Fouille de Données relationnelle

et

échantillonnage

Représentation relationnelle

- Description par attributs : une entité est décrite par ses caractéristiques
  $\Rightarrow$ base de données composée d'une seule table

- Description structurée : une entité est décrite par ses composants et les relations entre ces composants
  $\Rightarrow$ base de données relationnelle

Exemple de description par attributs

décrire un champignon par ses caractéristiques globales (forme, couleur du chapeau, taille, couleur du pied...)

CHAMPIGNONS				
forme-chapeau	couleur-chapeau	taille-pied	couleur-pied	. . .
cloche	marron	étroit	chocolat	
⋮	⋮	⋮	⋮	. . .

Figure 1: Représentation de champignons par une unique table

Exemple de description structurée - 1

décrire un champignon par ses composants : chapeau (forme, couleur...), pied (forme, couleur...)

CHAMPIGNONS
ident
c1
⋮

CHAPEAU			
ident	forme	couleur	. . .
c1	cloche	marron	
⋮	⋮	⋮	⋮

PIED			
ident	taille	couleur	. . .
c1	étroit	chocolat	
⋮	⋮	⋮	⋮

Figure 2: Représentation de champignons par une base de données relationnelle

Exemple de description structurée - 2

décrire une molécule par ses atomes (sous-composants) et les liaisons entre ses atomes (relations)

MOLÉCULES
mident
m1
⋮

ATOMES		
aident	élément	charge
a1	O	2−
a2	H	+
⋮	⋮	⋮

LIAISONS			
mident	lident	atome1	atome2
m1	l1	a1	a2
m1	l2	a1	a2
⋮	⋮	⋮	⋮

Figure 3: Représentation de molécules par une base de données relationnelle

Entrées/sorties d'un algorithme de FD

Entrées :

1. un ensemble de données $\mathcal{D}$, stockées dans une base de données
2. un langage de représentation des motifs (connaissances)
3. restrictions et préférences sur les motifs engendrés et le processus de recherche
4. un critère d'évaluation c des motifs

Sorties : un (ensemble de) motif(s) satisfaisant c

Approche générer-et-tester

Soient un ensemble de données $\mathcal{D}$ et un critère c

1. engendrer un motif m
2. évaluer m sur $\mathcal{D}$ compte tenu de c
3. si m ne satisfait pas c alors aller en 1

Exemple : GRIL, algorithme génétique pour l'apprentissage de concepts relationnels ([Braud, Vrain])

Evaluation des motifs

Exemple en classification :

évaluer la règle “une molécule est organique si elle a un atome de carbone et un atome d’hydrogène”

$\Rightarrow$ compter les molécules organiques ayant un atome de carbone et un atome d’hydrogène parmi les données

Forme des requêtes

SELECT COUNT (DISTINCT $(x_1, \ldots, x_i)$)
FROM $rel\_elements, rel_1, \ldots, rel_n$
WHERE $EnsSelections$ AND $EnsJointures$

avec

$rel\_elements(x_1, \ldots, x_p)$: relation établissant le lien entre les éléments à dénombrer et la règle

$EnsSelections/EnsJointures$: sélections/jointures sur les attributs de $rel\_elements, rel_1, \ldots, rel_n$

Exemple : compter les molécules ayant un atome de carbone et un atome d'hydrogène

SELECT COUNT(DISTINCT $m.id_{mol}$)
FROM $molecule\, m$, $atome\, a1$, $atome\, a2$
WHERE $m.id_{mol} = a1.id_{mol}$ AND $a1.elt =' C'$
AND $m.id_{mol} = a2.id_{mol}$ AND $a2.elt =' H'$

Evaluation des motifs

Soit les tables :

