François Goube (Cogniteev) : « Le Big Data et l'analyse sémantique doivent se rejoindre »

Interview réalisée par Olivier Andrieu

Domaine :	Recherche	**Référencement**
Niveau :	**Pour tous**	Avancé

François Goube est co-fondateur de Cogniteev, une start-up qui œuvre dans le monde du Big Data en mode SaaS. Il nous présente ici les outils qu'il propose aux webmasters et notamment OnCrawl, qui surfe sur la forte demande actuelle de systèmes capables d'auditer un site, notamment en termes de SEO. Mais pas que...

Bonjour François et merci de bien vouloir répondre à nos questions. Tout d'abord, peux-tu te présenter ?

Je suis le co-fondateur de Cogniteev, un éditeur de solutions Big data en mode SaaS. Avant cela, j'ai fondé JobiJoba.com un moteur de recherche d'emploi disponible dans une dizaine de pays en Europe, ainsi que l'agence de conseil SEO Propulseo.
Je siège au board de différentes start-ups, je suis l'ambassadeur de Majestic en France, et membre actif de la Frenchtech Bordeaux et du SEOcamp.

Et présenter la société Cogniteev ?

Nous éditons plusieurs services à destination des ecommerçants (Oncrawl.com) ou des PME (Docido.com). Nous nous efforçons de mettre à la disposition du plus grand nombre des technologies réservées jusqu'à present à des grandes entreprises. Nous avons créé une plateforme qui embarque le meilleur des deux mondes entre les technologies Big Data (Hadoop, notamment) et les algorithmes d'analyse sémantique. Notre conviction est que ces deux champs technologiques doivent croiser leurs chemins pour extraire du sens des datasets aujourd'hui à notre disposition.

Vous proposez plsuieurs types de produits, quels sont-ils ?

Docido est un moteur de recherche pour le cloud qui permet de rechercher pêle-mêle dans toutes les applications : Facebook, Linkedin, Twitter, mais aussi Gmail, Salesforce, Trello, Gdrive, Box ou Dropbox... Nous proposons une solution de recherche unifiée.

Oncrawl est un crawler SEO sémantique qui permet de suivre l'ensemble de ses paramètres SEO et de détecter d'éventuelles erreurs ou améliorations possibles. Nous nous efforçons de mettre dans ce produit tout notre savoir-faire en matière d'analyse sémantique. Aussi allons-nous proposer dans quelques semaines des fonctionnalités avancées sur l'analyse du contenu : qualité éditoriale, duplication de contenu, détection du *nearly duplicate*…

Nous préparons également un système d'alertes très avancé pour permettre aux webmasters de détecter rapidement des éléments néfastes à leur SEO : erreurs 4xx, Temps de chargements trop long, pages orphelines…

En quoi un site web peut-il profiter d'un outil comme OnCrawl, par exemple, qui était présenté lors du dernier SEO Campus ? A quel type de site s'adapte-t-il le mieux ?

L'idée d'Oncrawl est venue d'un consortium industriel que nous avions monté en 2013 avec Cdiscount notamment. Nous avons fourni la technologie de crawling et d'analyse de data dans le cadre de ce projet. Petit à petit, avec notre connaissance du métier du référencement naturel, nous nous sommes dit qu'il serait intéressant d'utiliser nos crawlers pour aider les référenceurs. Nous avons alors décidé d'industrialiser notre techno en partant d'un cas plutôt complexe : celui du premier ecommerçant de France. Notre service est donc capable de traiter de grosses volumétrie de pages (plusieurs dizaines de millions).

Nous avons défini une architecture et des tarifs permettant à chacun d'utiliser notre technologie : que vous soyez blogueur, ecommerçant petit ou grand, ou media en ligne, nous avons une solution.

Aujourd'hui, nous travaillons à 60% pour des acteurs de l'ecommerce, 30% de media en ligne et 10% de grandes marques.

Comment OnCrawl se positionne-til par rapport à des concurrents comme Screaming Frog ou Botify, par exemple ?

Tout d''abord, si on regarde l'ensemble du marché dans le monde, nous nous positionnons entre 1,5 et 5 fois moins chers que les autres acteurs du marché. Nous pouvons le faire pour deux raisons assez simples : d'une part notre infrastructure est mutualisée sur l'ensemble de nos produits, et d'autre part nous avons revu complètement notre technologie d'indexation qui nous permet d'exécuter notre service à moindre coût. Mon associé Tanguy Moal est l'un des tout premier ingénieurs R&D d'Exalead, cela fait plus de 10 ans que nous travaillons sur ces technologies.

