Le corpus comme guide d'optimisation des contenus (2ème 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. Après les bases le mois dernier, voici ce mois-ci une plongée dans les différents types de corpus avec l'exemple concret de l'un d'entre eux...

Nous reprenons ce mois-ci la suite de notre article sur les corpus et leur utilisation en référencement. Après les définitions et l'explication des grands concepts dans la première partie, le mois dernier, intéressons-nous ce mois-ci aux différents types de corpus utilisés dans le cadre d'une stratégie SEO.

Les différents types de corpus pour le SEO

Pour fabriquer un guide d'aide à la rédaction, tout commence, vous l'aurez compris maintenant, par la création d'un corpus. Mais, quels documents va-t-on récupérer pour fabriquer ce corpus ?
Il existe deux principaux cas d'usage :

- Si vous souhaitez faire en sorte que vos textes soient reconnus comme appartenant à une thématique très précise, alors le corpus à fabriquer doit intégrer des sites de référence de la thématique. Vous les connaissez certainement, il s'agit de sites particulièrement reconnus pour leurs qualités et leur popularité. Il s'agit alors de récupérer le contenu de ces sites, ce qui, de fait, créera un corpus thématisé. Une machine lisant des articles respectant les « règles » de ce corpus pourra sans difficulté en comprendre la thématique.
- Si vous souhaitez rédiger un texte sur une problématique précise, ou optimisé pour une requête particulière, il s'agira alors plutôt de récupérer des pages web qui évoquent cette même problématique ou des pages déjà optimisées pour la requête. Pour trouver de tels textes, l'utilisation d'un moteur de recherche est satisfaisante : on tape la requête ou la problématique dans la boite de recherche, tout simplement. L'essentiel est ensuite d'extraire des pages de résultats des URL et d'ajouter les pages liées à votre corpus.

Un corpus de requête est beaucoup plus simple à fabriquer qu'un corpus thématique complet car on ne va, dans le premier cas, récupérer qu'une poignée de documents, et dans le second cas analyser des sites complets.

Bien sûr, un corpus avec énormément de documents est plus fiable qu'un corpus ne comprenant que peu de textes.
Cependant, en pratique, pour le corpus de requête, quelques dizaines de textes suffisent à pouvoir obtenir un guide d'aide à la rédaction correct. Ce sera un guide approximatif bien sûr, mais comme on est de toute façon dans l'approximation de bout en bout (qui connaît les véritables formules des moteurs de recherche ?), on peut se le permettre… puisque ça marche !

C'est là où la science atteint ses limites en référencement web. On fait tout dans les règles de l'art, c'est mieux, mais on finira surtout par faire le tri pour n'appliquer et utiliser que ce qui marche.

Scraper

Une fois que l'on a récupéré une liste contenant les URL des documents que l'on souhaite intégrer au corpus, il faut « scraper » les contenus, c'est-à-dire utiliser un script qui va extraire les parties que l'on souhaite récupérer de chaque page.

Pour faire cela, de nombreuses options existent, à vous de choisir celle avec laquelle vous êtes le plus à l'aise :
• Utiliser Python et ses bibliothèques (Beautiful Soup, Requests, etc.) ;
• Utiliser PHP et des fonctions types *filegetcontents()*, ou ses modules (cURL est excellent pour cela) ;
• Utiliser un outil open source en ligne de commande (Scrapy http://scrapy.org/) ;
• Utiliser un outil propriétaire mais facile d'accès (80legs http://www.80legs.com/).
Les solutions pour récupérer des pages web ne manquent pas...

Une fois les pages récupérées, il faut les nettoyer : supprimer les balises html, le javascript, les commentaires html, le css, etc.
On peut vouloir aussi supprimer les accents car il n'est pas rare de constater que les mots accentués sont souvent écrits très différemment sur le web…

Racinisation

On peut aussi souhaiter aller plus loin en faisant de la racinisation sur les mots (cf. http://fr.wikipedia.org/wiki/Racinisation).

