Recherche en temps réel de préfixes massifs hiérarchiques dans un réseau IP à l'aide de techniques de stream mining

Pascal Cheung-Mon-Chan*, Fabrice Clérot*

* France Télécom R&D
2, avenue Pierre Marzin BP 50702
22307 Lannion Cedex -France
{pascal.cheungmonchan, fabrice.clerot}@francetelecom.com

Résumé. Au cours de ces dernières années, de nombreuses techniques de stream mining ont été proposées afin d'analyser des flux de données en temps réel. Dans cet article, nous montrons comment nous avons utilisé des techniques de stream mining permettant la recherche d'objets massifs hiérarchiques (hierarchical heavy hitters) dans un flux de données pour identifier en temps réel dans un réseau IP les préfixes dont la contribution au trafic dépasse une certaine proportion de ce trafic pendant un intervalle de temps donné.

1 Introduction

Les progrès techniques récents ont eu pour conséquence l'augmentation du nombre de flux d'information et la croissance rapide de leurs débits. L'architecture traditionnelle de l'analyse de données — où les données, préalablement stockées, sont analysées puis rafraîchies — étant inadaptée au traitement de ces flux, une nouvelle famille de techniques, dites de stream mining, se propose d'inverser radicalement cette architecture et de mettre en oeuvre des systèmes reposant sur des capacités de stockage minimales qui sont mises à jour à la vitesse du flux. L'objectif de cet article est d'expliquer comment nous avons utilisé des techniques de stream mining afin d'identifier en temps réel, dans un réseau IP, les préfixes dont la contribution au trafic dépasse une certaine proportion de ce trafic pendant un intervalle de temps donné.

2 La recherche d'objets massifs hiérarchiques dans un flux de données

2.1 La notion d'objet massif hiérarchique

Les flux de données que nous allons considérer ici sont de la forme $(i_t, c_t)_{t\in\mathbb{N}}$ où, pour tout instant $t \in \mathbb{N}$, l'*identifiant* i_t appartient à un ensemble fini $\boldsymbol{U}$ et la *marque* c_t est un nombre réel positif ou nul. Dans cet article, l'identifiant i_t correspondra à une adresse IP, par exemple l'adresse destination d'un paquet IP transitant en un point P donné d'un réseau, l'ensemble fini $\boldsymbol{U}$ correspondra à l'ensemble des adresses IP v4 (autrement dit chaque adresse comportera 32 bits) et la *marque* c_t correspondra au nombre d'octets transportés par le paquet