Des entrepôts à la fouille de données

Jean-Marc Petit
INSA de Lyon

jmpetit@liris.cnrs.fr

Laboratoire d'InfoRmatique en Image et Systèmes d'information

LIRIS UMR 5205 CNRS/**INSA de Lyon**/Université Claude Bernard Lyon 1/Université Lumière Lyon 2/Ecole Centrale de Lyon
Université Claude Bernard Lyon 1, bâtiment Nautibus
43, boulevard du 11 novembre 1918 — F-69621 Villeurbanne cedex
http://liris.cnrs.fr/

Plan du cours

- Fouille de données : une vision d'ensemble
- Un petit focus sur le passage à l'échelle
- Problèmes d'énumération des motifs intéressants
  - Cadre formel
  - Exemples
  - Algorithmes d'énumération
- Conclusions

que ce cours est/n'est pas

- Est :
  - Positionnement de la fouille de données
  - Un aperçu des enjeux et des techniques
- N'est pas :
  - Un descriptif de méthodes/algorithmes
  - Un aperçu du fonctionnement des systèmes de fouille de données

Bibliographie

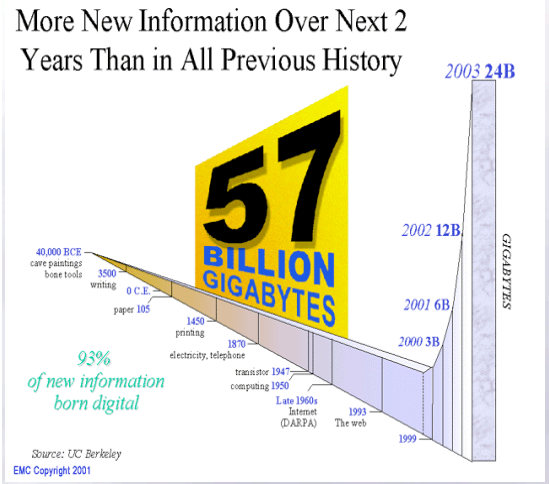
- Principale référence
  - « Data Mining: Concepts and Techniques » par Jiawei Han et Micheline Kamber, Morgan Kaufmann Publishers, 550 pages, 2004
- Conférences scientifiques
  - ACM KDD, IEEE ICDM, SIAM DM
  - + conférence DB et Apprentissage
- Sites web dédiés à la fouille de données
  - www.kddnuggets.com

duction à la fouille de données

« Nécessité est mère d'invention »

- Constat : déluge de données
  - dans des entrepôts de données (ou data warehouses (DW) ou data marts)
  - dans des bases de données (BD)
  - dans des fichiers
- Pénurie de connaissances sur ces données
- Les techniques de fouille de données peuvent être une solution

le de données : au delà des ED



Exemples :

- Téléphone AT&T : BD d'appel téléphonique 20TB, suivi des appels sans fil
- Grande consommation Wal-Mart: BD 70TB
- WEB crawl : 200M de pages et 2000M liens, 7 millards de clics par jour

ension des requêtes SQL

- Suite logique des bases de données
  - requêtes SQL classiques
    - idée simple : accéder efficacement aux données pour faire des analyses
  - requêtes OLAP (OnLine Analytical Processing)
    - Idée simple : améliorer l'analyse des données par agrégations « complexes »
  - Fouille de données (data mining)
    - aller au-delà : construction de modèles d'apprentissage pour la classification, regroupement de données en classes, …

D vs entrepôt et fouille

- construit différemment des BD (en général)
  - Intégration de sources de données (multiples, hétérogènes, BD relationnelles, fichiers)
  - Nécessite de nettoyer et d'intégrer ces données (normaliser, conventions de nommage, mesures utilisées, conversion des données)
- construit pour faciliter l'aide à la décision
  - exclu des données jugées non pertinentes par rapport à ces objectifs
- Technologie utilisé ?
  - Souvent basée sur les SGBDR !

st ce que la fouille de données ?

