De l'indexation des mots clés à l'indexation des concepts (1ère partie)

Retour au sommaire de la lettre

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

Depuis que les moteurs de recherche existent, le référencement existe. Mais les moteurs ont grandement évolué depuis près de 20 ans. D'un système "simple" d'indexation de mots clés isolés, ils sont passés à la détection de syntagmes, de synonymes puis aujourd'hui aux entités nommées et aux "index de concepts"... Nous faisons un retour sur cette évolution dans cet article en deux parties. Une évolution que les référenceurs devront prendre en compte à l'avenir pour s'adapter aux progrès toujours plus importants que font les moteurs de recherche pour mieux analyser le Web et rendre leurs résultats toujours plus pertinents...

Depuis l'apparition des premières techniques de référencement, le "mot-clé" a toujours été le centre des préoccupations des spécialistes du SEO. Le positionnement des pages sur un ou plusieurs mots clés constitue même l'unique objectif pour la plupart des référenceurs dans leur travail quotidien.

Cette obsession du mot-clé tire son origine de la structure même du fonctionnement d'un moteur de recherche. Mais cette structure a fortement changé au fil des ans : les "syntagmes" ont remplacé les mots-clés, sans que la plupart des référenceurs n'en tirent de conséquences dans leurs méthodes. Aujourd'hui, les moteurs commencent à manipuler des "concepts", et cette dernière évolution rend les optimisations SEO basées sur les mots clés un peu plus caduques encore.

Dans cet article, nous allons rappeler le fonctionnement de ces différents types d'index, avant d'expliquer ce que change l'utilisation des "concepts" pour le référencement, et préciser comment il faut s'adapter.

L'approche originelle : l'indexation des mots clés

A l'origine, les moteurs de recherche sur internet étaient surtout construits pour extraire, au sein d'une collection de documents web, ceux qui contenaient les termes présents dans la requête. Pour les trouver, on pourrait penser naïvement qu'à chaque requête, le moteur parcourt le contenu de toutes les pages du web pour trouver les documents qui contiennent les mots clés de la requête. Le problème, c'est que cette approche est désespérément lente et inefficace.

Pour accélérer la production de la page de résultats, les moteurs stockent à l'avance dans une "base de données" (je place l'expression entre guillemets pour indiquer que le mode de stockage et d'interrogation dans un moteur est assez différent de ce que l'on trouve dans une base de données SQL, même si les points communs sont assez nombreux) la liste des différents mots clés identifiés au sein des pages, accompagnés de la liste des documents qui les contiennent.

Cette "base de données" ressemble *grosso modo* dans sa structure, aux index figurant à la fin des livres. Cette utilisation des "**index inversés**" explique pourquoi par extension on a pris l'habitude d'appeler la "base de données" des moteurs de recherche un "**index**".

Notons au passage que le nombre d'entrées dans cet index est beaucoup plus important qu'on ne le pense souvent. Le nombre de termes différents que l'on trouve sur le web se compte en millions, alors que le nombre d'entrées dans un dictionnaire se compte habituellement en dizaine de milliers. La différence provient de la présence des "**formes fléchies**" (graphies qui changent selon le genre, le nombre, le cas, la conjugaison etc...), de beaucoup de termes de "jargon", et de tous les noms propres et autres entités nommées qui s'ajoutent à tout cela. N'oublions pas en plus qu'un moteur comme Google travaille sur des pages rédigées dans de nombreuses langues différentes et que cela multiplie d'autant la taille des index nécessaires.

