Extraction des k plus grandes tuiles dans un flux de données

Toon Calders[1], Elisa Fromont[2], Baptiste Jeudy[2],
Hoang Thanh Lam[1], Wenjie Pei[1], et Adriana Prado[2]

[1]TU Eindhoven,Department of Maths and Computer Science, Eindhoven, Netherlands
[2]Laboratoire Hubert Curien, UMR CNRS 5516, Saint-Etienne, France

Résumé

Les grandes tuiles dans les base de données sont les itemsets ayant l'aire la plus importante. L'aire est définie comme étant la fréquence de l'itemset dans la base multipliée par sa taille, i.e. le nombre d'items qu'il contient. Les grandes tuiles permettent de représenter une partie importante de la base de données, leur recherche est donc un problème important de fouille de motifs. Nous proposons de rechercher les k plus grandes tuiles apparaissant dans un flux de données en utilisant une technique de fenêtre glissante. Une approche naïve et inefficace puisque ce calcul est connu comme étant NP-dur et non-approximable, consisterait à recalculer ces k plus grandes tuiles à chaque fois que la fenêtre glissante se déplace. Pour s'attaquer à ce problème, nous proposons une autre approche basée sur le calcul incrémental d'un résumé du flux de données. Les k plus grandes tuiles sont alors calculées à partir de ce résumé. Cette technique est beaucoup plus efficace que dans le cas naïf quand la taille de la fenêtre est suffisamment petite. Nous proposons également un algorithme approximatif avec des garanties théoriques sur le taux d'erreur des résultats pour des tailles de fenêtre plus importantes. Des expériences sur deux jeux de données réelles montrent que notre algorithme approximatif est jusqu'à 100 fois plus rapide que notre solution exacte ou que les algorithmes de l'état de l'art adaptés pour répondre à notre problème. En outre, cet algorithme approximatif permet de retrouver les k plus grandes tuiles dans chaque fenêtre de manière précise avec un taux négligeable de faux positifs et faux négatifs sur les données réelles.

Mots-clef : Fouille de flux de données, Recherche de tuiles.