- Mauvais noms, analogie avec la recherche d'or
- Définition informelle : *"Extraction of interesting (non-trivial, implicit, previously unknown and potentially useful) information or patterns from data in large databases"*
- Synonymes :
  - Data mining, KDD pour knowledge discovery in databases, business intelligence, etc.
- Ce que n'est pas la fouille de données ?
  - Systèmes experts
  - Petits programmes statistiques/d'apprentissage
  - Bases de données déductives

Fouille de données

rincipales techniques de data mining

- Les techniques "classiques"
  - Association
    - corrélation et causalité
  - Classification et prédiction supervisée
    - arbre de décision,
    - réseaux bayésiens,
    - réseaux de neurones,
    - régression logistique, linéaire, non linéaire
  - Segmentation (Clustering) non supervisée
    - Analyse des cas particuliers (outliers)
- Et beaucoup d'autres …

M : un domaine pluridisciplinaire

- Analyse de données, Statistiques, Mathématiques
- Bases de données
  - Accès à des données volumineuses, entrepôts de données
  - Algorithmique en mémoire externe
- Intelligence Artificielle
  - Techniques d'apprentissage "Machine learning"
- Autres domaines :
  - Recherche opérationnelle, optimisation combinatoire
  - Imagerie
  - Visualisation
- Nouveau domaine à part entière ?

incipales applications

- A qui sert le data mining ?
  - à l'homme, e.g. pour faciliter la prise de décision
  - à des systèmes, e.g. pour s'adapter automatiquement en fonction des données collectées (e.g. systèmes d'exploitation, bases de données).
- Exemples pour l'aide à la décision
  - Analyse de marchés
    - cible marketing,
    - gestion clientèle,
    - analyse du panier de la ménagère,
    - segmentation de marché
  - Analyse de risques
    - prévision, contrôle qualité,
    - détection de fraudes, ...
- Domaines d'application : commerce, bio-informatique, WWW ...

emples du monde réel

- Les transactions dans les grandes surfaces
  - Comment mieux disposer les articles dans les rayons pour favoriser les ventes ?
  - Comment concevoir un catalogue ? une page Web ?
- Les achats sur Amazon.com
- Les transactions aux péages autoroutes
  - Quels sont les motifs « fréquents » d'utilisation du réseau autoroutier ?

éparation des données

- Nettoyage des données
  - valeurs manquantes,
  - valeurs aberrantes,
  - inconsistances
- Intégration des données
  - fichiers, BD, entrepôts
- Transformation des données
  - Normalisation
  - Agrégation
  - Réduction des données
  - Représentation plus "petites" des données sans altérer la qualité des résultats analytiques
  - Discrétisation des données

Etape critique et laborieuse qui conditionne la suite du processus

é de la connaissance extraite ?

- Problème du volume de la connaissance extraite
  - Par ex, plusieurs millions de règles d'association
- Mesure d'intérêt de la connaissance extraite
  - Mesure objective
    - basé sur des statistiques
      - *support, confiance d'une règle*
    - Mesure subjective
      - *basé sur l'utilisateur*
- Quand est ce que la connaissance est intéressante ?
  - Quand elle est inattendue
  - Quand elle sert à résoudre le problème de l'analyste

Problème majeur en pratique, pas de solutions à priori

s systèmes DM (SDM)

- Les systèmes supportant le Data Mining
  - Analogie avec les SGBD
- Industrie
  - Côté statistiques
    - SAS : Entreprise Miner
    - SPSS avec Clementine
  - Côté bases de données
    - Microsoft : OLE DB for DM
    - IBM : Intelligent Miner
    - Oracle : package ad-hoc, SQL-like data mining
- Académie
  - DMQL (J. Han, USA)
  - Logiciel libre WEKA (Nouvelle Zélande)

action des SDM avec les SGBD

- pas de couplage
  - déchargement d'une partie de la BD dans un fichier
- couplage faible
  - récupérer les données ligne à ligne avec SQL et curseurs via un LP
- couplage semi-fort
  - quelques primitives de DM intégrés dans un SGBD
- couplage fort
  - intégration totale des techniques de DM dans un SGBD
- Positionnement des outils dans ce canevas ?

