Requêtes Skyline en présence des données évidentielles

Sayda Elmi*, Karim Benouaret**, Allel Hadjali***
Mohamed Anis Bach Tobji*, Boutheina Ben Yaghlane*

*LARODEC, Université de Tunis, Tunisie
elmi.sayda@gmail.com, anis.bach@isg.rnu.tn
boutheina.yaghlane@ihec.rnu.tn
**LIRIS – UMR 5205 CNRS, Université Claude Bernard Lyon 1, France
karim.benouaret@liris.cnrs.fr
***LIAS, ENSMA, France
allel.hadjali@ensma.fr

Résumé. Dans cet article, nous nous intéressons à la recherche des points les plus intéressants au sens de l'ordre de Pareto, dans les bases de données évidentielles. Nous présentons le modèle skyline évidentiel qui est adapté à la nature des données incertaines. Ensuite, nous présentons une évaluation expérimentale de notre approche.

1 Introduction

À cause de l'explosion du nombre d'informations stockées et partagées sur Internet, et l'introduction de nouvelles technologies pour capturer ces données, l'analyse des données incertaines est devenue essentielle dans de nombreuses applications pour la prise de décision. Pour gérer et traiter l'incertitude des données, plusieurs modèles ont été proposés, ce qui a donné naissance à différents types de bases de données imparfaites. Nous pouvons citer les plus connues : les bases de données probabilistes présentées par Dalvi et Suciu (2007); Aggarwal et Yu (2009), les bases de données possibilistes introduites par Bosc et Pivert (2010) et les bases de données évidentielles basées sur la théorie de Dempster-Shafer proposées par Bell et al. (1996). L'utilisation des bases de données évidentielles offre plusieurs avantages à savoir : (i) Elles permettent de modéliser l'incertitude et l'imprécision des données ; et (ii) cela représente une généralisation des deux modèles ; probabiliste et possibiliste à la fois.
Dans cet article, nous nous nous intéressons aux requêtes Skyline sur des données incertaines où l'incertitude est modélisée par la théorie de l'évidence, ce qui constitue un travail pionnier.

Le reste cet article est organisé comme suit. La section 2 contient un rappel sur l'opérateur Skyline, les notions de base de la théorie de l'évidence et les bases de données évidentielles. Dans la section 3, nous définissons formellement la relation de dominance et modélisons le Skyline évidentiel. Nos expérimentations sont données dans la section 4. Enfin, la section 5 conclut l'article.

2 Notions de base

Dans cette section, nous présentons d'abord les requêtes Skyline sur des données classiques Borzsonyi et al. (2001). Ensuite, nous présentons les notions de base de la théorie de l'évidence et les bases de données évidentielles.

2.1 Les requêtes Skyline

Considérons un ensemble d'objets $\mathcal{O} = \{o_1, o_2, \ldots, o_n\}$ défini sur un ensemble d'attributs $\mathcal{A} = \{a_1, a_2, \ldots, a_d\}$. $o_i.a_k$ désigne la valeur du k^{ieme} attribut de l'objet o_i. Pour simplifier, nous supposons que la valeur la plus élevée, est la plus préférée.

Définition 1 *(Relation de Dominance) Étant donnés deux objets $o_i, o_j \in \mathcal{O}$, o_i domine o_j, noté $o_i \succ o_j$, si et seulement si o_i est au moins aussi bon que o_j dans tous les attributs et strictement meilleur dans au moins un attribut, i.e., $\forall a_k \in \mathcal{A} : o_i.a_k \geq o_j.a_k \wedge \exists a_\ell \in \mathcal{A} : o_i.a_\ell > o_j.a_\ell$.*

Définition 2 *(Skyline) Le Skyline de $\mathcal{O}$, noté $Sky_{\mathcal{O}}$, comprend les objets de $\mathcal{O}$ qui ne sont dominés par aucun autre objet, i.e., $Sky_{\mathcal{O}} = \{o_i \in \mathcal{O} \mid \nexists o_j \in \mathcal{O}, o_j \succ o_i\}$.*

2.2 La théorie de l'évidence

La théorie de l'évidence a été introduite par Shafer (1976) dont le but est d'évaluer subjectivement la vérité d'une proposition. Cette théorie, aussi connue sous le nom de "théorie des fonctions de croyance" est une généralisation de la théorie bayésienne Dempster (1968). Elle représente un ensemble d'hypothèses désigné par le cadre de discernement.

