Google peut-il se passer de l'analyse des « anchor text » dans son algorithme de classement ?

Retour au sommaire de la lettre

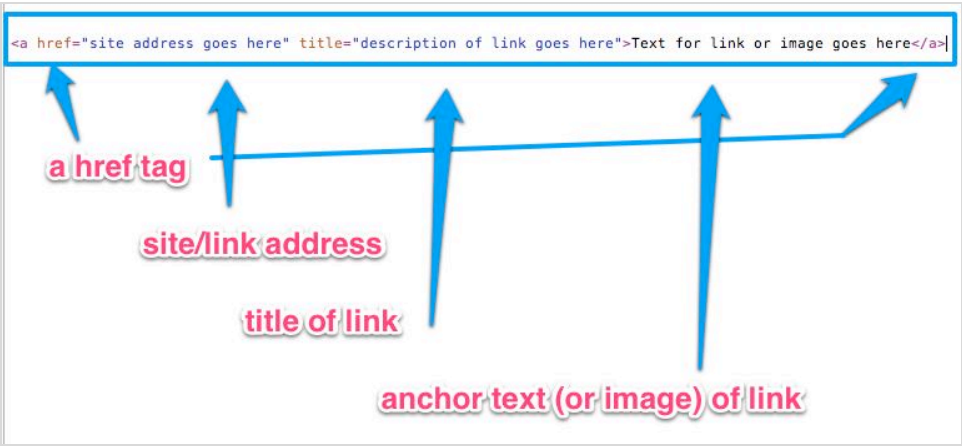
Domaine :	Recherche	**Référencement**
Niveau :	Pour tous	**Avancé**

L'analyse des textes d'ancre ou « anchor text » sont à la base des algorithmes de pertinence des moteurs de recherche, dont Google, depuis de nombreuses années. Certains disent que ce signal est en perte de vitesse et pourrait être remplacé par d'autres critères de pertinence. Qu'en est-il exactement en 2013 et quelles sont les solutions éventuellement envisagées par Google pour améliorer ce signal ? Voici quelques pistes de réflexion...

Depuis le lancement de Google en 1998, le moteur de recherche fait une utilisation intensive des textes d'ancre (anchor texts) des liens dans son algorithme de classement (notion de « réputation »). Mais Google n'est pas le seul dans ce cas. En fait, la plupart des moteurs de recherche grand public ont choisi cette approche. Tout se passe comme si ces petits bouts de texte étaient indispensables pour construire un algorithme opérationnel. Cet article a pour objectif d'expliquer pourquoi, et comment on peut tirer parti de cette information pour le référencement.

Les « anchor texts » dans le code html

Rappelons tout d'abord ce que sont exactement les « anchor texts ». Les liens hypertextes reliant deux URL se codent à l'aide d'une balise le texte du lien. Le « texte du lien » en bleu souligné dans le code type que nous venons de mentionner n'est autre que « le texte d'ancre » du lien hypertexte (« anchor text ») du lien. Le terme « ancre » vient du fait qu'à l'origine la balise <a> (pour « anchor ») était conçue pour donner accès à une zone d'une page web (« l'ancre ») depuis une autre zone de la même page web - dans ce cas d'usage, nous continuons à appeler cela une ancre - mais aussi depuis une autre page web (ce que les webmasters appellent aujourd'hui tout simplement : un lien).



Les « anchor texts » dans l'algorithme de Google

On trouve une description de l'utilisation des « anchor texts » dans le fonctionnement quotidien de Google dès le premier article scientifique publié en 1998 par Sergey Brin et Larry Page sur le moteur (« The Anatomy of a Large-Scale Hypertextual Web Search

Engine », cf. bibliographie). En voici un extrait traduit tiré de la version « courte » de l'article :