Ensuite, je pense qu'il existe de vraies différences dans les offres de chacun. Certains proposent un soft qui s'installe sur son poste, nous sommes, de notre côté, sur une solution 100% en ligne. D'autres fournissent un analyseur de logs, nous préférons travailler avec un réseau de consultants sur ce sujet en fournissant la plateforme technique car nous pensons que l'analyse de logs doit obligatoirement s'inscrire dans une démarche plus globale que simplement le SEO. C'est un travail d'expert que les softs ne règlent pas à 100%.

Enfin, nous sommes avant tout des spécialistes de l'analyse sémantique. Nous sommes donc plus concentrés sur le traitement du contenu : analyse d'entités nommées pour proposer un véritable knowledge graph de votre site, développement d'algorithmes de détection de contenus « presque similaires »…

As-tu quelques "cas d'école" à nous proposer où l'outil aurait aidé à résoudre certains problèmes sur un site ?

Sans rentrer dans des choses trop confidentielles, je peux parler de deux cas :

1er Cas : un media en ligne. Ce site avait jusqu'à 39 niveaux de profondeur dans son arborescence, comptant plusieurs centaines de milliers de pages. Soit 39 liens à parcourir pour accéder à certains articles. Autant dire que le robot de Google avait du mal à accéder à de telles profondeurs, ou en tous cas n'attribuait à ces pages qu'une très faible popularité. Difficile donc de les positionner.
Dans un premer temps, notre solution leur a permis d'identifier le problème. Ensuite, grâce au calcul du Inrank, sorte de pagerank interne permettant d'évaluer la popularité interne d'une page, ils ont pu tester l'ensemble de leurs optimisations pour corriger ce problème et savoir ce qui donnait le plus d'impact.
Après mise en production, ils ont réalisé plus de 30% de trafic en à peine 1,5 mois. Quand votre business est lié à la publicité, je peux vous dire qu'une telle progression est inespérée.

2eme cas : Un site de Ecommerce pénalisé. OnCrawl lui a permis d'identifier son contenu en « nearly duplicate ». Plus de 80% de ses pages avaient des contenus de balises Title, Description, H1 parfaitement uniques, mais un contenu de ses pages très similaires. Il importait les textes des catalogues fournisseurs… Mais jusque là, aucune alerte GWT. Nous lui avons fourni la liste des « pages similaires ». Il a ensuite croisé nos données à ses statistiques d'usage afin de prioriser la rédaction de contenu. Au bout de 3 mois, après mise en ligne d'environ la moitié de fiches produit corrigées, il a amélioré son trafic organique de +20%.

Quelle est la gamme de prix proposée ?

Nous démarrons à 9,90€/ mois pour 10.000 pages crawlées dans le mois et jusqu'à 249€/ mois pour 2 millions de pages. Au delà, nous préférons faire du sur-mesure.
Mais j'ai une exclusivité pour Abondance ! Nous venons de mettre en ligne une version gratuite (https://app.oncrawl.com/signup?plan=free) du produit vous permettant de tester l'analyse de 100.000 urls.
Ensuite, pour les lecteurs d'abondance, vous pouvez utiliser le code LETTREABONDANCE pour bénéficier de 50€ de réduction sur votre abonnement.

Quels sont les projets de développement de Cogniteev ?

Nous travaillons activement sur de nouvelles fonctionnalités autour de l'analyse de contenu pour Oncrawl. Du côté de nos autres produits, nous allons continuer à développer Docido, notamment aux US, et espérons fournir une application de « Social news » pour la fin de l'année.

Une question que j'aurais oubliée ?

Je pense que cela est déjà bien complet. Toutefois, je profite de l'occasion pour dire que nous recrutons des ingénieurs de talents et des product manager donc si les mots Hadoop, OpenNLP, et Mahout vous parlent, envoyez moi votre CV à mailto:fr@ncois.eu ☺

Merci, François, pour tes réponses.

Interview effectuée par Olivier Andrieu*, éditeur du site Abondance (http://www.abondance.com/)*