Tableau de Bits Indexé (TBI) pour la Recherche de Séquences Fréquentes

Lionel Savary, Karine Zeitouni

Laboratoire PRiSM, Université de Versailles, 45 Avenue des Etats-Unis, 78035 Versailles

{Lionel.Savary, Karine.Zeitouni}@prism.uvsq.fr

A la différence de la fouille d'articles fréquents, la recherche de sous-séquences fréquentes tient compte de l'apparition multiple et de l'ordre des articles. L'algorithme proposé parcourt la base de données une seule fois. Durant cette passe, il construit un vecteur *VS* contenant toutes les combinaisons de séquences présentes dans la base. A ce vecteur est associé un tableau de bit *TB* codant toutes les séquences de la base en correspondance avec les articles codés dans *VS*. Les bits à 1 indiquent les articles présents dans la séquence et les bits à 0 ceux qui ne le sont pas. Les séquences sont représentées dans chaque ligne du tableau et regroupées par taille dans l'ordre décroissant. Un index associé au tableau permet de pointer directement les séquences de taille choisie. Ce qui évite des comparaisons superflues et améliore les performances. Le tableau *NB* associé au *TB*, indique les fréquences associées à chaque séquence. Dans l'exemple de la figure 1, la séquence (M) de taille 1 se trouve à la première ligne dans le *TB* et a une fréquence de 500. Cette structure est construite dynamiquement au cours de l'unique passe dans la base de données. Un deuxième algorithme TBI2, basé sur un tableau de booléens, offre de meilleures performances mais nécessite plus d'espace mémoire. TBI et TBI2 affichent de meilleures performances que les algorithmes existants tel que Prefixspan [1].

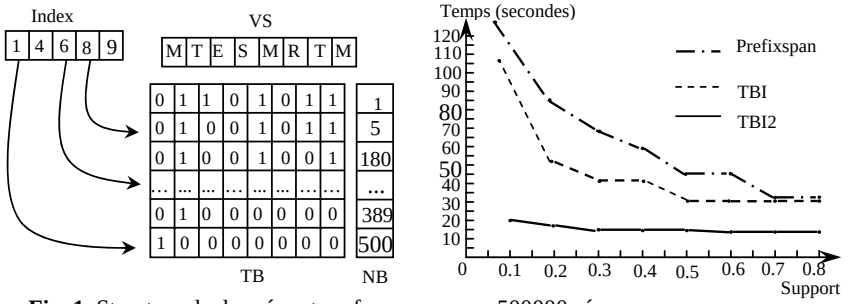

Fig. 1. Structure de données et performance pour 500000 séquences

Références

1. J. Pei, J. Han, B. Mortazavi, H. Pinto, Q. Chen, U.Dayal, and M-C. Hsu. PrefixSpan: mining sequential patterns efficiently by prefix-projected pattern growth. Proceedings of the 17th International Conference on Data Engineering, 215-224, Heidelberg, Germany, Apr. 2001.