WDF*IDF : la "formule magique" des référenceurs allemands

Retour au sommaire de la lettre

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

*Le microcosme SEO allemand est en effervescence actuellement autour du concept de "WDF*IDF", qui n'est rien d'autre qu'un dérivé du TF*IDF bien connu depuis de nombreuses années et qui fait partie intégrante des algorithmes de pertinence de Google et consorts. Alors qu'en est-il ? Est-ce vraiment nouveau et faut-il s'y attarder ou n'est-ce que de la poudre aux yeux ?... Voici quelques éléments de réponse...*

Nota : certaines URL correspondant aux sources citées dans cet article se trouvent à la fin.

Depuis fin 2012, une "nouvelle" approche du référencement fait fureur au sein de la communauté SEO allemande : optimiser son contenu à l'aide de WDF*IDF. Les conférences sur le sujet et les billets de blogs se sont multipliés outre-Rhin, et la communauté anglo-saxonne commence à en parler. Mais de quoi s'agit-il ?

Nous allons voir que ce qu'ont découvert les allemands n'a rien de nouveau, que cette approche est déjà dépassée depuis longtemps et que cette méthode n'apporte pas grand-chose dans la pratique. C'est néanmoins l'occasion de reparler des méthodes de linguistique statistique utilisées par les moteurs de recherche. Bien connaître ces méthodes est un passage obligé pour être en mesure de comprendre les algorithmes d'un moteur de recherche, et pour mettre au point des techniques de référencement avancées.

Les origines de la formule WDF*IDF

La formule a été introduite par un référenceur allemand, Karl Kratz *(photo ci-contre)*, en avril 2010, dans un billet de son blog à propos de la "densité de mots clés" *(Cf. bibliographie en fin d'article)*. Mais c'est en réalité une conférence de Karl Kratz à Berlin fin 2012 qui a véritablement fait découvrir le concept à la communauté SEO allemande. Depuis lors, toutes les conférences SEO allemandes parlent de la formule et de ses applications.

Mais qu'est-ce que la formule WDF*IDF ?

En réalité, cette formule est plus connue dans le monde anglophone et francophone sous l'appellation : TF*IDF.

Cet acronyme signifie : *Term Frequency * Inverse Document Frequency* (*Fréquence des termes * Inverse de la Fréquence dans les Documents*). Certains auteurs préfèrent utiliser le terme WDF (*Within Document Frequency : Fréquence au sein du Document*) à la place de *Term Frequency*, pour mieux marquer la différence entre les deux termes de la formule.

Il s'agit d'une "vieille" formule, issue des travaux du linguiste et chercheur en informatique Gerald Salton à partir de la fin des années 60, et des perfectionnements introduits par Karen Spärck Jones et Stephen Robertson dans les années 70. La définition de l'*Inverse document frequency* par Karen Jones date de 1972, et la formule TF*IDF apparait pour la première fois en 1975 dans un article de Gerald Salton.

L'objectif de cette formule est de calculer un "poids" pour un terme dans un document, afin de pouvoir mesurer facilement si un document était susceptible d'être intéressant à regarder lors d'une recherche par mots clés.

