Indexation dynamique pour la maintenance du skyline dans les flux de données

Rui Liu*, Dominique H. Li*,**

*Département d'Informatique, Université de Tours
**Laboratoire d'Informatique Fondamentale et Appliquée de Tours
{liurui, dominique.li}@univ-tours.fr

Résumé. Le calcul du skyline reçoit une attention intensive de la communauté des bases de données dont de nombreux algorithmes ont été développés au cours des deux dernières décennies. Cependant, la maintenance des skylines dans les flux de données est un défi car les mises à jour continues de skyline doivent tenir compte consécutivement de l'ajout de n-uplets entrants et la suppression des n-uplets expirés. Dans cet article, nous présentons RSS, une approche efficace basée sur l'indexation dynamique pour calculer des skylines dans les flux de données en fenêtre glissante. Notre analyse théorique prouve que la complexité temporelle de RSS est limitée par un sous-ensemble du skyline instantané ainsi que notre évaluation expérimentale montre l'efficacité de RSS sur les flux de données de haute et de faible dimension.

1 Introduction

Le calcul du skyline reçoit une attention intensive depuis la première introduction de la *requête skyline* (Borzsony *et al.* [2001]) qui vise à récupérer l'ensemble de n-uplets dominants dans des données multidimensionnelles, pour lequel de nombreux algorithmes efficaces ont été développés au cours des deux dernières décennies. Cependant, le problème de maintenance du skyline dans le contexte de flux de données est un vrai défi car il nécessite des mises à jour instantanées du skyline en concernant l'arrivée de nouvelles données et l'expiration de données trop précoces.

Dans cet article, nous présentons une approche efficace pour calculer les skylines dans les flux de données avec la *fenêtre glissante* (Patroumpas and Sellis [2006]). En général, le modèle de fenêtre glissante est basé soit sur le *comptage* qui couvre un nombre d'enregistrements (*n-uplets*) les plus récents à chaque instant, soit sur le temps (*temporelle*) qui est limité par un nombre d'unités de temps coïncidant avec les *timestamps* de données dans le flux. La figure 1 montre un exemple de maintenance du skyline en fenêtre glissante sur la relation prix-distance, où nous considérons les prix (axe Y) d'hôtels par rapport à leurs distances (X axe) vers un endroit, comme le centre-ville, la plage ou la gare, etc. Considérons une fenêtre de comptage $W = 5$ et l'ordre d'arrivée des enregistrements d'hôtel de a à f, alors il y a initialement 4 n-uplets skyline $\{a, c, d, e\}$ illustré dans la figure 1(a). Au moment où f arrive, le premier n-uplet a doit être éliminé afin de conserver la taille de la fenêtre. En conséquence, puisque a est le seul n-uplet qui domine b, b devient un n-uplet skyline lorsque a est écarté ; de plus, le