Requêtes skyline en présence d'exceptions

Hélène Jaudoin, Olivier Pivert, Daniel Rocacher

Université de Rennes 1 – Irisa
{jaudoin@enssat.fr, pivert@enssat.fr, rocacher@enssat.fr}

Résumé. Dans cet article, nous nous intéressons à la recherche des points les plus intéressants au sens de l'ordre de Pareto, i.e., à l'évaluation de requêtes « skyline » , dans des jeux de données présentant des anomalies. Il n'est pas rare que les données, de petites annonces par exemple, soient peuplées d'erreurs ou d'exceptions qui peuvent perturber la recherche des meilleurs points car celles-ci sont susceptibles de dominer les autres points. L'approche présentée vise à calculer les requêtes skyline malgré la présence de ces exceptions, sans pour autant les écarter définitivement, et à présenter graphiquement les résultats de façon à identifier rapidement les points d'intérêt et les anomalies potentielles.

1 Introduction

In database research, the last two decades have witnessed a growing interest in preference queries on the one hand. Motivations for introducing preferences inside database queries are manifold Hadjali et al. (2008). First, it has appeared to be desirable to offer more expressive query languages that can be more faithful to what a user intends to say. Second, the introduction of preferences in queries provides a basis for rank-ordering the retrieved items, which is especially valuable in case of large sets of items satisfying a query. Third, a classical query may also have an empty set of answers, while a relaxed (and thus less restrictive) version of the query might be matched by items in the database.

Approaches to database preference queries may be classified into two categories according to their qualitative or quantitative nature Hadjali et al. (2008).

Dans la dernière, les préférences sont exprimées quantitativement grâce à une fonction de score monotone, le score global étant positivement corrélé avec les scores partiels. Dans les approches qualitatives, les préférences sont définies au travers de relations binaires. Comme ces relations peuvent être définies en termes de fonctions de score, cette famille est plus générale que la précédente.

Dans cet article, une vision qualitative est adoptée, à savoir l'approche dite Skyline, introduite dans (Börzsönyi et al. (2001)). Étant donné un ensemble de points dans l'espace, une requête skyline retrouve les points qui ne sont dominés par aucun autre au sens de l'ordre de Pareto. Ce problème correspond à la recherche des extrema dans un ensemble de vecteurs (Kung et al. (1975)). Quand le nombre de dimensions sur lesquelles les préférences sont exprimées devient grand, de nombreux tuples peuvent être incomparables. Quelques approches ont été proposées pour définir un ordre entre deux tuples incomparables dans le contexte des requêtes skyline, fondées sur :

- le nombre de tuples que chacun de ces deux tuples domine (notion de dominance k-représentative proposée dans Lin et al. (2007),
- des ordres de préférence entre les attributs : par exemple les notions de k-dominance et de k-fréquence introduites dans Chan et al. (2006a,b), ou
- la représentativité : Tao et al. (2009) redéfinissent l'approche de Lin et al. (2007) pour retourner les points du skyline les plus représentatifs possibles en présentant uniquement un représentant par cluster de points présent dans le skyline.

D'autres approches ont cherché à flexibiliser le concept de skyline selon différentes directions, voir par exemple Hadjali et al. (2011).

Ici, nous nous intéressons à un problème différent, celui de la possible présence de points *exceptionnels* dans la relation à laquelle la requête skyline est adressée. De telles exceptions peuvent correspondre à du bruit ou à la présence de points *atypiques*, ou *non représentatifs* dans la collection considérée. L'impact de tels points sur le Skyline peut évidemment être important s'ils en dominent d'autres, plus représentatifs. Deux stratégies peuvent être envisagées pour gérer les exceptions. La première consiste à éliminer les anomalies par la mise en place d'une procédure de nettoyage des données ou de contraintes de saisie. Néanmoins, il n'est pas toujours aisé de distinguer entre des points erronés et des points qui représentent simplement des cas exceptionnels. Une meilleure solution est donc de définir une approche *tolérante aux exceptions*, i.e., qui mette en avant des points représentatifs de la base de données *non dominés* par d'autres éléments *représentatifs*, tout en signalant les éventuelles exceptions. Dans cet article, nous décrivons une telle approche et interprétons la notion de représentativité à l'aide de celle de typicité (Zadeh, 1984). Nous proposons une nouvelle définition du skyline basée sur la typicité et nous montrons que celle-ci permet i) de retrouver les meilleurs compromis sans pour autant évincer les points potentiellement intéressants, quoiqu'exceptionnels, et ii) d'offrir un outil flexible pour visualiser les réponses.

La suite de ce papier est organisée comme suit. La Section 2 fournit quelques rappels à propos de l'ordre de Pareto, des requêtes Skyline, et de la notion de représentativité tout en motivant l'approche proposée. La Section 3 présente notre solution et définit le concept de skyline graduel tolérant aux exceptions. La Section 4 donne les principaux éléments de mise en œuvre de notre approche tandis que la Section 5 présente les premiers résultats obtenus sur un jeu de données réelles. Enfin, la Section 7 rappelle les principales contributions et propose des améliorations possibles de ce travail.