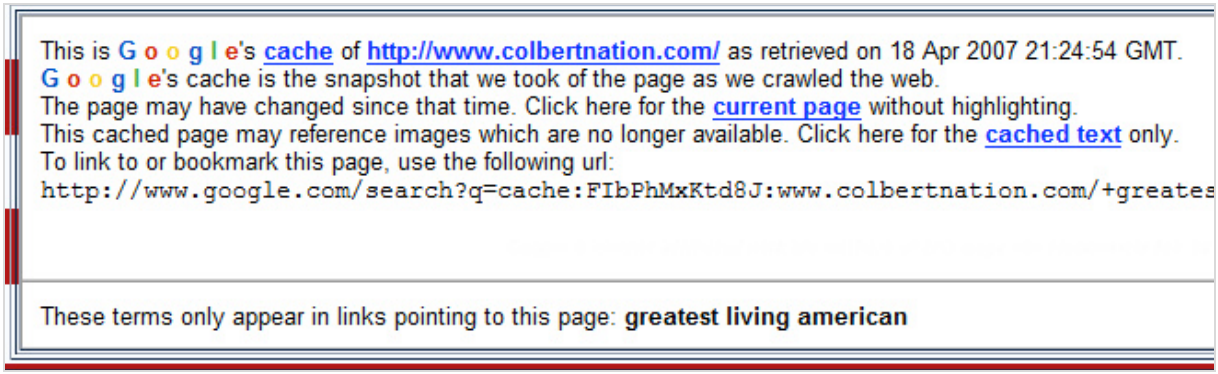
« Les textes de liens sont gérés d'une manière spéciale dans notre moteur de recherche. La plupart des moteurs de recherche associent le texte d'un lien avec la page sur laquelle ce lien est présent. En plus de cela, nous associons le lien avec la page vers laquelle ce lien pointe. Ceci a plusieurs avantages. Premièrement, des ancres peuvent exister pour des documents comme des images, des programmes ou des bases de données qui ne peuvent pas être indexés par un moteur de recherche basé sur les textes. Cela rend possible aussi de renvoyer des pages qui n'ont pas été réellement explorées… »

L'article décrit ensuite l'architecture du moteur, et explique que les anchor texts sont stockés avec tous les liens découverts, sous forme de « nœuds » reliant le document où ils apparaissent au document vers lequel il pointe.

Cette « table » des liens sert ensuite à créer un index inversé où les emplacements des occurrences d'un mot clé dans un document sont identifiées, à la fois dans la page web elle-même et dans les textes d'ancre.

L'indexation des anchor texts en action

Dans la pratique, cela signifie que les textes des liens sont indexés comme s'ils faisaient partie du texte de la page distante… Le résultat est qu'une page peut parfaitement se positionner sur un mot clé qui ne figure pas du tout dans le contenu de la page !
Ce comportement est révélé par un message qui apparait parfois lorsque l'on affiche les pages figurant dans le cache du moteur :



*Ces termes apparaissent seulement dans les liens qui pointent vers cette page : **greatest living american**. Ce message apparait dans l'en-tête des pages en cache quand la page retournée par le moteur ne contient pas les termes de la requête, mais est présente dans les liens pointant vers cette page*

Le poids des mots clés trouvés dans les anchor texts

Dans leur premier article décrivant Google, les fondateurs insistaient sur le fait que les « anchor texts » constituaient une description de la page de meilleure qualité que ce qui pouvait être trouvé sur la page elle-même (comme le contenu de la balise <title> ou de la balise <H1> par exemple). **Lorsqu'il s'agit de liens provenant d'autres sites que celui auquel la page de destination *appartient***, un autre facteur s'ajoute : ces liens sont potentiellement plus « neutres », moins « manipulés », moins « spammés » que ceux créés par le webmaster propriétaire de la page. **On peut en théorie accorder plus de confiance aux mini descriptions que constituent les *textes des* liens créés par des tiers, qu'à celles présentes dans des balises créées par le webmaster (comme la balise title par exemple).**

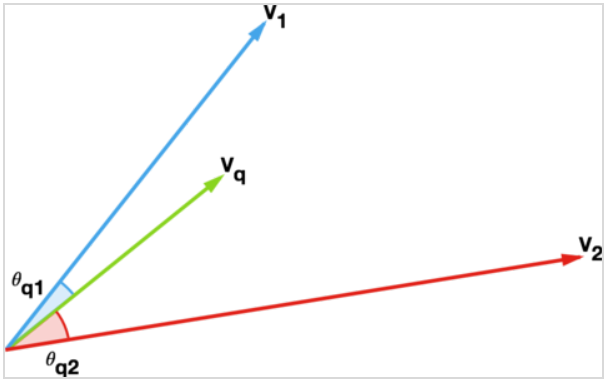
Ces deux facteurs ont conduit les créateurs de Google à attribuer un poids très important aux mots clés présents dans les anchor texts des liens pointant vers une page.

Pas de bon algorithme de classement sans les anchor texts

Dans l'extrait de l'article de Page et Brin mentionné plus haut, les auteurs rappelaient combien les anchor texts étaient utiles pour classer des documents impossibles à crawler, ou à indexer sur des mots clés comme les images.

Mais en fait, même quand les documents sont crawlables et indexables, les anchor texts sont indispensables pour « classer » correctement les résultats et faire remonter les plus pertinents en tête des résultats.

