Les expansions de requêtes à l'aide de synonymes dans Google

Retour au sommaire de la lettre

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

Google a dernièrement modifié et amélioré - encore une fois - la façon dont son moteur de recherche prend en compte les synonymes lors d'une recherche. Ce qui peut paraître simple - remplacer un mot clé par un autre - est en fait basé sur des algorithmes très pointus qui s'améliorent d'année en année pour amener une meilleure expérience utilisateur. Cet article vous propose un petit voyage dans le monde de la linguistique et de la sytaxe, entre "expansions de requêtes" et "formes fléchies", pour mieux comprendre la complexité des algorithmes actuels des moteurs de recherche et évaluer les impacts en termes de référencement...

Le 19 janvier 2010, un billet publié par Steven Baker sur le blog officiel de Google (http://googleblog.blogspot.com/2010/01/helping-computers-understand-language.html) a révélé que Google utilisait à présent un système perfectionné d'expansion de requêtes à base de synonymes pour améliorer la pertinence de ses résultats. C'est une évolution intéressante du moteur vers une solution expérimentée par de nombreux chercheurs en Recherche d'Informations depuis une vingtaine d'années.

En quoi consiste une "expansion de requête" ? Pourquoi est-ce utile dans un moteur comme Google ? Pourquoi est-ce qu'il a fallu attendre si longtemps pour voir apparaître cette fonctionnalité ? Et qu'est-ce que cela change d'un point de vue SEO ? Voila toute une série de questions auxquelles nous chercherons à répondre dans cet article...

Qu'est-ce qu'une expansion de requête ?

Dans un moteur de recherche traditionnel, l'utilisateur tape des termes dans un champ de recherche, et le système renvoie une page de résultats classée selon le niveau de pertinence supposé des documents par rapport à la requête. Mais le fait d'interroger le moteur en tapant des mots clés comporte en soi une difficulté majeure inhérente aux caractéristiques du langage.

Tout d'abord, un mot a plusieurs graphies possibles, c'est-à-dire qu'il existe plusieurs manières de l'écrire. Ces "graphies" différentes peuvent être dues à des fautes d'orthographe, mais aussi dans certains cas à l'existence de plusieurs graphies officielles (par exemple: clé / clef). Par ailleurs, certains mots ou expressions peuvent être abrégés, ou représentés par des symboles différents (2 kilos, deux kg, 2 kilogrammes). Certaines expressions peuvent être utilisées soit sous leur forme développée, soit sous la forme d'acronymes, (exemple : USA / United States of America). Etc.

On comprend donc que le fait de ne chercher qu'une seule graphie a pour conséquence immédiate d'éliminer toute une collection de documents qui comportent les graphies alternatives, alors qu'ils sont tout aussi pertinents !

Ensuite, un même mot peut prendre plusieurs formes sans changer de sens pour des raisons de syntaxe (la "grammaire" de la langue). Par exemple la marque du pluriel (régulier : toile / toiles, ou irrégulier : oeil / yeux), du genre (malin/maline, cheval/jument) ou de la conjugaison des verbes (aimer/aimes mais aussi est/sont). Dans les langues à déclinaison, on aura aussi la marque du cas. Ces formes dites "fléchies" sont plus ou moins variées et régulières selon les langues. On a même des cas en français de formes fléchies double ("asseyez-vous" et "assoyez-vous" : la deuxième forme était encore utilisée jusque dans les années 50 dans certaines régions et la plus répandue au Québec). Une fois de plus, comme il s'agit de "formes" d'un même mot, le fait de ne pas chercher ces formes "fléchies" peut conduire le système à "oublier" toute une série de documents pertinents.

Et enfin, pour couronner le tout, les langues contiennent des mots ou expressions synonymes, c'est à dire des termes ou groupes de termes qui ont des sens très proches. Les synonymes parfaits n'existent pas : les linguistes ont tendance à considérer que si deux mots avec le même sens subsistent dans la langue c'est parce qu'il existe une nuance, même subtile, qui les distingue encore entre eux. Ces nuances s'expriment par exemple par le fait que les mots sont parfois substituables dans un certain contexte, et pas dans d'autres : "explorateur" est synonyme de "navigateur" dans le domaine du web, mais les mots ne sont plus substituables dans d'autres contextes. Dans la plupart des cas, ces nuances sont liées aux connotations (par exemple péjoratives : noir/nègre) liées à certains mots, ou au registre du langage (flic / policier). Or les documents qui parlent de "policiers corrompus" peuvent être également pertinents sur la requête "flic ripoux".

