Extraction de motifs séquentiels dans les flots de données d'usage du Web

Alice Marascu, Florent Masseglia

INRIA Sophia Antipolis, 2004 route des Lucioles - BP 93, 06902 Sophia Antipolis, France
{Alice.Marascu,Florent.Masseglia}@sophia.inria.fr

Résumé. Ces dernières années, de nouvelles contraintes sont apparues pour les techniques de fouille de données. Ces contraintes sont typiques d'un nouveau genre de données : les "*data streams*". Dans un processus de fouille appliqué sur un data stream, l'utilisation de la mémoire est limitée, de nouveaux éléments sont générés en permanence et doivent être traités le plus rapidement possible, aucun opérateur bloquant ne peut être appliqué sur les données et celles-ci ne peuvent être observées qu'une seule fois. A l'heure actuelle, la majorité des travaux relatifs à l'extraction de motifs dans les data streams ne concernent pas les motifs temporels. Nous montrons dans cet article que cela est principalement dû au phénomène combinatoire qui est lié à l'extraction de motifs séquentiels. Nous proposons alors un algorithme basé sur l'alignement de séquences pour extraire les motifs séquentiels dans les data streams. Afin de respecter la contrainte d'une passe unique sur les données, une heuristique gloutonne est proposée pour segmenter les séquences. Nous montrons enfin que notre proposition est capable d'extraire des motifs pertinents avec un support très faible.

1 Introduction

Le problème de l'extraction de motifs séquentiels dans un grand ensemble de données statiques a été largement étudié ces dernières années (Agrawal et Srikant (1995), Masseglia et al. (1998), Pei et al. (2001), Wang et Han (2004), Kum et al. (2003)). Les schémas extraits sont utiles dans de nombreuses applications comme le marketing, l'aide à la décision, l'analyse des usages, etc. Depuis peu, des applications émergentes comme (entre autres) l'analyse du trafic réseaux, la détection de fraude ou d'intrusion, la fouille de clickstream[1] ou encore l'analyse des données issues de capteurs ont introduits de nouveaux types de contraintes pour les méthodes de fouille. Ces applications ont donné lieu à une forme de données connues sous le nom de "data streams". Dans le contexte des data streams l'utilisation de la mémoire doit être réduite, les données sont générées de manière continue et très rapide, les opérations bloquantes ne sont pas envisageables et, enfin, les nouvelles données doivent être prises en compte aussi vite que possible. Ainsi, de nombreuses méthodes ont été proposées pour extraire des items ou des motifs dans les data streams (Datar et al. (2002), Chang et Lee (2003), Cormode et Muthukrishnan

[1] clickstream : flot de requêtes d'un utilisateur sur un site Web