Résumé hybride de flux de données par échantillonnage et classification automatique

Nesrine Gabsi*,**, Fabrice Clérot **
Georges Hébrail*

*Institut TELECOM ; TELECOM ParisTech ; CNRS LTCI
46, rue Barrault 75013 Paris
PrénomAuteur.NomAuteur@telecom-paristech.fr,
** France Telecom RD
2, avenue P.Marzin 22307 Lannion
PrénomAuteur.NomAuteur@orange-ftgroup.com

Résumé. Face à la grande volumétrie des données générées par les systèmes informatiques, l'hypothèse de les stocker en totalité avant leur interrogation n'est plus possible. Une solution consiste à conserver un résumé de l'historique du flux pour répondre à des requêtes et pour effectuer de la fouille de données. Plusieurs techniques de résumé de flux de données ont été développées, telles que l'échantillonnage, le clustering, etc. Selon le champ de requête, ces résumés peuvent être classés en deux catégories: résumés spécialisés et résumés généralistes. Dans ce papier, nous nous intéressons aux résumés généralistes. Notre objectif est de créer un résumé de bonne qualité, sur toute la période temporelle, qui nous permet de traiter une large panoplie de requêtes. Nous utilisons deux algorithmes : CluStream et StreamSamp. L'idée consiste à les combiner afin de tirer profit des avantages de chaque algorithme. Pour tester cette approche, nous utilisons un Benchmark de données réelles "KDD_99". Les résultats obtenus sont comparés à ceux obtenus séparément par les deux algorithmes.

1 Introduction

Il existe actuellement plusieurs applications qui génèrent des informations en très grande quantité. Ces applications sont issues de domaines variés tels que la gestion du trafic dans un réseau IP. Lorsque le volume de données augmente, il devient très coûteux de stocker toutes les données avant de les analyser : il est judicieux d'adopter un traitement à la volée pour ces informations. Un nouveau mode de traitement de l'information émerge. Il s'agit du traitement de flux de données. Dans (Golab et Özsu, 2003), les auteurs définissent un flux de données comme étant une séquence d'items continue, ordonnée, arrivant en temps réel avec des débits importants.

Plusieurs travaux ((Babcock et al., 2002), (Golab et Özsu, 2003), (Ma et al., 2007), (Towne et al., 2007)) montrent que les Systèmes de Gestion de Base de Données (SGBD) sont inadaptés pour ce type d'applications. Ceci est essentiellement dû à la nature continue du flux