# Le corpus comme guide d'optimisation des contenus (1ère partie)

*Par Guillaume et Sylvain Peyronnet*

| Domaine : | Recherche | Référencement |
|---|---|---|
| Niveau : | Pour tous | Avancé |

*Une bonne stratégie SEO passe, on le sait, par un contenu de qualité écrit pour l'internaute mais également compréhensible par une machine et, donc, un moteur de recherche comme Google. Pour cela, la mise en place de corpus lexicaux est une phase indispensable et de plus en plus utilisée par de nombreux référenceurs. Mais comment ces corpus sont-ils créés et à quoi servent-ils ? Voici un article, en deux parties, qui a pour ambition de répondre à ces questions. Avec, ce mois-ci, une explication des concepts pris en compte dans ce domaine...*

Ces derniers mois, si vous suivez l'actualité du référencement web en France, vous n'avez pas pu passer à côté du concept de « corpus » pour le référencement.

Depuis quelques semaines, les américains se mettent à en parler également, sans utiliser forcément les mêmes termes. Si même eux, qui sont d'habitude plus adeptes du référencement orienté webmarketing que technique, commencent à s'y intéresser de près, c'est bien parce qu'il y a une carte à jouer dans ce domaine.

## Qu'est-ce qu'un corpus SEO ?

Un corpus SEO, malgré son nom qui peut laisser dubitatif, n'est rien de plus qu'un guide d'aide à la rédaction. C'est un document qui contient une liste de termes, issus de l'analyse d'un corpus documentaire, auxquels sont associés des valeurs ou d'autres mots.

Finalement, ce guide est une représentation simplifiée des contenus correspondants à une thématique ou une requête.

Un rédacteur motivé par l'optimisation d'un article pour le référencement web peut alors utiliser le guide comme recette. Charge à lui de saupoudrer son texte avec les bons ingrédients, dans les bonnes proportions.

Le terme de « corpus SEO », communément utilisé actuellement par les référenceurs, risque d'évoluer pour devenir plus précis, notamment en évitant de prendre comme nom celui de la première étape d'un processus plus complexe.

De notre côté, nous apprécions de parler de « guide d'aide à la rédaction », notamment parce que cette appellation est plus fidèle à la réalité, plus concrète, et assez imprécise pour permettre de produire des variantes du guide « classique » tel qu'on le conçoit généralement.

## L'analyse de la pertinence par un moteur de recherche moderne

Un moteur de recherche moderne, comme Google, possède de nombreux critères de classements et met en branle des algorithmes complexes. Cependant, il existe deux briques qui sont fondamentalement présentes et particulièrement importantes :

- Le moteur analyse les contenus des pages afin de pouvoir répondre à une requête d'internaute par des documents pertinents pour sa requête ;
- Le moteur cartographie le web pour savoir quelles sont les pages les plus populaires. Pour cela, Google utilise le Pagerank, du nom de son inventeur, Larry Page.

La pertinence, la partie qui nous intéresse dans le cadre de cet article, est difficile à définir. C'est en effet une notion un peu abstraite pour un moteur. Si, pour un humain, il est très facile de lire la requête d'un internaute dans un moteur de recherche et d'observer si les résultats renvoyés par la machine sont de bons candidats à une certaine pertinence (la recherche donne de bons résultats), ceci est très difficile pour un script informatique.

Si l'on entend parler souvent d'intelligence artificielle, c'est surtout la superficialité du concept qui est réelle. Les ordinateurs ne savent pas comprendre les choses. Ils peuvent éventuellement apprendre, voire comparer, et ainsi sembler très intelligents. Mais pour ce qui est de réfléchir comme un être humain, on en est encore bien loin...

Par exemple, lorsque qu'un ordinateur observe une requête et une page de résultats, il s'intéressera avant tout à la syntaxe des contenus. Si la requête tapée par l'internaute se retrouve dans des contenus qu'il a indexés, il pourra détecter une similarité, qu'il assimilera à une pertinence.