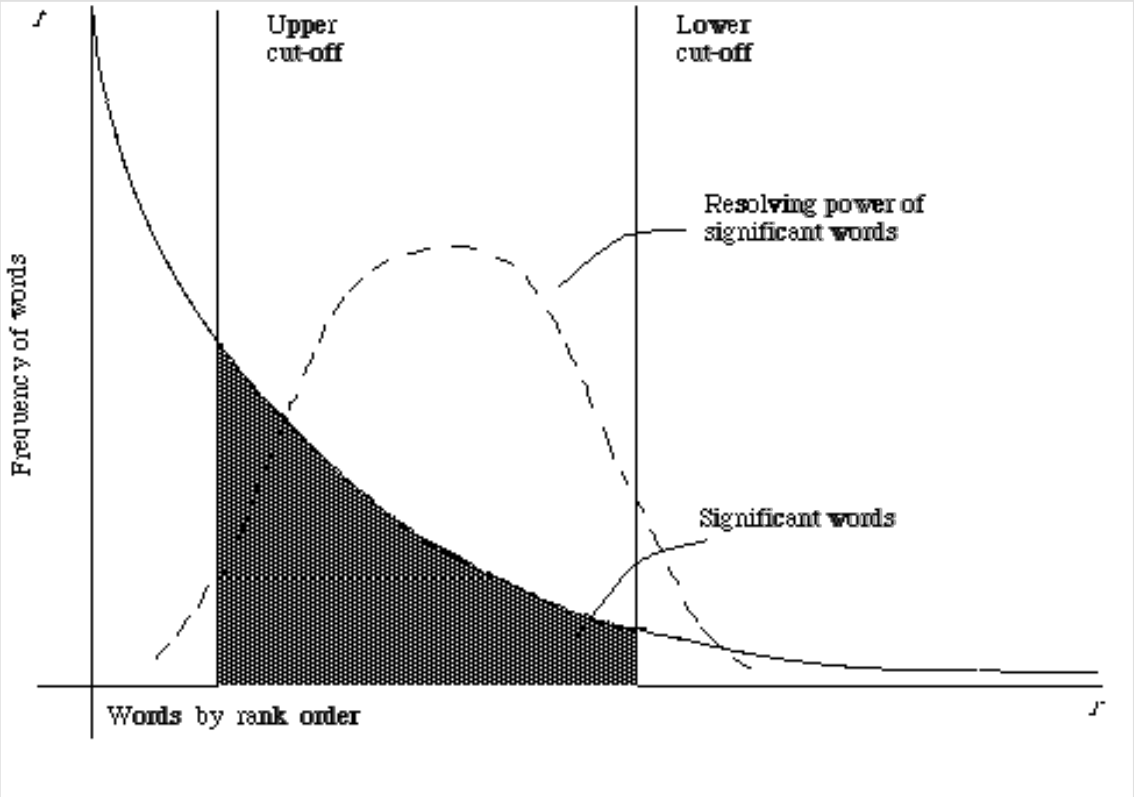


Figure 2.1. A plot of the hyperbolic curve relating f, the frequency of occurrence and r, the rank order (Adaped from Schultz[44] page 120)

L'analyse de fréquence d'occurrence des termes dans un document suit classiquement une loi de Zipf. L'expérience prouve que les mots clés les plus fréquents ne sont pas spécifiques au document et à son sujet mais résultent de la syntaxe, de la présence de « mots vides » et de termes creux. On doit donc éliminer de la liste la « top tail » et la « long tail », le problème c'est que les limites entourant la middle tail (« upper » et « lower cut off » sur le graphe ci-dessus) ne sont pas faciles à déterminer : cette approche n'est pas automatisable.

Pour calculer le "poids", l'"importance" d'un terme dans un document, la première idée qui vient à l'esprit est de mesurer le nombre d'occurrences d'un terme dans le document. On s'aperçoit assez vite que ce critère ne vaut rien si on ne tient pas compte du nombre de termes présent dans le document. On calcule donc la fréquence d'un terme dans le document (TF, term frequency), qui est improprement appelée "densité de mots clés" dans le jargon des référenceurs.

Document 1
Son nom est célébré par le bocage **qui** frémit, et par le ruisseau **qui** murmure, les vents l'emportent jusqu'à l'arc céleste, l'arc de grâce et de consolation que sa main tendit dans les nuages.

Calcul de TF (fréquence du terme) sur le Document1 ci-dessus pour le terme « qui ». Le texte comporte 38 termes, et 2 occurrences du mot « qui » :

$$\mathrm{tf}_{1,1} = \frac{n_{1,1}}{\sum_k n_{k,1}} = \frac{2}{38}$$

Remarque : La somme des occurrences pour chaque terme différent est égal au nombre total des termes dans le document.

