Relaxation des Requêtes Skyline : Une Approche Centrée Utilisateur

Djamal Belkasmi* **, Allel Hadjali**, Hamid. Azzoune ***

*DIF-FS UMBB Boumerdes, Algérie
djamal.belkasmi@ensma.fr - belkasmi.djamel@gmail.com

**LIAS-ENSMA Poitiers, France
allel.hadjali@ensma.fr

**LRIA-USTHB Alger, Algérie
azzoune@yahoo.fr

Résumé. Les requêtes skyline constituent un outil puissant pour l'analyse de données multidimensionnelles et la décision multicritère. Elles permettent d'extraire les meilleurs objets à partir d'un ensemble de données. Elles s'appuient sur le principe de dominance de Pareto. En pratique, le calcul du skyline peut conduire à deux scénarios : soit (i) un nombre important d'objets sont retournés, ce qui est généralement peu informatif du point de vue de l'utilisateur, soit (ii) un nombre réduit d'objets sont retournés, ce qui peut être insuffisant pour les décisions des utilisateurs. Dans cet article, nous abordons le second problème et proposons une approche permettant de le traiter. L'idée consiste à rendre le skyline plus permissive en lui ajoutant des objets qui, proprement dit, ne lui appartiennent pas mais qui sont proches des objets skyline. L'approche s'appuie sur une nouvelle relation de dominance floue appelée «Much Preferred». Un algorithme efficace pour calculer le skyline relaxé est proposé. Une série d'expériences sont menées pour démontrer la pertinence de notre approche et la performance de l'algorithme proposé.

1 Introduction

Les requêtes skyline Börzsönyi et al. (2001) constituent un outil puissant d'analyse de données en vue de prendre des décisions intelligentes face à des données à grande échelle. Elles permettent d'extraire l'ensemble des points les plus intéressants quand différents critères, souvent conflictuels, sont pris en compte. Elles s'appuient sur le principe de dominance de Pareto. Soit D un ensemble de points à d dimensions, une requête skyline calcule l'ensemble des points non dominés dans D. Un point p domine (au sens de Pareto) un point q si et seulement si p est meilleur ou égal à q sur toutes les dimensions et strictement meilleur que q sur au moins une dimension. Par conséquent, les points skyline sont incomparables. Plusieurs études ont

été menées pour développer des algorithmes efficaces et introduire des variantes pour les requêtes skyline Chomicki et al. (2013); Yiu et Mamoulis (2007); Khalefa et al. (2008); Pei et al. (2007); Hadjali et al. (2010). Toutefois, l'interrogation d'un ensemble de données multidimensionnelles à l'aide de l'opérateur skyline peut conduire à deux scénarios possibles : soit (i) un nombre important de réponses sont retournées, ce qui est généralement peu informatif et donc n'apporte pas assez de connaissances à l'utilisateur, soit (ii) un nombre réduit de réponses sont retournées, ce qui peut être insuffisant du point de vue utilisateur. Afin de résoudre ces deux problèmes, un certain nombre de travaux ont été proposés afin de mettre en place des méthodes permettant de raffiner le skyline et donc de réduire sa taille (cas i) Abbaci et al. (2013); Chan et al. (2006a,b); Endres et Kießling (2011); Hadjali et al. (2011); Hüllermeier et al. (2008); Lin et al. (2007); Papadias et al. (2003), par contre, relativement peu de travaux existent afin de relaxer le skyline dans le but d'augmenter sa taille (cas ii) Hadjali et al. (2011); Goncalves et Tineo (2007). Dans Goncalves et Tineo (2007), les auteurs proposent une relation de dominance flexible utilisant des opérateurs flous de comparaison. Ce type de dominance permet de faire augmenter le skyline avec des points qui sont seulement faiblement dominés par tout autre point. Dans Hadjali et al. (2011), quelques idées de relaxation ont été aussi proposées.
Dans cet article, inspirés par l'étude préliminaire de Hadjali et al. (2011), nous abordons d'une manière détaillée le problème de la relaxation du skyline. Plus précisément, nous développons une approche efficace, appelée $\mathcal{MP2R}$ [1], pour la relaxation du skyline. L'approche est fondée sur une nouvelle dominance graduelle *Much Preferred (MP)* (qui signifie fortement préféré à) qui conduit à une dominance plus exigeante entre les points de D. Dans ce contexte, un point appartient toujours au skyline tant qu'il n'est pas fortement dominé, au sens de la relation MP, par un autre point skyline. Ainsi, le nombre des points incomparables augmente, et par conséquent, la taille de la version relaxée du skyline (notée S_{relax}) augmente aussi. Il est à noter que ces points ne sont pas des éléments du skyline car ils sont écartés par la relation de dominance de Pareto. Par ailleurs, le calcul du skyline relaxé S_{relax} est explicitement formalisé via un algorithme optimisé et performant. En résumé les contributions essentielles de cet article sont :

— Une nouvelle variante de la relation de dominance floue basée sur la relation MP est introduite. Les propriétés sémantiques du skyline relaxé S_{relax} sont aussi examinées.
— Un algorithme efficace pour le calcul de S_{relax} est développé et implémenté.
— Une série d'études expérimentales pour étudier et analyser la pertinence et l'efficacité de S_{relax}, sont décrites et commentées.