Pour retourner des résultats pertinents sur une requête donnée, il est important de déterminer avant tout quelles pages sont proches sémantiquement de la requête. Cette proximité sémantique doit aller au-delà de la simple recherche d'occurrences des mots clés de la requête dans la page. S'appuyer sur les textes d'ancre permet de faire remonter facilement une page sur Barack Obama sur la requête « Président des USA ». Il suffit pour atteindre ces résultats d'indexer les textes d'ancre avec le contenu de la page. Un bon algorithme doit aussi détecter des pages qui sont pertinentes sur une requête sans contenir d'occurrences des termes recherchés, ni dans le texte de la page, ni dans les anchor texts… C'est la tache dévolue à des méthodes de calcul de similarité entre documents textuels (comme le « Cosinus de Salton »).



Dans la méthode du cosinus de Salton, chaque document est transformé en vecteur dont les coordonnées dans l'espace représentent les poids des termes contenus dans chaque document. On peut ensuite facilement mesurer la distance angulaire entre chaque vecteur en calculant le cosinus de l'angle: plus elle est faible, plus les documents sont proches sémantiquement. Ici, Vq représente une requête, V1 et V2 deux documents. V1 est plus proche que V2. Google n'utilise probablement pas cette formule, qui était utilisée par Altavista, mais une méthode de calcul de similarité plus sophistiquée.

Mais le problème est que dans la pratique, il faut comparer un signal très court (la requête de l'internaute, qui ne contient que quelques termes), et un document très long (la page web). Si les calculs de similarité entre pages web donnent de bons résultats, les calculs de similarité entre requête et pages web donnent des résultats… médiocres. Les documents à comparer sont de nature très différentes, et le faible nombre de termes dans la requête rend la mesure de similarité imprécise.

Ces approches fonctionnent mieux quand on compare les requêtes aux anchor texts. Les anchor texts ont dans la pratique une longueur et un format qui les rapproche beaucoup des requêtes tapées par les internautes. La solution consiste donc à donner une fois de plus beaucoup de poids dans l'algorithme de classement aux mots clés contenus dans les anchor texts et la qualité de l'algorithme de classement s'en trouve améliorée…

Le Google Bombing : la preuve de l'importance des anchor texts dans le classement

L'une des manifestations les plus spectaculaires de l'importance des anchor texts dans l'algorithme de classement de Google est certainement la technique dite du « Google Bombing ». Il suffit que quelques webmasters se concertent pour créer des liens comportant un « anchor text » particulier et pointant vers une page « cible » pour que cette page se mette à se positionner en tête des résultats sur la requête correspondant à l'anchor text utilisé.
L'efficacité de la « bombe » est renforcée par le fait que l'algorithme du moteur a tendance aussi à favoriser les pages ayant un fort pagerank, un critère qui se base sur l'analyse des liens entre les pages pour leur donner une note de popularité ! Un Google bombing massif donne en plus toutes les chances à la page cible de se positionner en tête des résultats…

Le plus célèbre « Google Bombing » de l'histoire est celui dont a été victime George Bush junior en 2003 sur la requête « miserable failure ».

Un autre exemple de « Google Bombing » dont a été victime Nicolas Sarkozy. Même si Google a mis en place des outils pour déclasser les pages victimes de Google Bombing sur les requêtes litigieuses, l'algorithme de classement s'appuie trop sur l'information venant des liens pour empêcher ce genre de manipulations de fonctionner. Aujourd'hui, il faut juste plus de liens pour obtenir le même résultat.

Le texte autour des liens compte !

Dans l'article de Page et Brin, dans la description de l'architecture du moteur (qui à cette époque ne s'appelait pas encore Google mais « Backrub »), on découvre aussi que le moteur indexe non seulement l'anchor text, mais aussi les termes qui « entourent » le lien ! En effet, ces mots clés peuvent être en rapport avec le contenu de l'anchor text. Par contre, les mots clés situés avant ou après le lien ne sont pas toujours pertinents, et de fait se voient attribuer un poids très faible dans l'algorithme.