2 Rappel sur les requêtes Skyline et motivations

Soit $\mathcal{D} = \{D_1, \ldots, D_d\}$ un ensemble de d dimensions. Notons par $dom(D_i)$ le domaine associé à la dimension D_i. Soit $\mathcal{S} \subseteq dom(D_1) \times \ldots \times dom(D_d)$, deux points p et q de $\mathcal{S}$ et $\succ_i$ une relation d'ordre sur D_i. On dit que p *domine* q sur $\mathcal{D}$ (p est meilleur que q selon l'ordre de Pareto), denoté par $p \succ_{\mathcal{D}} q$, ssi

$$\forall i \in [1,d] \ : \ p_i \succeq_i q_i \text{ et } \exists j \in [1,d] \ : \ p_i \succ_i q_i.$$

Une requête skyline sur $\mathcal{D}$ appliquée à un ensemble de points $\mathcal{S}$, notée $\text{SKY}_{\mathcal{D}}(\mathcal{S})$, selon des relations d'ordre $\succ_i$, retourne l'ensemble de points qui ne sont dominés par aucun autre point de $\mathcal{S}$:

$$\text{SKY}_{\mathcal{D}}(\mathcal{S}) = \{p \in \mathcal{S} \mid \nexists q \in \mathcal{S} : q \succ_{\mathcal{D}} p\}$$

Selon le contexte, on peut essayer, par exemple, de maximiser ou minimiser les valeurs de $dom(D_i)$, en supposant que $dom(D_i)$ est un domaine numérique.

Pour illustrer le principe de l'approche proposée, considérons le jeu de données issu de la base Iris (Fisher (1936)), représenté sous forme graphique dans la Figure 1.

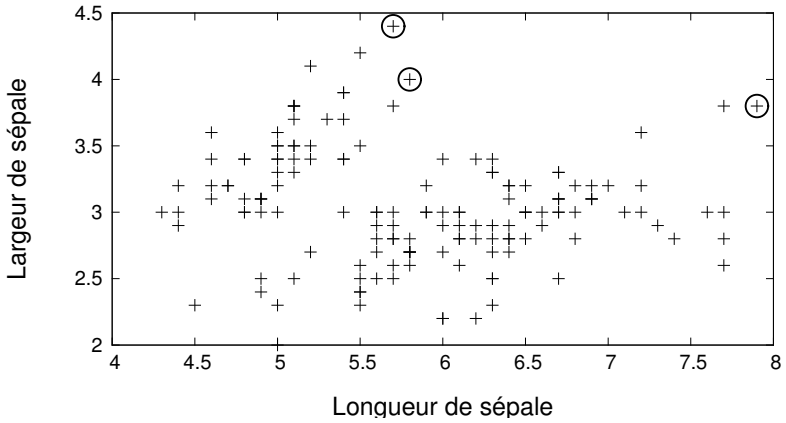

FIG. 1 – *Le jeu de données Iris*

En ordonnée, figurent les largeurs de sépale tandis qu'en abscisse apparaissent les longueurs de sépale. La requête skyline :

`select` * `from` ***iris***
`skyline of` ***sepallength*** `max`, ***sepalwidth*** `max`

recherche les points iris qui maximisent les dimensions *largeur* et *longueur* des sépales (points entourés de la Figure 1) .

Dans ce jeu de données, les points sont organisés en deux groupes qui correspondent respectivement aux intervalles sur les abscisses $[4,\ 5.5]$ et $[5.5,\ 7]$. Par définition, les points du skyline sont à la frontière de l'espace à deux dimensions décrits par les points du jeu Iris. Mais ces points sont très distants des zones décrites par les deux groupes et sont donc peu représentatifs du jeu Iris. Il pourrait être intéressant pour un utilisateur de pouvoir visualiser les points "quasi dominants", plus proches des groupes de points et donc plus représentatifs de la base. La notion de typicité introduite dans la section suivante va nous permettre de modéliser cette notion de représentativité.

2.1 Calculer un ensemble flou de valeurs typiques

La typicité d'un élément dans un ensemble indique dans quelle mesure cet élément est similaire à beaucoup d'autres points de l'ensemble. La notion de valeur floue typique d'un ensemble a été largement étudiée dans le domaine des résumés de données et dans celui du raisonnement approximatif. Zadeh (1984) définit x comme étant un élément typique d'un ensemble flou A ssi i) x a un haut degré d'appartenance à A et ii) *la plupart des éléments de* A *sont similaires à* x. Dans le cas où A est un ensemble non flou, comme ce sera le cas dans la suite, la définition devient : x est dans A et la plupart des éléments de A sont similaires à x.