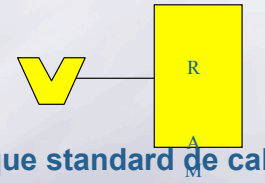
Plan du cours

- Fouille de données : une vision d'ensemble
- Un petit focus sur le passage à l'échelle
- Problèmes d'énumération des motifs intéressants
  - un cadre formel
  - Exemples
  - Algorithmes d'énumération
  - Structure de données pour la gestion d'ensemble d'ensemble
- Conclusions

assage à l'échelle

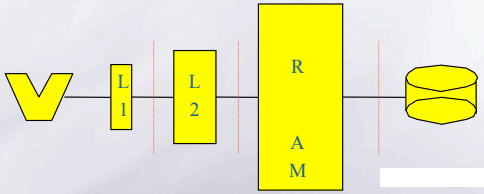
- Fouille de données : activité qui permet de découvrir des connaissances dans les données volumineuses
- Verrous ?
  - passage à l'échelle des techniques développées en analyse de données, machine learning, statistiques, optimisation combinatoire, …
- Passage à l'échelle ?
- Quel verrou technologique sous jacent ?

dèle de mémoire RAM



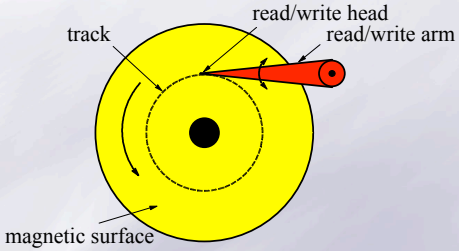
- Modèle théorique standard de calcul
  - Très simple
  - Mémoire virtuelle infini
  - Coût d'accès en temps constant
- Modèle crutial pour le succès de l'informatique

émoire hiérarchique



- Machines modernes
  - Plusieurs niveaux de mémoire
- Niveaux loin du processeur : plus grands et plus lents
- Mouvement des données entre les niveaux
  - Utilisation de "bloc" ou "page"
  - De plus en plus grand

E/S lente



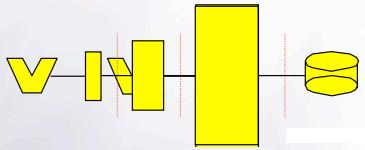
Accès disque ~ 10^6 plus lent que accès RAM

"The difference in speed between modern CPU and disk technologies is analogous to the difference in speed in sharpening a pencil using a sharpener on one's desk or by taking an airplane to the other side of the world and using a sharpener on someone else's desk." (D. Comer)

Important de stocker et d'accéder aux données de façon à tirer profit des blocs (contiguité)

blème de passage à l'échelle

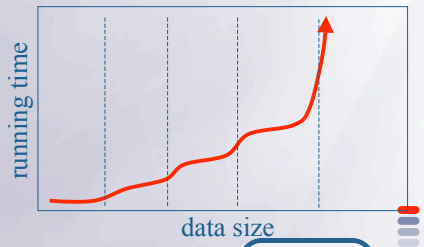
Programmes développés dans le modèle RAM
Fonctionne sur des grands jeux de données
OS déplace les blocs pour nous entre les niveaux !



