Marc Longo : "il est aujourd'hui impossible de créer un moteur de recherche capable de concurrencer Google de façon frontale"

Retour au sommaire de la lettre

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

Marc Longo est un développeur et entrepreneur, issu notamment du monde des annuaires web, qui est en train de développer, en phase expérimentale, un projet de moteur de recherche pour le début de l'année prochaine. L'occasion nous a semblé intéressante de lui poser quelques questions sur les principales difficultés qu'il rencontre actuellement dans la mise en chantier d'un tel outil...

Bonjour Marc, peux-tu te présenter à nos lecteurs ?

Bonjour, Je suis un ex-sportif, spécialisé dans la moto sur circuit (mondial Superbike 750cc 1992), avec seulement mon BEPC en poche. Je détestais l'informatique mais dans ma reconversion en 1993, j'ai voulu créer un service minitel (3615 Motocasion) et je me suis mis peu à peu à programmer.

Puis j'ai voulu en faire un journal en kiosque (1995, et il existe toujours), et j'ai donc plongé un peu plus dans l'informatique. C'est à ce moment-là que j'ai découvert Internet (1996) et que j'ai cru immédiatement en l'avenir de ce mode de communication.

J'ai donc été fournisseur d'accès sur mon département (Nièvre, 58) et j'ai créé un annuaire (il y en avait très peu à l'époque). Une expérience laborieuse d'ISP, mais j'étais dedans… J'ai donc 18 ans de pratique du Web, sans prétention bien sûr, avec un certain recul sur son évolution.

Peux-tu nous présenter l'expérimentation Premsgo ? Quelle est son ambition ?

C'est un petit secret… sourire. Il s'agit en effet et surtout d'une expérience. J'ai toujours mené mes projets à bien et jusqu'au bout (mais souvent avec du retard, car je prends mon temps pour essayer de bien faire les choses). L'ambition de Premsgo est simplement d'apporter un service manquant, un peu à tous les intervenants à la fois : internautes, établissements associatif ou entreprises, communauté web professionnelle... Combler des besoins, sur une petite échelle, ouvrir une nouvelle voie, c'est l'ambition de ce projet. Je sais, c'est vague, mais pour le moment, c'est préférable. Et puis, une mini-concurrence aux outils déjà en place, aussi petite soit-elle, représente probablement une petite bouffée d'oxygène pour tout le monde, y compris pour quelques géants je pense, en ces temps où l'hégémonie (légitime) de certains devient étouffante… Donc, Premsgo, c'est un petit moteur de recherche purement français et limité à l'Hexagone.

Où en est le projet ?

Il est fonctionnel sur une petite échelle. Les premiers tests sont motivants. On rêve, mais il faut souvent revenir sur terre pour aligner les lignes et les pages de programmation. on travaille sur une ébauche sur quelques millions de pages crawlées seulement, et sur un échantillon de 60 000 sites venant de l'annuaire (http://www.annuairefrancais.fr/).

Pourquoi baser ce moteur de recherche sur un annuaire ?

Je dois avouer que l'échec de l'auto-promotion de mon annuaire (la V2 n'a pas motivé les foules...) m'a conduit à ce projet. Les internautes ont beaucoup trop pris le pli de taper sur un moteur de recherche pour qu'un annuaire, aussi bon et exhaustif soit-il, puisse donner envie de devoir cliquer sur 5 ou 6 liens successifs et de plus, chercher parmi de longues listes de liens. Aujourd'hui, on passe par un moteur pour atterrir dans un annuaire... Il faut être

lucide… Rester dépendant de Google, devoir développer des stratégies SEO, tout cela ne me plait guère. Donc, un moteur indépendant me semble être LA solution.

En même temps, un annuaire bien entretenu est plus riche et qualitatif qu'un moteur qui ramasse tout et n'importe quoi sur le Web. Sur mon annuaire, tout est contrôlé et vérifié à l'inscription, manuellement, puis seulement validé. Rien ne passe, ni les tricheurs ni de fausses sociétés.

Associer les 2 me semble être une pure logique. Et si l'on intègre les établissements qui n'ont pas de site pour le moment, tout coule de source. D'ailleurs, Google a intégré sa base annuaire d'entreprise dans les résultats, Google Adresse puis Google MyBusiness désormais...

Quels ont été les principaux obstacles rencontrés ans ce projet ?