L'article est structuré comme suit. La section 2 introduit quelques notions préliminaires et donne un aperçu sur des approches existantes. Dans la section 3, nous définissons une nouvelle approche pour la relaxation du skyline basée sur la relation de dominance MP. Dans la section 4, l'algorithme de calcul de S_{relax} est présenté, alors que la section 5 est dédiée à l'étude expérimentale. Enfin, la section 6 conclut l'article et présente les perspectives et travaux futurs.

2 Notions Préliminaires

Nous présentons ici un bref aperçu sur la théorie des ensembles flous et les requêtes skyline.

1. $\mathcal{M}$uch $\mathcal{P}$referred $\mathcal{R}$elation for $\mathcal{R}$elaxation

2.1 Ensembles flous

Le concept des *ensembles flous* a été développé par Zadeh (1965) afin de représenter des classes ou des ensembles d'objets dont les frontières sont mal définies. Ces ensemble permettent de décrire des transitions graduelles entre l'appartenance totale et le rejet absolu. Des exemples typiques de ces classes floues sont celles décrites à l'aide d'adjectifs ou d'adverbes de la langue naturelle, comme *pas_cher*, *jeune* et *la plupart*. Formellement, un ensemble flou F sur l'univers X est décrit par une fonction d'appartenance $\mu_F : X \longrightarrow [0,1]$, où $\mu_F(x)$ représente le **degré d'appartenance** de x dans F. Par cette définition, si $\mu_F(x) = 0$ alors l'élément $x \notin F$, si $\mu_F(x) = 1$ alors $x \in F$, ces éléments représentent le **noyau** de F noté par $Noy(F) = \{x \in F | \mu_F(x) = 1\}$. Lorsque $0 < \mu_F(x) < 1$, on parle d'une **appartenance partielle**, ces éléments forment le **support** de F noté par $Supp(F) = \{x \in F | \mu_F(x) > 0\}$. Plus $\mu_F(x)$ est proche de la valeur 1, plus x appartient à F. Par conséquent, étant donné $x, y \in F$, on dit que x est préféré à y ssi $\mu_F(x) > \mu_F(y)$. Si $\mu_F(x) = \mu_F(y)$, alors x et y sont de même préférence. En pratique, la fonction d'appartenance associée à F est souvent représentée par une fonction d'appartenance trapézoïdale (f.a.t.) modélisée par un quadruplet $(\alpha, \beta, \varphi, \psi)$ où $[\alpha, \psi]$ est le support et $[\beta, \varphi]$ son noyau.

2.2 Requêtes skyline

Les requêtes skyline Börzsönyi et al. (2001) sont un exemple spécifique de requêtes à préférences. Elles s'appuient sur le principe de dominance de Pareto défini comme suit :

Définition 1. Soit D un ensemble de points à d dimensions et u_i et u_j deux points de D. On dit que u_i domine (au sens de Pareto) u_j (noté $u_i \succ u_j$) ssi u_i est meilleur ou égal à u_j sur toutes les dimensions et strictement meilleur que u_j sur au moins une dimension. On a :

$$u_i \succ u_j \Leftrightarrow (\forall k \in \{1,..,d\}, u_i[k] \geq u_j[k]) \wedge (\exists l \in \{1,..,d\}, u_i[l] > u_j[l]) \quad (1)$$

où chaque tuple $u_i = (u_i[1], u_i[2], u_i[3], ..., u_i[d])$ avec $u_i[k]$ représente la valeur du u_i pour la dimension k. Par souci de simplicité, et sans perte de généralité, nous considérons que plus la valeur de $u_i[k]$ est grande, meilleure elle est.

Définition 2. Le skyline de D, noté S, est l'ensemble des points qui ne sont dominés par aucun autre point de D.

$$(u \in S) \Leftrightarrow (\nexists u' \in D, u' \succ u) \quad (2)$$

Exemple 1. Pour illustrer le concept du skyline, considérons une base de données contenant des informations sur des candidats comme montré en table 1. Cette table comporte les informations suivantes : Code, Age, Expérience en management (man_exp en années), Expérience technique (tec_exp en années) et la distance séparant le travail au domicile (dist_td en Km). La procédure de recrutement mise en place par la direction des ressources humaines visent à choisir les candidats ayant plus d'expérience technique (max man_exp) et plus d'expérience en management (max tec_exp), tout en ignorant les autres critères. L'application du skyline traditionnel sur la liste 1 renvoie les candidats suivants : M_5, M_8 voir figure 1.