Il s'agit d'une opération qui consiste à ne conserver pour chaque mot que son suffixe. Par exemple, « unification » et « unifier » deviennent « unifi ».

Ceci est bien pratique pour regrouper des mots qui ne différeraient que par le temps de conjugaison par exemple, mais cela a le gros défaut de finalement produire un guide d'aide à la rédaction plus difficile à lire pour le néophyte (il faut comprendre ce que le terme « unifi » que l'on retrouvera dans le guide signifie). Il s'agit donc d'une option à tester selon votre familiarité avec l'utilisation d'un guide.

Calculer et mettre en forme

Finalement, une fois les contenus extraits, il faut procéder aux calculs. Pour cela, on va simplement compter des mots, vérifier la présence de mots dans des documents, faire quelques multiplications, divisions et additions. Tout cela pour obtenir, pour chaque terme, dans chaque document, la valeur de TF.IDF.

Il ne reste alors plus qu'à fusionner les informations pour obtenir un tableau avec les valeurs pour chaque terme, pour l'ensemble des documents (on peut

récupérer aussi la TF minimale et maximale ainsi que l'IDF).

Deux écoles s'affrontent alors :
• Soit on fait des moyennes des valeurs calculées précédemment, ce qui a le défaut de ne pas être très robuste aux éventuelles imperfections du corpus ;
• Soit on fait la somme des valeurs, ce qui permet, en donnant plus de poids aux répétitions de mots à travers le corpus, de moins être sensible aux défauts.
Et c'est ce tableau qui constitue le guide d'aide à la rédaction.

Comment utiliser le guide d'aide à la rédaction ?

Maintenant que le guide est constitué, il est temps de passer à la rédaction. Les indications chiffrées qu'il délivre permettent de prendre les bonnes décisions en termes de pertinence.

Prenons comme exemple un guide d'aide à la rédaction pour la requête « cours de cuisine » (fig. 1).

Le tableau est trié par somme des TF.IDF, c'est à dire que le premier terme du tableau est vraiment le plus déterminant au niveau des contenus de la requête « cours de cuisine », puis c'est « cours », puis « chef », etc.

Le rédacteur peut donc ici piocher parmi les mots de haut de tableau, il sait que ce sont des mots particulièrement utiles. Les utiliser donnera une puissance supplémentaire à son texte.

Cependant, utiliser uniquement les « bons » mots est une stratégie dangereuse. Il faut être plus prudent :
• Les mots ont chacun une fourchette de TF. Rester dans cet intervalle est préférable pour rester discret. Par exemple, le mot « chef » est tantôt le mot le plus répété dans la page (TF max à 100%), tantôt il n'est utilisé qu'à 5,6% du mot le plus présent dans la page.
• En utilisant l'IDF, on doit prendre des décisions de rédaction : il faut des mots rares et des mots plus courants. Par exemple, on peut se rendre compte que « restaurant » est plus rare que « cours » (3,08 > 1,27).

Utiliser un guide d'aide à la rédaction peut sembler intimidant à première vue, mais on s'adapte finalement très bien.

Cooccurrences

Après la création du corpus, il est très intéressant d'aller plus loin que le simple calcul de TF.IDF. En effet, on aimerait faire comprendre au moteur que l'on maîtrise notre sujet. Implicitement, cela veut dire que l'on doit montrer qu'on est capable

mot	TF (min)	TF (max)	IDF	tf.idf (sum)
cuisine	11,8 %	100,0 %	1,21	75,41
cours	16,7 %	100,0 %	1,27	69,49
chef	5,6 %	100,0 %	1,93	30,04
paris	11,1 %	100,0 %	2,86	24,78
atelier	5,6 %	100,0 %	2,23	23,48
recettes	7,7 %	100,0 %	2,31	15,37
restaurant	8,3 %	100,0 %	3,08	14,03
gourmand	5,9 %	100,0 %	3,37	12,73

Fig.1. Guide d'aide à la rédaction pour la requête « cours de cuisine »

d'embrasser l'ensemble des termes liés à la problématique.