Définition 3 *(Cadre de discernement) Un cadre de discernement, noté θ où $\theta = \{h_1, h_2, \ldots, h_m\}$ contient les hypothèses exhaustives et mutuellement exclusives.*

Définition 4 *(Fonction de masse) La fonction, $m : 2^{\Theta} \longrightarrow [0, 1]$, est appelée une masse de probabilité élémentaire sur un ensemble Θ si elle satisfait les deux conditions suivantes : $m(\emptyset) = 0$ and $\sum_{A \subseteq \Theta} m(A) = 1$.*

Définition 5 *(Fonction de Croyance) Pour chaque sous-ensemble A de θ, la croyance placée sur A, notée $bel(A)$, est définie comme la somme des masses affectées à chaque sous-ensemble B de A, i.e., $bel(A) = \sum_{B \subseteq A} m(B)$.*

2.3 Les bases de données évidentielles (BDE)

Les BDE permettent de représenter les données manquantes, incertaines ou imprécises. Il s'agit d'une collection d'objets $\mathcal{O} = \{o_1, o_2, \ldots, o_n\}$ définie sur un ensemble d'attributs $\mathcal{A} = \{a_1, a_2, \ldots, a_d\}$ où chaque attribut a_k a un domaine $dom(a_k)$, et chaque attribut a_k d'un objet o_i, noté $o_i.a_k$ contient une affectation d'une masse de probabilité élémentaire normalisée appelée fonction de masse, i.e, : $o_i.a_k = \{\langle A, m_{ik}(A)\rangle \mid A \subseteq Dom(a_k), m_{ik}(A) > 0\}$, avec $m_{ik} : 2^{Dom(a_k)} \longrightarrow [0, 1]$, $m_{ik}(\emptyset) = 0$ et $\sum_{A \subseteq Dom(a_k)} m_{ik}(A) = 1$. Chaque ensemble

Produit	Poids perdu / mois (kilogrammes)	Remboursement (%)
o_1	$\langle\{15,16,18\},0.1\rangle, \langle\{19,20\},0.9\rangle$	$\langle 90,0.3\rangle, \langle\{90,100\},0.7\rangle$
o_2	$\langle 7,0.7\rangle\ \langle\{8,9\},0.3\rangle$	$\langle\{70,80\},0.8\rangle, \langle 80,0.3\rangle$
o_3	$\langle\{1,4\},0.1\rangle, \langle 5,0.9\rangle$	$\langle\{70,80\},0.7\rangle, \langle 100,0.3\rangle$
o_4	$\langle 10,0.2\rangle, \langle 12,0.2\rangle, \langle\{13,14\},0.6\rangle$	$\langle 100,1\rangle$
o_5	$\langle\{12,13,14\},0.2\rangle, \langle 17,0.4\rangle, \langle 19,0.4\rangle$	$\langle\{20,30\},0.6\rangle, \langle 30,0.4\rangle$

TAB. 1: Exemple d'une BDE.

$A \subseteq Dom(a_k) : m_{ik}(A) > 0$ est appelé une hypothèse ; $\langle A, m_{ik}(A)\rangle$, est appelé un élément focal. Généralement, une base de données est obtenue en collectant différents avis d'experts.

Dans le tableau 1, on considère un ensemble de produits d'amaigrissement, défini sur deux attributs ; Poids perdu par mois et le remboursement si l'utilisateur n'est pas satisfait. Chaque produit peut avoir un ou plusieurs éléments focaux.