code	age	man_exp	tec_exp	dist_td
M1	32	5	10	35
M2	41	7	5	19
M3	37	5	12	45
M4	36	4	11	39
M5	40	8	10	18
M6	30	4	6	27
M7	31	3	4	56
M8	36	6	13	12
M9	33	6	6	95
M10	40	7	9	20

TAB. 1 – *Liste des candidats.*

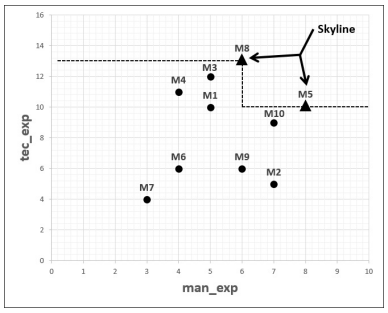

FIG. 1 – *Skyline des candidats*

2.3 Travaux apparentés

La majorité des travaux de recherche sur les requêtes skyline se sont focalisés sur le développement d'algorithmes efficaces, pour le calcul du skyline, suivant des conditions et contextes différents. Börzsönyi et al. (2001), Chomicki et al. (2013) et Godfrey et al. (2005) ont proposé des algorithmes séquentiels sans utilisation de structures de données complexes. Alors que, Tan et al. (2001), Kossmann et al. (2002) et Papadias et al. (2003) ont amélioré l'efficacité des algorithmes de calcul du skyline en proposant des algorithmes basés sur des structures de données d'index (arbres, indexes).
Peu de travaux se sont intéressés au problème de la relaxation du skyline. Goncalves et Tineo (2007) ont abordé le problème de la rigidité du skyline en introduisant une relation de dominance faible basée sur les opérateurs de comparaison floue. Cette relation permet d'enrichir le skyline par des points qui sont faiblement dominés par les autres points. Hadjali et al. (2011) ont introduit quelques idées afin de définir de nouvelles variantes du skyline. Tout d'abord, la première idée consiste à raffiner le skyline en introduisant un certain ordre entre ses points afin d'identifier les plus intéressants. La deuxième idée vise à rendre le skyline plus flexible en ajoutant quelques points qui à proprement dit ne lui appartiennent pas, mais qui sont fai-

blement dominés par tout autre point sur toutes les dimensions du skyline. La troisième, vise à simplifier le skyline par échelonnement des critères susceptibles de regrouper les points les plus similaires (classes de points). Enfin, la dernière idée aborde la question liée à la sémantique du skyline dans le contexte des données incertaines.

3 $\mathcal{MP2R}$: Une approche de relaxation du skyline

Soit la relation $R(A_1, A_2, ..., A_d)$ définie dans un espace à d dimensions $\mathbb{D} = (\mathbb{D}_1, \mathbb{D}_2, ..., \mathbb{D}_d)$, où $\mathbb{D}_i$ est le domaine de l'attribut A_i. On suppose que chaque domaine $\mathbb{D}_i$ est équipé d'une relation d'ordre total. Soit $U = (u_1, u_2, ..., u_n)$ un ensemble de n tuples appartenant à R. Soit aussi S le skyline de U et S_{relax} la version relaxée de S obtenue par l'approche $\mathcal{MP2R}$. Comme mentionné en introduction, $\mathcal{MP2R}$ s'appuie sur une nouvelle dominance graduelle permettant de récupérer les points les plus intéressants parmi ceux écartés par le skyline S. Cette dominance utilise la relation floue *"Much Preferred (MP)"* afin de comparer deux points u et u'. Ainsi, u appartient à S_{relax} s'il n'existe pas de point $u' \in U$ tel que u' est ***Much Preferred*** à u (noté $MP(u', u)$) sur toutes les dimensions du skyline. Formellement, on écrit :

$$u \in S_{relax} \Leftrightarrow \nexists u' \in U, \forall i \in \{1, ..., d\}, MP_i(u'_i, u_i) \tag{3}$$

où, MP_i est la relation ***Much Preferred*** définie sur le domaine $\mathbb{D}_i$ de l'attribut A_i. $MP_i(u'_i, u_i)$ exprime à quel point la valeur u'_i est ***Much Preferred*** à la valeur u_i. La nature graduelle de la relation ***MP*** permet d'associer à chaque élément u de S_{relax} un degré ($\in [0, 1]$). Ce degré exprime la mesure avec laquelle u appartient à S_{relax}. En utilisant les concepts des ensembles flous [2], la formule (3) s'écrit :

$$\mu_{S_{relax}}(u) = 1 - \max_{u' \in U} \min_i \mu_{MP_i}(u'_i, u_i) = \min_{u' \in U} \max_i (1 - \mu_{MP_i}(u'_i, u_i)) \tag{4}$$

où la sémantique de MP_i (définie sur $\mathbb{D}_i$) est donnée par la formule (5) (voir aussi Fig. 2). En termes de f.a.t., MP_i est représentée par $(\gamma_{i1}, \gamma_{i2}, \infty, \infty)$ et notée $MP_i^{(\gamma_{i1}, \gamma_{i2})}$. Il est facile de vérifier que $MP_i^{(0,0)}$ correspond à la relation de préférence classique exprimée par la relation d'ordre *"plus grand que"*.