C'est un bien bel abus que la machine fait. Mais il s'agit de la seule façon pour elle de pouvoir étudier efficacement des requêtes faces à des contenus. Surtout quand il faut traiter des requêtes face à des milliers de milliards de pages web.

## Représentation d'un contenu textuel et poids des termes

Le moteur, pour décider qu'une page est pertinente pour une requête, opère donc une comparaison entre les termes de la requête et les termes des pages qui sont dans son index. Il essaie de trouver les pages qui sont les plus semblables à la requête. Semblable, similaire, donc pertinent, oui, mais dans quel sens ?

L'approche la plus évidente est de penser que si une page contient les termes d'une requête, alors, il existe une similarité effective. Cela fonctionne bien sûr, mais plutôt mal. En effet, les pages sont alors soit similaires, soit pas du tout similaires. Aucune nuance ! Et il suffirait d'ajouter des mots bien choisis dans un texte pour le rendre d'un coup parfaitement pertinent…
Il est donc préférable que le moteur soit plus fin. Plutôt que de dire qu'une page est pertinente pour une requête, il va déterminer un degré de pertinence : très pertinent, moyennement pertinent, un peu pertinent… ce qui s'exprime facilement en données chiffrées.

Pour cela, le moteur va représenter le document sous forme de vecteur, un objet que l'on peut représenter dans l'espace. La

position de ce vecteur dépendra directement des mots contenus dans le document lui correspondant.

Mais là aussi, un mot est quelque chose d'assez abstrait pour un ordinateur, alors on va plutôt associer un poids, une quantité mathématique, plutôt que de réellement travailler sur des mots et sur leurs sens.

Par exemple, pour un document A, on pourrait imaginer avoir les poids suivants (on attribue comme poids le nombre de fois où les mots apparaissent) :

Document A : « Ours furie. Furie. ».
ours : 1
furie : 2

On peut alors tracer le vecteur dans l'espace.

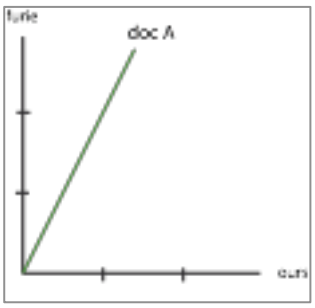


*Fig.1. Représentation du document A dans l'espace.*

On fait ensuite la même chose pour un second document B : « Furie. Ours. »
ours : 1
furie : 1
Et on trace le document dans l'espace.

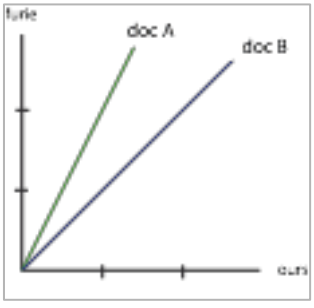


*Fig.2. Représentation des documents A et B dans l'espace.*

Maintenant, on étudie la requête « ours ours furie ». Une requête est un contenu textuel, et est donc assimilable à un document.
Document C : « ours ours furie »
ours : 2
furie : 1
On trace, de même, la requête (document C) dans l'espace.

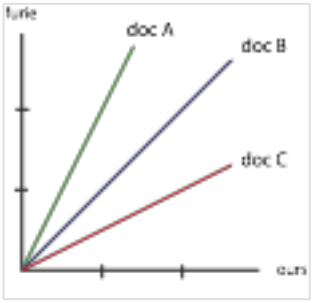


*Fig.3. Représentation des documents A, B et C dans l'espace.*

Les documents les plus proches sont ceux qui sont les plus alignés. C'est-à-dire ceux pour lesquels on passe de l'un à l'autre en effectuant la plus petite rotation possible.
Ici, le document C est plus proche de B que de A, donc on identifiera le document B comme document le plus pertinent pour la requête « ours ours furie ».

En réalité, les choses sont plus compliquées :
• Les documents présentent davantage de mots ;
• La requête subit un traitement invisible visant à agrandir sa taille (il est plus facile de faire des comparaisons efficaces entre documents de plus grandes tailles) ;
• La fonction de poids, ici le fait de compter le nombre d'apparitions du mot, est une fonction très naïve, facile à détourner.

