaires)
  - ✓ Séquence d'atomes et de liaisons
  - ✓ ‚o', ‚c', ‚cl', ‚n', 's',... Pour les éléments
  - ✓ ‚-' ... Liaison simple
‚=, ... liaison double
‚#' ... Liaison triple
‚:' ... Liaison aromatique
  - ✓ (hydrogène implicite)
- Codage Smarts

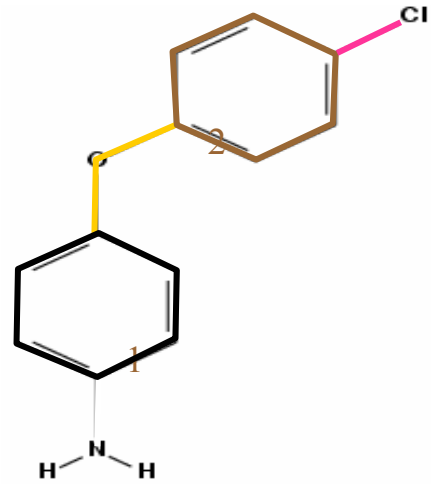
Codage Smiles



- Codage des structures de molécules utilisé par les "chimistes"
- Utilisé dans de nombreux outils (e.g., Daylight)
- Très compact
- Recherche de sous-structures efficace

$$N-c1:c:c:c(-O-c2:c:c:c(-Cl):c:c2):c:c1$$

Codage Smiles



$N-$

$N-c1:c:c:c:c:c1$

$N-c1:c:c:c(-O-):c:c1$

$N-c1:c:c:c(-O-$

$c2:c:c:c:c:c2):c:c1$

$N-c1:c:c:c(-O-c2:c:c:c(-Cl):c:c2):c:c1$,lt

Contraintes primitives pour Molfea

- Généralité
  - ✓ g est équivalent à s (*variante syntaxique*) quand ils sont l'inverse l'un de l'autre

    'C-O-S' et *'S-O-C'* désignent la même sous-structure
  - ✓ g est plus général que s ssi g est une sous-séquence de s ou g est une sous-séquence de l'inverse de s

    'Cl-O-S' ≤ *'Cl-O-S-c:c:c'*, *'O-Cl'* ≤ *'Cl-O-S'*

Contraintes primitives (suite)

- $f \leq P$, $P \leq f$, $not\ (f \leq P)$ et $not\ (P \leq f)$

  f ... Fragment cible à déterminer
  P ... Fragment spécifique

  Soit Freq(f,D) la fréquence relative du fragment f dans le jeu de données D

- $Freq(f,D1) \geq t$, $Freq(f,D2) \leq t$

  t ... nombre réel compris entre 0 et 1
  $D1$, $D2$ ... données

  E.g. $Freq(f, Pos) \geq 0.20$

Exemples de requêtes inductives

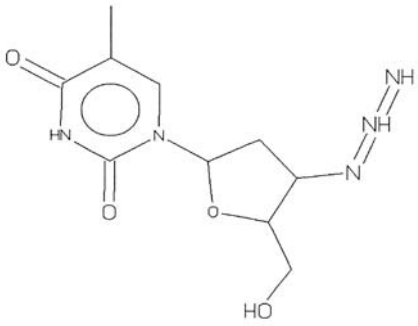
$(,N\text{-}O' \leq f) \wedge$
$(Freq(f, Act) \geq 0.1) \wedge (Freq(f, Inact) \leq 0.01)$

$not(,F' \leq f) \wedge not\ (,Cl' \leq f) \wedge$
$not\ (,Br' \leq f) \wedge not\ (,I' \leq f) \wedge$
$(Freq(f, Act) \geq 0.05) \wedge (Freq(f, Inact) \leq 0.02)$

The HIV Data Set Kramer & al 01 (sigkdd)

- Developmental Therapeutics Program's AIDS Antiviral Screen Database (http://dtp.nci.nhi.gov)
- Mesure de la protection des cellules humaines CEM contre une infection HIV-1
- 41768 composés sélectionnés parmi 43382
  - 40282 confirmés Inactifs
  - 1069 confirmés Modérément Actifs
  - 417 confirmés Actifs

AZT (Azidothymidine)



$N=N=N-C-C-C-n:c:c:c=O$

$N=N=N-C-C-C-n:c:n:c=O$

La majorité des fragments trouvés existent dans AZT.