$$\mu_{MP_i^{(\gamma_{i1}, \gamma_{i2})}}(u'_i, u_i) = \begin{cases} 0 & \text{si } u'_i - u_i \leq \gamma_{i1} \\ 1 & \text{si } u'_i - u_i \geq \gamma_{i2} \\ \frac{(u'_i - u_i) - \gamma_{i1}}{\gamma_{i2} - \gamma_{i1}} & \text{sinon} \end{cases} \tag{5}$$

Soit $\gamma = ((\gamma_{11}, \gamma_{12}), \cdots, (\gamma_{d1}, \gamma_{d2}))$ un vecteur de paramètres où $MP_i^{(\gamma_{i1}, \gamma_{i2})}$ représente la relation MP_i définie sur l'attribut A_i et $S_{relax}^{(\gamma)}$ représente le skyline relaxé en utilisant les paramètres du vecteur γ. Le skyline classique S correspond à $S_{relax}^{(\mathbf{0})}$ où $\mathbf{0} = ((0, 0), \cdots, (0, 0))$.

Définition 3. On dit que $MP_i^{(\gamma_{i1}, \gamma_{i2})}$ est plus forte que $MP_i^{(\gamma'_{i1}, \gamma'_{i2})}$ ssi $(\gamma_{i1}, \gamma_{i2}) \geq (\gamma'_{i1}, \gamma'_{i2})$ (i.e., $\gamma_{i1} \geq \gamma'_{i1} \wedge \gamma_{i2} \geq \gamma'_{i2}$).

Définition 4. Soit γ et γ' deux vecteurs de paramètres. $\gamma \geq \gamma'$ ssi $\forall i \in \{1, \cdots, d\}$, $(\gamma_{i1}, \gamma_{i2}) \geq (\gamma'_{i1}, \gamma'_{i2})$.

2. où le $\forall$ est modélisé par le *min* et le $\exists$ par le *max*

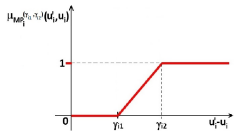

FIG. 2 – *La fonction d'appartenance* $\mu_{MP_i^{(\gamma_{i1},\gamma_{i2})}}$

Proposition 1. Soit γ et γ' deux vecteurs de paramètres. la propriété suivant est vérifiée : $\gamma' \leq \gamma \Rightarrow S_{relax}^{(\gamma')} \subseteq S_{relax}^{(\gamma)}$.
Preuve :
Soit $\gamma' \leq \gamma$. $u \in S_{relax}^{(\gamma')} \Rightarrow \nexists u' \in U, \forall i \in \{1, \cdots, d\}, MP_i^{(\gamma'_{i1},\gamma'_{i2})}(u'_i, u_i)$
$\Rightarrow \nexists u' \in U, \forall i \in \{1, \cdots, d\}, \mu_{MP_i^{(\gamma'_{i1},\gamma'_{i2})}}(u'_i, u_i) > 0$
$\Rightarrow \nexists u' \in U, \forall i \in \{1, \cdots, d\}, u'_i - u_i > \gamma'_{i1} \Rightarrow \forall u' \in U, \forall i \in \{1, \cdots, d\}, u'_i - u_i \leq \gamma'_{i1}$
$\Rightarrow \forall u' \in U, \forall i \in \{1, \cdots, d\}, u'_i - u_i \leq \gamma_{i1} \Rightarrow \nexists u' \in U, \forall i \in \{1, \cdots, d\}, u'_i - u_i > \gamma_{i1}$
$\Rightarrow \nexists u' \in U, \forall i \in \{1, \cdots, d\}, \mu_{MP_i^{(\gamma_{i1},\gamma_{i2})}}(u'_i, u_i) > 0$
$\Rightarrow \nexists u' \in U, \forall i \in \{1, \cdots, d\}, MP_i^{(\gamma_{i1},\gamma_{i2})}(u'_i, u_i) \Rightarrow u \in S_{relax}^{(\gamma)} \Rightarrow S_{relax}^{(\gamma')} \subseteq S_{relax}^{(\gamma)} \square$

Lemme 1. Soit $\gamma = ((0, \gamma_{12}), \cdots, (0, \gamma_{d2}))$ et $\gamma' = ((\gamma'_{11}, \gamma'_{12}), \cdots, (\gamma'_{d1}, \gamma'_{d2}))$, on a : $S_{relax}^{(0)} \subseteq S_{relax}^{(\gamma)} \subseteq S_{relax}^{(\gamma')}$