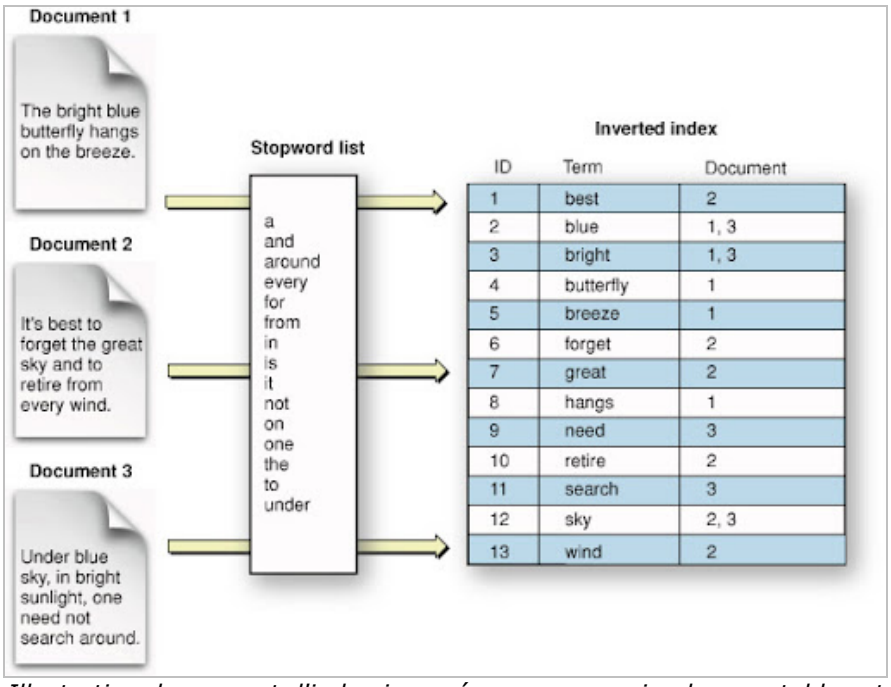
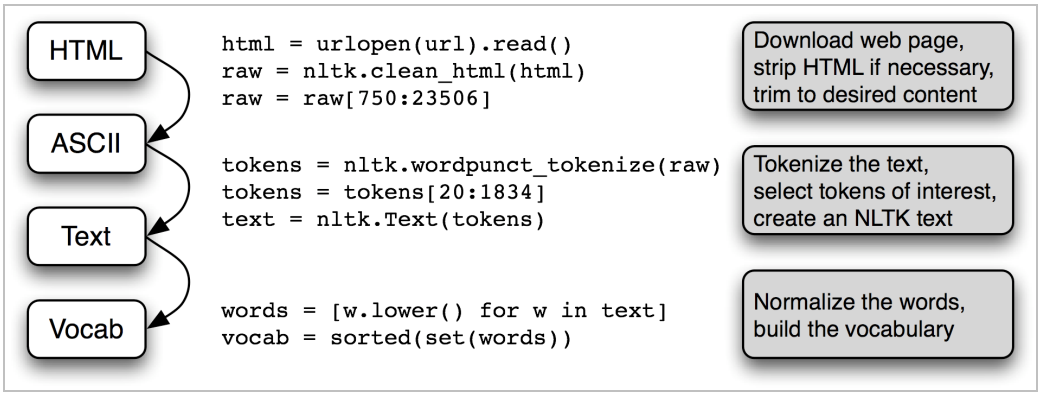


Illustration du concept d'index inversé pour un cas simple: une table est construite avec une entrée pour chaque terme identifié dans les documents, et en regard de chaque terme les identifiant des documents qui contiennent ce terme. On peut aussi stocker la position dans le document, ce qui permet de mesurer des distances entre les termes au sein d'un même document.