Vaste question … Il existe des tonnes d'obstacle, et mieux vaut ne pas y penser avant de se jeter à l'eau (comme dans une grande courbe à 300 km/h sur un circuit…). Faire un moteur se définit globalement en 7 phases :
- Le Crawl (récolter le contenu des pages web d'un site) et le suivi ;
- L'extraction des données utiles (parseurs en fonction des éléments; titre, meta description, balises Hn, texte, ancres, URL…) ;
- Le filtrage des données a conserver et bien classer selon des algorithmes de pertinence (et dans le futur filtre anti-spam) ;
- L'indexation de ces données dans la base qui va servir les résultats ;
- L'interprétations des requêtes afin de chercher ce qu'il faut dans l'index ;
- Les serveurs et leurs configuration pour le moteur de recherche (fournir les résultats) ;
- Le "back office" ou "service webmasters".

Le plus lourd est peut-être le crawl (récolter le contenu des pages web d'un site). Certains serveurs sont souvent inaccessibles quelques instants ou quelques heures, il faut donc beaucoup de machines (ce que je n'ai pas encore), et laisser des sites en attentes pour passer au suivant rapidement (surtout pour un site d'un million de page bloquées). Il peut y avoir une multitude de soucis liés au crawl, gérer toutes les réponses des serveurs (redirection, fichier robots.txt à interpréter, temporairement indisponibles ou interdit d'accès, panne DNS, etc.)

En temps de développement, le fait d'extraire les données de chaque page est de loin le plus lourd et le plus complexe. Car très souvent, les pages sont mal formées, les URL sont tronquées, les balises ne sont pas fermées, les encodages sont erronés (merci les développeurs chinois…), on trouve des pages avec des milliers de liens, ou des pages gigantesque de plusieurs Mo… C'est l'horreur… On a souvent des plantages en phase de test, il faut doubler les sauvegardes et on passe souvent des heures à réparer les bases ou réinstaller les sauvegardes… sourire.

Sérieusement, on passe son temps à se brûler les neurones (déjà que je ne suis pas ingénieur, les miens sont vite en fusion...) pour essayer de trouver des remèdes à des erreurs qui ne semblent jamais se tarir en diversité.

D'autre part, la pertinence est la clé d'un moteur. J'ai donc essayé d'éviter les erreurs de mes prédécesseurs largement plus illustres, et qui ont disparu (Altavista, Lycos…) dont les critères étaient la densité de mots clés ou des signaux du même genre, aussi basiques.

C'est là où le bât blesse, à vouloir faire très finement, chaque mot a sa propre fiche et plusieurs champs de calculs de pertinence, ce qui rend la base gigantesque sur un tout petit nombre de sites (60 000 seulement). Je ne suis pas loin de la saturation, mais j'ai encore des solutions. L'algorihtme actuel donne de bon résultats, et il me reste beaucoup d'améliorations à faire sur la "compréhension" de la requête .

Pour les requêtes, justement, il existe des usages, des habitudes communes à une majorité d'internautes, pour indiquer le sens de ce qu'ils recherchent. Par exemple : "avocat à paris" ou "avocat paris 11", on détermine que le premier mot est un thème ou une catégorie (donc tenter d'avoir des liens de tables de mots vers tables de thèmes/catégories pour vérifier si c'est bien un thème ou pas) et que le deuxième (ou dernier mot) est peut-être géographique (localité, arrondissement, département, pays). Là aussi, on a des tables de référence pour comparer. Il en va de même pour des fonctions (itinéraire, fonction; construire -étaler-escalader-cuisiner-comment faire-comment aller...), des noms communs, des adresses, des

noms propres, des actualités etc … Comprendre la requête est un élément vital pour arriver à une bonne pertinence. Avec des listes de comparaisons, on peu dégrossir pas mal cette compréhension, conjointement avec le classement de chaque mot d'une page.

Ce qui m'est impossible pour le moment, c'est de crawler les images, vidéos, interpréter les CSS, le Javascript,... C'est donc une expérimentation très simpliste. Google à ses débuts ne prenait même pas en compte les « le », « la », etc. dans ses requêtes ...

Reste les serveurs proprement dit, mais ça reste plus simple.

