Extraction de chaînes cohérentes en vue de reconstuire la Trajectoire de l'information

Charles Huyghues-Despointes[*,**], Leila Khouas[**], Julien Velcin[*] et Sabine Loudcher[*]

[*]Université de Lyon, Lyon 2, ERIC EA 3083, France
[**]Bertin IT

Résumé. Sur Internet, l'information se propage en particulier au travers des documents textuels. Cette propagation soulève de nombreux défis : identifier une information, suivre son évolution dans le temps, comprendre les mécanismes qui régissent sa propagation, etc. Étant donné un document parmi un grand corpus dans lequel de nombreuses informations circulent, pouvons-nous retrouver les chemins empruntés par l'information pour arriver à ce document ? Nous proposons de définir la notion de trajectoire comme l'ensemble des chemins le long desquels de l'information s'est propagée et nous proposons une méthode pour l'estimer. Nous avons mis en œuvre une évaluation humaine pour juger de la qualité des chemins calculés. Nous montrons que les évaluations concordent la plupart du temps et que notre algorithme est efficace pour retrouver les bons chemins.

1 Introduction

L'information se propage. Lorsqu'elle est reçue, une information est ingérée, nuancée, et reformulée pour être à nouveau transmise. Cette propagation se déroule à tous les niveaux de communication : lors d'une conversation, à la radio, à la télévision, mais aussi lorsque nous publions du contenu, par exemple sur Internet. Les documents que nous partageons, contiennent de multiples informations provenant d'autres documents, et qui seront, en partie, reprises dans le futur. Ainsi les informations présentes dans un document ont une histoire. Ce sont des suites d'événements de propagations qui les ont conduites à être présentes dans ce document. Nous appelons l'ensemble de ces lignées, pour chaque document, la Trajectoire de l'information.

Lorsqu'une information se propage, elle est sujette à des modifications, dans sa forme ou dans son fond. Certains travaux se sont intéressés à la traque de ces changements, comme Leskovec et al. (2009). Cependant, après de nombreuses mutations, il peut être difficile de trouver le lien entre l'information de départ et l'information actuelle, comme le soulignent des travaux cherchant à retrouver les sources d'une information, par exemple Farajtabar et al. (2015).

Nous proposons d'estimer la Trajectoire de l'information en calculant des chaînes de documents textuels le long desquelles il est plausible que de l'information se soit propagée. Pour ce faire nous n'explicitons pas l'information qui circule le long de la chaîne, mais nous intéressons à la manière dont se comportent les documents entre eux au sein de la chaîne. Aborder