Dans (Dubois et Prade (1984)), les auteurs définissent un indice de typicité basé sur la fréquence et la similarité. Dans la suite de cet article, nous adaptons leur définition comme suit. Considérons un ensemble $\mathcal{E}$ de points à deux dimensions, correspondant aux attributs *splength* (longueur du sépale) et *spwidth* (largeur du sépale). Nous dirons qu'un point est d'autant plus typique qu'il est proche de nombreux autres points. La relation de proximité sera basée sur la distance euclidienne. Soit par exemple deux points p_1 and p_2 de la base *Iris* (cf. page précédente). la distance $d(p_1,\ p_2)$ entre ces deux points est définie comme suit :

$$d(p_1,\ p_2) = \sqrt{(p_1.splength - p_2.splength)^2 + (p_1.spwidth - p_2.spwidth)^2}.$$

Nous considérons que ces deux points sont proches l'un de l'autre si $d(p_1,\ p_2) \leq \tau$ où τ est un seuil prédéfini. Pour le jeu *Iris*, ce seuil est fixé à $\tau = 0.5$. La fréquence d'un point est définie de la façon suivante :

$$F(p) = \frac{|\{p_i \in \mathcal{E},\ d(p,\ p_i) \leq \tau\}| - 1}{|\mathcal{E}|}. \quad (1)$$

Ce degré est ensuite normalisé en un degré de typicité dans $[0, 1]$:

$$typ(p) = \frac{F(p)}{\max_{p_i \in \mathcal{E}}\{F(p_i)\}}.$$

Nous utiliserons également les notations suivantes :

$$\textsc{Typ}(\mathcal{E}) = \{typ(p)/p \mid p \in \mathcal{E}\}$$

$$\textsc{Typ}_\gamma(\mathcal{E}) = \{p \mid p \in \mathcal{E} \text{ and } typ(p) \geq \gamma\}.$$

$\textsc{Typ}(\mathcal{E})$ représente l'ensemble flou des points un tant soit peu typiques de l'ensemble $\mathcal{E}$ tandis que $\textsc{Typ}_\gamma(\mathcal{E})$ rassemble les points de l'ensemble $\mathcal{E}$ dont la typicité dépasse le score γ. Un extrait du calcul de la typicité des points de la base *Iris* est présenté en Table 1.

TAB. 1 – *Extrait de la base iris et des valeurs de typicité*

Longueur	*Largeur*	*Représentativité*	*Typicité*
7.4	2.8	0.0600	0.187
7.9	3.8	0.0133	0.0417
6.4	2.8	0.253	0.792
6.3	2.8	0.287	0.896
6.1	2.6	0.253	0.792
7.7	3.0	0.0467	0.146
6.3	3.4	0.153	0.479
6.4	3.1	0.293	0.917
6.0	3.0	0.320	1.000

3 Skyline tolérant aux exceptions

Comme expliqué en introduction, notre objectif est de revisiter la définition du skyline afin de prendre en compte la typicité des points dans la base de données et ainsi d'éviter la perte des points qui seraient dominés par des exceptions ou des anomalies.

3.1 Vision booléenne

Une première idée consiste à restreindre le calcul du skyline à un sous-ensemble de $\mathcal{E}$ correspondant aux points qui sont suffisamment typiques. La définition correspondante donne :

$$\text{SKY}_{\mathcal{D}}(\text{TYP}_{\gamma}(\mathcal{S})) = \{p \in \text{TYP}_{\gamma}(\mathcal{S}) \mid \nexists q \in \text{TYP}_{\gamma}(\mathcal{S}) \text{ tel que } q \succ_{\mathcal{D}} p\} \quad (2)$$

Une telle approche réduit le coût d'évaluation du skyline puisque seuls les points typiques au degré γ sont considérés dans le calcul. Il n'est cependant pas possible avec cette définition de discriminer les points du résultat selon leur degré de typicité. En effet, le skyline obtenu est un ensemble classique (non flou). La Figure 2 illustre ce cas de figure et montre les maxima (croix entourées sur la figure) obtenus avec les points typiques à un degré ≥ 0.7 (croix simples).

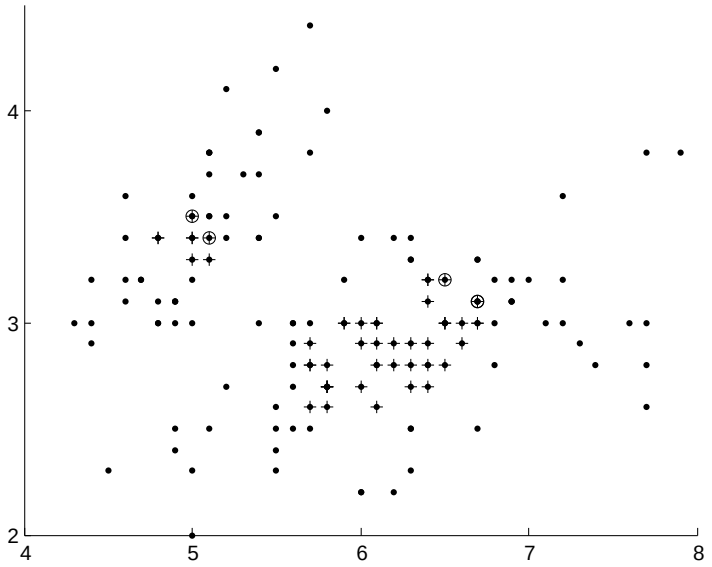

FIG. 2 – *Skyline des points de la base Iris typiques au delà du degré* 0.7.