Exemple 2. Reprenons le skyline calculé dans l'exemple 1. Supposons que les relations *"Much Preferred"* correspondant aux attributs du skyline (man_exp et tec_exp) sont données respectivement par :

$$\mu_{MP_{man\_exp}^{(1/2,2)}}(u', u) = \begin{cases} 0 & \text{si } u' - u \leq 1/2 \\ 1 & \text{si } u' - u \geq 2 \\ 2/3(u' - u) - 1/3 & \text{sinon} \end{cases} \quad (6)$$

$$\mu_{MP_{tec\_exp}^{(1/2,4)}}(u', u) = \begin{cases} 0 & \text{si } u' - u \leq 1/2 \\ 1 & \text{si } u' - u \geq 4 \\ 2/7(u' - u) - 1/8 & \text{sinon} \end{cases} \quad (7)$$

Mat	M5	M8	M3	M10	M1	M2	M4	M6	M7	M9
$\mu_{S_{relax}}$	1	1	0.85	0.85	0.66	0.66	0.57	0	0	0

TAB. 2 – *Degrés des éléments de* S_{relax}.

L'application de l'approche $\mathcal{MP}2\mathcal{R}$ pour relaxer le skyline S = {M_5, M_8} de l'exemple 1, conduit au skyline relaxé suivant S_{relax} = {(M_5, 1), (M_8, 1), (M_3, 0.85), (M_{10}, 0.85), (M_1, 0.66), (M_2, 0.66), (M_4, 0.57)}, voir Table 2. Il est important de noter que certains points qui n'appartenaient pas à S deviennent des éléments de S_{relax} (comme M_{10} et M_4), voir Fig. 3. Ce qui

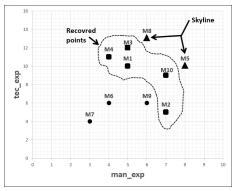

FIG. 3 – *Relaxation du skyline*

signifie que la taille de S_{relax} est plus grande que celle de S. Examinons, maintenant, en détails le contenu de S_{relax}, on observe que : (i) les éléments du skyline S appartiennent toujours à S_{relax} avec un degré égal à 1 ; (ii) l'apparition de nouveaux éléments récupérés par notre approche dont les degrés sont inférieurs à 1 (comme M_3). Ainsi, l'utilisateur peut sélectionner à partir de S_{relax} :

1. Les k meilleurs (*Top-k*) éléments de S_{relax} ayant le plus grand degré (où k est un paramètre défini par l'utilisateur), ou bien
2. Un sous ensemble d'éléments de S_{relax}, noté $(S_{relax})_\sigma$, dont les degrés sont supérieurs au seuil σ fourni par l'utilisateur.

A partir de l'exemple, on a $Top-6$ = {($M_5, 1$), ($M_8, 1$), ($M_3, 0.85$), ($M_{10}, 0.85$), ($M_1, 0.66$), ($M_2, 0.66$)} et $(S_{relax})_{0.7}$ = {($M_5, 1$), ($M_8, 1$), ($M_3, 0.85$), ($M_{10}, 0.85$)}.

4 Calcul de S_{relax}

Pour calculer S_{relax}, nous procédons en deux étapes (voir Fig. 4) : Premièrement, on cal-

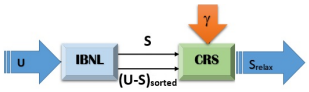

FIG. 4 – *Etapes de relaxation du skyline*

cule le skyline classique (S) en utilisant une version améliorée de l'algorithme BNL ($IBNL$), voir l'algorithme 1. La base de données U est triée dans l'ordre croissant en utilisant une fonction monotone Mf (ex., la somme des attributs du skyline en multipliant par -1 la valeur des attributs dont le critère est MAX). Cette fonction vérifie la propriété suivante :

$$\forall u, v \in U \mid Mf(u) \leq Mf(v) \Longrightarrow \neg(v \succ u) \tag{8}$$

La dominance au sens de Pareto entre u_i et u_j sur les dimensions du skyline est évaluée par la fonction **SkylineCompare** et renvoie le résultat dans la variable *status* qui peut être soit : 0 si $u_i = u_j$, 1 si $u_i \succ u_j$, 2 if $u_i \prec u_j$, 3 si u_i et u_j sont incomparables.

La deuxième étape consiste à appliquer notre algorithme de relaxation du skyline, appelé CRS

Algorithme 1 : IBNL

Input : A set of tuples U
Output : A skyline S

```
1  Sort(U);
2  for i := 1 to n − 1 do
3      if ¬u_i.dominated then
4          for j := i + 1 to n do
5              status = 0;
6              if ¬u_j.dominated then
7                  evaluate SkylineCompare(u_i,u_j,status);
8                  switch status do
9                      case 1
10                         u_i.dominated= true;
11                     case 2
12                         u_j.dominated= true;
13         if ¬u_i.dominated then
14             S = S ∪ {u_i};
15 return S;
```

(Computing Relaxed Skyline), en utilisant un vecteur de paramètres γ fourni par l'utilisateur (voir algorithme 2).