Mais les travaux de Spärck Jones ont prouvé très tôt (il y a 42 ans !) que la fréquence du terme était également un critère de poids sans intérêt si on ne tient pas compte de la rareté ou au contraire du caractère extrêmement répandu d'un terme dans l'ensemble des documents analysés (ce que l'on appelle dans le jargon des linguistes le "corpus"). En effet, plus un terme est rare, moins il a de chances d'apparaître souvent dans un document : il se verra donc attribuer systématiquement une fréquence faible (une "densité" faible), alors qu'une seule occurrence de ce terme rare devrait "qualifier" un document comme "intéressant" dans une recherche par mots clés. Il faut donc pondérer la fréquence des termes par le nombre de documents dans lequel le terme apparait.

$$\mathrm{idf}_t = \log \frac{N}{\mathrm{df}_t}.$$

La formule de l'inverse de la fréquence d'un terme dans le document est très simple à calculer : N représente le nombre total de documents analysés (toutes les pages web indexées pour un moteur). dft représente le nombre de documents contenant le terme. La fréquence serait donnée par dft / N : on prend l'inverse car IDF sert à pondérer la fréquence des termes. Et on utilise le logarithme de la valeur parce N/df peut donner des valeurs très grandes.

La formule complète de TF*iDF est la suivante :

$$w_{x,y} = \mathrm{tf}_{x,y} \times \log\left(\frac{N}{df_x}\right)$$

TF-IDF

Term x within document y

$\mathrm{tf}_{x,y}$ = frequency of x in y

df_x = number of documents containing x

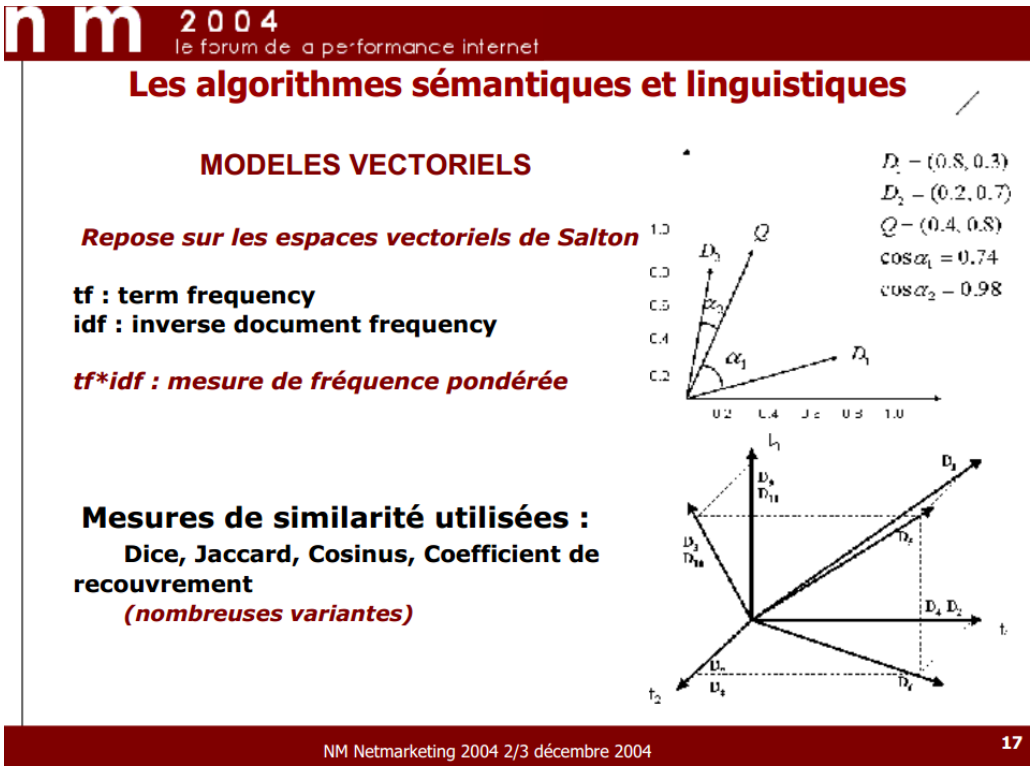
N = total number of documents

Nous évitons à dessein le terme "pertinent" dans ce contexte, car ces calculs de poids n'ont pas grand chose à voir avec la pertinence du document pour un terme donné : en fait, on mesure en quelque sorte l'écart entre la probabilité théorique qu'un terme soit présent dans un document (mesuré par le terme IDF), et sa présence réelle (mesuré par la fréquence d'apparition du terme dans le document : WDF). Certains chercheurs parlent de mesure de la "spécificité du terme" ou de "potentiel de discrimination du terme" pour décrire ce que représente vraiment ce poids.

En quoi est-ce nouveau ?