Cette première définition a également l'inconvénient d'exclure les points atypiques qui peuvent pourtant être des réponses valides. Une définition plus prudente consiste à garder les points atypiques dans le calcul du skyline et à transformer l'équation (2) en :

$$\text{SKY}_{\mathcal{D}}(\text{TYP}_{\gamma}(\mathcal{S})) = \{p \in \mathcal{S} \mid \nexists q \in \text{TYP}_{\gamma}(\mathcal{S}) \text{ tel que } q \succ_{\mathcal{D}} p\} \quad (3)$$

La Figure 3 illustre cette nouvelle définition. Elle représente les points (points entourés) de la base *Iris* qui ne sont pas dominés par des points typiques (croix) au degré $\gamma = 0.7$ au moins.

Avec l'équation (2), les points atypiques sont éliminés, alors qu'avec l'équation (3), le skyline devient plus large et englobe les extrema atypiques. Cette approche permet de relaxer les requêtes skyline de façon à transformer la ligne en un bandeau formé des points normaux du skyline et d'éventuels substituts.

Les principaux inconvénients de cette approche sont : i) le possible grand nombre de points retournés, ii) le fait qu'on ne puisse pas différencier, parmi les points du skyline, ceux qui ne sont pas du tout dominés de ceux qui le sont fortement.

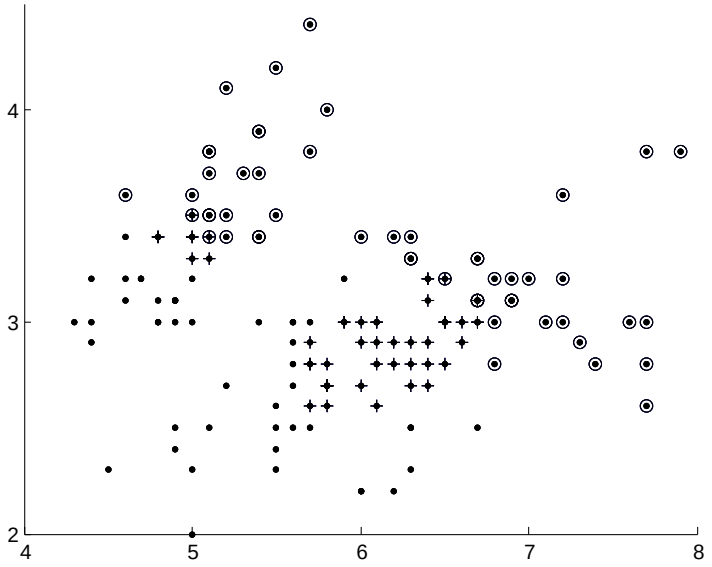

FIG. 3 – *Points non dominés par des points typiques ($\gamma = 0.7$).*

3.2 Vision graduelle

Une troisième version permet de calculer un skyline *graduel*, vu comme un ensemble flou, qui préserve la nature graduelle de la relation de typicité en entrée. Ainsi, aucun seuil (γ) n'est appliqué aux degrés de typicité. Un point appartient totalement au skyline (degré d'appartenance de 1) s'il n'est dominé par aucun autre point. Un point n'appartient pas du tout au skyline (degré d'appartenance de 0) s'il est dominé par au moins un point totalement typique. Un point dominé par des points peu typiques appartient fortement au skyline alors qu'un point dominé par des points très typiques appartient peu au skyline. L'appartenance d'un point p au skyline est donc dépendante du plus fort degré de typicité des points qui le dominent. Il en découle la définition suivante :

$$\textsc{Sky}_{\mathcal{D}}(\textsc{Typ}(\mathcal{S})) = \{\mu/p \mid p \in \mathcal{S} \wedge \mu = \min_{q \in \mathcal{S}} (\max(1 - \mu_{\textsc{Typ}}(q),\ deg(\neg(q \succ_{\mathcal{D}} p)\ ))\} \quad (4)$$

où $deg(\neg(q \succ_{\mathcal{D}} p)) = 1$ si q ne domine pas p (i.e., $(q \succ_{\mathcal{D}} p)$ est faux), 0 sinon. Ainsi, dans le cas où p est dominé par un point quelconque, son degré d'appartenance au skyline est fixé par la non typicité de ce point. L'équation (4) peut être réécrite comme suit :

$$\textsc{Sky}_{\mathcal{D}}(\textsc{Typ}(\mathcal{S})) = \{\mu/p \mid p \in \mathcal{S} \wedge \mu = 1 - \max_{q \in \mathcal{S} \mid q \succ_{\mathcal{D}} p} (\mu_{\textsc{Typ}}(q))\}. \quad (5)$$