3 Les requêtes Skyline en présence de données évidentielles

Dans cette section, nous introduisons la relation de dominance entre les objets dont l'incertitude est modélisée par la théorie des fonctions de croyance, par la suite, nous présentons la définition du Skyline évidentiel. Étant donné un ensemble d'objets $\mathcal{O} = \{O_1, O_2, \ldots, O_n\}$ défini sur un ensemble d'attributs $\mathcal{A} = \{a_1, a_2, \ldots, a_d\}$, avec $o_i.a_k$ représente l'ensemble des éléments focaux de l'objet o_i et l'attribut a_k ; par exemple [1], $o_1.wl = \{\langle\{15,16,18\},0.1\rangle, \langle\{19,20\},0.9\rangle\}$ et $o_1.r = \{\langle 90,0.5\rangle, \langle\{90,100\},0.5\rangle\}$. Le degré de croyance qu'une valeur incertaine de l'objet o_i définie sur l'attribut a_k est préférée ou égale à une autre valeur de l'objet o_j, est donné par Bell et al. (1996) :

$$bel(o_i.a_k \geq o_j.a_k) = \sum_{A \subseteq Dom(a_k)} (m_{ik}(A) \sum_{B \subseteq Dom(a_k), A \geq^{\forall} B} m_{jk}(B)) \quad (1)$$

Où $A \geq^{\forall} B$ représente $a \geq b$ pour tous $(a,b) \in A \times B$. Dans notre exemple, nous avons $bel(o_1.wl \geq o_3.wl) = 0.3 \cdot (0.1 + 0.9) + 0.7 \cdot (0.1 + 0.9) = 1$, et $bel(o_1.r \geq o_3.r) = 0.7 \cdot 0.7 + 0.3 \cdot 0.7 = 0.7$.

Étant donnés deux objets o_i et o_j dans $\mathcal{O}$ avec $o_i \neq o_j$, le degré de croyance qu'un objet o_i domine o_j est défini par :

$$bel(o_i \succ o_j) = \prod_{a_k \in \mathcal{A}} bel(o_i.a_k \geq o_j.a_k) \quad (2)$$

On a $bel(o_1.wl \geq o_3.wl) = 1$ et $bel(o_1.r \geq o_3.r) = 0.7$. Le degré de croyance que o_1 domine o_3 est $bel(o_1 \succ o_3) = 1 \cdot 0.7 = 0.7$. Le Tableau 2 présente les degrés de croyance que chaque objet en ligne, domine un autre objet en colonne.

Les objets o_i et o_j comparés sont supposés différents, ce qui fait que la relation $\geq$ suffit pour exprimer que o_i domine o_j (au sens de (2)). Remarquons que cette définition se réduit à la

1. Pour simplifier, nous utilisons wl et r pour désigner la perte de poids par mois et le remboursement, respectivement

Objects	o_1	o_2	o_3	o_4	o_5
o_1	1	1	0.7	0.3	0.92
o_2	0	1	0	0	0
o_3	0	0	1	0	0
o_4	0	1	1	1	0
o_5	0	0	0	0	1

TAB. 2: Croyance de Dominance.

relation de dominance de Pareto entre des objets certains, i.e., le degré de croyance est 1 si $o_i \succ o_j$ et 0 sinon. En s'appuyant sur cette relation de dominance, nous introduisons maintenant la notion de b-dominance (où b représente un seuil de croyance défini par l'utilisateur, $b \in]0, 1]$) :

Définition 6 *(La b-dominance) Etant donnés deux objets $o_i, o_j \in \mathcal{O}$ et un seuil de croyance b, o_i b-domine o_j, désigné par $o_i \succ_b o_j$, si et seulement si $bel(o_i \succ o_j) \geq b$.*

Par exemple, o_1 0.9-domine o_2 et o_4. Mais, il ne 0.9-domine pas o_4 car $bel(o_1 \succ o_4) = 0.3 < 0.9$. Intuitivement, un objet est dans le Skyline s'il n'est pas dominé par rapport à un certain seuil b.