En réalité, la formule WDF*IDF ne représente rien de nouveau : elle fait l'actualité en Allemagne car outre-Rhin, les SEO viennent juste de découvrir la linguistique statistique et les travaux de base sur la Recherche d'Informations (*Information Retrieval* en anglais). Qui plus est, le choix de l'appellation WDF*IDF (plutôt rare en dehors de l'allemagne) leur a fait croire que le monde entier ignorait ces travaux (alors qu'une recherche sur TF*IDF risquait de faire remonter pas mal d'articles de SEO sur le sujet).

En France, la première présentation de ces techniques à la communauté SEO remonte à 2004 : l'auteur de cet article a eu l'honneur de présenter les travaux de Gerald Salton lors du séminaire Netmarketing organisé par Raphaël Richard en 2004. Depuis, le concept a été abordé de nombreuses fois, notamment lors des SEO Camp'us (en particulier en 2009 et en 2010).



Une slide tirée de la conférence du Séminaire Netmarketing sur le thème « SEO et linguistique statistique, recherche d'information et traitement automatisé du langage.

La formule utilisée pour wdf*idf n'est que l'une des nombreuses variantes de la formule classique, et qui a été popularisée par Donna K. Harman, qui au sein de l'institut américain NIST (*National Institute of Standards and Technology*), a été la cheville ouvrière des ateliers TREC (Text Retrieval Conference) dont le principal objectif est d'évaluer la qualité des techniques déployées en *Information Retrieval*.

$$WDF(i) = \frac{log_2(Freq(i,j)+1)}{log_2(L)}$$

$$IDF_t = \log\left(1 + \frac{N_D}{f_t}\right)$$

*La formule de Donna K. Harman ci-dessus est l'une des nombreuses variantes de tf*idf inventées depuis les années 70.*

A quoi cela peut servir en théorie ?

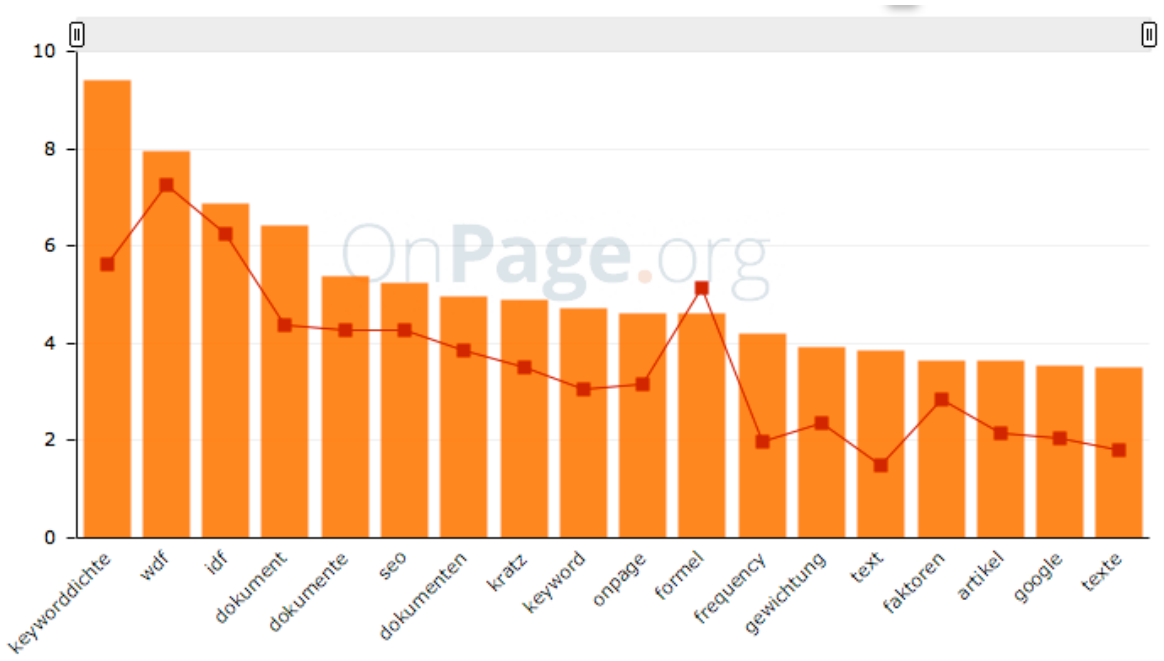
Connaître les poids des différents termes dans un document permet de :

- **Comparer des documents entre eux** pour savoir quel poids de termes caractérisent les premiers résultats fournis par un moteur sur une requête donnée. Cette retro-ingéniérie permet d'estimer grossièrement la fréquence de terme idéale (et bien mieux que l'approche traditionnelle se basant sur la densité de mots clés).