Avec la base de points *Iris*, on obtient le résultat présenté dans la figure 4. On observe que les points du skyline classique appartiennent totalement au skyline graduel. Néanmoins, on peut trouver des substituts intéressants qui sont plus ou moins typiques. Cette approche appa-

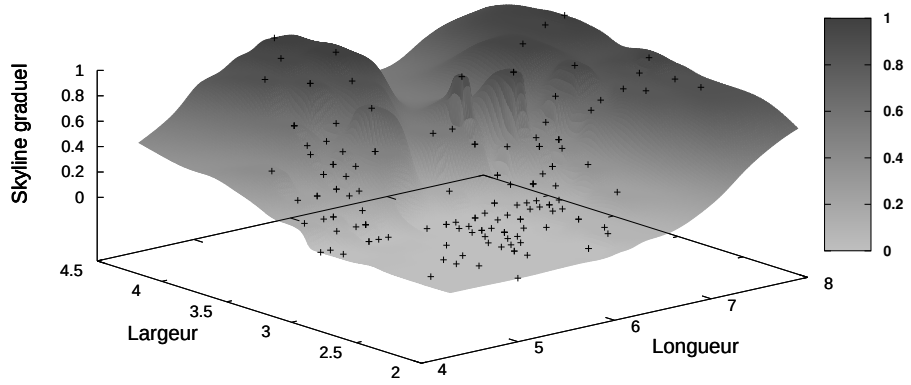

FIG. 4 – *Skyline graduel du jeu de données Iris*

raît intéressante en termes de visualisation. En effet, le score associé à chacun des points permet un affichage restreint des points selon leur degré d'appartenance (e.g., les points dont l'appartenance au skyline dépasse un degré α). Un affichage en 3 dimensions des points (comme le montre la Figure 4), où le degré d'appartenance au skyline donne la valeur en hauteur, permet d'accentuer l'effet "ligne d'horizon" du skyline. Une pente se dessine des points optimaux vers les points les plus typiques ou complètement dominés. En 2 dimensions, il serait également possible de distinguer les courbes de niveau, symbolisant des zones d'intérêt. Dans ces zones, l'utilisateur peut choisir des points qu'il considère intéressants. Même si ces points ne sont pas nécessairement optimaux, ils peuvent susciter l'intérêt de l'utilisateur dans la mesure où ils représentent de bonnes alternatives aux points optimaux dans le cas, par exemple, où ceux-ci apparaissent trop exceptionnels pour être crédibles. Enfin, un élément du skyline graduel possède deux degrés : un degré d'appartenance au skyline, et un degré de typicité qui permet de savoir dans quelle mesure il est exceptionnel.

On peut imaginer différentes formes de navigation dans les zones pour découvrir les points : un simple parcours de la zone pour afficher les caractéristiques des points, l'application de différents filtres dans la zone d'étude comme la recherche de la diversité des réponses (sur certains attributs à spécifier), la typicité, des zooms sur des zones d'intérêt, etc.

4 Éléments d'implémentation

La mise en œuvre effectuée vise à montrer l'intérêt de la notion de typicité dans le calcul d'une requête skyline. Deux phases sont nécessaires au calcul du skyline graduel : *i)* le calcul de la typicité et *ii)* le calcul du skyline. Il existe de nombreux algorithmes pour évaluer les requêtes skyline : l'algorithme Block-Nested-Loops (BNL) (Börzsönyi et al. (2001)) ; l'algorithme Divide and Conquer (Börzsönyi et al. (2001)) ; une proposition utilisant un B-tree ou un R-tree (Börzsönyi et al. (2001)) ; un algorithme basé sur des structures de type Bitmap (Tan et al. (2001)) ; une amélioration du BNL Sort-Filter-Skyline (Chomicki et al. (2003, 2005)), et aussi (Bartolini et al. (2008)) qui s'appuie sur un préclassement des tuples afin de limiter le

nombre de tuples à lire et à comparer. Nous avons choisi de suivre l'approche proposée dans (Tan et al. (2001)) qui permet d'implémenter facilement la formule (5).