Bref, dans de nombreux cas, la recherche des termes tapés par l'internaute est insuffisante pour identifier tous les documents pertinents. Le billet de Steven Baker de Google donne une information intéressante : selon lui, les chercheurs de Google ont évalué le pourcentage des requêtes "affectées" par ce problème à 70% des recherches effectuées sur le moteur !

Compte tenu de l'existence de tous ces cas, une solution peut être d'étendre la requête tapée par l'internaute en ajoutant automatiquement les variantes qui ont le même sens à cette requête. C'est ce que l'on appelle une "extension de requête" ou une "expansion de requête".

Les affres de l'expansion de requête

Très tôt (dès les années 60), les créateurs de moteurs de recherche ont donc cherché à utiliser les expansions de requête pour renvoyer plus de documents pertinents (c'est à dire améliorer le "rappel" selon le terme consacré). Mais ils se sont vite aperçus qu'il y avait un prix à payer : si le système renvoie plus de documents pertinents grâce aux expansions de requête, il renvoie aussi beaucoup plus de "bruit", c'est-à-dire de documents non pertinents ! Bref, l'augmentation du rappel obtenue se fait au détriment de la "précision" du moteur, donc le ratio "nb de documents pertinents envoyés" / "nb total de documents retournés"). Bref le rapport "signal sur bruit" du moteur n'était pas vraiment amélioré, et même souvent dégradé. Dans un cas, une étude à révélé qu'ajouter des termes pris au hasard ou des synonymes donnait les mêmes résultats en termes de précision du moteur.

Cette absence d'amélioration peut sembler surprenante, sauf si l'on prend en compte deux autres caractéristiques du langage : l'homonymie, l'homographie, et la polysémie. Les mots homonymes s'écrivent de la même façon, mais ont des sens totalement différents : ex "l'avocat mange un avocat", "le mousse mange de la mousse au chocolat", ou "faire le tour de la tour". Ensuite, la grammaire peut jouer des tours en créant des formes fléchies qui s'écrivent exactement comme d'autres mots de la langue : ex "les poules des soeurs du couvent couvent", "nous avions des avions", "je bois dans un godet en bois". C'est ce que l'on appelle des cas d'homographie. Enfin, il existe des mots dont le sens change radicalement selon le contexte, bien que restant rattaché, pour des raisons étymologiques par exemple, au même champ sémantique : par exemple "hôte", qui signifie *celui qui reçoit* ou *celui qui est reçu*, ou la "feuille" qui peut désigner la feuille de papier et la feuille d'un arbre. C'est le problème de la polysémie : les mots ont plusieurs sens.

Toutes ces caractéristiques du langage créent des ambiguïtés. Lorsque l'on procède à une expansion de requête, on est donc condamné à produire différents phénomènes qui à la longue dégradent la qualité (la "précision") du moteur de recherche :
- on étend la requête à des termes de natures différentes : par exemple verbe et nom "avions" et "avions" fait ajouter "posséder" et "aéroplanes".
- ou on étend la requête à des termes de champs sémantiques différents : par exemple "poulet" étendu à "policier" et "volaille".

Pour améliorer le système, il faut essayer de procéder avant toute chose à une "désambiguïsation", c'est-à-dire identifier la nature du terme et son sens dans le contexte.

Pour identifier si le terme est un nom, un adjectif, ou un verbe conjugué, il faut procéder à une analyse syntaxique, et s'appuyer sur les règles de grammaire pour identifier la nature du terme. L'exercice est déjà difficile dans un texte rédigé en langage naturel, il devient

impossible dans une requête : les internautes tapent des suites de mots clés, pas des phrases formulées en "bon français".

Pour identifier si le mot est pris dans une acception ou une autre, il faut s'appuyer sur le contexte. Or, dans une requête à deux ou trois mots, ce contexte est relativement limité (si je cherche "avocat mexicain", je cherche un défenseur à Mexico ou la recette du guacamole ? Impossible à dire). Bref, désambiguiser correctement les termes d'une requête est le plus souvent impossible...