Algorithme 2 : CRS

Input : A set of tuples U ; Skyline S ; γ a vector of parameters ;
Output : A relaxed skyline S_{relax} ;

```
1  begin
2      S_relax = S;
3      for i = 1 to n do
4          if u_i ∉ S then
5              for j = 1 to n do
6                  for k = 1 to d do
7                      evaluate μ_MP_k(u_i, u_j);
8                  compute min_k(μ_MP_k);
9              compute max_j(min_k(μ_MP_k)) ; μ_S_relax(u_i) = 1 − max_j(min_k(μ_MP_k));
10         if μ_S_relax(u_i) > 0 then
11             S_relax = S_relax ∪ {u_i};
12     rank u_i in decreasing order w.r.t. μ_S_relax(u_i);
13     return S_relax;
```

5 Etude expérimentale

Cette section présente l'étude expérimentale réalisée. Elle permet de valider l'efficacité et la pertinence de l'approche $\mathcal{MP2R}$ pour relaxer le skyline et aussi mesurer certaines performances liées au temps de calcul. Les tests ont été effectués, sous linux, sur une machine Intel core i7 2,90 GHz, 8 Go de RAM et 250 Go d'espaces disque. Les programmes sont codés en Java. Les données de tests ont été générées suivant la méthode décrite dans Börzsönyi et al. (2001). Les paramètres de tests utilisés sont : la taille du dataset [D] (10K, 50K, 100K), le schéma de la distribution des données [DIS] (indépendantes, corrélées, non-corrélées), le nombre de dimensions utilisées dans le skyline [d] (2, 4, 6, 8) ainsi que les seuils de relaxation $\gamma = (\gamma_{i1}, \gamma_{i2})$, pour i$\in \{1, \ldots, d\}$, avec $\gamma_{i1}, \gamma_{i2} \in$[0,1] et $\gamma_{i1} \leq \gamma_{i2}$. Ces paramètres sont initialisés comme suit : D = 10K ; DIS =" Corrélées" ; d = 2 ; γ = ((0.25,0.5),(0.25,0.5)). Nous rappelons que dans cette étude, plus la valeur d'un attribut est petite, meilleure est. Les points suivants sont ainsi abordés :

— Impact de [DIS] sur la taille et le temps de calcul du skyline relaxé,
— Impact de [d] sur la taille et le temps de calcul du skyline relaxé,
— Impact de [D] sur la taille et le temps de calcul du skyline relaxé,
— Impact de la relation de dominance ($MP_i^{(\gamma_{i1},\gamma_{i2})}$) sur la taille du skyline relaxé.

Impact de la distribution des données. La Figure 5 montre que la distribution des données n'affecte en aucun cas l'efficacité et la capacité de notre approche à enrichir le skyline. Toutefois, le cas des données corrélées présente une légère différence. Ceci est du à la nature de ces données. Nous remarquons que le skyline est fortement relaxé pour tous les types des données. De plus, le temps de calcul est pratiquement le même pour les 3 distributions.

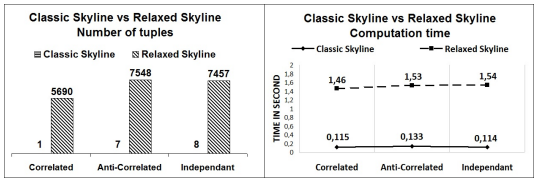

FIG. 5 – *Impact de [DIS].*

Impact du nombre de dimensions du skyline. il est bien connu que si le nombre de dimensions du skyline augmente, la taille du skyline classique augmente aussi. L'algorithme CRS conduite au même comportement surtout en présence des données fortement corrélées (coefficient de corrélation = 0.9952), voir Fig. 6. En passant de 2 dimensions à 8, la taille du skyline relaxé passe de 5690 à 7519 tuples et le temps d'exécution croit de 1.46 à 2.45 seconde.

Impact de la taille du Dataset. La taille du skyline relaxé est proportionnelle à la taille du dataset (Fig. 7), ce qui confirme la capacité de l'algorithme CRS à relaxer le skyline. Par contre, son temps d'exécution augmente considérablement.

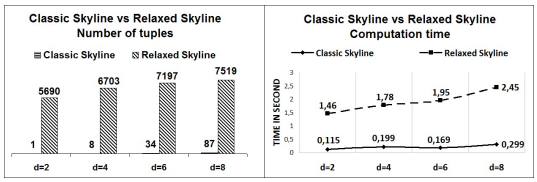

FIG. 6 – *Impact de [d].*

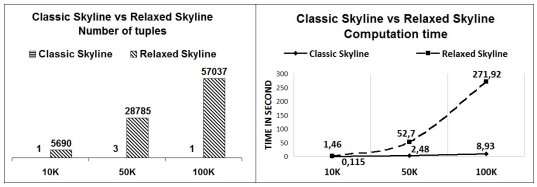

FIG. 7 – *Impact de [D].*

Impact de la relation de dominance $(MP^{(\gamma_1,\gamma_2)})$**.** Dans cette partie, nous présentons l'influence de la relation de dominance "Much Preferred" $(MP^{(\gamma_1,\gamma_2)})$ sur la taille du skyline relaxé. L'idée consiste à faire varier les valeurs des seuils de relaxation (γ_{i1} et γ_{i2}). Sachant que les données sont normalisées, et par souci de simplicité, nous appliquons les mêmes valeurs pour toutes les dimensions skyline. Notons que la taille du skyline classique est égale à 1. Les scénarios suivants sont réalisés :