- **OS récents utilise des stratégies de pagination et de "prefetching" sophistiquées**
  - **Inefficace si le programme fait des accès aléatoires**

⇓

problèmes de passage à l'échelle (scalability)



Retour sur le data mining

- Définition personnelle du contour des techniques de data mining

  « Toutes les propositions qui adressent le problème du passage à l'échelle pour leur technique »

Plan du cours

- Fouille de données : une vision d'ensemble
- Un petit focus sur le passage à l'échelle
- Problèmes d'énumération des motifs intéressants
  - Cadre formel
  - Exemples
  - Algorithmes d'énumération
- Conclusions

Théme large

- Motifs fréquents pour les règles d'association
  - Apriori [Sigmod 94] et ses innombrables améliorations
    - Souvent au niveau « structure de données »
  - Nombreuses applications
    - Requiert une discrétisation dans {0,1} des données
- Arbres fréquents dans des forêts ☺
  - données semi structurées
- Sous-séquences fréquentes dans une séquence
- Contraintes dans les bases de données
  - Clés via les dépendances fonctionnelles
  - Clés étrangères via les dépendances d'inclusion

Objectifs

- Identifier une ou des classes de problèmes que l'on pourra résoudre efficacement
- Tentative de définir un cadre commun
  - À partir des travaux de Mannila et Toivonen [DMKD'97, ACM TODS'05]
- La plupart des solutions proposées pour ces problèmes sont spécifiques

Cadre théorique

- Focus sur les problèmes d'énumération de motifs intéressants dans les grandes bases de données
- Définition des principaux termes
  - Bases de données d
  - Motifs
  - Motifs intéressants dans d : notion de prédicat
  - Relation d'ordre entre motifs
- Propriété du prédicat
- Cadre ensembliste

emple de jeux de données

- BD relationnelle
- Relation binaire ou BD de transactions
- Fichiers textes, documents structurés
- Données semi-structurés

Pas vraiment d'hypothèses fortes, doit décrire les données et les moyens pour y accéder

mples de motifs / langages

- Sous ensembles d'un ensemble
- Séquences
- Arbres
- Graphes
- Dépendances fonctionnelles
- Dépendance d'inclusion
- …

Forme des motifs qui nous intéressent, notion syntaxique

xemples de prédicats

- X est fréquent dans une BDT
- X->Y est satisfaite dans une relation
- Le sous arbre X apparait plus de n fois dans la collection d'arbres
- Formule mathématique

➡ Permet de définir formellement la notion de « motif intéressant »

Sens des motifs qui nous intéressent, notion sémantique

texte du travail : Notations

- Un modèle pour la découverte de connaissance
  - [Mannila & Toivonen, DMKD, 1997]
  - Cadre théorique, nombreuses applications
    - Motifs fréquents, séquentiels, DF, DI, …
- Cadre :
  - Etant donnés
    - Une base de données d
    - Un langage fini L (ou un ensemble de motifs)
    - Un prédicat Q définit sur d et un motif X de L
  - Trouver

$$Th(L,d,Q) = \{X \in L \mid Q(X,d) = VRAI\}$$

Relation d'ordre entre motifs

- Permet de structurer l'espace de recherche
  - si les motifs sont des ensembles, on retrouve l'inclusion ensembliste
  - Sinon, cela dépend des motifs
- Permet aussi de rechercher si le problème a de bonnes propriétés, typiquement la monotonie
- Permet d'éviter l'énumération « brute force » de tous les motifs intéressants

Notations (suite)

- Si de plus on suppose :
  - Un ordre partiel ≤ entre les motifs (relation de spécialisation/généralisation)
  - Q (anti-)monotone par rapport à ≤ :
    ∀ X,Y ∈ L tq X ≤ Y, Q(Y,d) vrai ⇨ Q(X,d) vrai
- Alors le problème peut se ramener à la recherche de :

$$MTh(L,d,Q) = \{X \in Th(L,d,Q) \mid \nexists Y \in Th(L,d,Q), X \leq Y, Y \neq X\}$$

Problèmes « représentables par des ensembles »

- Etude de plongement dans un treillis booléen
- Idées de base :
  - Faire correspondre à chaque motif un certain sous ensemble d'un ensemble
  - Préserver l'ordre partiel

Définit ainsi toute une classe de problèmes, qui d'un point de vue théorique, sont « équivalents » à un isomorphisme près.