L'algorithme proposé dans Tan et al. (2001) permet de retourner progressivement les points du skyline. La structure de données centrale à cette implémentation est celle d'un tableau de booléens ou bitmap. Un index bitmap est défini pour chacune des dimensions du skyline : chaque colonne désigne une valeur possible de la dimension et chaque ligne fait référence à un tuple de la base. La valeur de 1 à la croisée d'une ligne l et d'une colonne c indique que le tuple référencé en ligne l a comme valeur celle désignée par la colonne c. Ensuite, chaque point p de la base $\mathcal{S}$ est testé pour savoir s'il appartient ou non au skyline. Pour cela, deux autres structures de données sont créées. La première, appelée A, désigne *les tuples qui sont aussi bons que* p, la seconde B désigne *ceux qui sont meilleurs que* p sur une dimension. A et B sont définies comme des tableaux de booléens dont les colonnes font référence aux tuples de $\mathcal{S}$. Elles sont remplies à l'aide des index bitmap. L'organisation de cet index selon des valeurs de dimension décroissantes facilite la création de A et B. Une intersection entre A et B donne une suite de booléens. La présence d'un seul bit à 1 indique qu'il existe un tuple meilleur que p et que p n'appartient donc pas au skyline.

L'algorithme 1 qui donne une vue globale de notre implémentation suit l'approche expliquée ci-dessus. Pour permettre le calcul du skyline graduel, nous avons également utilisé des tableaux ($\mathcal{T}$, Sky_{grad}, A') dont chaque colonne désigne un tuple de $\mathcal{S}$. Les valeurs des tableaux sont des réels compris entre 0 et 1 qui indiquent pour $\mathcal{T}$ le degré de typicité de tout point de la base et pour Sky_{grad} le degré d'appartenance d'un point au skyline graduel. *AND* effectue le « et logique » entre toutes les paires de valeurs ($A[i]$, $B[i]$) indiquant ainsi si le point i est à la fois aussi bon que p et meilleur que p sur une dimension. Si c'est le cas, alors ce point domine p. *MULT* effectue une multiplication entre les paires de valeurs ($A[i]$, $\mathcal{T}[i]$), ce qui a pour effet d'affecter à un point i son degré de typicité s'il domine le point p étudié. Enfin, *MAX* désigne la valeur maximale du tableau A. Le degré d'appartenance du point p au skyline graduel est ensuite facilement obtenu par retranscription de la formule (5).

Algorithm 1 Algorithme principal du calcul du skyline graduel

Require: d distance, n cardinalité de la base $\mathcal{S}$, les points de la base $p \in \mathcal{S}$, l'ensemble des dimensions $\{d_i\}$
Ensure: skyline des points : $\forall p \in \mathcal{S},\ Sky_{grad}(p)$
  Prétraitement : création des index bitmap sur les d_i
  Prétraitement : Calcul de la typicité des points $\mathcal{T}$: $\forall p \in \mathcal{S},\ Typ(p)$
  for all $p \in \mathcal{S}$ **do**
    // Recherche des points qui dominent p
    Création de A
    Création de B
    $A := A$ AND B
    $A' := A$ MULT $\mathcal{T}$
    $Sky_{grad}(p) := 1 - Max(A')$
  end for

Le calcul de la typicité utilise une distance minimale d, évidemment dépendante des attributs sur lesquels porte la requête skyline. Si n est la cardinalité de la base, le temps nécessaire

au calcul de la typicité est au plus en n^2. En effet, l'Algorithme 2 nécessite pour chaque point p de la base, de rechercher parmi tous les points de la base ceux qui sont proches de p. Il repose sur la distance euclidienne entre deux points p et p', notée $dist(p, p')$.

Algorithm 2 Calcul de la typicité

Require: d distance, n cardinalité de la base $\mathcal{S}$,
Ensure: typicité des points : $\forall p \in \mathcal{S},\ Typ(p)$

```
max ← 0
for all p ∈ S do
  cpt ← 0
  for all p' ∈ S, p' ≠ p do
    if dist(p, p') ≤ d then
      cpt++
    end if
  end for
  Typ(p) ← cpt/n
  if max ≤ Typ(p) then
    max ← Typ(p)
  end if
end for
for all p ∈ S do
  Typ(p) ← Typ(p)/max
end for
```

On peut envisager deux façons de calculer la typicité : soit à la demande, sur les attributs spécifiés dans la clause *skyline*, soit au préalable, et dans ce cas un précalcul de la typicité sur différents ensembles d'attributs pertinents doit être effectué. Dans le premier cas, se pose la question du surcoût d'une telle opération. Le second cas soulève le problème de la mise à jour de la typicité en cas d'insertion/suppression d'un élément. Dans tous les cas, les index utilisés dans le cadre de la recherche des plus proches voisins (comme les kdtree) peuvent être exploités. En outre, comme précisé auparavant, le calcul du skyline graduel portant uniquement sur un fragment de la base, le surcoût du au calcul de la typicité ne devrait pas être pénalisant.

Enfin, l'algorithme proposé ici peut être adapté pour calculer en parallèle le skyline graduel en segmentant les tableaux A, B, A' et $\mathcal{T}$. De même, la création des structures A et B peut être parallélisée en segmentant leur remplissage (à condition de distribuer ou de partager les index et les valeurs de typicité).

