Une méthode pour la détection de thématiques populaires sur Twitter

Adrien Guille, Cécile Favre

Laboratoire ERIC, Université Lumière Lyon 2
http://mediamining.univ-lyon2.fr/people/guille/egc2014.php

Résumé. L'explosion du volume de messages échangés via Twitter entraîne un phénomène de surcharge informationnelle pour ses utilisateurs. Il est donc crucial de doter ces derniers de moyens les aidant à filtrer l'information brute, laquelle est délivrée sous la forme d'un flux de messages. Dans cette optique, nous proposons une méthode basée sur la modélisation de l'anomalie dans la fréquence de création de liens dynamiques entre utilisateurs pour détecter les pics de popularité et extraire une liste ordonnée de thématiques populaires. Les expérimentations menées sur des données réelles montrent que la méthode proposée est capable d'identifier et localiser efficacement les thématiques populaires.

1 Introduction

Twitter offre des fonctionnalités de microblogging qui sont utilisées par des millions de personnes à travers le monde pour publier des messages courts. Ces personnes créent et partagent de l'information liée à divers types d'évènements, allant d'évènements personnels banals à des évènements importants et/ou globaux, quasiment en temps-réel. L'explosion du nombre d'utilisateurs de ce réseau a entraîné l'apparition d'un phénomène de surcharge informationnelle. Pour lutter contre cela, il est nécessaire de doter les utilisateurs de moyens leur permettant d'identifier plus facilement les éléments d'information les plus intéressants et de se tenir au courant des derniers évènements significatifs.

L'information brute produite par Twitter est délivrée sous la forme d'un flux de messages. Par conséquent la manière dont ceux-ci arrivent au fil du temps recèle une part importante de leur signification. La dynamique temporelle des thématiques les plus populaires est constituée d'une succession de focus et dé-focus, autrement dits, une succession de *pics* de popularité. C'est pourquoi de nombreuses approches – allant de méthodes basées sur la fréquence des mots jusqu'à des méthodes plus complexes reposant sur des modèles de thématiques probabilistes dynamiques – ont été proposées dans le but d'identifier ce genre de thématiques. Ces méthodes reposent sur des stratégies variées de détection des pics et produisent des résultats très différents. Nos travaux s'intéressent au filtrage et à l'identification de thématiques à partir de l'information contenue dans un flux de messages produits par Twitter afin de, entre autres, fournir une vue rétrospective des thématiques les plus populaires ou bien recommander des éléments d'information intéressants en temps réel. Une bonne solution doit satisfaire deux critères : d'une part les thématiques identifiées doivent être précisément localisées dans le temps et intelligibles et d'autre part, la méthode doit pouvoir passer à l'échelle et traiter de grands