Le processus de création de l'index (**l'indexation**) se révèle en réalité particulièrement complexe, et passe, dans un moteur classique, par plusieurs étapes successives de traitement des pages web :

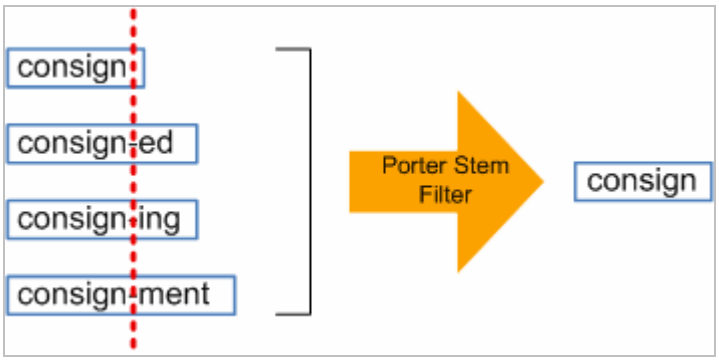
- le "**parsing HTML**" (analyse syntaxique HTML) et la "linéarisation" tout d'abord, qui ont pour objectif d'isoler le texte de la page du code qui l'entoure.

- la "**normalisation**" et la "**tokenisation**" ensuite. Le texte est débarrassé de tous les signes inutiles comme la ponctuation et certains caractères spéciaux, et converti en "bas de cas de casse" c'est à dire en caractères minuscules. A la fin de ce processus, il ne reste que des termes séparés par des espaces, et ces termes sont isolés ("tokenisés") pour être indexés.



Processus d'indexation simplifié

- la construction de l'index enfin : la position du terme est stockée en regard de l'entrée correspondant à ce terme dans l'index. Pour économiser de la place et gagner en pertinence, on peut procéder à une racinisation ("**stemming**") ou à une "**lemmatisation**", qui permet de stocker les formes fléchies (pluriels, conjugaisons etc...) en regard du terme racine ou du lemme, plutôt que de créer une entrée pour chaque graphie. On évite également d'enregistrer des entrées dans l'index pour les "**mots vides**" (""), qui sont des termes non porteurs de sens : savoir s'ils sont présents ou non dans un document n'apporte rien pour déterminer s'ils sont pertinents comme réponse sur une requête donnée.



Un exemple de « stemming » (racinisation en Français) : tous les termes commençant par « consign », quel que soit le suffixe, sont regroupés autour de la racine « consign ».

Les limites de cette approche

Cette approche basée sur des mots clés uniques était efficace dans un contexte de requêtes courtes, en particulier des requêtes comportant un seul mot. Mais les internautes ont appris à taper des requêtes comportant plusieurs termes pour faire apparaître directement les bonnes pages de résultats.

Dans une telle requête, l'ordre des mots peut avoir une importance : "*meilleur prix téléphone portable*" ne doit pas ramener les pages contenant la phrase "*le G28ZN est le meilleur téléphone de sa gamme, mais son prix le classe parmi les plus chers des portables de sa catégorie*". Pourtant c'est ce qui se produit le plus souvent, et cela dégrade clairement la pertinence des résultats.

La solution la plus fréquente avec un index de mots clés est de tenir compte de la **distance** entre les termes, ce qui favorise la remontée en haut des classements des pages plus probablement pertinentes. Concrètement, une requête sur "*cul de jatte*" renverra des pages contenant "*cul*" et "*jatte*" séparé par un mot (mais pas zéro mot ou plus d'un mot), sans que cela renvoie des pages de sites adultes, ou des pages sur de collectionneurs de récipients... Notons au passage que le "de" est ignoré dans le processus car il s'agit typiquement d'un mot vide (« stopword »), mais que sa présence est prise en compte dans le calcul de la distance.

Mais même cette approche est souvent prise en défaut. Sur la requête "*Champion de France de tennis de table*", les pages contenant "*champion de France de tennis à table*" peuvent sortir. Pour éviter cela, une approche capable d'identifier "*tennis de table*" comme un groupe nominal fréquent et formant un tout indissociable est nécessaire. Cela signifie que l'on ne peut plus "tokeniser" les textes de manière traditionnelle. La solution s'appelle le "**phrase indexing**", c'est à dire l'indexation des syntagmes.

L'indexation des phrases : attention au « faux ami » !