Apporte des indications sur les structures ayant une activité anti-VIH

Re-découverte intéressante

Abstraction cInQ

- Composante « Données »
  - ✓ Très spécifique dans le cas de Molfea
  - ✓ Problématique standard ?
- Composante « Motifs »

$$Th(L,r,q) = \{\phi \in L \mid q(r,\varphi) \text{ est vrai}\}$$

  - ✓ Langage de motifs
  - ✓ Fonctions d 'évaluation
  - ✓ Contraintes primitives
  - ✓ Langage de requêtes

Découverte de motifs

- Ensembles et règles d 'association

  Agrawal & al. 93 (sigmod) 96 (aaai Press)
  Quelques centaines d'articles …

- Motifs séquentiels

  Agrawal & Srikant 95 (icde)
  Quelques centaines d'articles …

Un bon point de départ : les publications des projets

QUEST (IBM Almaden, USA)
FDK (Université d'Helsinki, FIN)
DB Miner (Université Simon Fraser, Canada)

Ensembles et règles (1)

- Traitement de contextes booléens

A_1	A_2	A_3
1	0	0
1	1	1
1	0	1
0	1	1

paniers - produits

documents - mot-clés

sessions - ressources

cellules - gènes

Ensembles et règles (2)

- Ensembles et règles d'association

A_1	A_2	A_3
1	0	0
1	1	1
1	0	1
0	1	1

$A_2\ A_3$ [2/4, clos, ...]

$A_1\ A_2$ [1/4, non clos, ...]

$A_1 \Rightarrow A_2$ [1/4, 1/3, ...]

$A_1\ A_2 \Rightarrow A_3$ [1/4, 1, ...]

Extraction de règles d'association

- Processus standard Agrawal & al. 96 (aaai press)

  Calcul de toutes les règles d'association dont les fréquences et confiances excèdent des seuils fixés

  Calcul des ensembles fréquents

  Calcul de règles intéressantes parmi les règles fréquentes

  En pratique: des milliers d'attributs
  des millions de tuples

Extraction d'ensembles fréquents

- Problème

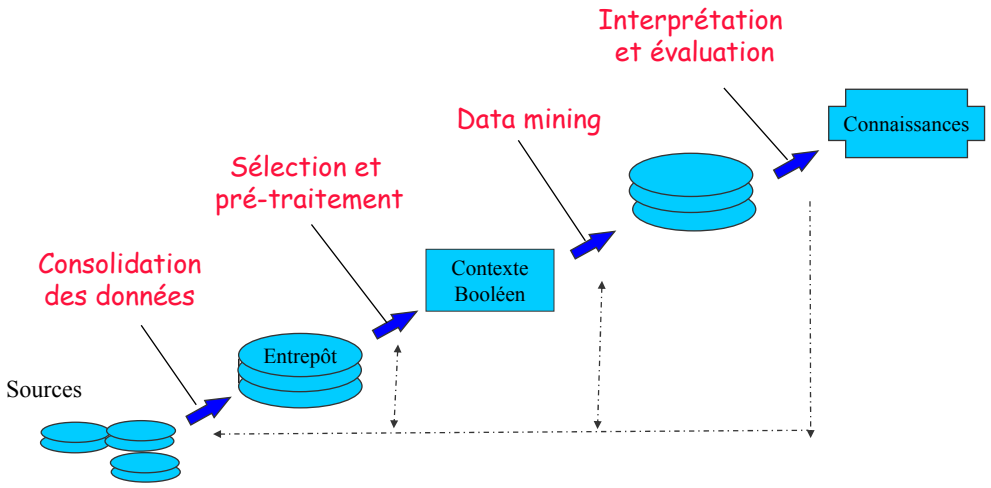
  – Calcul de théories « étendues » dans des données denses et très corrélées

  $Th(L \otimes E, r, q) = \{(\varphi, e) \in L \otimes E \mid q(r, \varphi) \text{ est vrai}\}$