5 Expérimentation sur un jeu de données réelles

L'approche proposée a été testée sur un extrait du site de vente de voitures d'occasion *Le bon coin* [1] de l'année 2012 qui comporte 845810 annonces. Dans ce contexte, il est naturel de vouloir à la fois minimiser le prix et le kilométrage. Nous avons considéré un certain nombre de requêtes dont celle qui sera détaillée ici. Nous nous focalisons sur les voitures *citadines* à

1. www.leboncoin.fr

moteur *essence*, ce qui correspond à 441 annonces. La figure 5 montre les résultats obtenus.

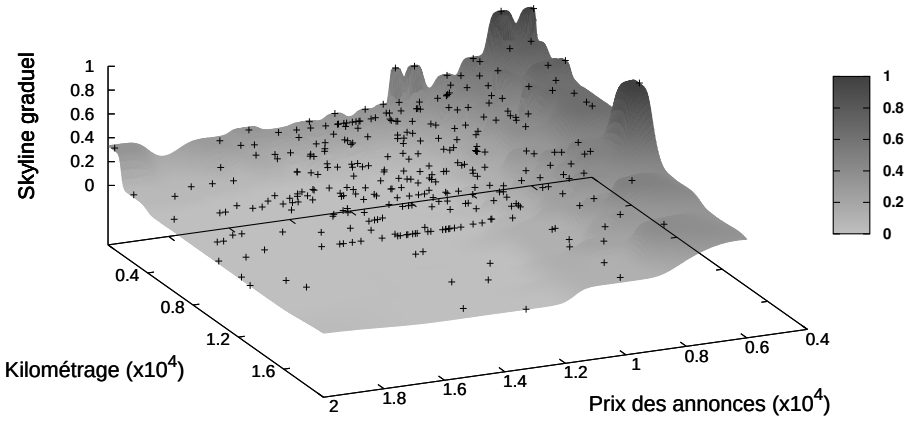

FIG. 5 – *Vue 3D des annonces de type citadines selon leur appartenance au skyline*

En gris foncé, apparaissent les points qui appartiennent le plus au skyline (avec un degré d'appartenance entre 0.8 et 1). Ces points sont détaillés dans la table 2. Vu la définition utilisée,

TAB. 2 – *Extrait de la base annonces et des valeurs d'appartenance au skyline et de typicité*

Id	*Prix*	*Km*	*Skyline*	*Typicité*
1156771	6000	700	1	0.24736843
1596085	5800	162643	1	0.005263158
1211574	7000	500	1	0.3526316
1054357	1800	118000	1	0
1333992	500	220000	1	0
1380340	800	190000	1	0
891125	1000	170000	1	0
1229833	5990	10000	1	0.12631579
1276388	1300	135000	1	0
916264	5990	2514000	0.8736842	0
1674045	6000	3500	0.75263155	0.3157895

des points dominés par d'autres un peu atypiques appartiennent à la réponse. C'est le cas de l'annonce 916264 dominée par les annonces 1229833 et 1054357. Les identifiants en gras correspondent aux points qui appartiennent au skyline classique. On constate que les points de la table 2 (zone $[0.8, 1]$) sont peu voire pas du tout typiques. De plus, certaines caractéristiques peuvent ne pas satisfaire l'utilisateur : le kilométrage peut être très élevé, le prix très faible (au point de sembler suspect). La Table 3 montre un extrait de la 0.6-coupe du skyline graduel, qui comporte des points plus typiques. Elle offre des compromis aux annonces de la table 2, plus représentatifs de la base, et plus rassurants, tout en restant attractifs (le kilométrage est plus faible et le prix, sensiblement plus élevé). Mentionnons aussi que le temps nécessaire au précalcul de la typicité des éléments sélectionnés est deux fois plus important (de l'ordre de 0,54

TAB. 3 – *Extrait de la zone [0.6,0.8]*

Id	*Prix*	*Km*	*Skyline*	*Typicité*
870279	6900	1000	0.71578944	0.35789475
981939	6500	4000	0.63684213	0.36315787
1022586	6500	7200	0.63684213	0.25789472
1166077	7750	2214	0.6421052	0.53157896
1208620	6500	3300	0.71578944	0.36315787
1267726	6500	100000	0.63684213	0
1334605	10500	500	0.64736843	0.6421053
1366336	7490	4250	0.63684213	0.51578945
1529678	7980	650	0.64736843	0.45789474
1635437	9900	590	0.64736843	0.6210526
1685854	7890	1000	0.6421052	0.45789474

seconde) que le temps nécessaire au calcul du skyline graduel (de l'ordre de 0,22 seconde). Mais ce résultat est à tempérer puisque le calcul de la typicité n'a pas été optimisé dans la version actuelle du prototype.