Même lorsque les termes ne sont pas ambigus, l'absence de synonymes "vrais" produit une dérive. Par exemple dans la requête "Président de la République", le terme "président" correspond au nom, et non à une forme conjuguée de "présider". Si j'utilise un dictionnaire de synonymes pour réaliser une expansion de requête, je vais trouver des termes comme "dirigeant" qui renverra des chefs d'entreprise, "gouvernant" qui renverra aussi des ministres, et surtout "monarque" ou "roi" qui m'éloigne un peu des résultats sur Nicolas Sarkozy, Jacques Chirac ou François Mitterrand.

Bref on comprend mieux pourquoi réaliser des expansions de requête n'améliore pas toujours la pertinence d'un moteur.

Les expansions de requête dans Google

Compte tenu de la difficulté de l'exercice, on comprend mieux pourquoi Google a pendant très longtemps recouru aux expansions de requêtes de manière limitée. En fait, on trouvait relativement peu de fonctionnalités linguistiques avancées dans Google dans les premières années de son existence, elles sont apparues progressivement ensuite.

Le stemming automatique

Historiquement, l'une des premières formes apparues dans Google est ce que l'on appelle le "stemming". Le stemming est une procédure qui vise à regrouper les mots sémantiquement proches à partir de ressemblances graphiques, en les associant à un mot racine (le "stem"). Le fait que les mots s'écrivent de manière similaire ne signifie pas toujours qu'ils appartiennent à la même famille, ou qu'ils sont rattachés sémantiquement, mais on peut réaliser des algorithmes qui reconnaissent les formes fléchies et créer des listes de graphies reliées au même "stem". En français ce processus se traduit par "racinisation".

La langue anglaise se prête facilement au "stemming" car elle génère peu de formes fléchies, est assez régulière, et fonctionne par suffixation essentiellement. Cela permet par exemple de relier facilement "fisher, fishing, fished" à la racine "fish".

Les premières observations du stemming dans Google datent de novembre 2003. Le stemming a été étendu un an plus tard mais de manière limitée dans d'autres langues.

Il faut dire que le stemming donne de piètres résultats dans la langue française, qui est beaucoup moins "régulière" dans sa logique de formation des formes fléchies. Pour le français, une autre approche est en principe plus appropriée : la lemmatisation, c'est à dire le regroupement autour du lemme, ou mot racine, sans tenir compte de la graphie. Par exemple, la lemmatisation permet d'associer "suis, sommes, sont, étaient, étions" au verbe être.

Inhiber les expansions de requête dans Google

Cette astuce n'est pas très connue, bien qu'elle ait toujours fonctionné depuis que Google a entrepris d'introduire des expansions de requête dans son moteur : il suffit d'ajouter un + devant un terme pour inhiber les expansions de requête sur ce terme. Logique puisque cet opérateur indique que vous exigez la présence du terme dans la page.

L'opérateur tilde

Google a introduit en 2003 également l'opérateur tilde ("~motclé") qui crée une expansion de requête à la demande de l'utilisateur : la requête qui contient "~motclé" est étendue à tous les termes "synonymes" du terme "motclé". L'opérateur tilde ne renvoie pas des synonymes pour tous les termes, et fonctionne mieux avec des termes génériques. Elle trouve également beaucoup plus de synonymes en anglais qu'en français. On peut remarquer aussi que la notion de synonyme gérée par cet opérateur est assez large : il s'agit de remplacer un terme par un terme jugé équivalent dans une requête.

Comment reconnaître l'utilisation d'une expansion de requête par Google ?

Il existe un moyen très simple pour identifier si le moteur a procédé à une expansion de requêtes : il suffit de regarder les termes mis en gras dans les pages de résultats pour voir si on ne trouve d'autres termes en gras que les mots de l'expression clé tapée dans le champ de recherche. Attention : il existe quelques situations où la mise en gras est boguée : les termes ne sont pas toujours correctement mis en gras.

La recherche des variantes orthographiques

Google, au fil du temps, a développé des approches de plus en plus sophistiquées pour corriger ou proposer des corrections pour les termes mal orthographiés. On trouve aujourd'hui plusieurs systèmes mis en oeuvre en fonction de la situation. Lorsqu'il existe plusieurs graphies possibles, mais pas d'erreur possible sur ce qui est recherché, Google procédera à une expansion de requête sur la bonne orthographe ("bri**tt**ney spears" renvoie des résultats étendus à la requête "bri**t**ney spears").

