Entrepôt de Données Spatiales basé sur GML: Politique de Gestion de Cache

Lionel Savary , Georges Gardarin, Karine Zeitouni

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

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

Motivation : Dans les entrepôts de données, la manipulation de gros volumes de données requière souvent un temps d'exécution important. En particulier, si les requêtes portent sur des données spatiales contenues dans des documents semi-structurés, les temps de réponse deviennent prohibitifs. Afin de réduire le temps de traitement imposé par l'utilisation d'opérateurs spatiaux dans ce type de document, nous proposons une politique de remplacement de cache adaptée aux documents GML. Cette politique prend en compte les données spatiales et non-spatiales, ainsi que le nombre d'opérateurs spatiaux présents dans les requêtes utilisateurs.

Politique de remplacement de cache : Soit Dq le nouveau document de taille Tq à insérer dans le cache. On désigne par Ti la taille du document i ($1 \leq i \leq n$) du cache et par C_{GMLi} le coût d'accès au document du cache. Soient $(Xi)_{i=1..n} \in \{0 ; 1\}^n$ tel que Xi = 1 si le document i est conservé dans le cache, 0 s'il est supprimé. Notons de plus D_{GMLj} le coût d'accès au document j sur disque (lorsqu'il n'est pas en cache). On recherche alors les documents i du cache à supprimer tels que la somme des coûts d'accès soit la plus petite possible:

$$\text{Minimiser Coût d'accès} = \sum_{j=1}^{n} Xj * C_{GMLj} + \sum_{j=1}^{n} (1-Xj) * D_{GMLj}$$

Une contrainte est que la somme des tailles des documents éliminés du cache soit supérieure ou égale à Tq: $\sum_{i=1}^{n} Xi * Ti \geq Tq$

C_{GMLi} représente le coût d'accès à un document GML i en cache, calculé selon la formule d'Arlitt [1], soit $C_{GMLi} = L + F_i * C_i / Si$, avec : L une constante ; Si la taille du document ; Fi la fréquence d'accès au document et Ci le coût pour une requête sur des données géographiques. Notre calcul de C_i tient compte du coût sur les données non-spatiales, du coût sur les données spatiales, ainsi que du nombre d'opérateurs spatiaux présents dans la requête. Le problème est de déterminer les Xi qui optimisent le coût total. Nous proposons une adaptation d'algorithmes classiques de recherche opérationnelle pour déterminer les documents à conserver.

Références

1. M. Arlitt, R. Friedrich L. Cherkasova,J. Dilley, and T. Jin. Evaluating content management techniques for web proxy caches. In HP Tec. report, Palo Alto, Apr. 1999.