Molécule

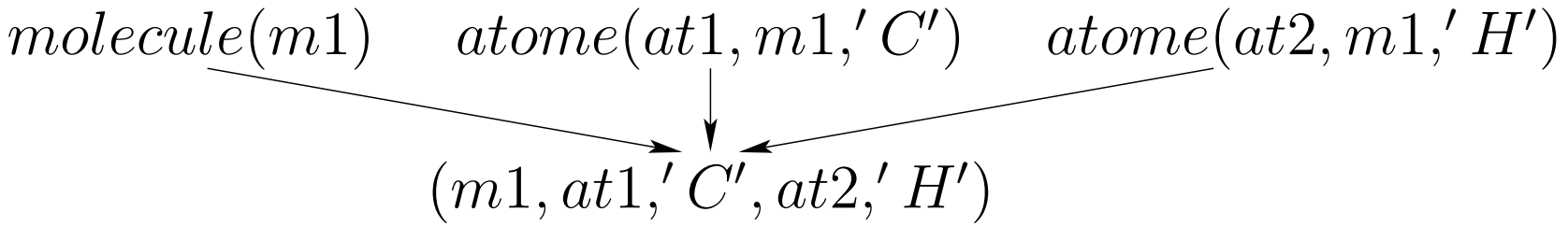
id_{mol}
$m1$
$m2$
$\vdots$

Atome

id_{at}	id_{mol}	elt
$at1$	$m1$	'C'
$at2$	$m1$	'H'
$at3$	$m1$	'C'
$at4$	$m2$	'C'
$at5$	$m2$	'C'
$\vdots$	$\vdots$	$\vdots$

Construction d'un tuple de la jointure :

$$a1.id_{mol} = m.id_{mol} \text{ AND } a2.id_{mol} = m.id_{mol}$$



Evaluation des motifs

Table issue des jointures :

id_{mol}	id_{at1}	$elt1$	id_{at2}	$elt2$
$m1$	$at1$	'C'	$at1$	'C'
$m1$	$at1$	'C'	$at2$	'H'
$m1$	$at1$	'C'	$at3$	'C'
$m1$	$at2$	'H'	$att1$	'C'
$m1$	$at2$	'H'	$at2$	'H'
$m1$	$at2$	'H'	$at3$	'C'
$m1$	$at3$	'C'	$a1$	'C'
$m1$	$at3$	'C'	$at2$	'H'
$m1$	$at3$	'C'	$at3$	'C'
$m2$	$at1$	'C'	$at1$	'C'
$m2$	$at1$	'C'	$at2$	'C'
$m2$	$at2$	'C'	$at1$	'C'
$m2$	$at2$	'C'	$at2$	'C'

$\Rightarrow$ problème multi-instances : compter des valeurs distinctes

Motivations

Le calcul d'un ensemble de jointures est

- coûteux
- répété car commun à de nombreuses requêtes

Etudier une approche pour accélérer (voire rendre possible) l'évaluation des motifs engendrés

- temps de calcul raisonnable
- compromis gain en temps / précision acceptable
- passage à l'échelle

Spécificités des requêtes

- utilisées lors de l'évaluation de connaissances engendrées par un processus de Fouille de Données,
- requêtes de dénombrement de valeurs distinctes,
- composées de :
  - jointures spécifiant des égalités entre attributs
  - sélections traduisant des conditions sur les valeurs des attributs

Différentes approches

- Réduction des requêtes (minimisation, simplification sémantique, . . .)
- Réduction des données (échantillonnage, ondelettes, histogrammes, . . .)
- Traitemement par lots

$\Rightarrow$ choix d'une approche par échantillonnage : souple et permet de traiter tous les types de données

Pré-calculs de jointures

Observation :

contraintes liées au schéma de la base

$+$

contraintes de forme pour les règles

$\Downarrow$

combinaison de jointures limité

[Acharya,Gibbons,Poosala,Ramaswamy SIGMOD 1999]
utilisent le schéma uniquement

Approche proposée

En phase initiale :

pour chacun des différents schémas de jointures ($SJoin_i$) possibles, calcul et stockage d'un échantillon (Ech_i)

Puis pour chaque requête :

SELECT COUNT(DISTINCT $(x_1, \ldots, x_i)$) FROM $rel\_elts, rel_1, \ldots, rel_n$

