Analyse et classification d'urls à des fins de segmentation automatique par groupes de ciblage

Matthieu Brito Antunes
Data Scientist
Tradelab - Paris
avril 2020

Table des matières

Liste des tableaux

Table des figures

1. Introduction

L'ensemble des adresses web (URL) des pages visitées sur Internet peut constituer une source pertinente lorsque l'on souhaite réfléchir à une méthode de ciblage publicitaire. Les informations extraites à partir du traitement des URL permettent de comprendre les mécanismes qui régissent le regroupement d'utilisateurs, ou *segmentation*.

La segmentation d'utilisateurs est une étape cruciale dans le processus de ciblage publicitaire en ligne et peut s'avérer très chronophage. Il apparaît donc essentiel de chercher à utiliser toute information qui permettrait de comprendre comment tendre vers son automatisation totale.

Les informations contenues dans la structure des URL nous renseignent principalement sur l'organisation globale d'un site web, et la manière dont sont créées les URL associées aux nouvelles pages. Ceci permet de construire des regroupements de pages web sans avoir à analyser leur contenu. Ces regroupements pourraient idéalement être réalisés de manière automatique, à partir d'un modèle de *clustering* qui recevrait une URL en entrée et fournirait en sortie le groupe auquel elle appartient après avoir analysé sa structure.

Ce rapport présente le travail de construction d'un modèle d'analyse et de classification automatique d'URL. La première partie du rapport est centrée sur l'étude de la problématique et des enjeux associés. Diverses pistes d'analyse de structure d'URL sont proposées. La seconde partie du rapport offre une vision théorique de la solution proposée pour automatiser la création de segments à partir des données d'URL collectées chaque jour chez Jellyfish France. La dernière partie du rapport présente la mise en place de la solution et les résultats obtenus sur des données récoltées sur une période fixe.

2. Analyse de la structure des URL et premiers essais de segmentation automatique

2.1. Structure classique d'une URL

Sur Internet, les pages web regroupent divers formats de données (texte, image, son...) et sont repérées par une chaîne de caractères renseignant deux choses :

— leur emplacement

— le protocole internet permettant d'y accéder [1]

Cette chaîne de caractères, appelée *adresse web* ou encore URL (pour *Uniform Resource Locator*) est générée pour chaque nouvelle page ajoutée à un site, et permet donc à l'utilisateur de savoir où il se trouve. Encore faut-il qu'il sache la lire ! On retrouve en effet plusieurs éléments constitutifs d'une URL, qui peuvent être présent ou absents selon le type de site sur lequel l'utilisateur se trouve. Globalement, une URL peut être décrite comme suit :

coucou Spiky P

1. Les protocoles les plus connus sont ceux que l'on rencontre tous les jours dans la barre d'adresse de notre navigateur : `http` et `https` pour *HyperText Transfer Protocol (Secure)*.

coucou Spiky P

3. Approche sémantique de l'analyse des URL et segmentation automatique par la méthode des n-grams