Scénario 1 : Dans ce scénario, nous fixons la valeur de γ_{i1} et varions celle de γ_{i2} afin d'augmenter la zone de relaxation. Nous observons les cas suivants :

Case 1 : $\gamma_{i1} = 0$ **et** $\gamma_{i2} \in \{0.25; 0.5; 0.75; 1\}$**.** La Figure 8 montre les résultats obtenus. L'analyse de ces courbes indique, d'une part, que la taille du skyline relaxé est proportionnelle à la taille de la zone de relaxation et, d'autre part, qu'il n'existe aucun tuple relaxé dont le degré est égal à 1 (ceci est du à la valeur de $\gamma_{i1} = 0$).

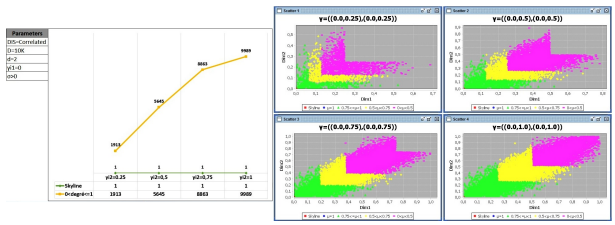

FIG. 8 – *Fixer* $\gamma_{i1} = 0$ *et faire varier* γ_{i2} ***(cas 1)***

Case 2 : $\gamma_{i1} = 0.25$ **et** $\gamma_{i2} \in \{0.25; 0.5; 0.75; 1\}$**.** Plus la zone de relaxation augmente, plus la taille du skyline augmente. On note, par conséquent, l'apparition de tuples relaxés avec un degré égal à 1 (voir Fig. 9).

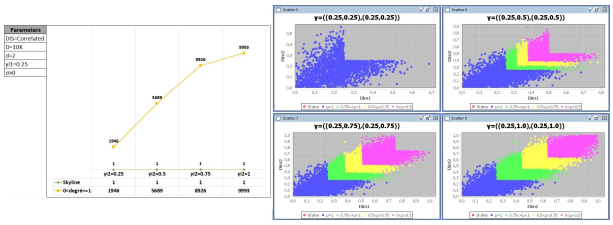

FIG. 9 – *Fixer* $\gamma_{i1} = 0.25$ *et faire varier* γ_{i2} ***(cas 2)***

Case 3 : $\gamma_{i1} = 0.5$ **et** $\gamma_{i2} \in \{0.5; 0.75; 1\}$**.** Le même résultat est observé que précédemment sauf que le nombre de tuples, dont le degré est égal à 1, est plus important (voir Fig. 10).

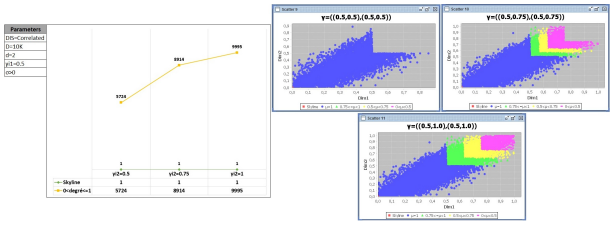

FIG. 10 – *Fixer* $\gamma_{i1} = 0.5$ *et faire varier* γ_{i2} ***(cas 3)***

Case 4 : $\gamma_{i1} = 0.75$ **et** $\gamma_{i2} \in \{0.75; 1\}$**.** Le même résultat est obtenu dans ce cas, mais on remarque que le nombre de tuples, dont le degré est égal à 1, est plus important que celui des cas 2 et 3 (voir Fig. 11). Ceci explique que plus la zone de relaxation est proche de 1, plus le nombre de tuples de degré égal à 1 est important.

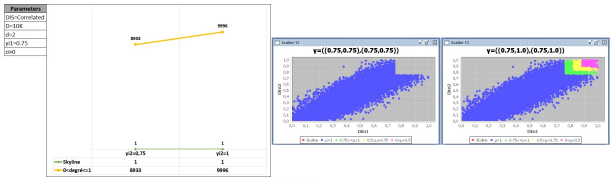

FIG. 11 – *Fixer* $\gamma_{i1} = 0.75$ *et faire varier* γ_{i2} ***(cas 4)***

Scénario 2 : Dans ce scénario, nous varions la valeur des deux seuils. Le résultat obtenu est illustré par la Figure 12. L'analyse de ces courbes montre que la fonction de relaxation devient plus permissive lorsque les seuils s'approchent de 1.