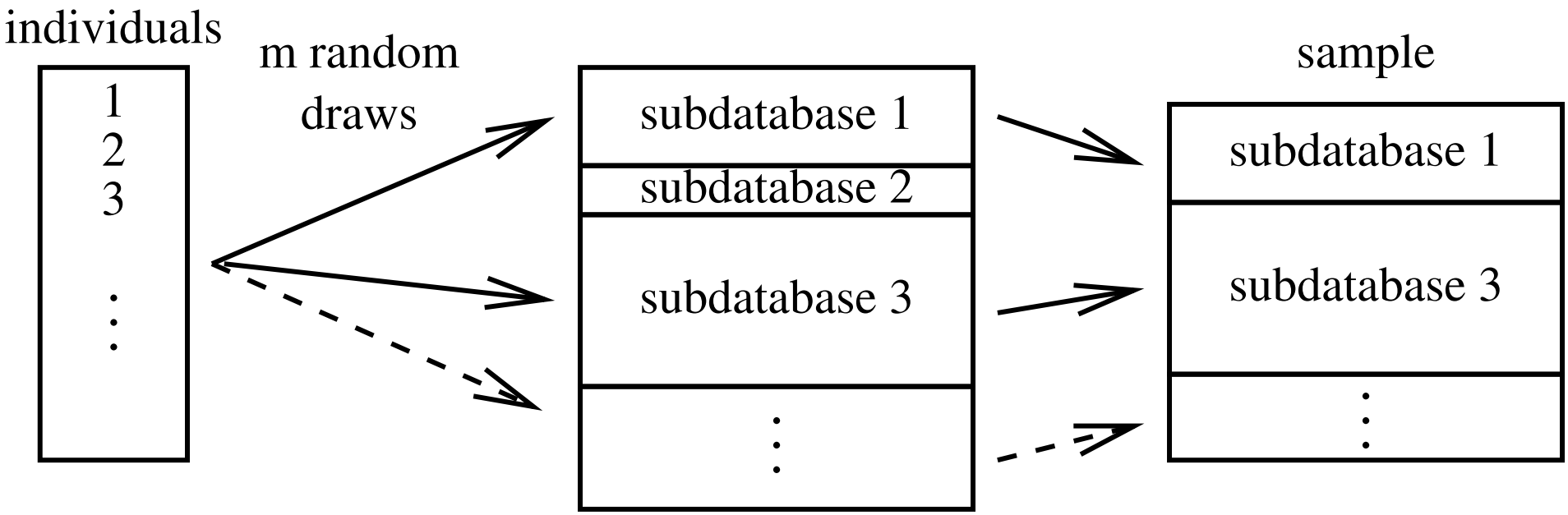
WHERE $EnsSelections$ AND $EnsJointures$

- sélectionner l'échantillon Ech_r tel que $SJoin_r = EnsJointures$
- exécuter la requête

  SELECT COUNT(DISTINCT $(x_1, \ldots, x_i)$) FROM Ech_r

  WHERE $EnsSelections$
- utiliser un estimateur pour extrapoler le résultat

Echantillonnage au niveau des entités



Echantillonnage au niveau des entités

Classiquement :
P proportion totale estimée à p proportion dans l'échantillon

Intervalle de confiance à 95% :

$$P \in [p - 1.96\sqrt{\frac{p(1-p)}{n}}, p + 1.96\sqrt{\frac{p(1-p)}{n}}]$$

Dans le pire des cas :

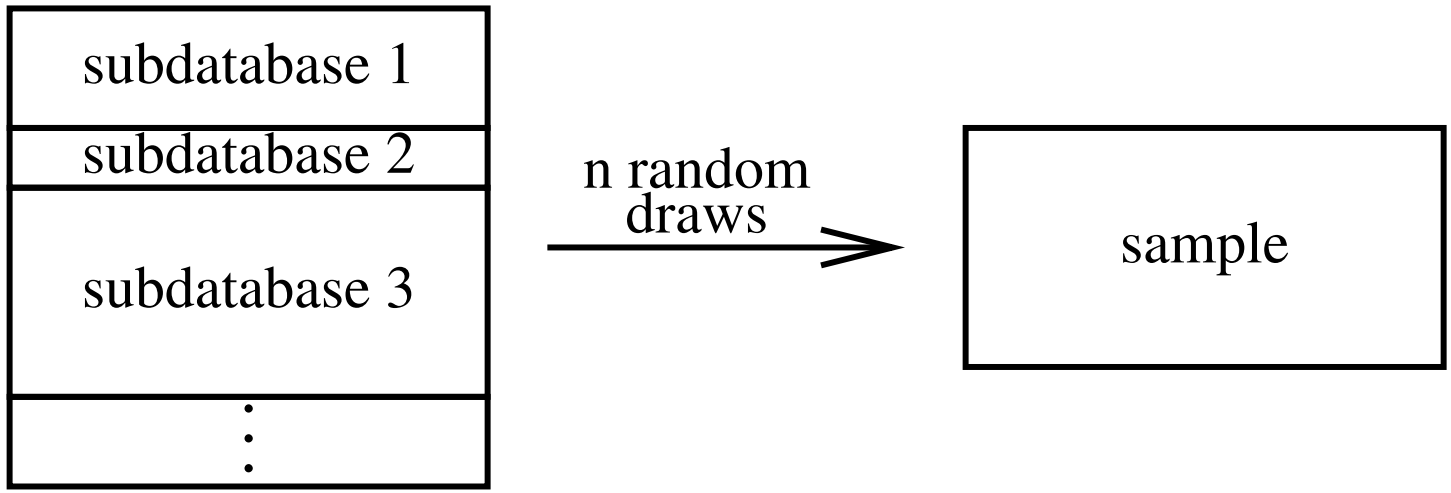
Nombre d'individus dans échantillon	erreur sur estimation
96	$\pm$ 10%
385	$\pm$ 5%
9604	$\pm$1 %

