Extraction d'outliers dans des cubes de données : une aide à la navigation

Marc Plantevit, Anne Laurent, Maguelonne Teisseire

LIRMM, Université Montpellier II, France

EDA'07
7-8 Juin 2007
Poitiers

Plan

OLAP et Utilisateurs

- Utilisateurs sont désormais des décideurs
- Prendre les décisions les mieux adaptées le plus vite possible
- « maximiser les gains »
- Naviguent dans le cube de données à l'aide des opérateurs

OLAP Mining

- Fournir aux utilisateurs des connaissances
- Mieux Appréhender les données
- Améliorer le processus d'aide à la décision

Extraction de motifs

- Permet d'avoir des comportements généraux
- Ne permet pas d'identifier directement les atypicités.

Stratégie Commerciale

- Orientations définies par un décideur
- Motifs : $X\%$ des individus suivent la politique commerciale définie.
- Quid des anomalies ?

Anomalie

- Règle des 10%-90%
- Identifier les événements anormaux.
- Identifier les causes de ces événements anormaux.

Proposition

Objectifs : Voir ce qui se cache derrière l'agrégation

- Si une séquence agrégée est « outlier », quelles en sont les causes ?
- Une séquence particulière à un niveau plus fin ?
- Un ensemble de séquences à des niveaux plus fins ?

Exemple

Les ventes *anormales* dans la région X sont en partie dues aux ventes anormales des pôles x et y.

Proposition

- Rechercher des séquences outliers à un niveau d'agrégation donné.
- Détecter les causes d'atypicité à des niveaux plus fins.
- Proposer des chemins de navigation dans le cube de données.

Plan

Contexte général

Outliers [Hawkins, 1980]

Des objets tellement différents des autres observations qu'ils en sont suspicieux et ont dû être générés par un autre mécanisme.

- règle des 3σ

En français ?

Anomalies, fraudes, exceptions, intrusions, erreurs . . .

- Approches paramétriques
- Approches "Data Mining"

De nombreuses approches data mining

Recherche d'objets outliers

Knorr et Ng, 1997 : outlier basé sur la distance.

Breuning et al., 2000 : outlier local.

Fan et al., 2006 : outlier basé sur la résolution.

Séquences outliers

Sun et al., 2006 : Construction d'un arbre probabiliste des suffixes pour approximer des probabilités.

Etat de l'Art

OLAP

Knorr et Ng

Les outliers basés sur la distance dans un contexte OLAP - Un *outlier : une cellule*.

Sarawagi

- Exploration guidée sur la découverte.
- cellule *"exception"* si la mesure est différente de la valeur attendue (2.5σ)
- "version olap" de la règle des 3σ.

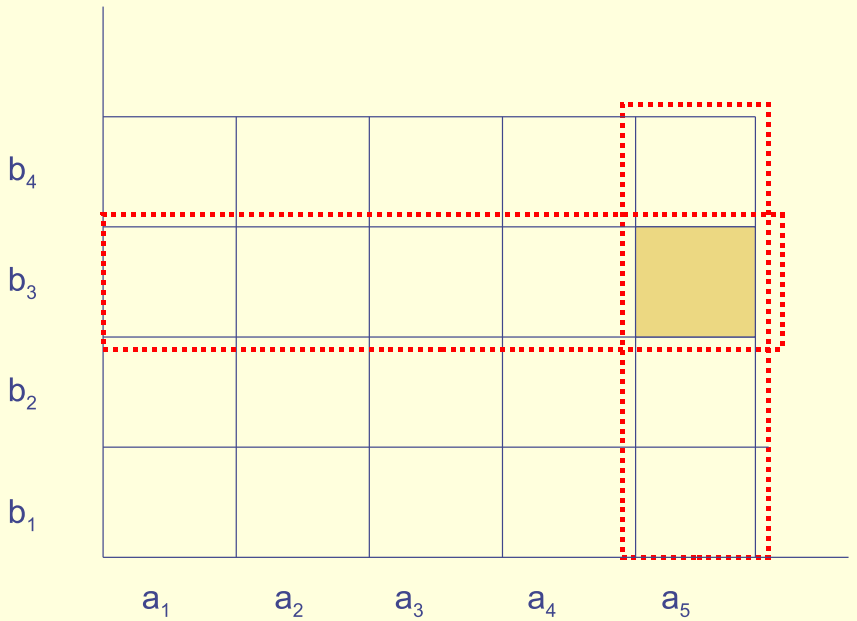
Lin et Brown

- Recherche les cellules avec des mesures extrêmes
- Associe les positions des cellules outliers.
- Corrélations entre événements.

