JOSEP M. PUJOL, CHIEF OF SEARCH, BRAVE : « BRAVE EST AUX MOTEURS DE RECHERCHE CE QUE TESLA EST À L'AUTOMOBILE »

Posted on 15 juillet 2021

Le moteur de recherche Brave Search a débarqué sur le Web il y a quelques semaines de cela et s'est fait remarquer par des résultats d'excellente qualité dès le départ. Pour en savoir plus à son sujet, nous avons posé une douzaine de questions à Josep M. Pujol, Chief of Search du moteur, au sujet de l'index, des technologies utilisées, du robot, du modèle économique et de la genèse de ce nouveau moteur qui pourrait faire parler de lui dans les mois qui viennent. Voici ses réponses...

Le moteur de recherche Brave Search est ce que l'on peut appeler un « moteur bien né ». Peut-être même l'un des seuls depuis bien longtemps... Et en tout cas depuis plus de 10 ans. Après les échecs de Cuill, Blekko, Wikia search, Qwant et bien d'autres, il est presque surprenant de voir débarquer un nouveau moteur (basé sur un index et des algorithmes propres, donc pas un metamoteur ou une interface de recherche comme DuckDuckGo, Lilo ou Ecosia). Avec des résultats très pertinents dès le départ. Annoncé en mars 2021, puis disponible auprès d'une petite communauté de bêta-testeurs début juin, il a finalement été lancé fin juin en version bêta pour tous les internautes désireux de le tester. Aussi, sa bonne tenue nous a amené à contacter les équipes de Brave pour en savoir plus au sujet de ce moteur : origine, historique, robot, mixité des résultats, modèle économique, avenir envisagé, etc.

Merci donc à **Josep M. Pujol**, *Chief of Search* du moteur, d'avoir bien voulu répondre à nos questions (traduites ici de l'anglais et validées par les équipes de Brave) :

Pouvez-vous nous donner la genèse du projet Brave Search. Quelles technologies ont été utilisées ? Quel est l'historique et le calendrier de ce projet ?

Brave search est le résultat de la vision de Brave selon laquelle les navigateurs et la recherche sont dépendants les uns des autres, comme la voiture et la roue. Pour pouvoir agir comme un véritable agent utilisateur pour les utilisateurs,

Brave doit offrir les deux. En fait, les personnes qui ne sont pas des techniciens ont souvent du mal à faire la distinction entre le navigateur et le moteur de recherche, d'où l'intérêt de proposer les deux. La découverte par la recherche est au cœur de l'expérience de navigation, et Brave a toujours été intéressé par l'amélioration de la recherche.

Malgré cette vision, le développement d'un moteur de recherche est une entreprise très coûteuse que Brave ne pouvait pas entreprendre dès sa genèse, d'autant plus que Brave a lancé son navigateur en 2016. À peu près à la même époque, en 2014, il existait une autre entreprise avec une vision très similaire (servir les utilisateurs en premier et protéger la vie privée) : Cliqz, basée à Munich, fondée par JP Schmetz et Marc Al-Hames et entièrement financée par Burda. Ils partageaient la même mission et, grâce à un bon financement, avaient la capacité et le courage de commencer à construire un moteur de recherche. Cependant, Cliqz ne s'est pas développé autant que Brave et ses coûts sont devenus difficiles à assumer. De nombreuses personnes souhaitent voir une alternative à Google, mais très peu sont prêtes à risquer leur argent pour en faire une réalité.

Après que Cliqz ait cessé d'être opérationnel, pour éviter que la recherche ne disparaisse, Burda a autorisé une scission de la propriété intellectuelle, des données et de la technologie de recherche auprès d'une petite équipe, appelée Tailcat. Lorsque Brave l'a appris, il était clair qu'il fallait se diriger dans cette direction, et nous y voilà. Avec un moteur de recherche indépendant entièrement fonctionnel et un navigateur fantastique, nous sommes prêts à offrir au monde une alternative à la Big Tech. Un pas après l'autre, ce qui était autrefois considéré comme une tâche impossible commence à devenir une réalité, d'autant que des millions d'utilisateurs recherchent désormais des outils de préservation de la vie privée après avoir perdu leur confiance en Google.

Quel est le nom du robot Brave Search ? Cliqzbot, Bravebot ou autre ?