A l'opposé, lorsque la probabilité d'une faute d'orthographe ou de frappe est importante mais pas assez pour que l'erreur soit considérée comme certaine, Google suggérera juste l'existence d'une graphie alternative possible ("Essayez avec cette orthographe:" ou "Did You Mean"). Et dans les cas où deux graphies semblent équivalentes en terme de fréquence d'occurrence, Google présente une page de résultats composite comportant des résultats issus des deux requêtes possibles.

La "traduction" automatique de la requête

Autre forme d'expansion de requête présente dans Google : la traduction automatique. Si on recherche "*furnished apartment Paris*" dans Google, la page de résultats comporte des pages qui correspondent à la requête "*appartement meublé Paris*" (en français). Cette expansion sur les termes traduits se produit quand la requête dans la langue de départ risque de ne pas ramener tous les documents pertinents sur la requête.

L'identification des acronymes

Depuis quelques mois, Google réalise une expansion de requête entre les versions développées et non développées d'un acronyme. Par exemple, PPC renvoie des pages qui contiennent effectivement l'acronyme PPC, mais aussi "*pay per click*" (et "*Pocket PC*" également, les acronymes sont souvent ambigus.)

Comment Google "oublie" des termes dans les requêtes

C'est un peu moins connu, mais Google peut également "oublier" volontairement des termes dans une requête. Dans certains cas, il s'agit purement et simplement d'un mot vide ("*stop word*") en anglais. La plupart des moteurs ne tiennent pas compte dans leurs recherches de termes qui ne sont pas porteurs de sens.

Par exemple, les termes "le" ou "de" en français sont classés dans les mots vides. Google a changé sa manière de gérer les mots vides depuis environ deux ans. Avant cette date, une requête sur "*a room with a view*" aurait produit des résultats contenant "*room*" et "*view*" uniquement, plus un avertissement signalant que les termes "*a*" et "*with*" ont été ignorés dans la requête. Mais la gestion des requêtes s'est sophistiquée, et les expressions toutes faites comme "*a room with a view*" sont détectées : la requête renvoie bien des pages contenant l'expression exacte "*a room with a view*". Les stopwords sont donc, en fonction des circonstances, soit ignorés, soit conservés. Bing a également adopté cette gestion sophistiquée des stopwords.

Mais Google procède également à de véritables "réductions" de requête, en faisant disparaître de la requête des termes porteurs de sens (ce ne sont pas des "stopwords") qui risquent de diminuer inutilement le nombre de résultats renvoyés.

La reformulation automatique des requêtes sous forme de questions en requêtes plus appropriées

En 2007, Udi Manber (*VP of Engineering* chez Google) a expliqué dans une présentation que l'analyse des requêtes tapées par les internautes avait permis à Google d'imaginer une solution pour reformuler certaines requêtes exprimées sous forme de question en requêtes plus adaptées et ceci, automatiquement.

C'est ainsi que la requête : "*overhead view of bellagio pool*" est reformulée en "*bellagio pool pictures*" (mais aussi en "*view of bellagio poo*l"). La requête "*fedora 5 losing network connections*" devient "*fedora 5 network configuration*". Ou encore "*How much does it cost for an exhaust system*" qui est reformulé en "*cost of an exhaust system*".

Dans ces cas, la requête réellement "cherchée" par le moteur diffère sérieusement de la requête tapée par l'internaute...

La recherche de synonymes

Le 19 janvier 2010, Google a donc communiqué sur la mise en place depuis décembre d'un système étendu d'expansion de requêtes aux synonymes des termes de la requête. Il s'agit d'une expansion de requête automatique, par opposition aux effets de l'opérateur tilde évoqué plus haut. Ce nouveau système d'expansion de requête produit des résultats différents en fonction du contexte, ce qui permet d'éviter au maximum la génération du bruit inhérente à la plupart des systèmes d'expansion traditionnellement utilisés.

Comment Google pourrait-il identifier les synonymes ?