Définition 7 *(b-dominant skyline) Le skyline de $\mathcal{O}$, désigné par b-$Sky_{\mathcal{O}}$, comprend les objets dans $\mathcal{O}$ qui ne sont pas b-dominés par aucun autre objet, i.e., b-$Sky_{\mathcal{O}} = \{o_i \in \mathcal{O} \mid \nexists o_j \in \mathcal{O}, o_j \succ_b o_i\}$.*

Par exemple, le 0.4-dominant skyline comprend les objets o_1 et o_4, car ils ne sont pas 0.4-dominés par aucun autre objet, par contre, le 0.2-dominant skyline contient seulement o_1 comme o_4 est 0.2 dominé par o_1. Partant de cette observation, nous illustrons une propriété clé du b-dominant Skyline.

Théorème 1 *Étant donnés deux seuils de croyance b et b', si $b < b'$ alors le b-dominant skyline est un sous-ensemble du b'-dominant skyline, i.e., $b < b' \Rightarrow b$-$Sky_{\mathcal{O}} \subseteq b'$-$Sky_{\mathcal{O}}$.*

Preuve 1 *Supposons qu'il existe un objet o_i tel que $o_i \in b$-$Sky_{\mathcal{O}}$ et $o_i \notin b'$-$Sky_{\mathcal{O}}$. Comme $o_i \notin b'$-$Sky_{\mathcal{O}}$, il existe un autre objet o_j qui b'-domine o_i. Ainsi, $bel(o_j \succ o_i) > b'$. Mais, $b < b'$. Donc, $bel(o_j \succ o_i) > b$, d'où, $o_j \succ_b o_i$, ce qui conduit à une contradiction car $o_i \in b$-$Sky_{\mathcal{O}}$.*

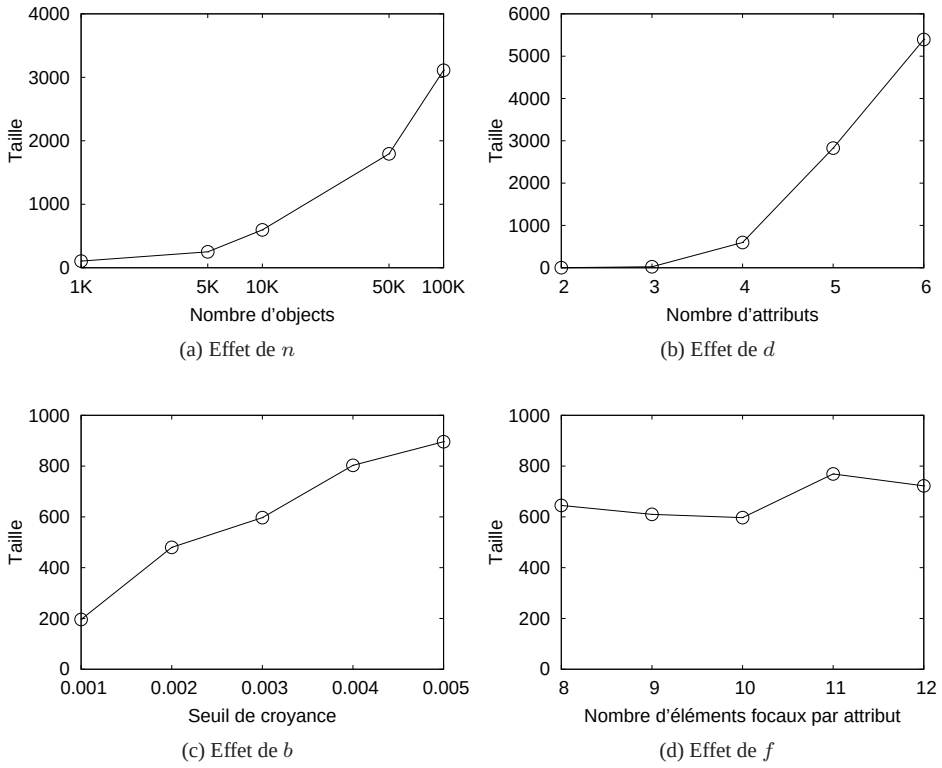
Le théorème 1 indique que la taille de b-dominant skyline est plus petite que celle de b'-dominant skyline si $b < b'$. Les utilisateurs ont donc la possibilité de contrôler la taille des objets que contient le Skyline évidentiel en faisant varier le seuil de croyance.