Echantillonnage au niveau des entités

MAIS

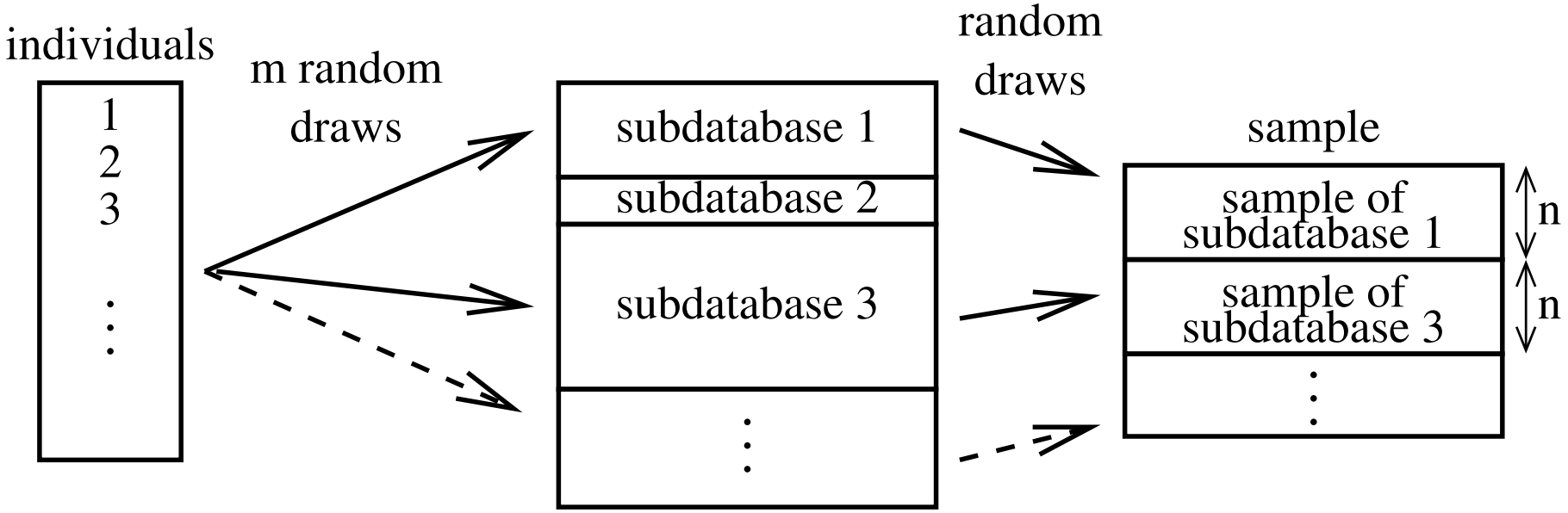
on ne contrôle pas la taille de l'échantillon