Or, il est difficile d'accumuler beaucoup de mots sans être artificiel. On entre rapidement dans des schémas où l'on passe du coq à l'âne en quelques lignes.

Ainsi, pour s'assurer d'une cohérence thématique et d'une bonne maîtrise de la question soulevée par l'article, on va faire un analyse de cooccurrences dans le corpus.

L'idée est de repérer quels sont les mots qui sont fréquemment associés entre eux. Si on les voit souvent ensemble dans un contexte « naturel », c'est qu'ils se complètent. Utiliser un terme sans l'autre serait presque une hérésie.

Par exemple, un corpus de requête « programme tv » fournit, pour le mot-clé « france », la liste de termes suivante :

maintenant => 0.6
voir => 0.3125
programme => 0.30232558139535
soir => 0.29411764705882
tele => 0.29166666666667
jour => 0.28571428571429
episode => 0.28571428571429
direct => 0.23529411764706
saison => 0.22222222222222
paris => 0.21428571428571

On se rend bien compte qu'il s'agit de cooccurrences propres à la thématique de la télévision.

Il existe différents types de calcul de cooccurrence, mais sans entrer dans les détails, on peut tout à fait se satisfaire, dans un premier temps, de prendre tous les termes du corpus, deux par deux, et de compter le nombre de document qui fait apparaître la paire.

mot	TF (min)	TF (max)	IDF	tf.idf (sum)
rediger un cv	12,5 %	100,0 %	2,37	37,03
faire un cv	16,7 %	100,0 %	2,52	35,09
lettre de motivation	11,8 %	100,0 %	2,33	30,30
comment faire un	25,0 %	100,0 %	2,58	29,61
comment rediger un	9,1 %	100,0 %	2,58	28,60
rediger son cv	4,2 %	100,0 %	2,63	23,17
de votre cv	14,3 %	100,0 %	2,03	21,81
un bon cv	12,5 %	100,0 %	2,90	20,58
cv comment rediger	37,5 %	100,0 %	3,39	16,45

Fig.2. Guide d'aide à la rédaction pour la requête « comment faire un cv »

Les n-grams

Tout au long du processus de création d'un guide d'aide à la rédaction, nous sommes partis du principe que l'on travaillait mot à mot. En réalité, on peut tout à fait décider, arbitrairement, qu'un terme est en réalité une suite de deux mots, voire trois mots, ou plus. Quand on travaille sur des ensembles de **n-mots**, on parle de **n-grams**.

L'avantage de travailler avec n > 1 est que l'on conserve ainsi une certaine structure grammaticale que l'on perdait auparavant et que les entités constituées de deux mots sont maintenant considérées (« être vivant » était « être » et « vivant »).

Cependant, il devient plus difficile de filtrer les termes peu intéressants et les calculs sont beaucoup plus longs. Mais cela reste tout de même une bonne pratique que de faire plusieurs guides, pour plusieurs valeurs de n, afin de repérer les ensembles de mots déterminants, ce qu'on ne peut pas faire avec les simples 1-grams.
Par exemple, la confection d'un guide à partir des 3 grams, pour la requête « comment faire un cv » donne le type de résultats indiqué sur la figure 2, page précédente.

Pour conclure

Utiliser un corpus, puis un guide d'aide à la rédaction est très satisfaisant pour optimiser au mieux la pertinence d'un texte.

C'est un guide qui éclaire le chemin vers la pertinence sans pour autant obliger à fournir une écriture stricte et peu humaine.

Quelques règles à appliquer, des libertés dans les proportions, et l'article se fera unique, complet et à la fois plus intéressant pour le moteur de recherche et le lecteur humain. Il serait donc dommage de négliger un tel outil.

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/*