Il existe un robot, mais il utilise l'agent utilisateur du navigateur, comme le ferait un utilisateur normal. Les robots provenant de moteurs de recherche inconnus (ce qui était le cas de Brave Search avant sa mise à disposition du public) sont fortement discriminés, nous avons donc dû miser sur l'anonymat. Il va sans dire que notre robot respecte la netiquette et que nous ne provoquons jamais d'attaque DoS ou de surcharge sur un site. Nous respectons le fichier robots.txt, mais nous nous basons sur les directives pour googlebot plutôt que sur les "user-agent:*". Là encore, c'est nécessaire et pragmatique, car de nombreux sites réservent un traitement spécial à googlebot, ce qui, à notre avis, constitue une violation de la neutralité du réseau.

Au fur et à mesure que Brave Search se fera connaître, nous nous apprêtons à évoluer vers une approche plus conventionnelle, mais pour le moment, nous sommes occupés à construire et nous préférons ne pas toucher à ce qui fonctionne déjà. La construction d'un moteur de recherche est une tâche extrêmement complexe et coûteuse. Nous ne pouvons pas survivre sans être à la fois guidés par des principes et et des fonctionnements pragmatiques.

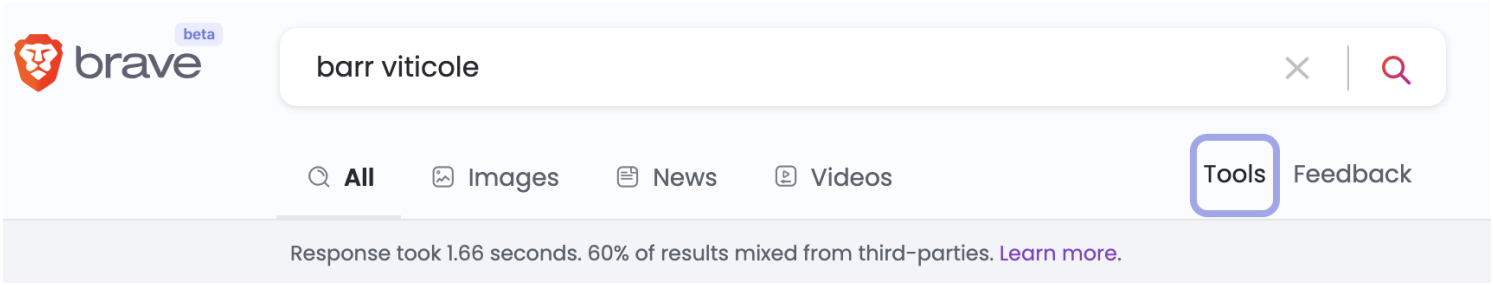
Brave Search utilise-t-il d'autres briques déjà connues (comme celles de Bing ou de Yandex par exemple) pour son index et ses algorithmes ? Tout dans Brave Search est-il vraiment "fait maison" ?

Pas encore tout : nous sommes indépendants dans le sens où nous avons notre propre index, dont la couverture augmente au fur et à mesure. Mais nous faisons appel à des tiers (Bing et Google), bien que cela soit fait dans un faible pourcentage de requêtes.

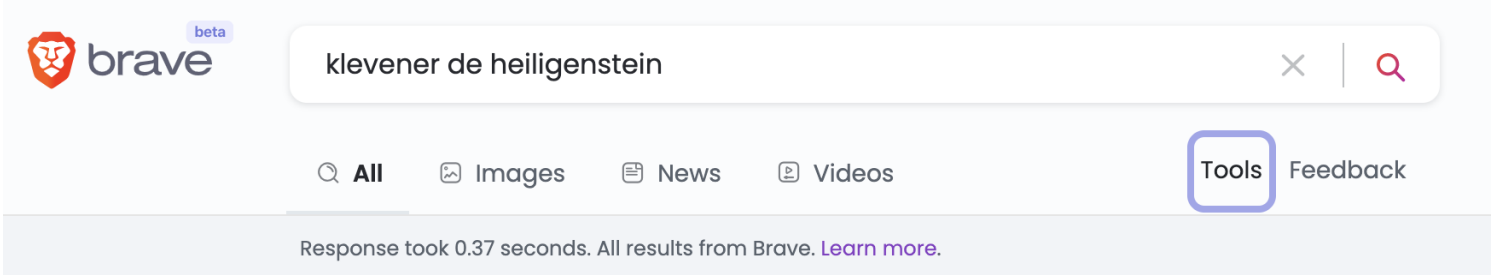
Contrairement à d'autres, nous sommes transparents sur ce point. Nous fournissons même une métrique d'indépendance (https://search.brave.com/help/independence). À ce jour, notre mesure d'indépendance globale est de 87 %, ce qui signifie que 13 % des résultats que nous affichons proviennent de tiers (Bing et Google donc). Cela ne signifie pas que nous ne pouvons répondre qu'à 87 % des requêtes, notre recherche peut répondre à presque toutes les requêtes ; toutefois, pour répondre aux attentes des utilisateurs en matière de qualité, nous mélangeons des résultats provenant de tiers lorsque notre confiance dans nos résultats est faible. Dans les 13 % de résultats provenant de tiers, nous devons actuellement inclure les onglets image, vidéo et actualités. Ceux-ci sont 100% Bing pour le moment. Nous disposons de nos propres moteurs de recherche verticaux. Par exemple, lorsque vous voyez des images, des actualités ou des vidéos sur notre page de résultats du moteur de recherche (SERP), ces éléments sont 100 % Brave, mais nous n'avons pas encore une couverture suffisante pour effectuer une recherche verticale à part entière.

