WEB SCRAPING. COMMENT ETRE PROTEGES CONTRE LA COLLECTE AUTOMATIQUE DE DONNEES

Mihai MALAIRAU[1],
Denis REDKO[1],
Sandu RAȘ[1*]

[1]*Université Technique de Moldavie, Faculté Ordinateurs, Informatique et Microélectronique, Département Génie Logiciel el Automatique, Groupe FI-191, Chișinău, République de Moldavie*

*Auteur correspondant : Sandu Raș, ras.sandu@isa.utm.md

Résumé : *La transformation automatique des ressources Web dans un format spécifique consiste à extraire des données web ou web scraping. L'extraction des données Web se fait à l'aide des langages de programmation backend, en accédant au site Web avec un client http et en extrayant les données en adressant les balises, les classes, les identifiants du document qui décrit la page. Plusieurs fois, le web scraping est effectué à des fins d'analyse et de collecte d'informations à partir de plusieurs sources en un seul endroit. Le but principal de cet article était de familiariser et d'expliquer ce qu'est le scraping Web, comment il est utilisé, les techniques, quels objectifs et comment les administrateurs de sites Web peuvent être protégés contre la collecte automatique de données.*

Mots-clés : *extraction de données, analyse HTML, structures de données, demande GET, automatisation, headless browser.*

Introduction

Avec l'avènement des applications Internet et Web, le besoin de transformer les informations des pages HTML en formats plus pratiques est apparu, ce qui permet une gestion plus flexible. En réponse à ce besoin, une technologie de grattage Web a vu le jour. Au fil du temps, la communauté des développeurs, les développeurs Web ont créé de nombreux outils, bibliothèques, packages qui permettent un grattage Web plus facile en utilisant l'API de ces outils.

Le Web scraping a trouvé son application dans la recherche de certaines informations, dans l'indexation de pages web, l'analyse et le suivi des données, il a également trouvé son application dans la lutte contre la concurrence des entreprises, il peut être utilisé par les utilisateurs pour capturer des offres plus rentables, analyser plus de magasins Internet, etc.

Web Scraping

L'extraction de données Web est le processus de transformation automatique des ressources Web en un format structuré spécifique. Par exemple, si une collection de pages Web HTML décrit des détails sur différentes sociétés (noms, emplacements, etc.), l'extraction de données Web signifierait transformer ce format HTML natif en structures de données traitables par ordinateur, telles que des entrées dans des tables de base de données relationnelle.

Le but de l'extraction de données Web est de rendre les données Web disponibles pour les étapes de manipulation ou d'intégration ultérieures.

Historique

Alors que le Web proliférait dans les années 1990, les chercheurs dans le domaine de l'informatique d'horizons divers (bases de données, systèmes, intelligence artificielle, récupération d'informations, etc.) ont réalisé que la capacité d'intégrer des données provenant de sources hétérogènes donnerait naissance à une grande variété d'applications attrayantes, telles que les assistants d'achat qui comparent les produits sur plusieurs sites de vente au détail.

De nombreux chercheurs qui ont étudié les formulations traditionnelles du problème d'intégration de données ont tourné leur attention vers l'intégration de données Web. Cette attention a révélé de nombreux nouveaux défis. L'extraction de données est rapidement devenue l'un des principaux défis à relever.

Les chercheurs n'ont tout simplement pas pu prouver leurs algorithmes d'intégration de données Web de manière convaincante avant d'avoir développé des moyens systématiques d'accéder automatiquement à un grand nombre de documents Web, puis d'extraire des données structurées à partir des formats natifs de ces documents. .

La première approche pour extraire des données Web consistait simplement à coder la récupération d'URL et à extraire les données nécessaires dans les langages de programmation conventionnels. En effet, de nombreuses applications d'intégration de données Web continuent d'utiliser cette approche aujourd'hui. Cependant, les programmes d'extraction de données mis en œuvre de cette manière ont tendance à être relativement importants, ce qui rend difficile leur conception, leur débogage, leur réutilisation et leur maintenance.

