

Le Web Scraping au service de l'intelligence économique



Auteurs : Jean-Guillaume Dujardin
Pierre Vaidie

Version 1 - Juillet 2013

Le web Scraping, qu'est-ce que c'est ?

Le Web Scraping est un ensemble de techniques pour extraire le contenu d'un site Web. L'objectif est de transformer les données récupérées afin de les utiliser :

1. Soit dans un autre contexte, par exemple, pour faire une intégration rapide entre deux applications (lorsqu'aucune API n'est disponible) ;
2. Soit pour stocker ces données en base pour qu'elles soient analysées.

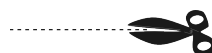
Note : ce n'est pas le premier aspect qui nous intéresse dans ce document.

Faire de la veille économique...

De nombreuses informations sur la concurrence sont disponibles sur le web. Aussi, récupérer et analyser ces informations peut vous permettre de :

- Mieux positionner le contenu de votre offre ou de vos produits ;
- Etudier le positionnement d'un prix ;
- Analyser le réseau de distribution de la concurrence ...

Un projet de Web Scraping concerne souvent le marketing en particulier ceux qui s'occupent de la veille (c'est même de l'intelligence économique : produire des connaissances servant un but économique à partir de sources ouvertes).



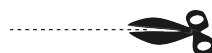
Est-ce que c'est légal ?

Question importante... En fait, il n'y a pas de réponse simple. Cela dépend du pays d'origine, des conditions générales du site et même de la nature des informations collectées.

Bon, Google utilise ces techniques intensément pour son moteur de recherche ou bien les Actualités. Par ailleurs, dans ce document, nous ne parlons que de données publiques.

Toujours est-il que cette question est délicate et qu'il vaut mieux le faire avec discrétion : adopter un rythme de mise à jour pas trop élevé, ne pas utiliser d'IP associé à la société qui récolte ces informations, voire diversifier les IP et enfin accéder à des fichiers directement (qui ne sont pas ou peu monitorés par les outils d'analyse des sites web).

Note : des choses étonnantes sont même possibles. Il nous est arrivé d'avoir accès à la totalité des produits d'une marque en analysant la réponse à une requête alors que seuls certains produits étaient présentés sur le web.



Un projet de Web Scraping ?

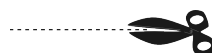
Dans la plupart des cas, un projet de Web Scraping doit « crawler » (parcourir et analyser) de nombreux sites web, organiser, trier et stocker les données récupérées puis présenter des interfaces d'agrégation et de restitution adaptés.

Nous espérons que ce livre blanc déclenchera chez vous des idées de développement d'un projet de web scraping. En fait, le web scraping ouvre de nouveaux champs d'expérimentation et d'analyse. Les projets de ce type peuvent étayer avantageusement une stratégie d'intelligence économique mais aussi, dans une approche purement pragmatique, favorisent la réduction des coûts de collecte, facilitent l'accès à des données à jour et améliorent la qualité des données comparativement à une saisie manuelle.

Pour un voyageur, cela permet par exemple de savoir comment évolue le prix d'une semaine de vacances en juillet en Espagne. Pour un groupe de luxe, il peut s'agir de savoir où la concurrence est implantée. Pour un retailer, il peut s'agir de surveiller les promotions de ses concurrents.

Pour mener ce genre de projet, deux points méritent particulièrement votre attention :

1. Ces outils peuvent être complexes à développer aussi ils doivent être bien définis en amont afin d'apporter rapidement des réponses à des problématiques clés.
2. Ne pas négliger les coûts de maintenance de ce type de projet. En effet, si l'on travaille sur beaucoup de sites cibles, il n'est pas rare de devoir changer des éléments.



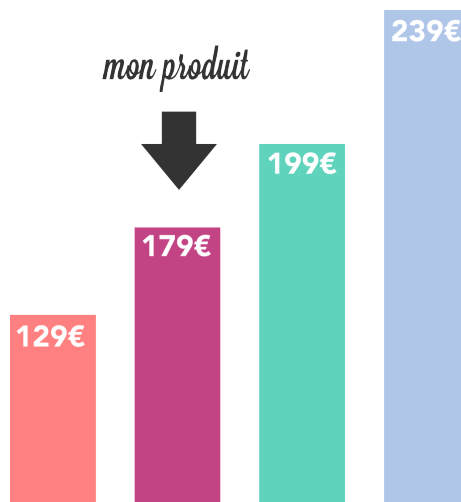
Qu'est-ce que l'on peut faire avec toutes ces données ?

Les outils de restitution dépendent étroitement de la nature des informations représentées.

1. Positionnement d'un prix

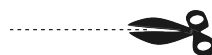
S'il s'agit d'un prix par exemple, l'utilisateur devra choisir l'univers dans lequel il veut représenter son prix ou bien identifier les produits comparables de la concurrence.

Cet outil établira des comparaisons en fonction de critères qui peuvent être : géographiques, type de positionnements, gamme de prix (par exemple, qu'est-ce que je peux obtenir pour un prix donné).