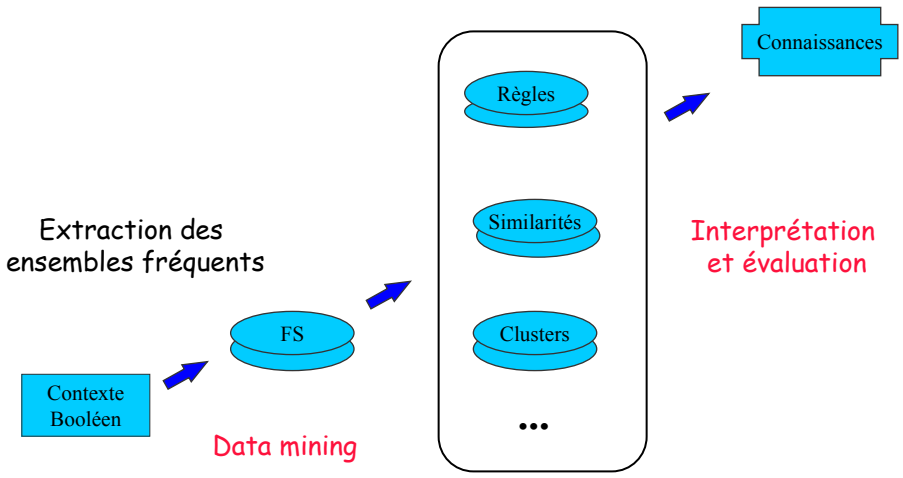
  Taille de l'espace de recherche

  Taille de la solution

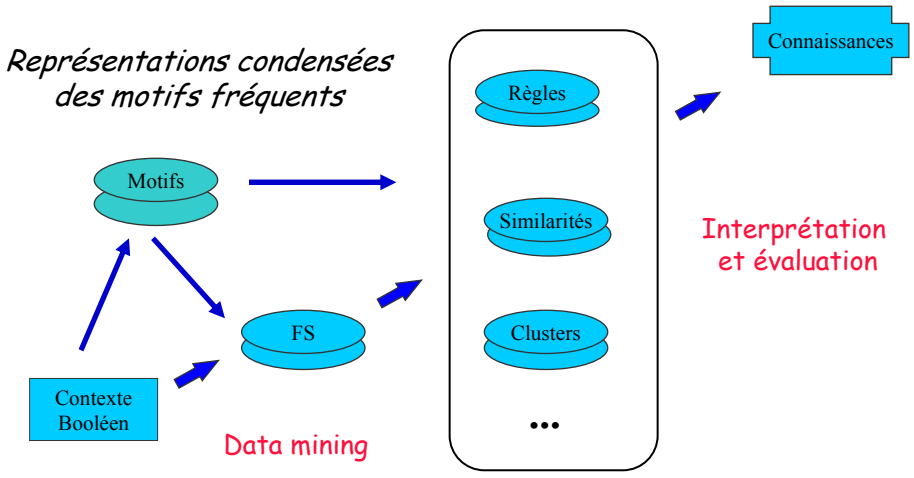
Retour sur le processus ECBD



… Usages multiples (1)



… Usages multiples (2)



Fonctions d'évaluation : la fréquence

A	B	C	D
1	0	1	0
1	1	1	0
0	1	1	1
0	1	0	1
1	1	1	0

Freq(S, r)

$C_{minfreq}(S)$ $C_{maxfreq}(S)$

Fréquence minimale/maximale par rapport à des seuils fixés par les utilisateurs

Autre fonction d'évaluation : la cloture

- La cloture de X est le sur-ensemble de X maximal qui a la même fréquence que X

cloture (X,r) = Items (Objets (X,r),r)

A	B	C	D
1	0	1	0
1	1	1	0
0	1	1	1
0	1	0	1
1	1	1	0

cloture{A} ={A,C}

NB. $A \Rightarrow C$ à la confiance 1

Ensembles clos

- Un ensemble clos est égal à sa clôture. C'est un ensemble de colonnes maximal qui supporte les mêmes lignes.

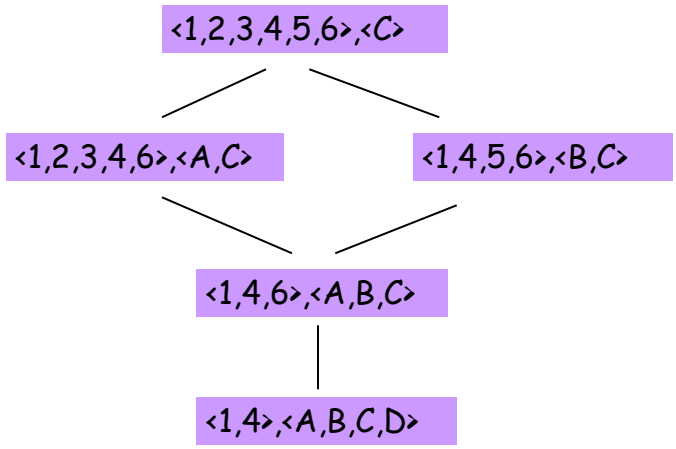
A	B	C	D
1	0	1	0
1	1	1	0
0	1	1	1
0	1	0	1
1	1	1	0

{A,C} est clos {A,B} n 'est pas clos

$C_{Close}(S)$

… treillis de concepts (Wille & al.)

