Ecole Polytechnique de l'Université François Rabelais de Tours
Département Informatique
64 avenue Jean Portalis
37200 Tours, France
Tél. +33 (0)2 47 36 14 14
polytech.univ-tours.fr

Projet Architecture, Système et Réseaux

2018-2019

Crawling web et requête HTTP par serveur proxy

Tuteur académique
Mathieu Delalandre

Étudiants
Thomas Couchoud (DI5)
Victor Coleau (DI5)

21 septembre 2018

Liste des intervenants

Nom	Email	Qualité
Thomas Couchoud	thomas.couchoud@etu.univ-tours.fr	Étudiant DI5
Victor Coleau	victor.coleau@etu.univ-tours.fr	Étudiant DI5
Mathieu Delalandre	mathieu.delalandre@univ-tours.fr	Tuteur académique, Département Informatique

Avertissement

Ce document a été rédigé par Thomas Couchoud et Victor Coleau susnommés les auteurs.

L'Ecole Polytechnique de l'Université François Rabelais de Tours est représentée par Mathieu Delalandre susnommé le tuteur académique.

Par l'utilisation de ce modèle de document, l'ensemble des intervenants du projet acceptent les conditions définies ci-après.

Les auteurs reconnaissent assumer l'entière responsabilité du contenu du document ainsi que toutes suites judiciaires qui pourraient en découler du fait du non respect des lois ou des droits d'auteur.

Les auteurs attestent que les propos du document sont sincères et assument l'entière responsabilité de la véracité des propos.

Les auteurs attestent ne pas s'approprier le travail d'autrui et que le document ne contient aucun plagiat.

Les auteurs attestent que le document ne contient aucun propos diffamatoire ou condamnable devant la loi.

Les auteurs reconnaissent qu'ils ne peuvent diffuser ce document en partie ou en intégralité sous quelque forme que ce soit sans l'accord préalable du tuteur académique et de l'entreprise.

Les auteurs autorisent l'école polytechnique de l'université François Rabelais de Tours à diffuser tout ou partie de ce document, sous quelque forme que ce soit, y compris après transformation en citant la source. Cette diffusion devra se faire gracieusement et être accompagnée du présent avertissement.

Pour citer ce document

Thomas Couchoud et Victor Coleau, *Crawling web et requête HTTP par serveur proxy*, Projet Architecture, Système et Réseaux, Ecole Polytechnique de l'Université François Rabelais de Tours, Tours, France, 2018-2019.

```
@mastersthesis{
    author={Couchoud, Thomas and Coleau, Victor},
    title={Crawling web et requête HTTP par serveur proxy},
    type={Projet Architecture, Système et Réseaux},
    school={Ecole Polytechnique de l'Université François Rabelais de Tours},
    address={Tours, France},
    year={2018-2019}
}
```

Document mis en forme sous LaTeX
grâce à la classe de document `polytech.cls` V2.2.5 (2017/05/02)

Table des matières

Table des figures

Table des figures

Première partie

Introduction

Dans notre société moderne l'Internet occupe une place très importante. Il offre une quantité pharaonique d'information en libre accès. Parmi ces sources d'information, on trouve notamment des « wikis » qui sont des encyclopédies collaboratives permettant la large diffusion de données.

Malgré leur apparente générosité et leur connaissance des pratiques, ces sites n'apprécient que peu que les données qu'ils fournissent en soient extraites.

Cette nouvelle mane d'information attire les convoitises et polarise les comportements. D'un côté nous retrouvons les collecteurs de données cherchant à en agréger et stocker de plus en plus de leur propre chef. De l'autre les sites mettant à disposition l'information dont le but paradoxal est de fournir gratuitement tout en conservant l'exclusivité.

Cela entraine une guerre technologique entre crawlers et sites web. Les premiers développent des technologies de plus en plus efficaces, rapides et discrètes. Les seconds cherchent à contrecarrer les premiers grâce à des techniques de détection de plus en plus sofistiquées.

Le but de ce projet est d'étudier à la fois les techniques mises en place par les crawler pour se rendre invisible et celles mise en place par les sites pour se défendre. Cette recherche se concrétisera par la réalisation d'un crawler effectuant ses connexions au travers d'un proxy.

Figure 1 – *Crawler*

Deuxième partie

Veille technique

1 Les proxy

2 Stratégies de masquage

Rotation d'IP Rotation des UserAgent Gestion de sessions Limitation du nombre de requêtes Liste noire d'IP Réessais

Troisième partie

Application : Réalisation d'une crawler avec proxy

Quatrième partie

Conclusion

Annexes

Crawling web et requête HTTP par serveur proxy

Résumé

Mots-clés

Abstract

Keywords

Tuteur académique
Mathieu Delalandre

Étudiants
Thomas Couchoud (DI5)
Victor Coleau (DI5)