4 Étude Expérimentale

Dans cette étude, nous nous focalisons sur la taille (cardinalité) du skyline évidentiel. Pour ce faire nous avons conduit un ensemble d'expérimentations. La figure 1 montre les résultats

Paramètre	Symbole	Valeurs
Nombre d'objets	n	1K, 5K, **10K**, 50K, 100K
Nombre d'attributs	d	2, 3, **4**, 5, 6
Seuil de Croyance	b	0.001, 0.002, **0.003**, 0.004, 0.005
Nombre d'éléments focaux par attribut	f	8, 9, **10**, 11, 12

TAB. 3: Paramètres utilisés.



(a) Effet de n (b) Effet de d

(c) Effet de b (d) Effet de f

FIG. 1: Taille du skyline évidentiel.

obtenus. Dans chaque expérimentation., nous varions un seul paramètre, tandis que les autres paramètres prennent leurs valeurs par défaut. Le Tableau 3 montre ces paramètres et leurs symboles ; les valeurs par défaut sont en gras.

La figure 1.a montre que la taille du skyline évidentiel augmente avec l'augmentation de n. Dans la figure 1.b, on montre que la taille du skyline évidentiel augmente aussi de façon significative avec l'augmentation de d. Figure 1.c illustre que la taille du skyline augmente en augmentant b (ce qui signifie que le b-dominant skyline contient le b'-dominant skyline si $b > b'$). Contrairement à n, d et b, f n'a pas d'effet apparent sur la taille du skyline évidentiel.

5 Conclusion

Dans cet article, nous avons abordé le problème des requêtes skyline dans le cadre des bases de données évidentielles et nous avons introduit un nouveau type de skyline. Notre étude expérimentale a démontré la faisabilité et la flexibilité du skyline évidentiel. Comme perspective, nous envisageons de développer des techniques de classement des objets retournés par l'opérateur skyline évidentiel.

Références

Aggarwal, C. C. et P. S. Yu (2009). A survey of uncertain data algorithms and applications. *IEEE Trans. Knowl. Data Eng. 21*(5), 609–623.

Bell, D. A., J. W. Guan, et S. K. Lee (1996). Generalized union and project operations for pooling uncertain and imprecise information. *Data Knowl. Eng. 18*(2), 89–117.

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

Bosc, P. et O. Pivert (2010). Modeling and querying uncertain relational databases : a survey of approaches based on the possible worlds semantics. *UFKBS J*, 565–603.

Dalvi, N. N. et D. Suciu (2007). Efficient query evaluation on probabilistic databases. *VLDB J.*, 523–544.

Dempster, A. P. (1968). A generalization of bayesian inference. *Journal of the Royal Statistical Society 30*(B), 205–247.

Shafer, G. (1976). *A Mathematical Theory of Evidence*. Princeton : Princeton University Press.

Summary

The skyline operator is a powerful means in multi-criteria decision-making since it retrieves the most interesting objects according to a set of attributes. On the other hand, uncertainty is inherent in many real applications. One of the most powerful approaches used to model uncertainty is the evidence theory. Databases that manage such type of data are called evidential databases. In this paper, we tackle the problem of skyline analysis on evidential databases. We introduce a skyline model that is appropriate to the evidential data nature and we present an experimental evaluation.