	A	B	C	D
1	1	1	1	1
2	1	0	1	0
3	1	0	1	0
4	1	1	1	1
5	0	1	1	0
6	1	1	1	0



Représentations condensées des ensembles fréquents

- Ensembles maximaux *e.g.*, Bayardo 97 (sigmod) Max-Miner
- Ensembles clos Pasquier & al. 99 (icdt) - Boulicaut & Bykowski 00 (pakdd) - Han & Pei 00 (wdmkd) - Zaki 00 (sigkdd) Close - Closet - Charm
- Ensembles libres
- Ensembles δ-libres

  Boulicaut & al. 00 (pkdd) 03 (dmkd) - Bastide & al. 00 (sigkdd explorations) Min-Ex - Pascal
- Ensembles v-libres Bykowski & Rigotti 01 (pods) 02 (is) Lin-Ex
- NDI *Calders & Goethals 02 (pkdd)*

Liberté

- Un ensemble libre (motif clé) est tel qu'aucune règle d'association logique n'existe entre ses sous-ensembles

A	B	C	D
1	0	1	0
1	1	1	0
0	1	1	1
0	1	0	1
1	1	1	0

{A,B} est libre

{A,C} n 'est pas libre

$C_{Free}(S)$

Sa fréquence est différente des fréquences de ses sous-ensembles

r

ABCDE
ABCD
ACD
ABE
CD
CE

BD BC

ABC ABD BCD

ABCD

DE

ACE BCE ADE BDE CDE

ABCE ABDE ACDE BCDE

ABCDE

δ-liberté

- Un ensemble δ-libre est tel qu'il n'existe pas de règle δ-forte entre ses sous-ensembles

  $X \Rightarrow_\delta Y$ est δ-forte si elle accepte au plus δ exceptions

A	B	C	D
1	0	1	0
1	1	1	0
0	1	1	1
0	1	0	1
1	1	1	0

{A,B} était libre mais pas 1-libre

$C_{\delta\text{-Free}}(S)$

∨-liberté

- Un ensemble ∨-libre X est tel qu'il n'existe pas $A \subset X$, $B, C \in X$ de sorte que $A \Rightarrow B \vee C$ est valide dans r

A	B	C	D
1	0	1	0
1	1	1	0
0	1	1	1
0	1	0	1
1	1	1	0

A est ∨-libre

$C_{\vee\text{-free}}(S)$

Exemples de représentations

1	ABCD
2	AC
3	AC
4	ABCD
5	BC
6	ABC

Seuil 2

16 ensembles

1 maximal

5 clos

C, AC, BC, ABC, ABCD

5 libres

∅, A, B, D, AB

3 1-libres

∅, B, D

4 $\vee$-libres

∅, A, B, D

Régénération : exemples des clos

- $C_{close}(S,r) \wedge C_{minfreq}(S,r)$ $(C_{free}(S,r) \wedge C_{minfreq}(S,r))$

  Quand S est fréquent (inclus dans un clos fréquent)

  On considère le clos X tel que $S \subseteq X$ qui a une fréquence maximale et Freq(S,r) = Freq(X,r)

  Extractions faisables dans les contextes denses où l 'utilisation de $C_{minfreq}(S,r)$ seule est insuffisante

  Voir les doctorats de Pasquier (99), Bastide (00), Bykowski (02)

r	ABCDE ABCD ACD ABE CD CE

BD BC

ABC ABD BCD

ABCD

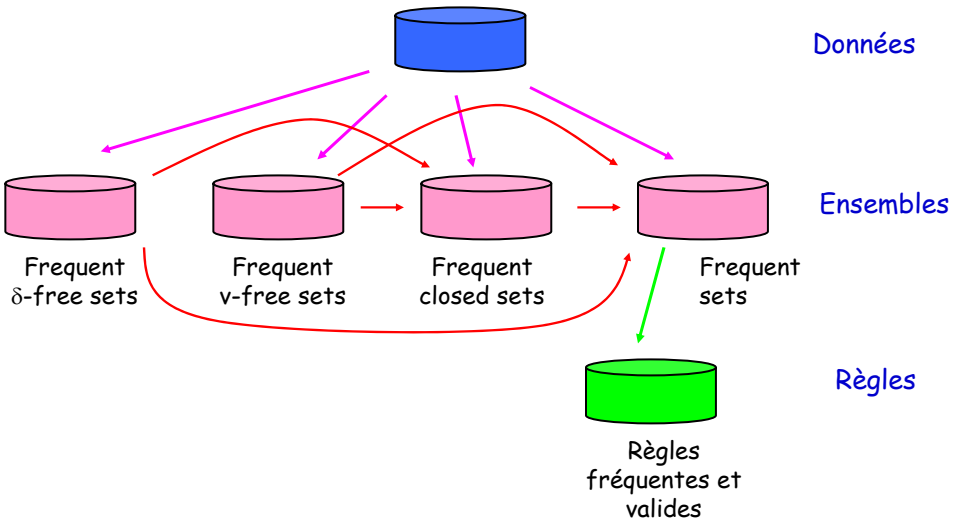
DE

ACE BCE ADE BDE CDE

ABCE ABDE ACDE BCDE

ABCDE

Exemples de prototypes Bykowski 02 (Ph.D)