Opportunité pour faire de l'optimisation de requêtes!

Notations (fin)

- Si de plus on suppose :
  - L est « représentable » par un ensemble si il existe un ensemble E et une fonction f : L→2^E tel que
    - f soit bijective
    - X ≤ Y ⇔ f(X) ⊆ f(Y)
- Alors, le problème initial peut se ramener à :

$$Bd^+(I) = \{X \in Th(L,d,Q) \mid \forall Y \supset X, Q(Y,d) = FAUX\}$$

$$Bd^-(I) = \{X \notin Th(L,d,Q) \mid \forall Y \subset X, Q(Y,d) = VRAI\}$$

Notions de bordures

- Soit I un ensemble de motifs intéressants de E
- La bordure positive de I (les motifs les plus spécifiques de I), noté $Bd^+(I)$, est :

$$Bd^+(I) = \{X \in I \mid \forall Y \supset X, Y \notin I\}$$

- La bordure négative de I (les motifs les plus généraux qui ne sont pas dans I), noté $Bd^-(I)$, est :

$$Bd^-(I) = \{X \in (2^E \setminus I) \mid \forall Y \subset X, Y \in I\}$$

- Il existe un lien fort entre les bordures positive et négative
  - Transformation de l'un à l'autre et vice et vers ça
- Objet combinatoire connu :
  - transversaux minimaux d'un hypergraphe

Déclaration des problèmes

- Grande unité de définition des problèmes
  - Même si une grande partie de la difficulté réside dans cette formulation
  - "un problème bien posé est à moitié résolu" ☺
- Trois formes
  - Enumération des éléments I de E
  - Enumération des éléments de $Bd^-(I)$
  - Enumération des éléments de $Bd^+(I)$

Ces problèmes « ensemblistes »

- A la base d'algorithmes de parcours de l'espace de recherche
- Structure de données pour manipuler de grandes collections d'ensembles d'ensemble.

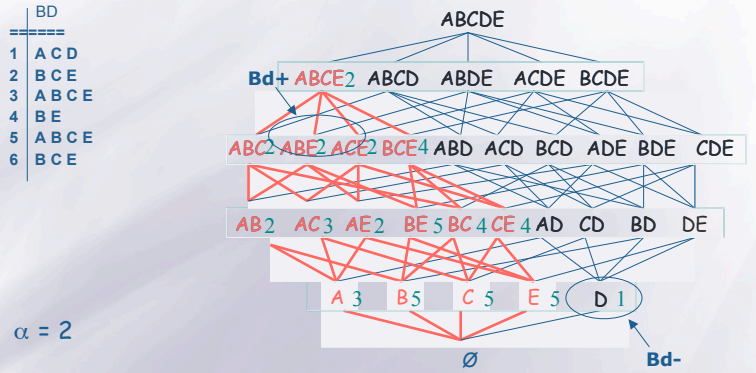
Cadre restrictif néanmoins …

Possibilité d'appliquer des techniques d'optimisation de requêtes en BD

lication aux motifs fréquents

- Ensemble de propriétés P
- Une ligne t est un sous-ensemble de P, i.e. t ⊆P
- bd : multi-ensemble de lignes
- Q(X,bd,α) : « être fréquent » par rapport à un seuil α, i.e. |{t | t∈BD et X⊆t}|≥ α
  - Si Q(X,bd,α) = VRAI alors X est un **motif fréquent**
- Propriété :
  - Q est anti-monotone par rapport à ⊆
- Fonction f : fonction identité