Les linguistes anglo saxons emploient facilement le terme anglais "*phrase*" pour décrire un groupe de mots. Le terme "*phrase*" en français se traduit mieux par "*sentence*" en anglais. Pour éviter toute ambiguïté, on emploiera plutôt dans cet article, soit le terme "*groupe de mots*", soit le terme technique "*syntagme*". Un syntagme constitue l'étape intermédiaire entre le mot et la phrase : c'est un groupe de mots qui forme une unité par son sens et par sa fonction, à l'intérieur de la phrase.

Par exemple dans la phrase "*Le chien du voisin a aboyé toute la nuit*" on pourra isoler les syntagmes "*Le chien du voisin*" "*a aboyé*" et "*toute la nuit*".

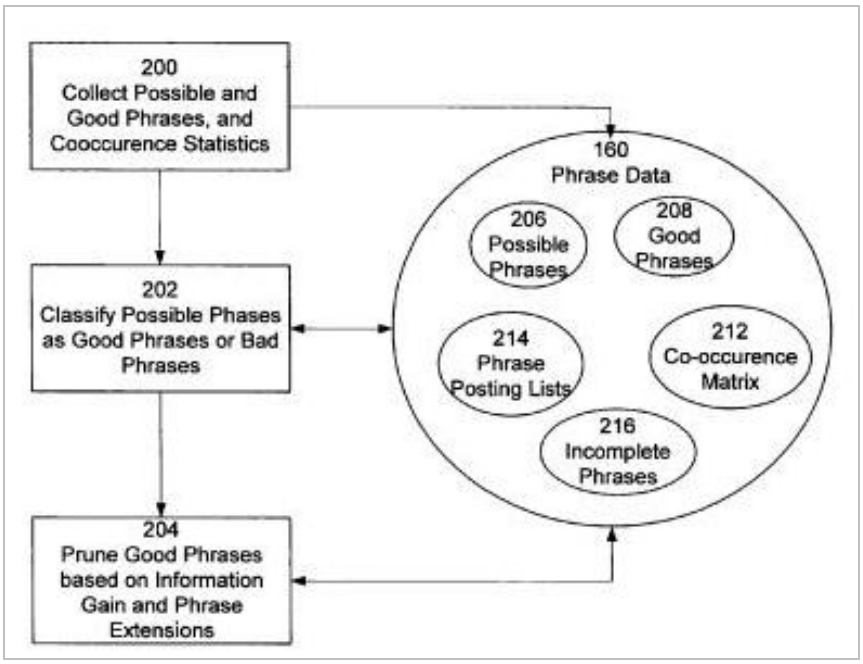
Donc il faut faire attention lorsqu'on lit des articles en anglais sur ce sujet : quand les auteurs parlent d'indexer des phrases, en réalité leurs ambitions sont bien plus modestes et ils cherchent à indexer des groupes de mots, parfois très réduits en nombre.

Comment indexer des syntagmes ?

Pour permettre à un moteur de reconnaître des groupes de mots, on analyse les **fréquences de cooccurrence** (c'est-à-dire le nombre de fois où ces termes apparaissent ensemble dans le même ordre) des termes. Les termes qui ont des fréquences de cooccurrence élevées (comme "*carte blanche*" ou "*voeu pieux*") sont de bons candidats, et sont placés dans l'index.

Le problème avec cette approche, c'est que décider d'indexer des groupes de mots conduit à une **explosion combinatoire**, qui pourrait faire grossir l'index dans des proportions ingérables. On se référera à l'article de juin 2010, dans cette même lettre, sur l'indexation des syntagmes pour avoir plus de précisions sur ce problème et ses solutions (cf Bibliographie le mois prochain).

De la même façon, le traitement des requêtes est rendu plus complexe, puisqu'il faut aussi identifier les syntagmes présents dans les requêtes et décider de ce que l'on doit chercher dans l'index. Cet exercice n'est pas trivial : dans la requête "*champion de France de tennis de table*", il faut éviter de reconnaître "*champion de France de tennis*", qui a une toute autre signification. En fait, "*champion de France*" et "*tennis de table*" sont deux syntagmes qui auront de fortes probabilités d'être reconnus individuellement. Le problème est plutôt de reconnaître le syntagme complet "*champion de France de tennis de table*" qui est le seul susceptible d'être associé avec des pages pertinentes !