Exemples de régénérations

1	ABCD
2	AC
3	AC
4	ABCD
5	BC
6	ABC

Seuil 2

16 ensembles

1 maximal

5 clos

C, AC, BC, ABC, ABCD

5 libres

∅, A, B, D, AB

3 1-libres

∅, B, D

4 $\vee$-libres

∅, A, B, D

1	ABCD
2	AC
3	AC
4	ABCD
5	BC
6	ABC

Comme ABCD est maximal

$F(AB) \approx (0.33 + 1)/2 \approx 2/3$

Connaissant les clos

$F(AB) = F(ABC) = 3/6 = 1/2$

Connaissant les libres

$F(AC) = F(A) = 5/6$

$F(C) = F(\varnothing) = 6/6 = 1$

Conaissant les 1-libres

$F(A) \approx F(\varnothing)$ car $\varnothing \Rightarrow_1 A$

$F(C) \approx F(\varnothing)$ car $\varnothing \Rightarrow_0 C$

Connaissant les $\vee$-libres

$F(C) = F(\varnothing)$ car $\varnothing \Rightarrow C \vee C$

Contraintes primitives sur les ensembles

- $C_{minfreq}$ (S) $C_{maxfreq}$ (S)
- $A \notin S$ $A \in S$
- $\{A,B,C,D\} \supset S$ $\{A,B,C,D\} \subseteq S$
- $S \cap \{A,B,C\} = \varnothing$ $S \cap \{A,B,C\} \neq \varnothing$
- sum(S.weight) $\leq$ v sum(S.weight) > v
- $\nexists A \in$ Items, Interest(A) > Interest (S)
- Constraintes sur les clôtures

Voir Ng & al. 98 (sigmod) Jeudy 02 (Ph.D)

Exemples de requêtes

- Sur les ensembles

$S \cap \{A,B,C\} = \varnothing \wedge C_{minfreq}(S,r)$

$C_{close}(S,r) \wedge C_{minfreq}(S,r)$

$C_{minfreq}(S,r1) \wedge C_{maxfreq}(S,r2)$

- Sur les règles d'association

$C_{minfreq}(X \Rightarrow Y,r) \wedge C_{minconf}(X \Rightarrow Y,r) \wedge sum(X.weight) \leq v$

$C_{minfreq}(X \Rightarrow Y,r) \wedge C_{free}(X,r) \wedge C_{close}(X \cup Y,r)$

Un exemple d'application

- Utilisation du prototype ac-miner sur des matrices d 'expression 89 par 822

  Bykowski 02 (Ph.D) Becquet & al. 02 (genome biology)

  4 225 , 775 _ 320 503 664 : 78
  3 149 , 125 _ 412 504(-1) 537 : 78

  X, Y : n $X \Rightarrow Y$

  C_{close} $(X \cup Y)$
  $C_{minfreq}$ (X)
  $C_{\delta\text{-free}}$ (X)

3. Des langages de requêtes concrets

- Langages SQL-like pour l 'extraction de règles d 'association
  - ✓MINE RULE — Meo & al. 96 (vldb) 98 (icde,dmkd)
  - ✓MSQL — Imielinski & al. 96 (kdd) 99 (dmkd)
  - ✓DMQL — Han & al. 96 (kdd) 00 (book)

  Voir également étude comparative Botta & al. 02 (dawak)

Exemples de requêtes MINE RULE (1)

- Une extension de SQL Meo & al. 96 (vldb) 98 (icde,dmkd)

Table Purchase

Tid	Customer	Item	Date	Price	Qty
1	c1	ski-pants	12/1	55	1
1	c1	beer	12/1	4	2
2	c2	shirts	12/1	21	1
2	c2	jackets	12/1	115	1
3	c1	diapers	12/1	18	1
…	…	…	…	…	…

Exemples de requêtes MINE RULE (2)

```
MINE RULE exemple as
SELECT DISTINCT 1..n Item as BODY, 1..1 Item as HEAD,
                SUPPORT, CONFIDENCE
WHERE HEAD.Item=« umbrellas »
FROM Purchase
GROUP BY Tid
HAVING COUNT(*) < 6
EXTRACTING RULES WITH SUPPORT: 0.02, CONFIDENCE: 0.9
```

E.g., jacket flight_Dublin ⇒ umbrellas (0.02,0.93)

Exemples de requêtes MINE RULE (3)

```
MINE RULE WordOfMouth as
SELECT DISTINCT 1..1 Customer as BODY, 1..n Customer as HEAD,
                SUPPORT, CONFIDENCE
WHERE BODY.Date <= HEAD.Date
FROM Purchase
GROUP BY Item
EXTRACTING RULES WITH SUPPORT: 0.01, CONFIDENCE: 0.9
```