Nous nous attendons à ce que le pourcentage de 87 % continue de croître au cours des prochains mois pour atteindre finalement 100 %. Nous tenons à souligner que l'index de recherche est indépendant, même si des tiers sont utilisés. Si demain Bing et Google devaient fermer leurs portes ou cesser de fournir des services, Brave Search continuerait de fonctionner. Bien sûr, la qualité baisserait un peu, mais nous serions toujours un bon moteur de recherche.

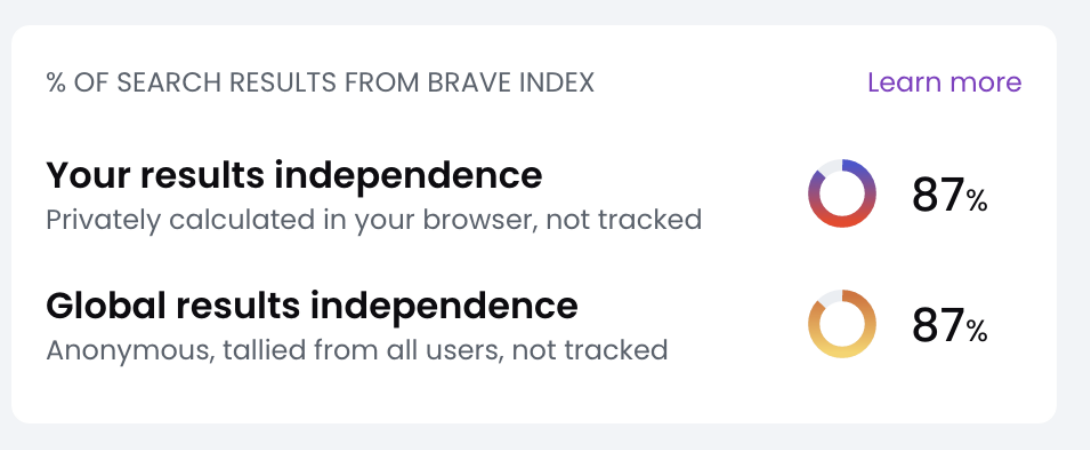
Encore une fois, le pragmatisme est la clé. Nous aurions pu essayer de nous lancer avec une indépendance à 100 %, mais cela aurait signifié attendre plus longtemps, et peut-être ne pas survivre entre-temps. Nous pensons que 87 % est le point idéal. Moins d'indépendance aurait été dangereux, car les incitations à cesser le développement et à s'en remettre exclusivement à la tierce partie auraient été trop fortes, ce qui est exactement ce qui est arrivé à DuckDuckGo et à Qwant : ils ont lancé avec une indépendance à un seul chiffre (voire aucune), en espérant que l'indépendance de la tierce partie ne serait pas compromise. La logique financière empêche cette voie, pourquoi dépenser beaucoup d'argent pour construire quelque chose qui est plus rentable à louer ? Par conséquent, les espoirs de construction sont rapidement, et logiquement, enterrés. Le prix à payer dans ce cas signifie que vous finissez par avoir la peau d'un moteur de recherche, et non un vrai.