Parents

La cellule (a_5, b_3) a 2 parents :

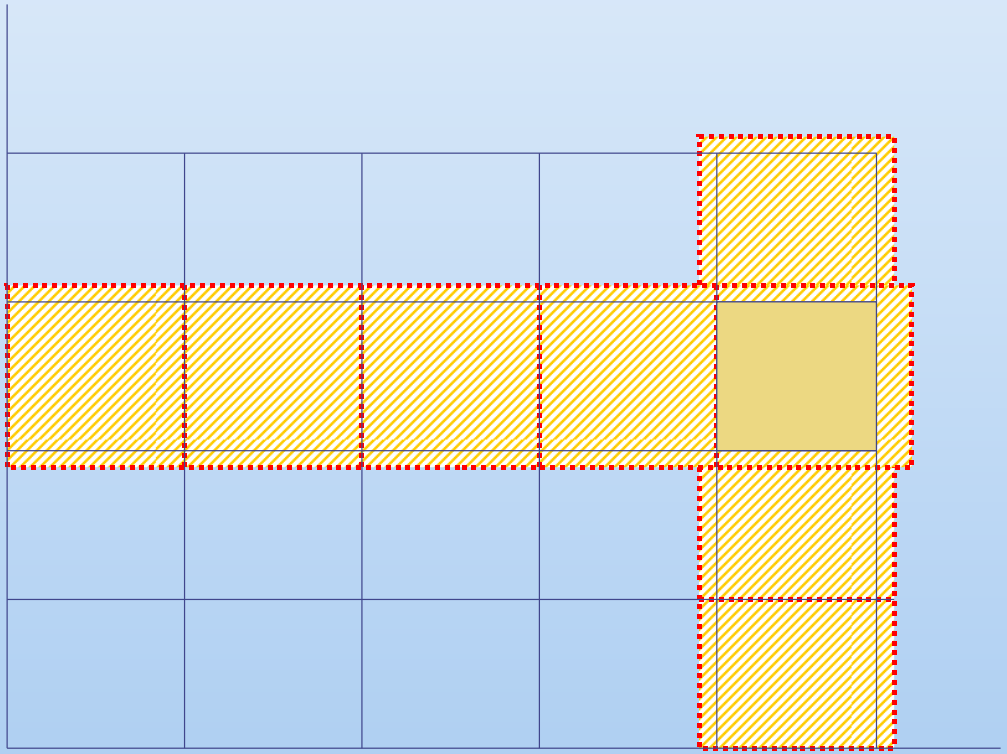
- $(a_5, *)$
- $(*, b_3$



Voisinage

Voisinage de *x*

L'ensemble des cellules qui partagent un parent de la cellule *x*.



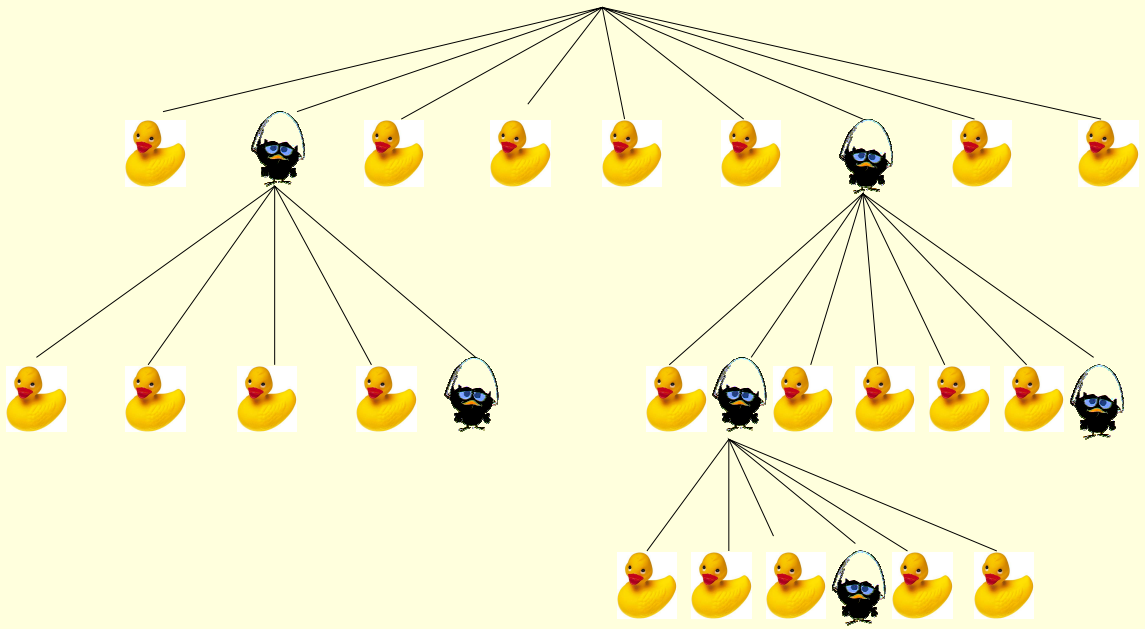
Limites des approches existantes

Quid des séquences ?