E.g., c7 ⇒ c3 c12 (0.02,0.95)

MINE RULE (4)

++

- ✓ Utilisation complète de SQL
- ✓ L'évaluation des requêtes peut être efficace (ad-hoc)

--

- ✓ Limitation aux règles d'association
- ✓ Peu de possibilités d'exploiter la connaissance du domaine
- ✓ Pas de mécanismes dédiés au post-traitement (règles stockées dans des tables relationnelles)

Exemples de requêtes MSQL (1)

- Une extension de SQL Imielinski & Virmani 96 (kdd) 99 (dmkd)

  job=researcher $\wedge$ age = [26,38] $\Rightarrow$ position=AssProf (0.3,0.9)

  Emp(Id, Age, Job, Salary, Position)

  GET_RULES (Emp)
  INTO Rules
  WHERE … and support > 0.1 and confidence > 0.8

  SELECT_RULES (Rules)
  WHERE body has { (Age=*) (Job=*)} and head is { (Position=*)}

Exemples de requêtes MSQL (2)

- Recherche des enregistrements qui ne vérifient pas une règle de la forme « Age $\Rightarrow$ Salary » ayant une confiance de plus de 70%

  Emp(Id, Age, Job, Salary, Position)

  SELECT *
  FROM Emp
  WHERE violates all (GET_RULES (Emp)
  WHERE body is {(Age=*)}
  and head is {(Salary=*)}
  and confidence > 0.7)

Problèmes ouverts

- Quelles primitives ?
  - ✓ Insuffisance des propositions actuelles
    - Pré-traitement – Extraction – Post-traitement
- Comment optimiser ces requêtes ?
  - ✓ Notion de requête inductive complexe
    - Au delà des conjonctions de contraintes monotones et anti-monotones
    - Au delà de domaine uniques et simples
- Comment optimiser des séquences de requêtes ?
  - ✓ Compromis difficiles

Développement de solveurs ?

- Calcul de

$$Th(L \otimes E, r, q) = \{(\varphi, e) \in L \otimes E \mid q(r, \varphi) \text{ is true}\}$$

quand q est une combinaison de contraintes primitives

« Générer et tester » est généralement impossible

« Pousser » des contraintes peut être très efficace

« Pousser » des contraintes peut être très difficile

Propriétés des contraintes

- Anti-monotonicité de q par rapport à $\leq$
  - ✓q est anti-monotone w.r.t. $\leq$ ssi pour tout g,s $g \leq s$ et s satisfait q implique g satisfait q

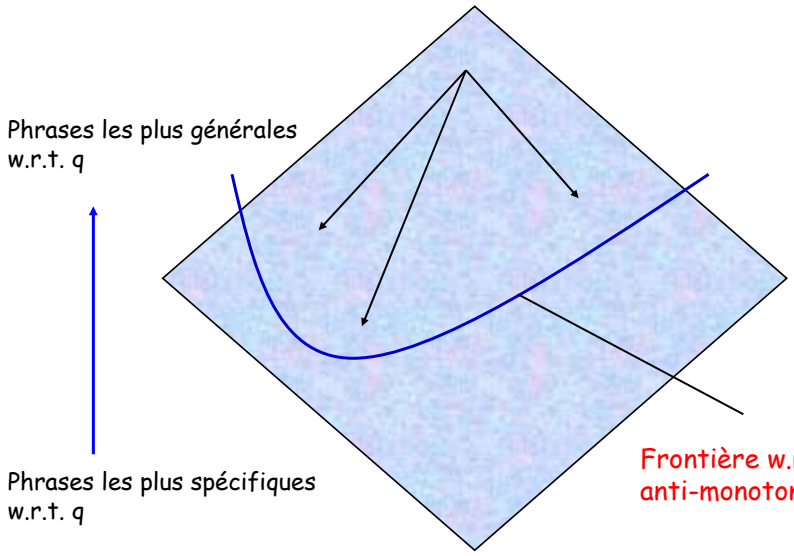
    E.g., la fréquence minimale est anti-monotone w.r.t. généralité (inclusion ensembliste)

    Possibilités d'élagage

    Mannila & Toivonen 97 (dmkd)
- Frontière (S-set, Mitchell 82 (ai))

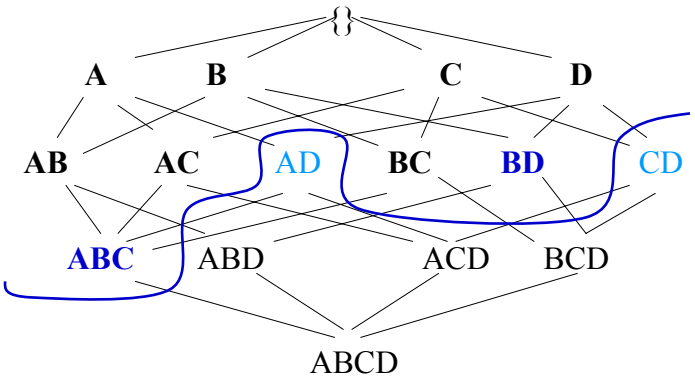
Contraintes anti-monotones sur les ensembles

