REGLO : une nouvelle stratégie pour résumer un flux de séries temporelles

Alice Marascu*, Florent Masseglia* and Yves Lechevallier**

*INRIA Sophia-Antipolis, 2004 route des lucioles - BP 93 , 06902 Sophia-Antipolis
**INRIA Paris-Rocquencourt, Domaine de Voluceau , 78150 Rocquencourt
Prenom.Nom@inria.fr

Résumé. Les flux de séries temporelles sont aujourd'hui produits dans de nombreux domaines comme la finance (Zhu et Shasha (2002)), la surveillance de réseaux (Borgne et al. (2007); Airoldi et Faloutsos (2004)), la gestion de l'historique des usages fréquents (Giannella et al. (2003); Teng et al. (2003)), etc. Résumer de tels flux est devenu un domaine important qui permet de surveiller et d'enregistrer des informations fiables sur les séries observées. À ce jour, la majorité des algorithmes de ce domaine s'est concentrée sur des résumés séparés et indépendants (Giannella et al. (2003); Zhu et Shasha (2002); Chen et al. (2002)), en accordant à chaque série le même espace en mémoire. Toutefois, la gestion de cet espace mémoire est un sujet important pour les flux de données et une stratégie accordant la même quantité de mémoire à chaque série n'est pas forcément appropriée. Dans cet article, nous considérons que les séries doivent être en compétition vis à vis de l'espace mémoire, selon leur besoin de précision. Ainsi, nous proposons : (1) une stratégie de gestion de l'espace mémoire optimisée et (2) une nouvelle méthode de résumé des séries temporelles par approximation. Dans ce but, nous observons à la fois l'erreur globale et les erreurs locales. La répartition de la mémoire suit les étapes suivantes : (1) recherche de la séquence la mieux représentée et (2) recherche de la partie à compresser en minimisant l'erreur. Nos expérimentations sur des données réelles montrent l'efficacité et la pertinence de notre approche.

1 Introduction

Les résumés de séries temporelles sont l'objet de travaux de plus en plus importants en raison de leurs nombreuses applications mais aussi de leur complexité. Parmi les applications liées à ce domaine citons la recherche d'information (Ding et al. (2008); Palpanas et al. (2008)), la fouille (Chen et al. (2002); Papadimitriou et al. (2005); Lin et al. (2003)) ou encore la prévision (Cheng et Tan (2008); Borgne et al. (2007)) de données. La difficulté associée à la production de ces résumés est due aux très grands volumes de données (qui sont produites en continu et à grande vitesse). En effet, ces données arrivant sous forme de flux, le temps de traitement de chaque donnée doit être le plus faible possible. De plus, la taille d'un flux étant potentiellement illimitée, il est inévitable de traiter des problèmes liés à la mémoire nécessaire pour le résumer.