- **Calculer un "Cosinus de Salton"** entre une requête et un document, pour voir si le classement en fonction de ce critère reflète les SERP. Si oui, on a un moyen simple de savoir si un changement dans le contenu textuel de la page peut permettre de remonter sur une requête.

Nous avions en son temps appelé cette approche l'"alignement sémantique" : un changement dans le contenu textuel permet en effet théoriquement de "rapprocher" le document de l'idéal attendu par l'algorithme du moteur (mais uniquement sur les critères de proximité sémantique, qui font partie des très nombreux signaux utilisés par les moteurs pour identifier les documents les plus pertinents sur une requête).

La méthode de Karl Kratz est très exactement une méthode d'alignement sémantique. Une méthode qui est apparue tellement révolutionnaire outre-Rhin que nombre de sociétés de création contenus travaillant pour le marché allemand se sont senties obligées d'annoncer fièrement qu'elles fabriquaient des textes optimisés selon la méthode WDF*IDF. La formule de Karl Kratz est donc devenue un argument marketing.



*Une analyse des poids wdf*idf pour les termes d'une page web réalisée par l'outil Onpage.org*

Que vaut la formule des SEO allemands ?

La formule de Donna Harman, comme toutes les autres variantes, a fait l'objet d'évaluation lors des conférences TREC successives. Il s'avère que l'approche qui a donné les meilleurs résultats et qui fait (presque) consensus aujourd'hui, est plutôt Okapi BM25 (voir bibliographie). Donc, quitte à vouloir calculer un TF*IDF, on peut faire mieux que la version allemande de WDF*IDF. Il existe d'ailleurs sur Internet des articles décrivant comment implémenter Okapi BM25 dans des moteurs de recherche open source comme Xapian, ou Lucène SolR (qui utilise par défaut une version basique de TF*IDF).

Une formule approchée et "corrigée" selon des critères très contestables

Néanmoins, analyser le poids des termes avec TF*IDF représente déjà un saut qualitatif énorme par rapport à une analyse de type "densité de mots clés". Et si la formule WDF*IDF de Donna Harman n'est pas la meilleure, c'est l'une des meilleures.

Besser. **Aber schwieriger.**

Linkless Outranking WDF*IDF

Wenn man es ein bisschen genauer nimmt:

$$w_{i,j} = \frac{lb\left(TF_{i,j}+1\right)}{lb\left(L_j-(TF_{i,j}*(tc-1))\right)} * ld\left(1+\frac{R_{1,j}\cup R_{2,j}\cup\ldots\cup R_{n,j}}{N_i}\right)$$

Gewichtung eines Terms i im Dokument j

Anzahl aller Terme im Dokument j.

Häufigkeit des Term i im Dokument j.

Anzahl der Einzelterme der Term-Kombination

Anzahl aller Dokumente in denen der **Term i** vorkommt.

Vereinigungsmenge aller Resultate in der Datenbank, die alle Terme i_n enthalten, welche im Dokument j vorkommen.

© karl kratz online marketing 2012 19

La formule de Donna K. Harman telle que Karl Kratz la présente dans la langue de Goethe.

Mais les choses se gâtent lorsqu'on regarde de près comment les référenceurs allemands l'utilisent... Le principal problème avec une formule de type TF*IDF, c'est de savoir comment on peut calculer le terme IDF (la fréquence des termes dans le document) alors qu'on ne connait pas le corpus ! En effet, il faudrait, comme le ferait un moteur comme Google, crawler tout le web, et compter la fréquence d'apparition de chaque terme dans toutes les pages web.