- $C_{minfreq}$ (S)
- $A \notin S$
- $\{A,B,C,D\} \supset S$
- $S \cap \{A,B,C\} = \varnothing$
- sum(S.weight) $\leq$ v
- C_{free} (S) … mais pas C_{close} (S)



$C_{minfreq}(S,r)$

A	B	C	D
1	0	1	0
1	1	1	0
0	1	1	1
0	1	0	1
1	1	1	0



Seuil 0.3

Extractions

$C_{minfreq}\ (S,r)$

$C_{close}\ (S,r) \wedge C_{minfreq}\ (S,r)$

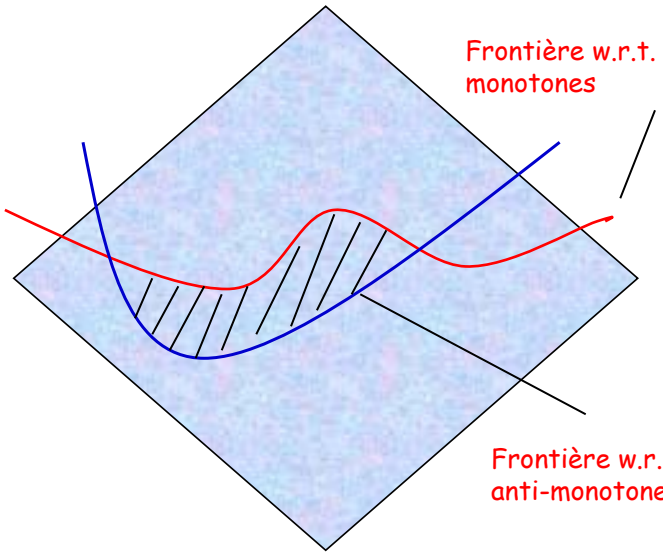
$C_{free}\ (S,r) \wedge C_{minfreq}\ (S,r)$
$C_{\delta\text{-}free}\ (S,r) \wedge C_{minfreq}\ (S,r)$
$C_{v\text{-}free}\ (S,r) \wedge C_{minfreq}\ (S,r)$

anti-monotones

$Cam\ (S,r) \wedge Cm\ (S,r)$

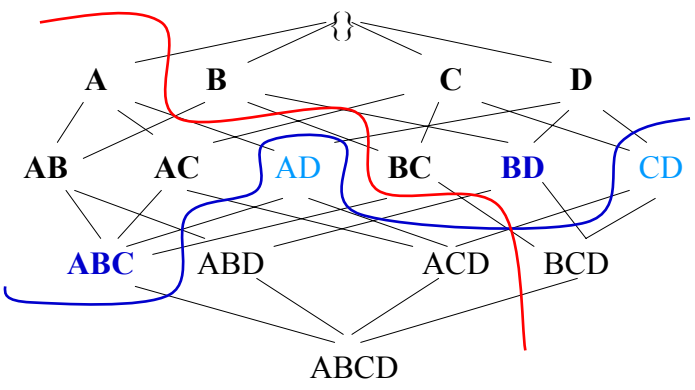
Représentation des solutions w.r.t. contraintes monotones

- Les contraintes monotones imposent une frontière G sur l'espace solution
  - ✓ q est monotone w.r.t. $\leq$ ssi not(q) est anti-monotone w.r.t. $\leq$



$C_{minfreq}(S,r) \wedge (A \in S)$

A	B	C	D
1	0	1	0
1	1	1	0
0	1	1	1
0	1	0	1
1	1	1	0



Seuil 0.3

Algorithmes génériques sous des conjonctions de contraintes

- CAP Ng & al. 98 (sigmod)

  Contraintes anti-monotones et succinctes

- Levelwise version space algorithm De Raedt & al. 01

  Molfea

- Generic levelwise algorithm Jeudy 02 (Ph. D)
- Dual Miner Bucila & al. 02 (sigkdd)

  Contraintes anti-monotones et monotones

Analyse de données "séquentielles"

- Exemple: données issues d'un serveur "intranet"
  - ✓Trace des connexions

    CIPC125 Bob pc107a-14 12/02/99 23:19:00 In
  - ✓WWW logs

    222.222.222.222, -, 13/02/99, 00:53:46, W3SVC, CIPCSERVEUR, 111.111.11.111, 0, 544, 111, 404, 3, GET, /intranet/it2793.html, -
  - ✓Nslookup dictionary

    pc107a-14 12/07/98 222.222.222.222
- Autres sources

Types de contextes

Evt	Date	P1	P2	P3	...
E	1	0	0	1	...
D	2	0	1	1	...
F	2	1	0	0	...
A	3	0	0	0	...
B	3	1	1	0	...
C	3	0	1	0	...
...	...	...	...	...	...

