Des fonctions d'oubli intelligentes dans les entrepôts de données

Aliou Boly*, Sabine Goutier**, Georges Hébrail*·**
*46, Rue Barrault, 75634 PARIS Cedex 13 - FRANCE
boly@enst.fr, hebrail@enst.fr
**1, Av. du Général de Gaulle, 92141 CLAMART Cedex - FRANCE
georges.hebrail@edf.fr, sabine.goutier@edf.fr

Résumé. Les entrepôts de données stockent des quantités de données de plus en plus massives et arrivent vite à saturation. Un langage de spécifications de fonctions d'oubli est défini pour résoudre ce problème. Dans le but d'offrir la possibilité d'effectuer des analyses sur l'historique des données, les spécifications définissent des résumés par agrégation et par échantillonnage à conserver parmi les données à 'oublier'. Cette communication présente le langage de spécifications ainsi que les principes et les algorithmes pour assurer de façon mécanique la gestion des fonctions d'oubli.

1 Introduction

De nos jours, bien que les moyens de stockage soient de plus en plus performants et de moins en moins chers, les entrepôts de données arrivent vite à saturation et la question des données à conserver sous forme d'historique va se poser rapidement. Il faut donc choisir quelles données doivent être archivées, et quelles données doivent être conservées actives dans les entrepôts de données. La solution qui est appliquée en général est d'assurer un archivage périodique des données les plus anciennes. Cette solution n'est pas satisfaisante car l'archivage et la remise en ligne des données sont des opérations coûteuses au point que l'on peut considérer que des données archivées sont des données perdues (en pratique inutilisables dans le futur) du point de vue de leur utilisation dans le cadre d'une analyse des données.

Dans cette communication, nous proposons une solution pour éviter la saturation des entrepôts de données. Un langage de spécifications de fonctions d'oubli des données anciennes est défini pour déterminer les données qui doivent être présentes dans l'entrepôt de données à chaque instant. Ces spécifications de fonctions d'oubli conduisent à supprimer de façon mécanique les données à 'oublier', tout en conservant un résumé de celles-ci par agrégation et par échantillonnage. L'agrégation et l'échantillonnage constituent deux techniques standard et complémentaires pour résumer des données. Considérons un entrepôt de données d'analyse des click-stream sur les sites web. Avec le temps, les données détaillées anciennes deviennent de moins en moins 'utiles' et peuvent donc être agrégées par jour ou par mois par exemple. En plus d'agréger des données, on peut conserver certaines données jugées intéressantes ou choisies de façon aléatoire dans le but de pouvoir effectuer des analyses sur les données de l'entrepôt.

Le langage de spécifications est défini dans le cadre du modèle relationnel : sur chaque table, est défini au moyen de spécifications un ensemble de n-uplets à archiver. Pour des raisons applicatives, parmi les n-uplets à archiver, des échantillons peuvent être conservés dans le cadre de l'utilisation de l'entrepôt. De plus, des algorithmes pour mettre à jour le