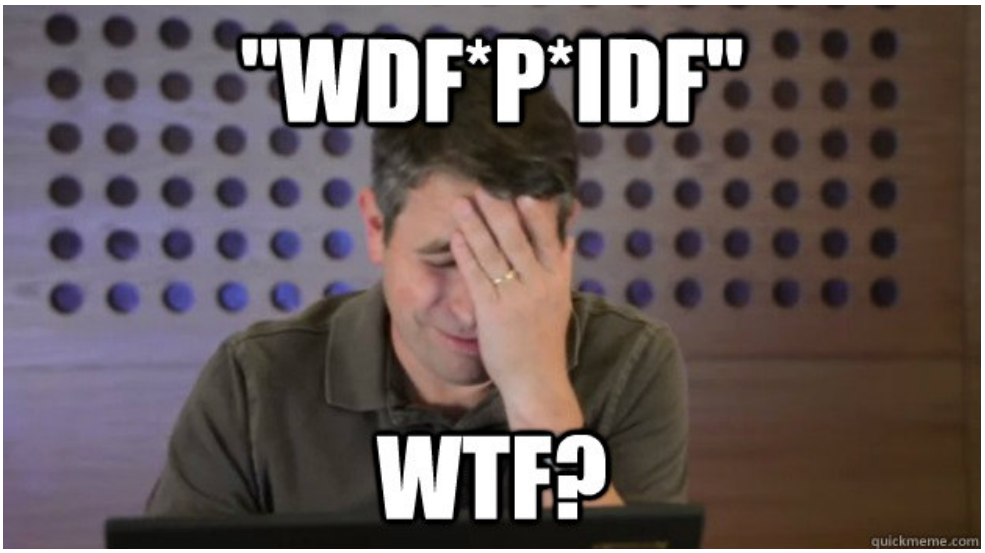
Une première idée appliquée par nos cousins allemands consiste à remplacer la fréquence des documents par le nombre de résultats fournis par Google pour un terme donné. Mais ce n'est pas une fréquence et la formule ainsi modifiée donne des résultats absurdes.

La seconde solution consiste à calculer une fréquence approchée en extrayant tous les termes d'un document, d'interroger Google pour obtenir le nombre de résultats pour une requête sur chaque terme du document, et d'en déduire la fréquence IDF. En théorie, on obtient effectivement un poids approché, mais selon le type de document, on peut être très proche ou très éloigné de la vérité !!

On pourrait certes sérieusement améliorer la précision et la pertinence des calculs dès lors que l'analyse des corpus est possible. C'est notamment vrai si on analyse des pages techniques, et si on peut "aspirer" toutes les pages qui parlent du sujet, soit un sous ensemble limité du World Wide Web. On peut aussi se baser sur des statistiques disponibles pour des gros corpus web mis à la disposition des chercheurs en traitement automatisé du langage.

Ceux qui ont réellement essayé et notamment ceux qui ont développé les plugins ou les applications de calcul en ligne de WDF*IDF ont donc transformé la formule en WDF*P*IDF ou P est un coefficient de correction qui permet d'éviter que la hiérarchie des poids donnée par la formule sur un texte donné ne soit totalement absurde. Notons que le coefficient P est empirique et qu'il est croustillant de constater que les auteurs des

méthodes opérationnelles de calcul restent toujours très évasifs sur ce que cache le coefficient P et comment il est déterminé.



*Ce que pense probablement Matt Cutts de la formule wdf*p*idf…*

*L'alignement sémantique à l'aide de TF*IDF a ses limites…*

L'un des défauts la méthode est qu'elle ne tient aucun compte de trois réalités :

- La fréquence d'apparition des termes est fortement influencée par la syntaxe d'une langue et les habitudes de langage. Bref, on peut croire que certains termes sont importants dans un document, alors que la plupart ne sont pas réellement porteurs de sens. *A contrario*, on ne verra pas que certains termes rares ont une forte probabilité d'apparition dans un document parce qu'ils sont très fréquemment présents à côté d'autres termes (c'est à dire qu'ils ont une forte fréquence de coocurrence).

- L'approche ne fait aucun cas de l'existence d'expression toutes faites, comme dans "Président de la République" : les termes "président" et "république" sont analysés séparément alors qu'ils forment un tout.

- Enfin, certaines pages web sont « multi-sujets » comme des pages d'actualité ou des pages de catégories ou de rubriques. Dans ce cas, la présence d'un mot clé spécifique à un sujet donné ne garantit pas que la page soit pertinente pour ce terme. Ce cas est tellement fréquent que les moteurs en tiennent compte, et c'est pourquoi leurs algorithmes prennent en compte d'autres facteurs que tf*idf, wdf*idf, ou Okapi BM25.