Sur la base de cette expérience, les chercheurs ont rapidement remarqué que les programmes d'extraction de données Web écrits dans des programmes conventionnels ont de nombreux modèles logiciels courants. Cette observation a conduit à des efforts pour encapsuler ces modèles, soit en tant que bibliothèques réutilisables pour les langues existantes, soit en tant que primitives dans des langues spécialisées pour l'extraction de données Web. Par exemple, de nombreux programmes d'extraction Web ont un comportement d'exploration, comme « analyser un document HTML pour trouver tous ses hyperliens ; puis récupérez toutes les URL nouvellement découvertes et répétez ou "soumettez plusieurs fois un formulaire Web, en reliant à chaque fois l'un des paramètres d'entrée à l'une des valeurs".

Pour faciliter la construction de robots d'exploration Web, ce comportement peut être codé comme un ensemble de fonctions / classes dans une bibliothèque ou comme éléments primitifs dans un langage de récupération de données Web spécialisé.

Utilisation

Le scraping web se fait en exécutant un programme, écrit dans un langage de programmation backend. Bien que cela ne puisse pas être fait par de simples utilisateurs, il existe des ressources, des services et des applications qui offrent aux utilisateurs simples les privilèges de cette technologie.

À l'aide de nodeJS, une plate-forme logicielle basée sur le moteur V8 qui transforme JavaScript d'un langage hautement spécialisé en un langage à usage général, et des packages dans le gestionnaire de packages npm - cherioo - l'implémentation rapide et flexible de la base jQuery conçue spécifiquement pour serveur, axios - client http pour les navigateurs et nodeJS, le web scraping est très facile.

Le programme fait une requête get sur la page utm.md, extrait les dernières nouvelles et les place dans un fichier qui peut ensuite être importé dans un autre programme.

```
const Axios = require("axios").default;
const cheerio = require("cheerio");
const fsPromise = require("fs").promises;
const baseURL = "https://utm.md/";
const axios = Axios.create({
   baseURL,
   method: "GET"
 });
(async ()=> {
   let news = [];
   const { data } = await axios.get();
   const $ = cheerio.load(data);
   const h3 = $(".entry-title");
```

```
    for (let i = 0; i < h3.length; i++) {
      news.push(h3[i].children[0].data);
    }
    await fsPromise.writeFile(
      `./noutati_UTM.json`,
      JSON.stringify(news, null, 2),
      "utf8"
    );
})();
```

Les données collectées sont écrites dans un fichier JSON et peuvent ensuite être manipulées et facilement gérées, et grâce au format JSON, elles peuvent être utilisées par n'importe quel langage de programmation de serveur.

[

"UTM va beneficia de expertiza a 3 profesori din SUA prin intermediul programului Fullbright Specialist.",

"Studenta FCIM, Diana MARUSIC: exemplu de perseverență la „Mold SEF”",

"„MOLD SEF” – 2020",

"Studiu de politici publice: Securitatea energetică a RM în contextul funcționării pieței concurențiale",

"UTM a dat start cursurilor de pregătire pentru BAC",

"STAFF SELECTION CONTEST for international credit mobility within ERASMUS + Programme at Slovak University of Agriculture in Nitra",

"HOTĂRÂREA ȘEDINȚEI CONSILIULUI DE ADMINISTRAȚIE NR. 12 DIN 10 februarie 2020",

"HOTĂRÂREA ȘEDINȚEI CONSILIULUI DE ADMINISTRAȚIE NR. 11 DIN 27 IANUARIE 2020",

"STUDENT SELECTION CONTEST FOR ICM WITHIN ERASMUS+ PROGRAMME AT Slovak University of Agriculture in Nitra, Slovakia",

"INIVITATION – lectures on nanobiotechnology from researchers at the Royal Institute of Technology and Joint Research Center of the EC"

]

Sécurité

Certains sites Web utilisent des méthodes pour empêcher le grattage Web, tels que la détection et le blocage de l'exploration (visualisation) par des robots sur leurs pages. En réponse à cela, il existe des systèmes de grattage Web qui reposent sur l'utilisation de méthodes d'analyse DOM, de vision par ordinateur et de traitement du langage naturel pour simuler la visualisation humaine afin de fournir la collection de contenu de page Web pour une analyse hors ligne.