Echantillonnage au niveau des instances



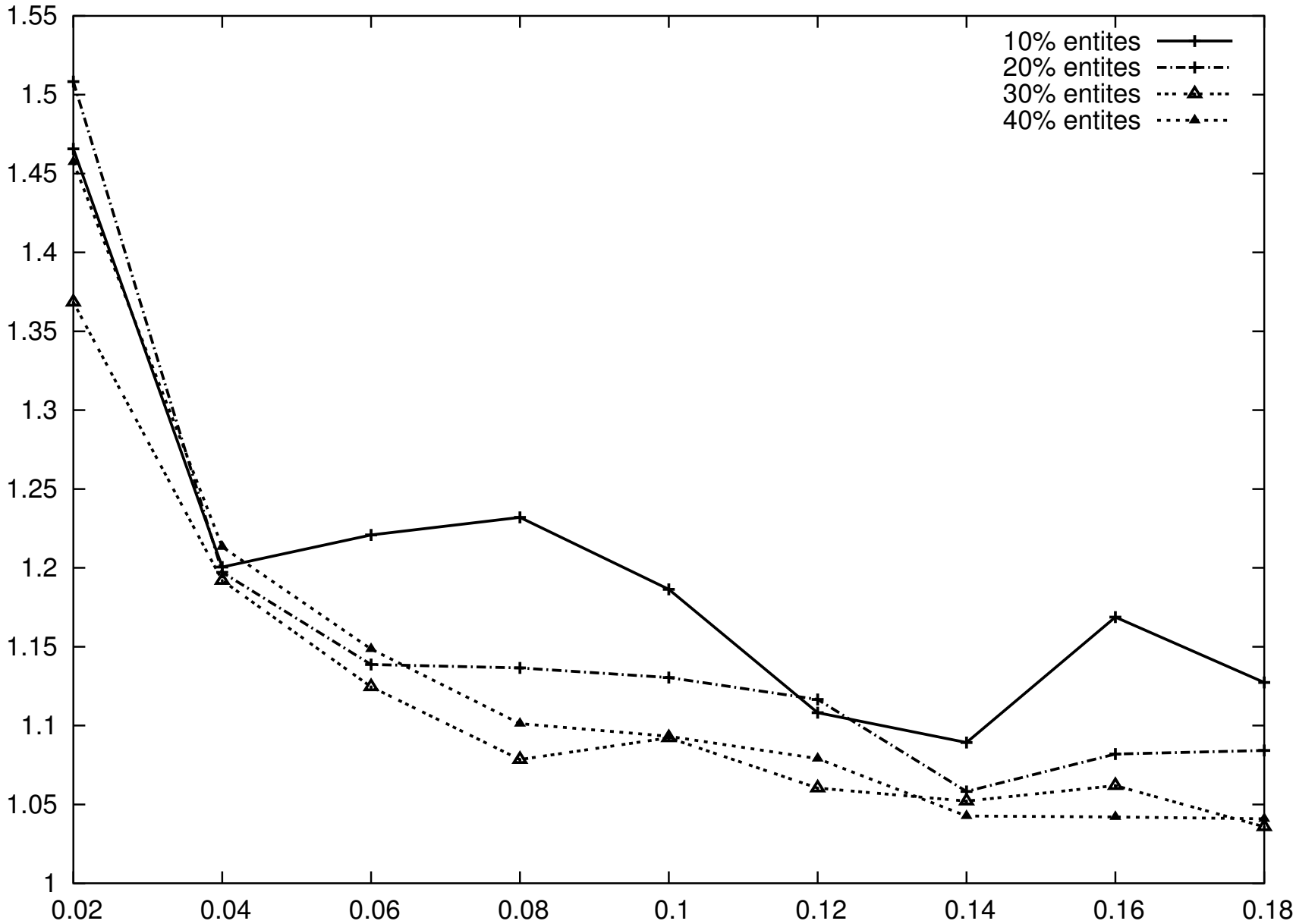
$\Rightarrow$ pas d'estimateur avec erreur garantie pour le nombre de valeurs distinctes [Charikar, Chaudhuri, Motwani, Narasayya]

Echantillonnage à deux niveaux



$\Rightarrow$ proportion totale qui passe la sélection estimée à p
proportion ayant au moins une instance qui passe la sélection dans l'échantillon

Exemple de résultats expérimentaux



Conclusions et perspectives

Conclusions

- échantillonnage pour l'évaluation des motifs
  - $\Rightarrow$ enjeux importants : traiter de plus grandes bases de données, rendre des algorithmes applicables
  - $\Rightarrow$ travaux en Fouille de Données relationnelle peu nombreux et peu avancés
  - $\Rightarrow$ exploitation des specificités du problème : marge d'imprécision tolérée, pré-calcul possible
  - $\Rightarrow$ tests tendent à montrer l'intérêt d'un échantillonnage à deux niveaux

Conclusions et perspectives

Objectifs actuels :

- étudier une adaptation de l'estimateur classique pour l'échantillonnage à deux niveaux
- étudier les paramètres de l'échantillonnage (nombre d'individus et d'instances par individu retenus), établir un seuil pour ces paramètres garantissant en probabilité une erreur d'estimation fixée