Une illustration issue du brevet déposé par Anna Paterson de Google sur le « Phrase Based Indexing ».

Dans la pratique, on n'indexe que les syntagmes qui présentent une forte **probabilité de cooccurrence**, c'est à dire les groupes de mots les plus susceptibles de former une expression toute faite. Cela permet de limiter la taille de l'index à un niveau acceptable.

Notons tout de même que Google aura eu besoin de pratiquement dix ans de travail pour réussir à implémenter un index de syntagmes dans son moteur. Cette évolution semble avoir été rendue possible par la mise à jour de l'infrastructure de Google qui a été appelée « Big Daddy » par les SEO et qui date de début 2006. Les premiers « symptômes » de la présence d'un index de phrases ont été détectés bien plus tard, vers 2008, mais cela ne signifie pas qu'un tel index n'existait pas plus tôt.

Les index de syntagmes et les M. Jourdain du SEO

Dans la pratique, l'indexation des syntagmes a amélioré la pertinence du moteur, mais n'a pas changé la manière de faire du SEO.

En réalité, ce que les référenceurs appellent souvent des "mots clé" sont en fait des "expressions clé". On cherche depuis longtemps par exemple à optimiser une page sur "*bon plan*" (dans cet ordre) et non sur "*bon*" et "*plan*". La même remarque s'applique au SEA (les Adwords). Ceci est logique, puisque ces "expressions clé" sont censées correspondre à des requêtes tapées par les internautes, et que les moteurs basés sur des index de mots isolés réussissent, par le jeu de la prise en compte des distances entre les mots, à faire remonter les pages contenant les expressions dans les premiers résultats.

L'indexation des syntagmes a donc renforcé la nécessité de raisonner par "expression clé", mais la plupart des SEOs ont fait du référencement sur des syntagmes sans le savoir... Mais l'introduction des approches sémantiques depuis quelques années, et la mutation progressive de Google en moteur sémantique, réclament cette fois-ci un changement subtil d'approche.

La révolution sémantique

Pour le moment, les processus d'indexation que nous avons évoqués ne tiennent absolument pas compte du "sens" des groupes de mots. Les mécanismes mis en oeuvre reposent sur la recherche d'une correspondance (*matching*) entre une requête (éventuellement retraitée et modifiée) avec un index.

Sur la requête "*Président des USA*", un moteur disposant d'un index de mots cherchera les pages contenant "*Président*" et "*USA*". Un moteur disposant d'un index de syntagmes cherchera quant à lui "*Président des USA*" et "*Président des Etats Unis*". Mais un moteur sémantique sera capable d'associer "*Barack Obama*" et "*Président Obama*" à "*Président des USA*". On augmente donc le rappel du moteur (le nombre de pages retournées) sans diminuer la précision du moteur (le ratio entre le nombre de pages pertinentes retournées et l'ensemble des pages affichées dans les résultats).

La première approche a consisté à identifier des termes équivalents et synonymes (par exemple "*bateau*" et "*navire*") et à étendre la requête aux termes équivalents. Par exemple, si on cherche "*bateau de commerce*", Google renvoie des pages contenant soit "*bateau de commerce*" soit "*navire de commerce*".

Mais ce sont les techniques de reconnaissance des entités nommées qui ont peu à peu révolutionné la manière de stocker l'information dans un moteur, ainsi que la manière de l'interroger. C'est ce que nous évoquerons le mois prochain dans la seconde partie de cet article, avec les notions de "concept" et de synonymes, qui devront certainement être pris en compte de façon importante dans les stratégies SEO de demain...

***Philippe YONNET** , Directeur SEO international, Twenga.*