

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)



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}
}
```

Table des matières

Liste des intervenants	a
Avertissement	b
Pour citer ce document	c
Table des matières	i
Table des figures	iii
I Introduction	1
II Veille technique	2
1 Les proxy	3
1 Késako ?	3
1.1 Avantages	3
1.2 Inconvénients.....	3
2 Description technique	4
2 Stratégies de masquage	5
3 Stratégies de défense	6

III	Application : Réalisation d'une crawler avec proxy	7
IV	Conclusion	8
Annexes		9



Table des figures

Table des figures

1	Crawler	1
---	---------------	---

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 sophistiquées.

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

1 Késako ?

Un proxy est un composant logiciel servant d'intermédiaire entre deux hôtes afin de faciliter ou de surveiller leurs échanges.

1.1 Avantages

Les proxys peuvent par exemple servir à contourner certains filtrages. Supposons le cas d'un pays qui bloque l'accès à certains sites, en se connectant à un proxy non bloqué, l'utilisateur pourra accéder à son site au travers de ce dernier car le proxy ne dispose pas des mêmes règles de filtrage.

A l'inverse, certains établissements scolaires ou entreprises limitent l'accès à certains sites grâce à un serveur proxy. En effet, toutes les requêtes effectuées par les utilisateurs du réseau passe par ce serveur intermédiaire qui bloque les sites dont l'adresse a été spécifiquement interdite.

Un autre avantage de l'utilisation d'un proxy est de pouvoir surfer anonymement. Les sites visités n'ont conscience que de l'adresse du proxy et non de(s) utilisateur(s) caché(s) derrière.

De plus un proxy permet le masquage de son lieu de connexion. En effet le proxy peut ne pas être situé dans le même pays que l'utilisateur. Si un site se base sur un système de géolocalisation pour afficher son contenu (YouTube, Google, Google Maps, etc.), sera prise en compte la géolocalisation du proxy.

1.2 Inconvénients

Bien que les proxys offrent de nombreux avantages, ils ont aussi certains inconvénients. S'agissant d'une plateforme reliant un utilisateur au web et effectuant les requêtes du premier, celui-ci voit passer tous les échanges. Certains proxys pourraient alors les enregistrer à des fins malveillantes.

L'intérêt d'un proxy étant important, il centralise toutes les requêtes d'une structure et peut donc être saturé ralentissant par la même occasion la connexion de tous les utilisateurs. De manière générale on peut dire qu'une connexion internet passant par un proxy sera toujours plus lente qu'une connexion directe.

2 Description technique

2

Stratégies de masquage

Rotation d'IP Rotation des UserAgent Gestion de sessions Limitation du nombre de requêtes
Réessai

3

Stratégies de défense

Liste noire d'IP/comptes => Analyse du trafic => Attention à ne pas bloquer les utilisateurs normaux CAPTCHA Changer la structure du DOM régulièrement Remplacer des parties du site par des images

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)