- Aucune notion de temporalité
- Utiliser pleinement le contexte OLAP (navigation dans les hiérarchies).

Plan

Extraire des séquences outliers à différents niveaux d'agrégation



- Notion de séquence
- Comparaison de séquences
- Mesures de distances

$$D = D_t \oplus D_A \oplus D_R \oplus D_F$$

Partition des dimensions

Pour tout cube défini sur les dimensions D, on considère une partition de D en quatre sous-ensembles notés respectivement :

- D_t pour la ou les dimensions temporelles
- D_A pour les dimensions dites d'*analyse*
- D_R pour les dimensions dites de *référence*
- D_F pour les dimensions oubliées.

Bloc

On appelle bloc B un ensemble de cellules dont les positions sur D_R et D_T sont fixes. B est l'ensemble des n-uplets prenant leurs valeurs sur D_A :

$$B = \{\langle({d_{i_1}}^1, \ldots, {d_{i_m}}^1), {\mu}^1\rangle, \ldots, \langle({d_{i_1}}^p, \ldots, {d_{i_m}}^p), {\mu}^p\rangle\}$$

On notera $B = \{c_1, \ldots, c_p\}$.

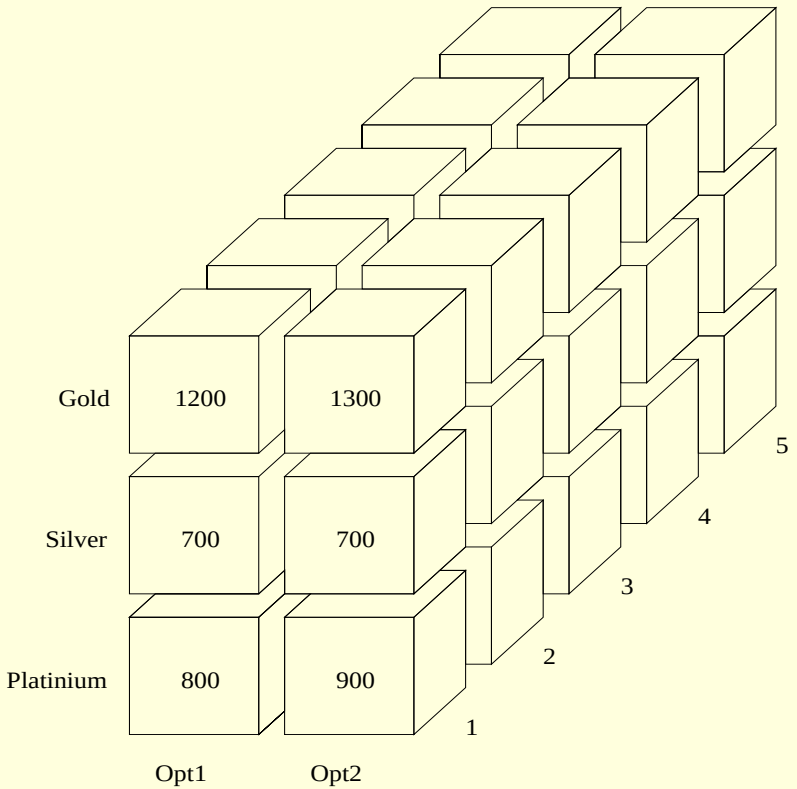
Séquence

On appelle séquence, une liste ordonnée non vide de blocs de la forme :

$$\varsigma = \langle B_1, \ldots, B_l \rangle$$

Chaque valeur sur D_R identifie une séquence de blocs.

Séquence pour GEO=Sud



Distance entre deux séquences

Distance d'édition

- nombres d'opération (insertion, suppression, déplacement) nécessaires pour *transformer* une séquence en une autre ;
- la plus utilisée ;

Limites

- $dist(s_1, s_2)$ faible (1 opération) ;
- Séquences en *opposition de phase* ;
- *Une antenne régionale qui suit le comportement national mais avec 6 mois de retard.*

Cellules comparables

Deux cellules $c_1 = \langle (d_1, \ldots, d_n), \mu \rangle$ et $c_2 = \langle (d'_1, \ldots, d'_n), \mu' \rangle$ sont comparables si et seulement si $c_1.D_A = c_2.D_A$.