Depuis lors, de nombreux auteurs ont publié des papiers montrant l'intérêt d'exploiter le contexte du lien pour améliorer la qualité des résultats d'un moteur de recherche. Le « link context » comporte aussi bien les termes qui entourent un « anchor text » que des choses comme :
• L'emplacement du lien sur la page (header, footer, menu, colonne de gauche, droite ou au milieu ?) ;
• La nature du lien (iframe/frame), image etc…) ;
• Le caractère interne ou externe du lien (reliant deux sites différents, ou deux pages du même site) ;
• La nature, le niveau de confiance accordé et/ou qualité de la page support du lien ;
• La hiérarchie dom de la balise HTML qui encapsule le lien ;
• L'âge de la page qui fait le lien (page récente, ancienne) ;
• L'importance de la page qui fait le lien (pagerank et autres notes similaires) ;
• La présence d'autres liens similaires :
  - un brevet de Google explique par exemple comment, à partir d'un certain seuil de liens identiques, les liens supplémentaires peuvent être ignorés.
  - cette information peut également être utilisée pour deviner que des pages sont similaires quand des ancres identiques pointent vers des pages différentes.

• Etc.

Remarquons qu'au-delà de l'importante littérature scientifique sur l'exploitation des anchor texts, on trouve aussi de nombreux brevets déposés par Google, Yahoo ! et Bing sur ce thème.

Exploiter les anchor texts pour le SEO

L'utilisation intensive de l'information extraite à partir des anchor texts par les moteurs de recherche, ainsi que le poids très important des mots contenus dans les textes d'ancre en font évidemment une cible de choix pour des actions SEO efficaces.

La plupart des référenceurs l'ont compris et cherchent donc à « manipuler » l'algorithme en influant, directement ou indirectement, sur les anchor texts des liens pointant vers leurs pages (les actions de Google Bombings en sont de bons exemples). Cela peut se manifester par :

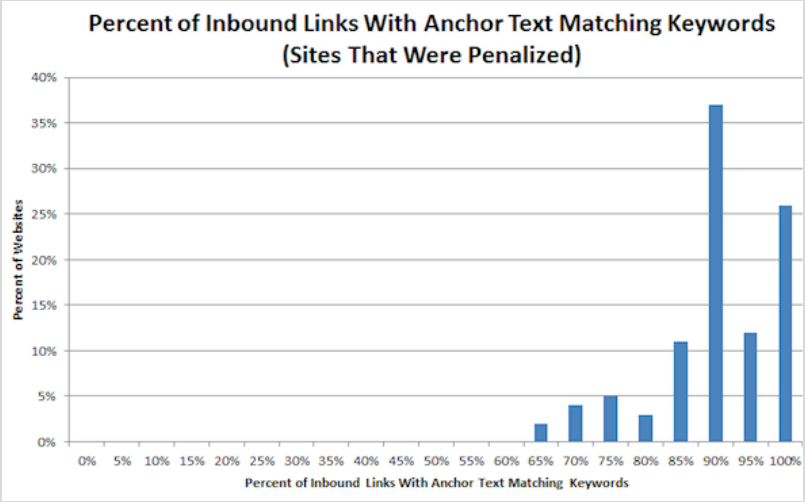
• L'optimisation des ancres des liens internes (figurant sur des pages du même site). On peut :
- modifier les textes des ancres soit pour renforcer le poids de certains mots clés par ailleurs présents sur la page de destination ;
- soit « ajouter » un mot clé non présent dans la page de destination ;
- et même faire « tourner » les ancres pour positionner les pages sur de multiples expressions clés, présentes ou non sur les pages de destination.

• L'optimisation des ancres des backlinks externes (les liens en provenance d'autres sites que celui de la page de destination). Parmi ces backlinks externes, certains peuvent en réalité être directement posés par le webmaster propriétaire de la page de destination (quand il poste sur un forum ou rédige un commentaire de blogs), d'autres doivent être négociés (et donc parfois même… achetés) auprès d'autres webmasters.

Dans la pratique, un moteur de recherche comme Google s'avère particulièrement sensible à ce type de manipulation. Google n'a pas d'autres choix que d'élaborer des contremesures pour éviter ces manipulations. Depuis deux ans en particulier, Google a réussi à faire des progrès remarquables dans la détection de ce qu'ils appellent des « liens non naturels », dont le caractère artificiel est détecté notamment par l'analyse des anchor texts.

L'analyse des anchor texts joue également un rôle majeur dans l'algorithme de détection de sites « spammy » (baptisé Penguin). Les référenceurs ont pu se rendre compte que l'un des principaux critères qui pouvaient faire « tomber » un site dans le filtre Penguin, était une répartition non naturelle des ancres des liens pointant vers un site.

L'exploitation des informations provenant des liens et des ancres sert probablement aussi dans le filtre « Panda » et dans de nombreux algorithmes de déclassement automatiques.