Recherche de motifs séquentiels

- Nature des données
  - ✓ Séquences ou bases de séquences
  - ✓ Ordre temporel ou ordre spatial
- Langage des motifs

  A > B > C *A B C* *(A B) (A C)*
- Contraintes primitives
  - ✓ Contraintes syntaxiques, temporelles, sur les fréquences ...

Solveurs

- GSP

  Agrawal & Srikant 95 (icde) 96 (edbt)

- Minepi/Winepi

  Mannila & al. 95 (kdd)

- SPADE/cSPADE

  Zaki 99 (cikm) 00 (cikm) 01 (ml)

- FreeSPAN

  Han & al. 00 (sigkdd)

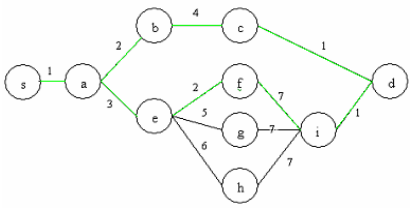
- SPIRIT

Une famille d'algorithmes

- SPIRIT Recherche de motifs séquentiels fréquents satisfaisant une expression régulière

  Garofalakis & al. 99 (vldb,tkde)

  - ✓ Représentation de l 'expression par un AEF
  - ✓ 4 algorithmes
    - SPIRIT (N)
    - SPIRIT (L)
    - SPIRIT (V)
    - SPIRIT (R)



Autres solveurs (exemples)

- Doctorat Masseglia (2001)
- Algorithme Galibot

  Capelle & al. 02 (egc,ideal)
- Algorithme RE-Hackle

  Albert-Lorincz 02 (dea)
- SPIRIT-Log

  Masson & Jacquenet 02 (ilp)
- SEQLOG

  Lee & De Raedt 02 (kdid)

cInQ IST-2000-26469
consortium on discovering knowledge with Inductive Queries (05/01-05/04)

- Objectif à long-terme
  - ✓ Identifier les primitives fondamental pour la fouille de bases de données, i.e., une étape vers la définition de langages de requêtes généralistes pour l 'extraction de connaissances


Méthodologie cInQ

- Etudier diverses extensions inductives de langages de requêtes pour différents domaines (e.g., ensembles, séquences, équations) et démontrer l'intérêt de l'approche « Bases de données inductive » sur des applications ciblées.
- Plan de travail Année 1
  - ✓Depuis des instances spécifiques de bases inductives à une première théorie
  - ✓Primitives et solveurs pour les domaines ITEM, DD et DLOG

Plan de travail cInQ

WP1 The inductive database framework

WP2 Design of pattern domains
WP3 Design of solvers
WP4 Implementation
WP5 Application
WP6 Evaluation and assesment
WP7 Dissemination
WP8 Project management

Principaux résultats à mi-parcours

- De Molfea à une première théorie
  - ✓ Extraction de graphes linéaires et de chaînes (MolFea, SEQLOG)
  - ✓ Une première théorie des bases de données inductives

    De Raedt 02 (dtdm,rr) De Raedt & al. 02 (icdm)
- Représentations condensées et ε-adéquates
  - ✓ Recherche d 'instances de BDI permettant de répondre efficacement (et plus ou moins exactement) à certains types de requêtes
- Langages de requêtes SQL-like (MINE RULE, XMINE)

  Meo & al. 02 (rr), Braga & al. 02 (dawak,rr)

4. Perspectives

- Robustesse des concepts identifiés ?
  - ✓ E.g., concept de représentations condensées
- Evaluation sous contraintes : étude de stratégies pour les requêtes inductives complexes
- Descriptif vs. Prédictif ?
  - ✓ Extraction sous contraintes de modèles (voir numéro spécial de SIGKDD explorations en juin 2002)
- La recherche doit être dirigées par les applications … et les données réelles

Remerciements

Consortium cInQ

- INSA Lyon (C. Rigotti, B. Jeudy, C. Masson)
- University of Torino (R. Meo, M. Botta)
- Politecnico di Milano (S. Ceri)
- Albert-Ludwigs University Freiburg (L. De Raedt, S. Kramer)
- Nokia Research Center Helsinki (H. Mannila, M. Klemettinen)
- Institute Jozef Stefan (S. Dzeroski)

http://www.cinq-project.org