Exemples

- $c_1 = \langle (Ouest, 1, Gold, opt1), 1200 \rangle$ et $c_2 = \langle (RA, 1, Gold, opt1), 900 \rangle$ sont comparables
- c_1 et $c_3 = \langle (Sud, 1, Silver, opt2), 600 \rangle$ sont incomparables étant donné que leurs restrictions sur D_A sont différentes.

Algo

Calcul de la distance entre deux blocs

- Construction de vecteurs de mesure.
- Chaque dimension sur les deux vecteurs, correspond à des valeurs de mesures entre deux cellules comparables.
- Représentation permet d'appliquer plusieurs mesures de distance ou similarité.

Ex :

- $b_1 = \{\langle (a,b), 100\rangle, \langle (a,c), 150\rangle\}$
- $b_2 = \{\langle (a,b), 35\rangle, \langle (a,c), 78\rangle\}$
- $v_1 = (100, 150), v_2 = (35, 78)$

Distance entre deux séquences

Distance

- $s_1 = \langle b_1, b_2, \ldots, b_k \rangle$ et $s_2 = \langle b'_1, b'_2, \ldots, b'_k \rangle$ deux séquences multidimensionnelles
- *dist* une mesure de distance
- *Op* un opérateur d'agrégation

La distance entre s_1 et s_2 se définit de la façon suivante :

$$d(s_1, s_2) = Op(dist(b_j, b'_j)) \; pour \; j = 1 \ldots k$$

Mesures

Mesures Utilisées

Distance de Manhattan :

$$Man(v_1, v_2) = \sum_{k=0}^{m} |v_{1_k} - v_{2_k}|$$

Distance euclidienne :

$$Euclid(v_1, v_2) = \sqrt{\sum_{k=0}^{m} (v_{1_k} - v_{2_k})^2}$$

Cosinus :

$$cos(v_1, v_2) = \frac{v_1 . v_2}{||v_1|| ||v_2||} = \frac{\sum_{k=0}^{m} (v_{1_k} v_{2_k})}{\sqrt{\sum_{k=0}^{m} v_{1_k}^2 \sum_{k=0}^{m} v_{2_k}^2}}$$

Opérateurs d'agrégation

- moyenne
- médiane
- min/max

Sequence_Id	1	2	...	l
1	1	$sim(1,2)$	...	$sim(1,l)$
2	•	1	...	$sim(2,l)$
...	•	•	1	...
l	•	•	•	1

(a) Matrice de similarités

Sequence_Id	1	2	...	l
1	0	$d(1,2)$	...	$d(1,l)$
2	•	0	...	$d2,l)$
...	•	•	0	...
l	•	•	•	0

(b) Matrice de distances

FIG.: Comparaison d'une séquence par rapport aux autres

$$d(s_\alpha, S) = \frac{\sum_{i=1}^{i<\alpha} d(s_\alpha, s_i) + \sum_{j=\alpha+1}^{|s|} d(s_j, s_\alpha)}{|S| - 1}$$

Top *n* outliers

top *n*

- Difficile de déterminer un seuil δ pour déterminer outlier ;
- *n* permet de déterminer le degré de l'arbre de navigation ;

top *n* outlier

Une séquence s_α est un top *n* outlier s'il n'existe pas plus de $n-1$ séquences telles que $d(s_i, C_{D_R}) > d(s_\alpha, C_{D_R})$

Algorithme

Data : C_{v_R} Cube de données, n entier, L ensemble, $dist$ une mesure de distance

Result : Séquences outliers à chaque niveau de granularité bis

Fonctionnement général

begin
  Calculer la matrice de distance
  foreach *séquence* $S_{v_{R_i}} \in C_{v_R}$ *top n outlier* **do**
    $add(v_{R_i}, L)$
    if v_{R_i} *is not leaf* $\wedge$
    $Normal(S_{v_{R_i}})$ *is not top n outlier in* $C_{roll\ up(v_R)}$
    then
      $RechTopn(C_{DrillDown(v_{R_i})}, n, L, dist)$
  return L
end

Plan

Protocole expérimental

Cube de données

- 35000 cellules
- opérateur d'agrégation : la somme
- $|D_R| = |D_T| = 1$
- $|D_A| = 4$

Mesures : Manhattan, Euclidienne, Cosinus

Opérateur d'agrégation : moyenne, médiane, min/max.