Un graphe issue d'une analyse effectuée par le site www.apexpaficic.com sur un échantillon de sites impactés par Penguin. En ordonnée, le pourcentage de sites concernés, et en abscisse, le pourcentage d'ancres de liens pointant vers ces sites contenant un mot clé ciblant une requête précise (par opposition à « cliquer ici » ou « rien » ou « n'importe quoi »). La distribution est claire : les sites impactés ont tous des profils montrant des ancres suroptimisées !

Vers une perte d'importance des anchor texts dans l'algorithme de Google?

La communauté SEO bruisse depuis un ou deux ans de rumeurs selon lesquelles les anchor texts auraient perdu du poids dans l'algorithme de Google.

Ceci est probablement excessif. Il est possible qu'au fil du temps, l'algorithme s'étant sophistiqué, le poids relatif des ancres ait diminué avec l'ajout de nouveaux signaux. Qui plus est, la nécessité de rendre l'algorithme plus résilient et plus résistant à des manipulations a conduit Google à sophistiquer la prise en compte des textes d'ancre pour éviter qu'il soit trop facile de « spammer » le moteur. Le résultat, c'est que les référenceurs trouvent plus difficile qu'avant d'obtenir des résultats en optimisant simplement les ancres.

Mais cela ne veut pas dire pour autant que Google a eu besoin de diminuer le poids absolu des anchor texts pour améliorer son algorithme. La vérité est sans doute plus subtile que ça.

Les anchor texts remplacés par les co-citations ?

Un article récent (fin 2012) de Rand Fishkin (Moz.com) a été largement commenté sur le Net. Dans cet article, Rand Fishkin annonçait avoir détecté un cas où une autre méthode que l'analyse des anchor texts semblait expliquer la présence en tête de résultats de certaines pages. Il avait appelé cette méthode la « co-citation ». L'idée serait que l'analyse de textes (sans liens) contenant des allusions à un site web puisse aboutir à la création d'un lien virtuel, dont l'anchor text (virtuel) serait établi à partir des mots clés les plus fréquemment associés dans ces textes à la page web cible.

En fait, au stade actuel de nos connaissances, il est assez difficile de savoir si ces exemples ne s'expliquent pas par d'autres mécanismes, comme l'indexation des phrases, ou la reconnaissance des entités nommées et autres formes d'indexation de dernière génération (comme celle qui sert à réaliser le « knowledge graph » de Google). Mais il semble impossible que cette approche remplace totalement l'exploitation des anchor text : elle est en réalité complémentaire, elle améliorerait le fonctionnement du moteur, mais ne pourrait se substituer à l'approche traditionnelle.

Faut-il encore optimiser ses « anchor texts » en 2013 ?

Les stratégies d'optimisation des anchor texts continuent d'être redoutablement efficaces pour se positionner dans Google. Ce qui a changé est que certaines méthodes trop grossières et artificielles sont à proscrire : elles sont détectées (facilement ?) par Google, et peuvent conduire au déclassement de votre page ou même d'un site entier. Mais ceci est vrai surtout pour les backlinks externes. Le webmaster est beaucoup plus libre de faire ce qu'il veut sur les liens internes. Néanmoins, il est également devenu plus difficile d'avoir des résultats avec des méthodes d'optimisation des anchor texts dans le maillage interne d'un site. Mais cela n'empêche pas d'obtenir encore des résultats spectaculaires avec ces méthodes.

Les techniques d'optimisation des anchor texts sont-elles durables ? Pour certaines d'entre elles certainement pas, Google devient de plus en sévère et de plus en plus sensible à ce qui n'est pas selon lui « naturel » et à ce qui ressemble de près ou de loin à de la manipulation de son algorithme. Mais gageons que comme Google ne peut pas se passer de l'information extraite des anchor texts pour classer correctement ses pages de résultats, les méthodes d'optimisation SEO basées sur les anchor texts ont encore de beaux jours devant elles.

Bibliographie

The Anatomy of a Large-Scale Hypertextual Web Search Engine
Sergey Brin and Lawrence Page
http://infolab.stanford.edu/~backrub/google.html

Liens divers

« Miserable failure » : compte rendu sur Abondance en 2003 :
http://actu2.abondance.com/2003-50/miserable-failure.html

Google bombing : article de Wikipedia :
http://fr.wikipedia.org/wiki/Bombardement_Google

L'article de Moz sur les anchor texts et la co-citation :
http://moz.com/blog/prediction-anchor-text-is-dying-and-will-be-replaced-by-cocitation-whiteboard-friday

Philippe YONNET , ***Directeur de l'agence Search-Foresight / Groupe MyMedia. Président de l'association SEO Camp (http://www.seo-camp.org/)***