Les sites Web peuvent utiliser différents mécanismes pour détecter un grattoir / araignée d'un utilisateur normal. Certaines de ces méthodes sont répertoriées ci-dessous :

1. Trafic inhabituel / taux de téléchargement élevé, en particulier à partir d'un seul client / ou d'une seule adresse IP dans un court laps de temps.
2. Tâches répétitives effectuées sur le site Web - sur la base de l'hypothèse qu'un utilisateur humain n'exécute pas les mêmes tâches répétitives tout le temps.
3. Détection par les pots de miel - ce sont généralement des liens qui ne sont pas visibles pour un utilisateur normal, mais uniquement pour une araignée. Lorsqu'un grattoir / araignée essaie d'accéder au lien, les alarmes sont déclenchées.

Pour éviter le blocage, les scrapers Web doivent effectuer des actions sur le site aussi près que possible du comportement de l'utilisateur. Par conséquent, vous devez périodiquement faire pivoter les adresses IP, changer d'agent utilisateur et définir la vitesse de compression Web sur optimale et entre les appels pour créer des actions aléatoires sur le site qui ne provoqueront pas de suspicion.

Conclusion

Cet article parle de l'extraction de données Web, qui s'est développée avec le développement des technologies de l'information. L'extraction de données Web est utilisée avec de bonnes pensées ainsi qu'avec de mauvaises pensées sans prendre en compte le respect du droit d'auteur. Pour cela, différents langages de programmation sont utilisés qui permettent la transformation des données en des formats plus flexibles et plus faciles à utiliser, le plus souvent c'est du *JSON*. Différentes méthodes qui détectent l'action de raclage sont utilisées pour se protéger contre l'extraction malveillante de données.

Comme Internet s'est développé de façon astronomique et que les entreprises sont devenues de plus en plus dépendantes des données, il est désormais indispensable d'avoir accès aux dernières données sur chaque sujet. Les données sont devenues la base de tous les processus décisionnels, que ce soit une entreprise ou une organisation à but non lucratif. Par conséquent, le web scraping a trouvé ses applications dans tous les efforts des notes de l'époque contemporaine. En outre, il devient de plus en plus clair que ceux qui utiliseront l'outil de grattage Web de manière créative et avancée auront une longueur d'avance sur les autres et gagneront un avantage concurrentiel.

Bien qu'il existe des méthodes qui protégeraient l'extraction automatique des données, elles n'ont qu'un effet de ralentissement, car davantage de code est nécessaire et la vitesse de collecte des données est nécessaire.

Nous remercions: Mme Daniela Istrati, lecteur universitaire au Département Génie Logiciel et Automatique, Université Technique de Moldova, pour l'aide à l'élaboration de cet article

Bibliographie

1. Shaumik Daityari, *Protect Your Site Against Web Scraping*, https://blog.jscrambler.com/protect-your-site-against-web-scraping/, accedé le 18/02/2020
2. ScrapeHero, *How to prevent getting blocked while scraping*, https://www.scrapehero.com/how-to-prevent-getting-blacklisted-while-scraping/ , accesat accedé le 18/02/2020
3. Nicholas Kushmerick, *Languages for Web Data Extraction*, https://link.springer.com/referenceworkentry/10.1007%2F978-0-387-39940-9_1156, accedé le 18/02/2020
4. Jeffrey Neuburger, *QVC Sues Shopping App for Web Scraping That Allegedly Triggered Site Outage,* https://newmedialaw.proskauer.com/2014/12/05/qvc-sues-shopping-app-for-web-scraping-that-allegedly-triggered-site-outage, accedé le 18/02/2020
5. Adler, Kenneth A. ,*Controversy Surrounds 'Screen Scrapers': Software Helps Users Access Web Sites But Activity by Competitors Comes Under SCrutiny*, https://corporate.findlaw.com/law-library/controversy-surrounds-screen-scrapers-software-helps-users.html, accedé le 18/02/2020