GDR I3 – Groupe Fouille de Données

"Quelques problématiques de gestion des données pour la fouille des données"

Georges Hébrail

TELECOM Paris

12 juin 2003

Plan

Positionnement

Oubli des données dans les DW

Résumés de données pour la fouille

Environnements de gestion de la fouille

Positionnement

Gestion des données pour la fouille

- Bases et entrepôts de données
- Fouille de données
- Interactions, apports mutuels

Plan

Positionnement

Oubli des données dans les DW

Résumés de données pour la fouille

Environnements de gestion de la fouille

Oubli des données dans les DW

Engorgement des entrepôts de données

Deux approches principales

- Échantillonnage
- Fonctions d'oubli par agrégation

Combinaison des approches

Oubli par échantillonnage

Opérateurs d'échantillonnage dans les BD

Échantillonnage de plusieurs populations ?

Mise à jour des échantillons

Méta-données à conserver (cube)

Fonctions d'oubli par agrégation

Langage de spécification de l'oubli

Agrégations successives avec l'ancienneté

Mécanisation de l'oubli

Fonctions d'oubli par agrégation (2)

Id_client	Nom	Ville	Dept	Région	Sexe	Age	Revenu
125	Dupont	Issy	92	Paris	M	40	25000
….	….	….	….	….	….	….	….

SUMMARY TABLE Clients {

UPDATESTEP = MONTH (15:19:30:00);

DISCRETISE (Age) = Discretise_Age : (]0, 25['Jeune' ; [25, 60['Adulte' ; [60, 90['Agé' ;)

HIERARCHY (Géographie) : Ville → Dept → Région ;

TO – 30 DAY : DETAIL ;

TO – 3 MONTH : SUM (Revenu) BY Ville, Sexe, Discretise_Age, DAY ;

TO – 1 YEAR : SUM (Revenu) BY Dept, Discretise_Age, MONTH;

TO – 10 YEAR : SUM (Revenu) BY Région, YEAR ;

TO – 15 YEAR : SUM (Revenu) BY YEAR;

}

END SUMMARY ;

Plan

Positionnement

Oubli des données dans les DW

Résumés de données pour la fouille

Environnements de gestion de la fouille

Résumés de données pour la fouille

Les méthodes d'analyse n'ont pas besoin des données détaillées

Ex : ACP avec matrice des corrélations

Quels résumés suffisent à quelles méthodes ?

- Échantillonnage
- Classification automatique (individus, variables)
- Extraction de règles, corrélations, propriétés, …

Etude de la robustesse des analyses vis à vis des résumés

Augmente l'interopérabilité des BD

IO-Net

Plan

Positionnement

Oubli des données dans les DW

Résumés de données pour la fouille

Environnements de gestion de la fouille

Environnements de gestion de la fouille

Quelques pistes possibles à creuser

- Mémorisation/historisation des analyses effectuées
- Capitalisation des interactions sur un tableau de données
- Evolution des résultats de la fouille (exploratoire/décisionnel)
- Couplage avec les logiciels collaboratifs
- Gestion de la confidentialité
- …