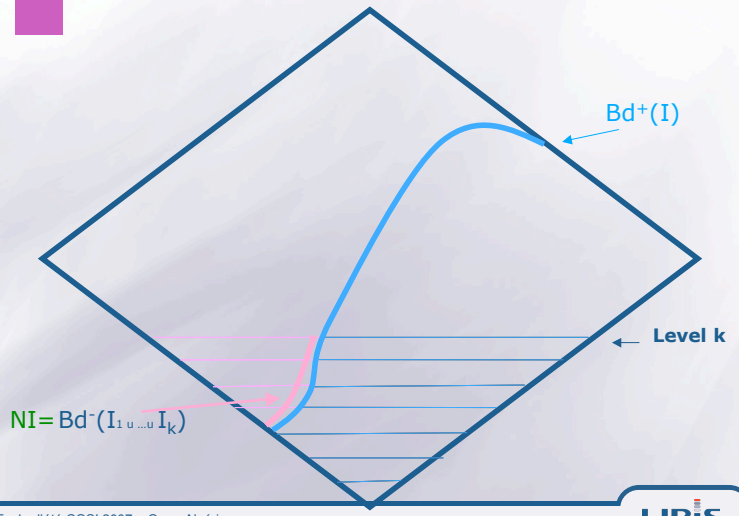
Exemple parcours par niveaux (Apriori)



orithmes d'énumération

- Par niveau: connu en fouille de données sous le nom d'Apriori (Agrawal et al 1993, ACM Sigmod)
- Par dualisation (transversaux minimaux)
  - moins connu mais très élégant
  - Idée : passer d'une bordure à l'autre en évitant l'explosion combinatoire des approches par niveaux quand de grands éléments existent

Exemple pour bd+



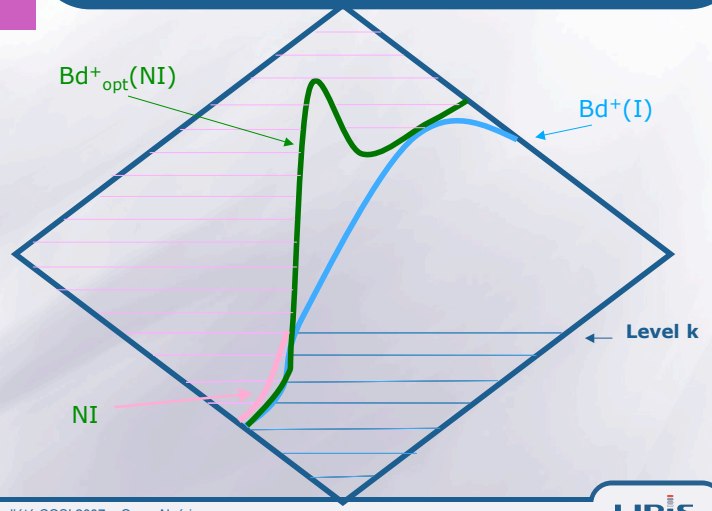
solution est peut etre là …

- ***Bordure positive optimiste*** *= plus "grands" éléments pas disqualifiés par NI*
- **On a donc :**

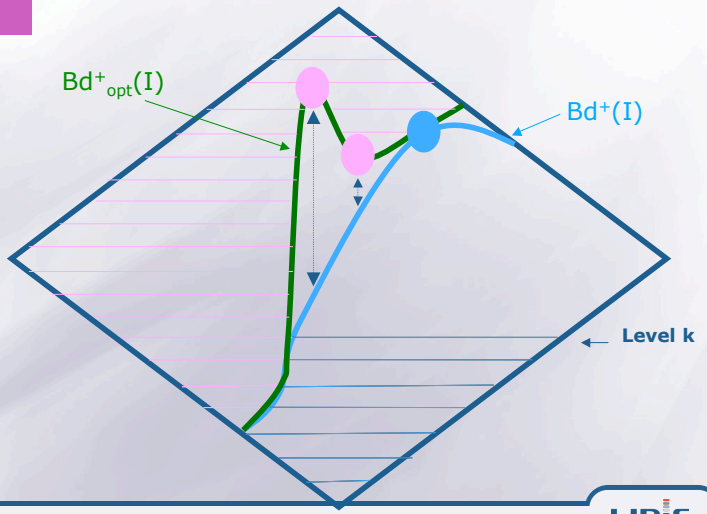
$$Bd^{+}_{opt}(\mathrm{I}) = \overline{\mathrm{MinTr}(NI)}$$