Bref, une méthode d'alignement sémantique efficace nécessite d'aller plus loin :

- Comme utiliser LSA/LSI pour identifier les termes qui sont "corrélés" dans un document, afin de filtrer le bruit dans les fréquences de termes (WDF) et identifier les termes qui comptent d'un point de vue sémantique.

- Avoir une approche tenant compte des entités nommées et des syntagmes importants (approche par groupe de mots clés, ou n-grammes).

Le problème, c'est que toutes ces analyses demandent beaucoup de temps et de compétences, et que dans bien des cas, le jeu n'en vaudra pas la chandelle. Car la supériorité de cette méthode sur un bon brief éditorial et un bon contrôle qualité sur le travail des rédacteurs web n'est pas toujours assurée.

Cette méthode était valable en 2004 ! Nous sommes en 2014 ...

Qui plus est, Google a évolué, et nous sommes à peu près sûr qu'il utilise des critères bien plus sophistiqués que TF*IDF pour identifier les documents à faire remonter dans ses pages de résultats. Cela fait trois ans que l'on parle d'indexation de syntagmes (et donc la fin du paradigme en "sac de mots" ou tous les termes sont analyses individuellement sans tenir compte de leur ordre et de ce que l'on trouve comme autres termes autour d'eux).

Et maintenant avec Hummingbird, nous en sommes à l'indexation de concepts, ce qui sonnera à terme le glas du paradigme précédent où le mot-clé était tout puissant !

Faut-il oublier ces techniques, et ne pas s'y intéresser ? Pas du tout, il est important qu'un référenceur de 2014 connaisse ces fondamentaux, pour comprendre comment les algorithmes des moteurs ont fonctionné jadis, car après tout, les versions actuelles ne diffèrent en réalité que par leur sophistication plus élevée. Et elles peuvent aider à faire du référencement avancé et à évaluer correctement le contenu d'une page et d'un site web. Mais il faut aussi garder en tête deux réalités : ces techniques sont datées, et elles ont leurs limites.

Et nous espérons que cet article vous permettra de distinguer le vrai du faux et surtout les argumentations marketing pseudo scientifiques et opportunistes quand, immanquablement, la mode du WDF*IDF va commencer à influencer les méthodologies du monde du SEO francophone.

Bibliographie

Un article de Wikipedia sur TF*IDF :
http://fr.wikipedia.org/wiki/TF-IDF

L'article séminal de Karl Kratz (en allemand) :
http://www.karlkratz.de/onlinemarketing-blog/seo-keyword-density/

Le premier article sur le sujet dans l'univers SEO anglo-saxon :
http://pagerankchecker.com/blog/seo/that-s-wdf-idf/

Un autre article en anglais:
http://sno.pe/b/seo-content-quality-and-mathematical-determination-with-wdfidf/

Tf*idf : l'une des premières évocations en France devant la communauté SEO :
http://www.webmaster-hub.com/publication/IMG/pdf/slides-semantique-netmarketing2004.pdf

L'origine de la formule : extrait en ligne d'un livre de Donna K. Harman :
http://www.ai.wu.ac.at/~koch/courses/wuw/archive/inf-sem-ws-00/zoechling/statistik.htm

Okapi BM25 : la "meilleure" formule de calcul de poids de termes dans un document :
http://en.wikipedia.org/wiki/Okapi_BM25

Présentation au SEO Camp'us 2009 : TF*IDF, Cosinus de Salton et alignement sémantique :
http://www.seo-camp.org/wp-content/uploads/2009/02/apport_semantique.pdf

Calculateur de WDF*IDF pour Wordpress :
http://codecanyon.net/item/wordpress-wdfidf-seo-calculator/5122990

Calculateur online :
http://onpage-tool.de/

Et pour en savoir plus : Cours de Recherche d'Information par A. Rozenknop (pour les étudiants de l'Université de Paris XIII)
http://lipn.univ-paris13.fr/~rozenknop/Cours/MICR_REI/Seance6/modeles-RI-1.pdf

Bonne lecture à tous !!

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