

Détection de changement dans les profils en ligne d'utilisateurs

Parisa Rastin, Basarab Matei, Guénaél Cabanes

LIPN-CNRS, UMR 7030, Université Paris 13
rastin@lipn.univ-paris13.fr

Résumé. L'analyse des données dynamique est difficile. En effet, la structure de telles données évolue dans le temps, potentiellement à une vitesse très rapide. De plus, les objets dans ces ensembles de données sont souvent complexes. Dans cet article, notre motivation pratique est d'analyser l'évolution des profils en ligne d'utilisateurs, c'est-à-dire de suivre la localisation géographique des utilisateurs ainsi que leurs traces de navigation en ligne afin de détecter des changements dans leurs habitudes et leurs intérêts. Nous proposons un nouveau cadre dans lequel nous créons d'abord, pour chaque utilisateur, des signaux de l'évolution de ses intérêts et des localisations physiques enregistrées au cours de sa navigation. Ensuite, nous détectons automatiquement les changements d'intérêt ou de lieu grâce à un nouvel algorithme de détection de sauts dans les signaux.

1 Introduction

L'un des défis principaux de l'exploration de données est la détection de changement dans les ensembles de données dynamiques. Ce phénomène est connu sous le nom de "dérive de concept" (Gama, 2010; Silva et al., 2013). Une des applications directes, qui constitue notre intérêt pratique dans le présent document, est la détection de changement dans l'intérêt des utilisateurs sur la base des données enregistrées au cours de leur navigation en ligne. Cette tâche, appelée "profilage de l'utilisateur", revêt une grande importance économique pour les entreprises du secteur de la publicité en ligne. Les tâches de profilage visent à reconnaître les "états d'esprit" des utilisateurs à partir de leurs navigations sur différents sites Web ou leurs interactions avec des "points de contact" numériques (différentes façons dont une marque interagit et affiche des informations aux utilisateurs). Il est très important de pouvoir détecter les changements d'intérêt d'un utilisateur ou son déménagement dans une autre ville ou un autre pays afin d'ajuster la stratégie publicitaire le concernant. Ces profils sont calculés à partir d'une très grande base de données de navigation sur Internet, qui liste les séquences d'URL ou les points de contact visités par un grand nombre de personnes. Chaque URL d'un "point de contact" est caractérisée par des informations contextuelles et sémantiques. Dans ce contexte, chaque utilisateur est décrit comme une série temporelle de catégories d'URL et d'emplacements physiques. Les catégories d'URL sont calculées à l'aide d'une approche de classification adaptée aux données complexes (Rastin et al., 2016; Rastin et Matei, 2018). Les emplacements sont enregistrés à l'aide des informations de géolocalisation collectées lors de la navigation de l'utilisateur, mais sont limités à une série de codes postaux. La détection de changements dans les