Ces résultats intermédiaires montrent que l'approche proposée est intéressante pour proposer des substituts à des points exceptionnels quand ces derniers semblent peu crédibles. De même, elle permet une visualisation intuitive des résultats et ouvre de nouvelles possibilités en termes de navigation dans les données.

6 Conclusion

Dans cet article, nous nous sommes intéressés à une version graduelle du skyline dont l'objectif est d'empêcher les exceptions de « masquer » des points a priori moins satisfaisants mais plus typiques. Notre approche permet une visualisation en trois dimensions des points de la base mettant en valeur la ligne de crête (skyline) et la pente en direction des vallées formées des points totalement dominés. Une amélioration de cette approche consisterait à utiliser une technique plus fine pour caractériser les points de la base selon leur niveau de représentativité. Par exemple, il serait intéressant d'exploiter les travaux de clustering basés sur la notion de typicité, voir notamment (Lesot (2006)). Nous envisageons à court terme une implémentation parallélisée de l'approche et la mise en place d'index pour améliorer le temps de calcul de la typicité. Une autre amélioration de ce travail serait d'intégrer le prototype au sein d'une plateforme permettant une navigation graphique dans les données, guidée par les courbes de niveau.

Références

Bartolini, I., P. Ciaccia, et M. Patella (2008). Efficient sort-based skyline evaluation. *ACM Trans. Database Syst. 33*(4), 1–49.

Börzsönyi, S., D. Kossmann, et K. Stocker (2001). The skyline operator. In *Proc. of ICDE'01*, pp. 421–430.

Chan, C., H. Jagadish, K. Tan, A. Tung, et Z. Zhang (2006a). Finding k-dominant skylines in high dimensional space. In *Proc. of SIGMOD 2006*, pp. 503–514.

Chan, C., H. Jagadish, K. Tan, A. Tung, et Z. Zhang (2006b). On high dimensional skylines. In *Proc. of EDBT 2006, LNCS 3896*, pp. 478–495.

Chomicki, J., P. Godfrey, J. Gryz, et D. Liang (2003). Skyline with presorting. In U. Dayal, K. Ramamritham, et T. M. Vijayaraman (Eds.), *ICDE*, pp. 717–719. IEEE Computer Society.

Chomicki, J., P. Godfrey, J. Gryz, et D. Liang (2005). Skyline with presorting : Theory and optimizations. In M. A. Klopotek, S. T. Wierzchon, et K. Trojanowski (Eds.), *Intelligent Information Systems*, Advances in Soft Computing, pp. 595–604. Springer.

Dubois, D. et H. Prade (1984). On data summarization with fuzzy sets. In *Proc. of IFSA'93*, pp. 465–468.

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics 7*(2), 179–188.

Hadjali, A., S. Kaci, et H. Prade (2008). Database preferences queries – a possibilistic logic approach with symbolic priorities. In *Proc. of FoIKS'08*, pp. 291–310.

Hadjali, A., O. Pivert, et H. Prade (2011). On different types of fuzzy skylines. In M. Kryszkiewicz, H. Rybinski, A. Skowron, et Z. W. Ras (Eds.), *ISMIS*, Volume 6804 of *Lecture Notes in Computer Science*, pp. 581–591. Springer.

Kung, H. T., F. Luccio, et F. P. Preparata (1975). On finding the maxima of a set of vectors. *J. ACM 22*(4), 469–476.

Lesot, M. (2006). Typicality-based clustering. *Int. J. of Information technology and Intelligent Computing 12*, 279–292.

Lin, X., Y. Yuan, Q. Zhang, et Y. Zhang (2007). Selecting stars : the k most representative skyline operator. In *Proc. of the ICDE 2007*, pp. 86–95.

Tan, K.-L., P.-K. Eng, et B. C. Ooi (2001). Efficient progressive skyline computation. In P. M. G. Apers, P. Atzeni, S. Ceri, S. Paraboschi, K. Ramamohanarao, et R. T. Snodgrass (Eds.), *VLDB*, pp. 301–310. Morgan Kaufmann.

Tao, Y., L. Ding, X. Lin, et J. Pei (2009). Distance-based representative skyline. In Y. E. Ioannidis, D. L. Lee, et R. T. Ng (Eds.), *ICDE*, pp. 892–903. IEEE.

Zadeh, L. A. (1984). A computational theory of dispositions. In Y. Wilks (Ed.), *COLING*, pp. 312–318. ACL.

Summary

This paper deals with the issue of retrieving the most interesting objects in the sense of Pareto ordering, i.e., of answering skyline queries, in the presence of anomalies. Indeed, numerous datasets, as for example sales ads websites, can be populated with odd data. They disturb the skyline computation as odd data may hide interesting points. The proposed approach exploits the fuzzy relation of typicality to make it possible to get, in addition to exceptional objects, some more typical substitutes. Furthermore, it allows for a graphical representation of the data set.