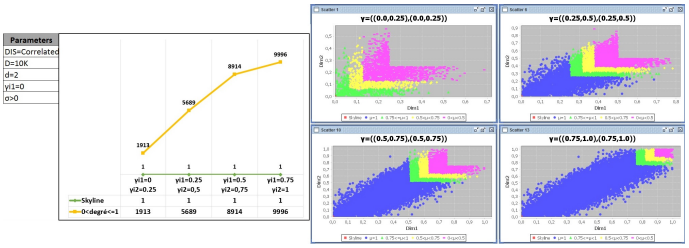

FIG. 12 – *Varier* γ_{i1} *et* γ_{i2}

Finalement, nous constatons, au travers de cette étude expérimentale, que le choix des valeurs de $\gamma = (\gamma_{i1}, \gamma_{i2})$ est très important dans le processus de relaxation du skyline. Ce choix constitue un très bon critère pour le contrôle de la taille du skyline.

6 Conclusion

Dans cet article, nous avons abordé le problème de la relaxation du skyline dont la taille est assez réduite. Une approche de relaxation, appelée $\mathcal{MP2R}$, est proposée. Le concept clé de cette approche est une relation spécifique, notée *Much Preferred*, dont la sémantique est définie par l'utilisateur. Un nouvel algorithme, appelé CRS, pour le calcul du skyline relaxé est développé. L'étude expérimentale réalisée a montré, d'une part, que l'approche $\mathcal{MP2R}$ est une très bonne alternative pour résoudre le problème de la relaxation du skyline et, d'autre part, que le coût de calcul de S_{relax} est assez raisonnable. De plus, $\mathcal{MP2R}$ dépend de divers paramètres qui permettent de contrôler la taille du skyline relaxé. En perspective, nous comptons examiner la question liée au temps de calcul de S_{relax} par l'utilisation des indexes multidimensionnels avancés (sous forme des R-arbres et leurs variantes).

Références

Abbaci, K., A. Hadjali, L. Lietard, et D. Rocacher (2013). A linguistic quantifier-based approach for skyline refinement. In *IFSA/NAFIPS*, pp. 321–326.

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

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

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

Chomicki, J., P. Ciaccia, et N. Meneghetti (2013). Skyline queries, front and back. *SIGMOD Record*, 6–18.

Endres, M. et W. Kießling (2011). Skyline snippets. In *FQAS*, pp. 246–257.

Godfrey, P., R. Shipley, et J. Gryz (2005). Maximal vector computation in large data sets. In *VLDB*, pp. 229–240.

Goncalves, M. et L. Tineo (2007). Fuzzy dominance skyline queries. In *DEXA*, pp. 469–478.

Hadjali, A., O. Pivert, et H. Prade (2010). Possibilistic contextual skylines with incomplete preferences. In *SoCPaR, Cergy Pontoise / Paris, France*, pp. 57–62.

Hadjali, A., O. Pivert, et H. Prade (2011). On different types of fuzzy skylines. In *ISMIS*, pp. 581–591.

Hüllermeier, E., I. Vladimirskiy, B. Prados-Suárez, et E. Stauch (2008). Supporting case-based retrieval by similarity skylines : Basic concepts and extensions. In *ECCBR*, pp. 240–254.

Khalefa, M. E., M. F. Mokbel, et J. J. Levandoski (2008). Skyline query processing for incomplete data. In *IEEE ICDE*, pp. 556–565.

Kossmann, D., F. Ramsak, et S. Rost (2002). Shooting stars in the sky : An online algorithm for skyline queries. In *VLDB*, pp. 275–286.

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

Papadias, D., Y. Tao, G. Fu, et B. Seeger (2003). An optimal and progressive algorithm for skyline queries. In *ACM SIGMOD*, pp. 467–478.

Pei, J., B. Jiang, X. Lin, et Y. Yuan (2007). Probabilistic skylines on uncertain data. In *VLDB*, pp. 15–26.

Tan, K., P. Eng, et B. C. Ooi (2001). Efficient progressive skyline computation. In *VLDB*, pp. 301–310.

Yiu, M. L. et N. Mamoulis (2007). Efficient processing of top-k dominating queries on multi-dimensional data. In *VLDB*, pp. 483–494.

Zadeh, L. A. (1965). Fuzzy sets. *Information and Control*, 338–353.

Summary

Skyline queries have gained much attention in the last decade and are proved to be valuable for multi-criteria decision making. They are based on the concept of Pareto dominance. When computing the skyline, two scenarios may occur: either (i) a huge number of skyline which is less informative for the user or (ii) a small number of returned objects which could be insufficient for the user needs. In this paper, we tackle the second problem and propose an approach to deal with it. The idea consists in making the skyline more permissive by adding points that strictly speaking do not belong to it, but are close to belonging to it. A new fuzzy variant of dominance relationship is then introduced. Furthermore, an efficient algorithm to compute the relaxed skyline is proposed. Extensive experiments are conducted to demonstrate the effectiveness of our approach and the performance of the proposed algorithm.