2. Positionnement géographique

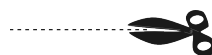
S'il s'agit d'une donnée géographique, la représentation graphique peut, par exemple, faire ressortir les dernières implantations de magasins ou de distributeurs afin de montrer les nouvelles zones investies par la concurrence, les zones potentielles d'implantation, les zones saturées etc.



3. Montrer une tendance

L'outil peut aussi historiser les données collectées afin de montrer l'évolution du prix d'un produit. Si des réajustements sont effectués par la concurrence, des alertes peuvent être lancées.

Ce type d'évolution peut également être couplé à d'autres types d'informations, par exemple, géographiques. Ainsi, il est possible de déterminer des prix par zones.



Techniques

Si, si, vous aurez droit à un petit paragraphe technique...
C'est plus fort que nous ! Les éléments qui doivent être maîtrisés pour ce type de projet sont :

- La gestion des requêtes HTTP: récupération de pages, formulaires, appels AJAX, ..
- Les fonctions de parsing du DOM, du JSON, du XML
- L'utilisation de Regex (pour regular expression) pour traiter les chaînes de caractères.

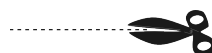
Outils existants

Certains outils vous permettront certainement d'aller plus vite sur votre projet :

- Scrapy (en Python) : malheureusement, nous ne l'avons pas testé mais il semble qu'il remporte certain suffrage ;
- Node.js avec jQuery, phantom.js (webkit) : il ne s'agit pas à proprement parler d'un outil de Web Scraping mais il permet de manipuler facilement les URL et le DOM (cf. notre livre blanc sur le Web Temps réel).

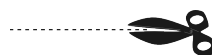
Autres éléments techniques

Les performances comptent lorsque l'on a à faire à un nombre de données conséquent. Dans ce cas, il ne faut pas hésiter à passer en NoSQL (cf. notre livre blanc sur les performances).



Quelques astuces ...

1. En fonction des technologies des sites ciblés, les "crawlers" seront différents : par exemple, il est difficile d'analyser un site en flash, il faudra essayer dans ce cas de « parser » (en français, analyser) le fichier XML source. Si ce n'est pas possible, vous pouvez tenter les sites mobiles (ou tablettes), ceux-là peuvent souvent être « parsés » car ils sont rarement en flash.
2. Il est possible aussi de gérer des mécanismes d'alerte pour étendre sa veille au web tout entier : la mise en place d'alerte de type google permet d'enrichir les données intégrées dans l'application.
3. Ces outils doivent être facilement paramétrables par les administrateurs. En effet, les urls ou les technologies des sites ciblés pouvant changer, il doit être simple d'adapter les règles de parsing des sites.



Alternatives au Web Scraping

Parfois le Web Scraping ne suffit pas ! Il faut alors proposer des mécanismes collaboratifs de veille qui sont soit complémentaires (ils ajoutent des informations à celles déjà collectés automatiquement), soit indépendants.

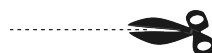
Les outils déployés doivent permettre aux utilisateurs de devenir acteur de la veille concurrentielle qu'ils effectuent naturellement.

Les fonctionnalités doivent alors être simples et ergonomiques. Elles permettent de remonter des informations qu'ils ont pu collecter : indiquer un nouveau produit, un prix relevé dans une boutique voisine ou de préciser la nouvelle implantation d'un magasin...

Quelques préconisations :

L'outil ne doit pas nécessiter de formation et, dans la plupart des cas, être intégrés dans l'intranet de l'entreprise afin de favoriser la prise en main.

Ces outils doivent valoriser ceux qui font l'effort de renseigner ces informations. Il est même possible de mettre en œuvre des mécanismes de « peers-review ». Cela réduit les coûts de traitements de ces informations. Les participants les plus actifs pouvant devenir, par exemple, ceux qui valident et remontent les informations dans l'outil.

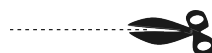


Conclusion

Le web est une mine d'informations : actualités, innovations dans votre secteur mais aussi veille sur la concurrence... On y trouve de tout. Encore faut-il chercher !

Il est critique de repérer ces gisements d'informations, de les traiter et de les exploiter. Vous pourrez ainsi gagner :

- Une plus grande réactivité par rapport aux actions effectuées par la concurrence grâce à une information en quasi temps-réel ;
- Une meilleure sensibilité au positionnement prix de la concurrence avec, à l'arrivée, un impact positif en bout de chaîne sur les ventes (TheCodingMachine vous apporte ici un argument majeur pour défendre ce type de projet auprès de votre hiérarchie !)
- Une meilleure adéquation au marché en permettant d'ajuster de nouveaux placements de produits ou de nouvelles stratégies grâce à l'information recueillie et classée par catégorie, type de produit ou zone géographique. Ce sont souvent des projets pionniers qui peuvent faire office de pilote pour d'autres filiales, d'autres branches de la société, et même dans d'autres pays à l'international.



**Si vous aussi, vous voulez Web Scraper,
n'hésitez pas à nous solliciter !**

**Ces projets nécessitent pas mal
d'astuces et puis qui n'aime pas se rêver
dans la peau d'un pirate ?**

**Contact@thecodingmachine.com
01 71 18 39 73**