Pour le moment, une partie du projet seulement est développée, car la vraie difficulté, ce sont d'abord les moyens. Je n'ai encore pas demandé d'aide pour le moment, mais ça ne me semble pas le plus urgent tant que c'est expérimental. L'autre difficulté, c'est la stratégie d'ensemble à mettre en œuvre : le commercial, les partenariats français que je souhaite, etc. Car un moteur, il faut aussi le rentabiliser !

Voilà les 3 difficultés en résumé ; le construction du moteur dans son ensemble, l'aspect financier, la stratégie commerciale. Ca fait beaucoup pour un seul homme, mais j'y travaille depuis 3 ans au moins...

Selon toi, et suite à cette expérience, est-il possible de concurrencer Google aujourd'hui et si oui, avec quels moyens ?

NON, clairement. Sur son terrain mondial, sans l'ombre d'un doute, il est impossible de concurrencer Google. Idem sur une approche nationale exhaustive, c'est-à-dire pour couvrir tout ce qui intéresse les Français, à 100%, comme le fait Google. Du moins en bloc, d'un coup, comme l'a tenté Bing, c'est impossible.

Mais en France, Google possède plusieurs "parts de marchés" (sur les types de recherches), plus ou moins importantes : la recherche sur les images, les vidéos, le "X" peut aussi être considéré comme une part de marché. Le moteur leader tente d'investir la réservation d'hôtel, de billet d'avion, le shopping etc… Sa part de marché de 92% sur le "search" dans l'Hexagone est constituée d'une multitude de "secteurs".

Il y a donc une concurrence possible sur certains secteurs. Dans cette approche sectorisée, il est fort possible de concurrencer Google en étant plus spécialiste. La grande difficulté de Google, c'est le détail, le terrain, individu par individu (ou entité). Google travaille de loin, sur la masse, et fait déjà des miracles, on ne peut être qu'émerveillé par la technologie et le parcours de cette société, même si on la diabolise aujourd'hui. N'empêche, on l'utilise !

Eventuellement, on utilisera un autre moteur s'il apporte vraiment une plus-value sur l'information recherchée. Peut-être viendra-t-on sur Premsgo pour chercher ? Ce n'est pas gagné ! Il y a aussi Qwant qui a sa spécialité même si ce n'est pas (encore) un vrai moteur. D'autres sont peut-être à venir. La concurrence arrivera, par petit bout, comme toujours, sauf innovation miraculeuse. La concurrence Française aurait dû venir d'Exalead, je trouve dommage qu'il ait abandonné la partie...

Pour répondre a ta question des moyens, cela dépend de la part de marché à laquelle on s'attaque et l'approche que l'on a. Les Français ont un peu trop tendance a vouloir faire "payer cher" d'entrée de jeu au lieu de croître et évoluer en surfant sur la base du besoin, contrairement aux entreprises américaines qui font quasi tout gratuit et cherchent le réel besoin+service pour ensuite faire payer et rentabiliser.

Dans l'idéal, 3 développeurs et un chef de projet, c'est une bonne structure humaine pour commencer. Sur 2 ans, en comptant les salaires et les machines, un budget d'1 à 2 millions est le strict minimum pour être solide. Il faudra que j'en passe par là si ça plait et que ça marche, tout re-développer, car sous 4D, je vais vite atteindre les limites même si je suis très satisfait de ce SGBD Français, un peu délaissé ou décrié, mais qui a aussi ses atouts en V14 et pour ma dimension expérimentale modeste.

Pour ma part, j'ai dépensé 15 000 euros de matériel pour le moment… Je fais avec les moyens du bord ! Je viens de créer une SAS pour faire appel a des fonds quand la machine commencera...

Le mot de la fin ? Quand Premsgo verra-t-il le jour ?

Je dirais hier ! ^^ je ne cesse de taper des requêtes ou de demander a mes amis ou familles de tester Premsgo. Les résultats sont en majorité satisfaisants. Mais je vais me donner le temps de peaufiner encore un peu, le 1er Janvier 2015 me semble être une bonne date pour la sortie de cette expérimentation !

Je crains d'autre part terriblement le "bad buzz" et la saturation au début, mais j'aurai prévenu. Ca s'arrangera, promis ! Même si il y a peu d'utilisateurs, les mots clés du « referer » seront, au moins, bien transmis !!! Pas de "not provided" chez Premsgo :-)))

Merci Marc, pour tes réponses !

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