Le menu "Info"(qui se transforme en "Tools" une fois cliqué) en haut à droite de la SERP donne le taux de mixité des résultats fournis Brave / Bing-Google.



Dans ce cas, Brave s'est débrouillé tout seul, comme un grand :-)



Dans le menu Burger, le taux d'indépendance des résultats est affiché, d'une façon générale et pour vous-même.

Comment connectez-vous votre moteur de recherche à Google et Bing pour obtenir des résultats supplémentaires ? Via leur API ?

À Bing, par le biais de leur API, et le mélange des résultats (s'il y en a) se fait du côté du serveur. Pour les utilisateurs du navigateur Brave, s'ils y consentent par le biais d'un opt-in, nous pouvons également mélanger les résultats avec ceux de Google lorsque nous détectons que nos résultats se situent sous un certain niveau de confiance. Ces résultats sont récupérés par le biais d'une requête http anonyme depuis le navigateur de l'utilisateur lui-même, donc côté client (uniquement possible sur le navigateur Brave). La plupart des mélanges (environ 13 % des résultats proviennent de ces deux tiers, selon notre mesure d'indépendance) sont attribuables à Bing.

Combien de temps a-t-il fallu pour construire ce moteur de recherche ?

Cela est lié à la question 1. Si vous comptez depuis la création de Cliqz, cela fait 7 ans ; cependant, chez Cliqz, nous ne nous sommes pas concentrés sur un véritable moteur de recherche SERP et nous avons commencé par une recherche dans le navigateur. Le savoir-faire, les données, les modèles et la construction de l'index étaient toujours les mêmes. La recherche SERP présente toutefois certaines particularités qui doivent être prises en compte, principalement en ce qui concerne la couverture. L'adaptation pour être entièrement et uniquement SERP a pris quelques années. Cette partie a commencé tardivement chez Cliqz, et s'est épanouie chez Brave.

Pouvez-vous nous donner quelques chiffres et statistiques clés sur Brave Search : taille de l'index, nombre de requêtes traitées chaque jour, taille du datacenter, etc. Et où est situé le datacenter ?

L'index contient près de 9 milliards de pages, et bien que la taille ne soit pas si grande, nous faisons notre crawl sans tomber sur des pages non pertinentes. Ainsi, bien qu'il ne s'agisse que d'une fraction du Web entier, c'est une fraction significative du Web qui vaut la peine d'être visitée. Le datacenter est dans le cloud (AWS), à l'exception du récupérateur/crawler qui est distribué.

En ce qui concerne le crawl, le fetcher, puisque vous avez posé la question auparavant, il récupère en moyenne 40 millions de pages par jour à partir d'un large éventail d'adresses IP.

La charge n'est pas si élevée, car nous venons de sortir du programme de prévisualisation de notre liste d'attente. Nous avons eu des pics de 70qps. Nous constatons un flux constant de nouveaux utilisateurs avec une bonne rétention. Nous sommes ravis de constater cette rétention, car il est généralement difficile pour les utilisateurs de changer leurs habitudes et leurs paramètres par défaut. Google a une forte emprise sur les utilisateurs, ils dépensent des milliards chaque année pour s'assurer qu'ils sont le moteur par défaut, et ils ne sont pas un mauvais moteur de recherche, donc nous sommes impatients de voir plus de nos utilisateurs faire le changement.

Pourquoi limitez-vous l'accès aux 20 premiers résultats de la SERP ? Pour de nombreuses requêtes, il est intéressant d'avoir plus de 20 résultats, non ?

Là encore, la réponse est le pragmatisme. Nous pourrions renvoyer plus de résultats, mais cela affecterait notre infrastructure, nos opérations et notre évaluation de la qualité. Est-il normal de doubler les coûts pour satisfaire un cas d'utilisation de 1% ? Oui, mais en temps voulu. À ce jour, nous sommes un David, et les personnes qui nous soutiennent comprennent que nous ne pouvons pas rivaliser à armes égales avec un Goliath comme Google. Si nous voulons avoir une chance de gagner, nous devons choisir les bonnes batailles, et nous pensons que le fait de proposer plus de 20 résultats n'en fait pas partie à ce jour.