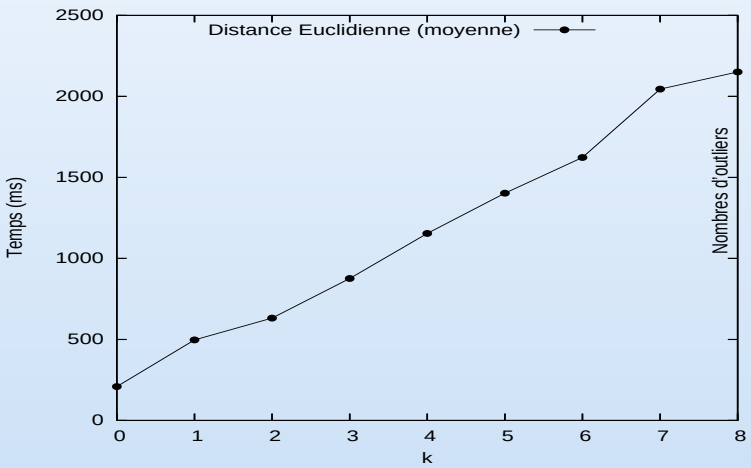


FIG.: Temps d'exécution en fonction du nombre de top *k* outliers recherchés

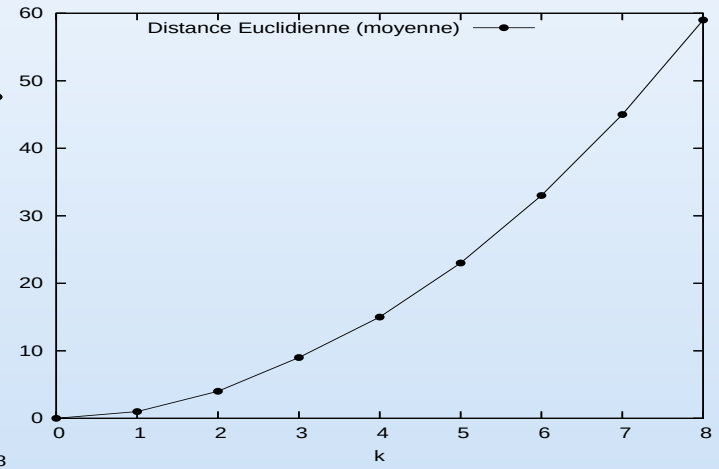


FIG.: Nombre d'outliers extraits en fonction du nombre de top *k* outliers recherchés.

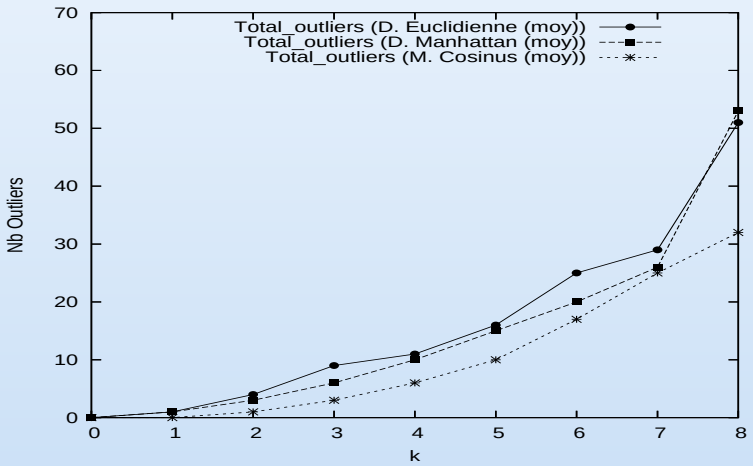


FIG.: Nombre d'outliers "totalement outliers" en fonction du nombre de top *k* outliers recherchés(moy.)

Plan

Conclusions

Une première approche

- Séquences
- Contexte OLAP
- Arbre de navigation de degré *n*

Perspectives

Toutes les séquences outliers

- Une approche *bi-directionnelle*

Séquences outliers :

- Quelle partie précisément ?
- *« Actions X et Y entre juin et juillet très atypiques. »*
- Drill Down sur les dimensions d'analyse ?
- Mettre une dimension d'analyse dans l'ensemble des dimensions de référence ? Laquelle ?

Perspectives

Inégalités triangulaires

- Pas de cellule vide : OK (toutes les cellules sont comparables)
- Cellules vides et $\emptyset \neq 0$: AÏE

- B_2 et B_3 ne partagent aucune cellule comparable
- B_1 et B_3 ne partagent aucune cellule comparable
- B_1 et B_2 partagent des cellules.
- alors on n'a pas :
  $d(B_1, B_2) \leq d(B_1, B_3) + d(B_2 + B_3)$
- $\rightarrow$ *malus* $\nexists$ cellules comparables.