Relaxation des Requêtes Skyline : Une Approche Centrée Utilisateur

Djamal Belkasmi[*,**], Allel Hadjali[**], Hamid Azzoune[***]

[*]DIF-FS UMBB Boumerdes, Algérie
djamal.belkasmi@ensma.fr

[**]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. En pratique, le calcul du skyline peut conduire à deux scénarios : soit (i) un nombre important d'objets sont retournés, soit (ii) un nombre réduit d'objets sont retournés, ce qui peut être insuffisant pour la prise de décisions. 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 les objets, non skyline, les plus préférés. 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 l'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

n'apporte donc pas assez d'informations 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). 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 aussi été 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 $\mathcal{MP}2\mathcal{R}$ [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 faisaient pas partie du skyline car ils étaient 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 analysées.

L'article est structuré comme suit. La section 2 introduit les requêtes skyline. Dans la section 3, nous discutons la nouvelle approche, pour la relaxation du skyline, basée sur la relation de dominance MP et nous présentons l'algorithme de calcul de S_{relax}. La section 4 est dédiée à l'étude expérimentale. Enfin, la section 5 conclut l'article et présente les perspectives et travaux futurs.

2 Les 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 :

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

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

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.

$$(u \in S) \Leftrightarrow (\nexists u' \in D, u' \succ u) \tag{2}$$

Plus de détails sur le calcul du skyline sont donnés dans Belkasmi et al. (2015).

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 de 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. 1). En terme de fonction d'appartenance trapézoïdal, 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 $(\gamma_{i1}, \gamma_{i2}) \geq (\gamma'_{i1}, \gamma'_{i2})$

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

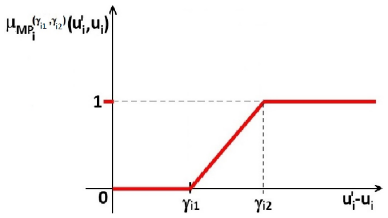

FIG. 1 – *La fonction d'appartenance* $\mu_{MP_i^{(\gamma_{i1},\gamma_{i2})}}$

(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 $\forall i \in \{1, \cdots, d\}, (\gamma_{i1}, \gamma_{i2}) \geq (\gamma'_{i1}, \gamma'_{i2})$.

Proposition 1. Soit γ et γ' deux vecteurs de paramètres. la propriété suivant est vérifiée : $\gamma' \leq \gamma \Rightarrow S_{relax}^{(\gamma')} \subseteq S_{relax}^{(\gamma)}$.

La preuve de la proposition peut-être consultée dans Belkasmi et al. (2015).

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

Le calcul de S_{relax} consiste à appliquer notre algorithme de relaxation du skyline, appelé CRS (Computing Relaxed Skyline), en utilisant un vecteur de paramètres γ fournis par l'utilisateur (voir algorithme 1).

Algorithme 1 : CRS

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

```
begin
    S_relax = S;
    for i = 1 to n do
        if u_i ∉ S then
            for j = 1 to n do
                for k = 1 to d do
                    evaluate μ_MP_k(u_i, u_j);
                compute min_k(μ_MP_k);
            compute max_j(min_k(μ_MP_k)) ; μ_S_relax(u_i) = 1 − max_j(min_k(μ_MP_k));
        if μ_S_relax(u_i) > 0 then
            S_relax = S_relax ∪ {u_i};
    rank S_relax in decreasing order w.r.t. μ_S_relax(u_i);
    return S_relax;
```

4 Etude expérimentale

Cette section présente l'étude expérimentale réalisée. Elle permet de valider l'efficacité et la pertinence de l'approche $\mathcal{MP}2\mathcal{R}$ pour relaxer le skyline. Nous présentons l'influence de la relation de dominance "Much Preferred" ($MP^{(\gamma_1,\gamma_2)}$) sur la taille du skyline relaxé. Les scénarios suivants ont été réalisés :

Scénario 1 : Dans ce scénario, nous fixons la valeur de γ_{i1} et varions celle de γ_{i2}. L'analyse de ce scénario est discutée en détails dans Belkasmi et al. (2015).

Scénario 2 : Dans ce scénario, nous varions la valeur des deux seuils. Le résultat obtenu est illustré par la Figure 2. L'analyse de ces courbes montre qu'on ne peut avoir de points relaxés avec un degré égale à 1 si $\gamma_{i1} = 0$. De plus, la fonction de relaxation devient plus permissive lorsque les seuils s'approchent de 1. 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.

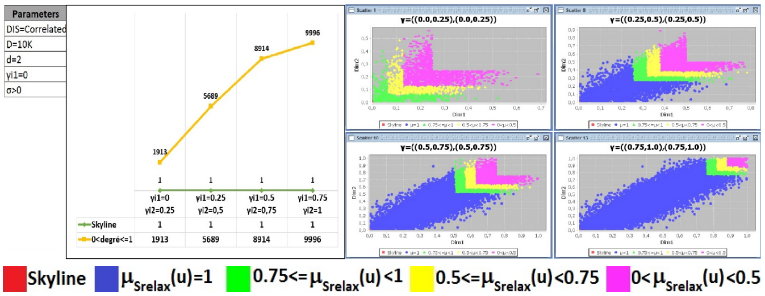

FIG. 2 – *Varier γ_{i1} et γ_{i2}*

5 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 $\mathcal{MP}2\mathcal{R}$, 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 $\mathcal{MP}2\mathcal{R}$ 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. En perspective, nous comptons améliorer le temps de calcul de S_{relax} par l'utilisation d'indexes multidimensionnels avancés.

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.

Belkasmi, D., A. HadjAli, et H. Azzoune (2015). Relaxation des requêtes : Une approche centrée utilisateur. Technical Report 103, LIAS/ENSMA. "http://www.lias-lab.fr/publications/20385/egc2016%20version%20longue.pdf".

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.

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.

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.

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

Summary

Skyline queries have gained much attention in the last decade and are proved to be valuable for multi-criteria decision making. When computing the skyline, two scenarios may occur: either (i) a huge number of skyline 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 and to make the skyline more permissive. 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.