Une question directe : comment est-il possible de construire un si bon moteur de recherche en si peu de temps, alors que Cuill, Blekko, Wikia search et en France Qwant et bien d'autres ont échoué dans le passé ? Quel est votre secret ? :)

Je pense que la question a déjà été abordée dans la réponse à la question 1 : Brave Search est en gestation depuis plusieurs années, mais sous une forme différente. Des millions de dollars et beaucoup de talent y ont été consacrés, mais le fait qu'il soit "si bon", comme vous le notez, n'est rien d'autre qu'un petit miracle dans le paysage actuel de la recherche. On pensait que la construction d'un moteur de recherche, juste l'index, était une tâche que seules les grandes entreprises pouvaient entreprendre et qui coûtait des milliards. Brave et Cliqz avant lui sont la preuve qu'il y a toujours eu une autre façon de construire un moteur de recherche compétitif.

Quel est le modèle économique de Brave Search ? Comment gagnez-vous de l'argent en ce moment ? Comment payez-vous les factures à la fin du mois ? :-D Allez-vous bientôt afficher des publicités ?

Nous afficherons des publicités (tout en respectant la vie privée, chez Brave cela va de soi) et nous proposerons également un abonnement pour ceux qui préfèrent ignorer les publicités. Brave génère déjà des revenus via sa plateforme publicitaire préservant la vie privée, qui reconnecte les utilisateurs et les annonceurs avec un système de publicité opt-in.

Combien de personnes travaillent sur le moteur de recherche Brave Search ? Où sont-elles situées ?

Sur les 120 employés de Brave, une vingtaine de personnes travaillent sur le moteur de recherche Brave Search, qu'il

s'agisse du moteur de recherche principal ou de l'assistance. Elles sont réparties dans le monde entier, puisque Brave possède des bureaux aux États-Unis, au Canada, en Europe et au Japon.

Ne pensez-vous pas que, pour l'instant, Brave search ressemble à un clone de Google ? Il lui manque peut-être une touche d'originalité ou une "killer app" pour l'utilisateur ? Il est si difficile de changer les habitudes (et notre moteur de recherche par défaut)...

Les habitudes sont très importantes, surtout dans un domaine aussi critique et banalisé que la recherche. Ainsi, être très original peut être loué par les spécialistes, mais va déplaire aux utilisateurs. Nous avons été critiqués et félicités sur un pied d'égalité pour l'expérience similaire à celle de Google.

La principale originalité de Brave Search est qu'elle existe de manière indépendante, car elle possède son propre index, qui sert la grande majorité des résultats que les gens voient. Comme une Tesla, une voiture d'apparence normale avec 4 roues et un volant, mais avec un moteur différent, qui change tout. À l'avenir, nous envisagerons de nous écarter de notre position actuelle, mais pour l'instant, cela semble être l'approche la plus sensée pour se développer rapidement.

Quelle est la prochaine étape pour Brave Search ? Quels sont vos objectifs ? Pour 2021 par exemple, et plus tard ?

L'une des choses que nous attendons avec impatience est le lancement de Goggles, notre exploration des filtres de classement pilotés par la communauté et basés sur notre index. Goggles permet aux communautés d'intérêt de créer leurs propres moteurs de recherche afin de répondre à des besoins que nous ne pouvons ni prévoir ni entreprendre. Nous pensons que la recherche est trop importante pour être gardée sous clé, et nous sommes convaincus que Goggles est la meilleure approche pour la rendre ouverte et accessible, pour combattre les préjugés et la censure, et bien sûr, pour encourager l'innovation.

Merci, Josep, pour vos réponses à nos questions (merci à Catherine Corre pour son aide) !

Josep M. Pujol est l'actuel chef de la recherche (*chief of search*) chez Brave. Il travaille activement à la construction d'un moteur de recherche alternatif depuis 2014, d'abord en tant que *chief scientist* chez Cliqz et en dernier lieu en dirigeant le projet Tailcat, acquis par Brave. Outre la recherche, Josep M. a été impliqué dans d'autres start-ups ainsi que dans la recherche dans différents laboratoires à travers le monde. Il a obtenu son doctorat en intelligence artificielle et en informatique en 2003. Il est l'auteur de plus de 20 publications de premier plan et de plusieurs brevets américains. Josep M. aime répéter qu'« *il faut tomber amoureux du problème, pas de la solution* », une citation honteusement volée à l'un de ses nombreux grands mentors.

Interview proposée par **Olivier Andrieu** Rédacteur en chef de la lettre "Réacteur".