La lecture d'un brevet déposé par John Lamping et Steven Baker (*Determining query term synonyms within query context*) donne quelques indications sur la méthode utilisée par Google. Comme d'habitude, l'approche privilégiée par Google est statistique, et ne s'appuie pas sur des bases linguistiques ou sémantiques. D'une manière générale, Google comme tous les grands moteurs commerciaux doivent être capables de créer des algorithmes universels, qui fonctionnent quelle que soit la langue ou le pays. Constituer des bases de synonymes ou de termes équivalents constituerait un travail long, et coûteux compte tenu de toutes les versions pays et versions linguistiques qu'il faudrait ainsi améliorer. En outre, l'approche par thésaurus de synonymes a pour défaut de produire du bruit, parce qu'elle ne tient pas compte du caractère ambigu de certaines termes.

L'approche statistique, en croisant les données issues de l'analyse du contenu de milliards de pages web, et d'années de logs de requête, permet par contre d'obtenir trois résultats intéressants :
- l'analyse des termes de la requête permet, en se basant sur les scores obtenus, de mieux désambiguiser le sens des termes contenus dans ces requêtes.
- cette désambiguation permet de savoir quelle famille de synonymes chercher, en excluant les

autres.
- l'approche contextuelle permet également de trouver des synonymes qui ne sont des synonymes que dans un contexte particulier.

Pour illustrer le 3e cas, Steven Baker cite le terme "*music*", qui n'est pas en principe un synonyme de "*loops*", sauf dans le cas de certaines requêtes comme "*free loops for flash movie*"

Quelles sont les indicateurs statistiques exploités pour identifier les synonymes ? Le brevet en cite trois :
1. Le nombre de fois ou le pourcentage de cas où les deux termes apparaissent dans les requêtes sur une période donnée.
2. Le nombre de fois ou le pourcentage de cas où les deux termes apparaissent dans une session utilisateur donnée.
3. La similarité entre les résultats retournés avec la requête originale, et une requête dans laquelle on substitue à certains termes un terme candidat en tant que synonyme.

Quelles conséquences pour le SEO ?

Pour le SEO, il est important de bien comprendre comment un moteur gère les requêtes d'un internaute. Il ne sert à rien, en particulier, de chercher à optimiser un site pour une expression spécifique, si on s'aperçoit en fait que les requêtes des internautes sont automatiquement reformulées pour chercher aussi d'autres expressions.

Beaucoup de stratégies basées sur la création de pages optimisées pour toutes les variantes possible dans la graphie d'une expression clé sont rendues caduques (ou presque) par la multiplication des cas d'expansion de requêtes. L'extension récente de ces expansions de requête à la recherche de synonymes rend aussi la création de pages répondant à toutes les expressions figurant dans un champ sémantique donné moins pertinent (par exemple : une page pour "véhicule", une page pour "automobile", une page pour "voiture"...).

Pour le moment, les cas traités ainsi sont encore statistiquement peu fréquents, mais Google semble vouloir progressivement étendre ce fonctionnement à beaucoup plus d'expressions. Globalement, ces évolutions rendent le travail du référenceur plus subtil. Cela rend également les résultats moins prévisibles car il devient difficile de "prédire" les expressions les plus cherchées sur les moteurs de recherche.

Bibliographie

The Official Google Blog - 19 janvier 2010
Helping computers understand language.

Translating Queries into Snippets for Improved Query Expansion
Stefan Riezler, Yi Liu, Alexander Vasserman, Proceedings of the 22nd International Conference on Computational Linguistics (COLING'08), 2008.

Statistical Machine Translation for Query Expansion in Answer Retrieval (pdf).
Stefan Riezler, Alexander Vasserman, Ioannis Tsochantaridis, Vibhu Mittal and Yi Liu

Brevets

Machine Translation for Query Expansion
Invented by Stefan Riezler, Alexander L. Vasserman
Assigned to Google
US Patent Application 20080319962
Published December 25, 2008

Determining query term synonyms within query context
Invented by John Lamping and Steven Baker
Assigned to Google
US Patent 7,636,714
Granted December 22, 2009
Filed: March 31, 2005

Philippe Yonnet, ***Directeur Technique @Position (http://www.aposition.com) et président de l'association SEO Camp (http://www.seo-camp.org/)***

Réagissez à cet article sur le blog des abonnés d'Abondance :
http://blog-abonnes.abondance.com/2010/03/les-expansions-de-requetes-laide-de.html