Pourtant, le concept est bien celui-ci. Dès lors, pour peu que l'on décide de muscler la fonction de poids en faisant appel à quelque chose de plus réaliste et robuste,

on peut alors tout à fait obtenir de bons résultats.

La fonction de poids la plus usitée est la TF.IDF (dont les variantes sont nombreuses). Il s'agit d'une fonction qui attribue à chaque terme d'un document un poids dépendant du contenu du document lui-même (facteur TF) ainsi que du reste du web (facteur IDF).

Grâce à la prise en compte d'informations extérieures nombreuses et hors de portée du rédacteur du document, il devient dès lors bien plus difficile de falsifier la pertinence.

## Calculer la TF

La TF, ou fréquence du terme (*Term frequency* en anglais), est généralement exprimée par la formule : ***(nombre d'apparition du terme dans le document) / (nombre d'apparition du terme le plus fréquent dans le document).***

Cette mesure prend donc en compte les répétitions de mots. Pour évoquer un concept dans un article, on a besoin de le répéter. Ceci est logique et normal.

## Calculer l'IDF

L'IDF, ou fréquence inverse du document (*Inverse Document Frequency* en anglais), est généralement exprimée par la formule ***(nombre de documents du corpus) / (nombre de documents contenant le terme).***

L'IDF est une mesure qui s'attache à déceler la rareté d'un terme et à donner plus d'importance aux termes rares.

Pour calculer l'IDF, il est nécessaire d'avoir une base documentaire, autrement dit un corpus, pour pouvoir statuer sur la rareté d'un mot.

## Calculer la TF.IDF

Il n'y a rien de plus simple que de calculer la TF.IDF une fois que l'on a d'un côté la TF et de l'autre l'IDF, puisqu'il s'agit simplement d'une multiplication de l'un par l'autre.

## Fabriquer son propre guide d'aide à la rédaction

Ici, l'enjeu est de chercher à fabriquer un guide d'aide à la rédaction. On devine donc qu'une solution pour rendre un document très pertinent pour une requête donnée est de tracer le vecteur de cette requête et celui du document, puis de modifier le contenu jusqu'à obtenir deux vecteurs quasiment alignés. C'est bien l'idée qui préside au guide d'aide à la rédaction.

Mais faire ainsi est trop flou : parfois, en changeant un mot, une phrase, le vecteur change complètement de direction, tandis que d'autres fois il restera globalement stable. Ainsi, pour espérer aligner deux vecteurs, il va falloir tâtonner énormément. A moins d'être particulièrement rompu à l'exercice, pour un rédacteur, utiliser la représentation visuelle du vecteur est trop difficile - même si l'efficacité est forte.

Heureusement, il existe une solution plus simple d'accès. En se concentrant sur les mots et les calculs de TF.IDF attachés, on va pouvoir déterminer quels sont les mots importants d'un corpus documentaire et on va pouvoir rédiger en imitant la « normalité » du corpus. Il ne restera alors plus qu'à s'écarter de temps à autre de

cette normalité pour obtenir un peu d'optimisation supplémentaire (l'enjeu est tout de même de rendre une page plus pertinente que d'autres dans un contexte compétitif !).

Puisque ce sont les termes et les valeurs de la fonction de poids qui dictent la représentation vectorielle, cela revient à travailler indirectement sur le vecteur, en se focalisant sur les parties plutôt que sur le tout.

Après cette introduction, nous nous intéresserons le mois prochain aus différents types de corpus utilisés dans le cadre d'une stratégie SEO et nous verrons un exemple d'un tel guide. A très bientôt !

Guillaume Peyronnet *est gérant de Nalrem Médias.* Sylvain Peyronnet *est co-fondateur et responsable des ix-labs, un laboratoire de recherche privé. Ensemble, ils font des formations, pour en savoir plus : http://www.peyronnet.eu/blog/*