- **Notion de "saut" dans l'espace de recherche**

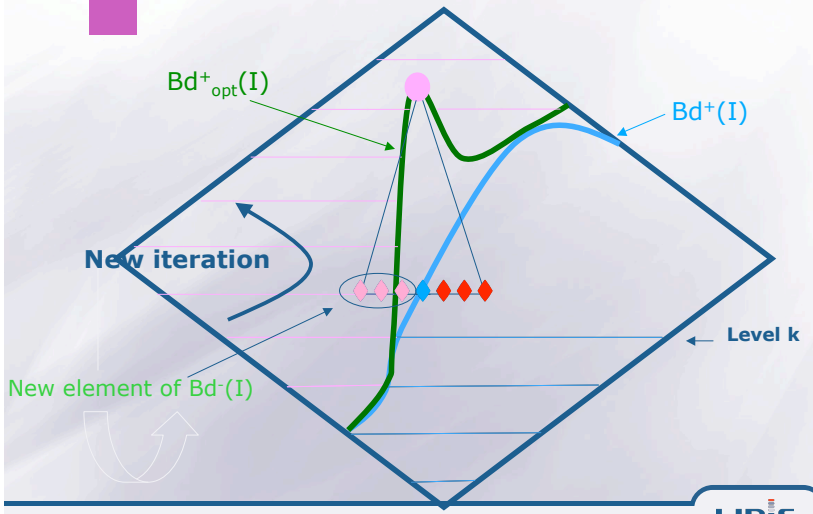
Intuition



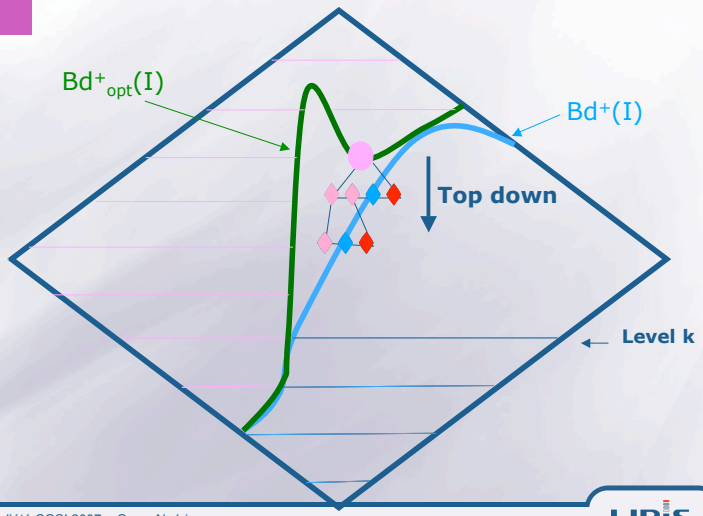
Estimation de l'erreur



Stratégies adaptatives 1/2



Stratégies adaptatives 2/2



Structure de données

- Quelles structures de données pour gérer des ensembles d'ensemble très volumineux ?
- Dépend des opérations à faire sur ces structures
- Pour des algorithmes type Apriori
  - Développement de structure de « trie » développée en recherche d'information (information retrieval)
    - Arbre de préfixe ou dictionnaire
    - Recherche d'un élément en temps constant
  - Et variantes … Patricia trees,
- Beaucoup de travaux dans la littérature

Conclusion et perspectives

- Fouille de données :
  - guidée par les applications
  - à la croisée de nombreuses branches de l'informatique
    - Constitue ses forces et ses faiblesses
- Beaucoup de business, réel besoin des entreprises
- Difficulté de la validation avec des experts

- A terme : nécessité d'intégration et de rapprochement avec les SGBD => passage à l'échelle