Le PageRank thématique, à l'origine du cocon sémantique

Par Guillaume et Sylvain Peyronnet

Domaine :	Recherche	Référencement
Niveau :	Pour tous	Avancé

La notion de PageRank est aussi ancienne que la naissance de Google (il doit d'ailleurs son nom à l'un des deux cofondateurs du moteur). Mais , au fil des ans et de rachat de start-ups (notamment Kaltix en 2003), une nouvelle notion de PageRank thématique est venue compléter la vision initiale de la popularité d'une page dans l'algorithme de Google. Comment définir ce PageRank thématique et comment est-il calculé et intégré par les moteurs, pour arriver aujourd'hui à la notion de cocon sémantique, notion chère à Laurent Bourelly ? Voici quelques explications.

Dans cet article nous allons aborder la notion souvent évoquée mais rarement expliquée de PageRank thématique. Tout le monde a déjà entendu parler du PageRank de Google, mais il est rare de trouver des personnes qui savent réellement ce que c'est. Le PageRank thématique est une évolution naturelle du PageRank, ceci est évident dès que l'on sait ce qu'est le PageRank standard.

Après avoir rapidement évoqué le PageRank usuel, nous verrons le principe de sa version thématique, ainsi que ses implications, avec notamment la notion de cocon sémantique, très populaire aujourd'hui chez les référenceurs.

Le PageRank, qu'est-ce que c'est ?

Il n'est pas rare de découvrir, au détour des sites web spécialisés, une belle formule pour définir ce qu'est le PageRank, couplée à des phrases du type « le PageRank considère qu'un lien vers un site est un vote pour ce site » ou encore « le PageRank est une mesure de l'autorité du site ». Nous allons maintenant voir qu'il existe une manière beaucoup plus simple et intuitive d'expliquer ce qu'est le PageRank.

Pour mieux comprendre, remontons le temps en 1998, date de la création de Google par Sergey Brin et Larry Page (voir [1]). L'objectif est alors de présenter aux utilisateurs du moteur de recherche des pages pertinentes pour ses requêtes. Le problème principal du moteur est qu'il existe beaucoup de pages qui sont pertinentes pour une requête donnée. Comment arbitrer entre plusieurs pages quand on ne peut en présenter que quelques unes ? L'idée de Larry Page (d'où le nom - assorti d'un jeu de mots - de PageRank) est de quantifier la popularité des pages : plus une page pertinente est populaire, plus on la présentera avant les autres.

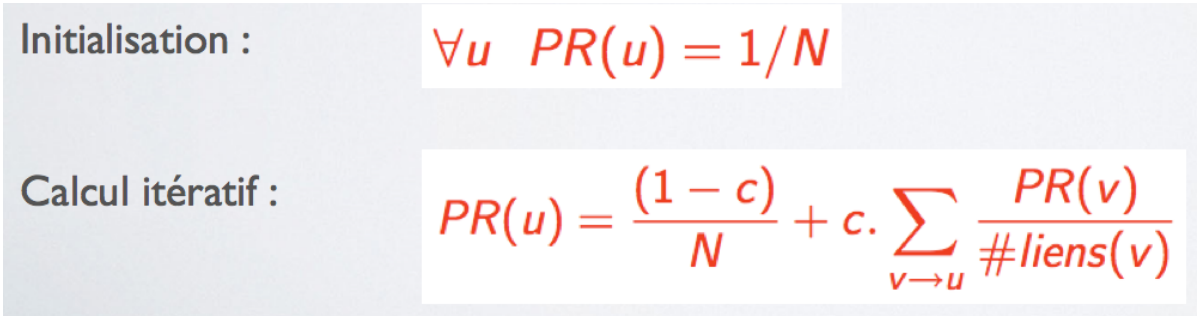

Fig.1. Calcul itératif du PageRank d'une page.

Comment déterminer, dans ce cas, qu'une page est populaire ? Le seul moyen indiscutable est de monitorer les flux de visiteurs : si une page est beaucoup visitée, elle est populaire. Mais en 1998, Google ne peut monitorer le flux de visiteurs de chaque site, et donc il faut trouver une autre astuce. Cela va être la modélisation d'un visiteur virtuel : le **surfeur aléatoire**. Intuitivement, il s'agit d'un internaute qui parcourt le web en cliquant aléatoirement sur les liens sortants de la page sur laquelle il est à un moment donné. Parfois, le surfeur aléatoire peut décider de changer totalement de voisinage, ce qui est implicitement modélisé par une probabilité de continuer, ou non, à suivre les liens. Lorsqu'il décide de ne plus suivre les liens, il recommence son parcours à partir d'une page tirée aléatoirement sur le web. On parle de « téléportation » pour désigner cet évènement.

La définition du PageRank est intimement liée à la notion de surfeur aléatoire. En effet, la probabilité que le surfeur aléatoire se trouve sur une page donnée à un moment précis est le PageRank, par définition. Cette probabilité permet de quantifier exactement la popularité : si une page est populaire, il y aura en permanence beaucoup de surfeurs aléatoires en train de la visiter, et donc la probabilité qu'un surfeur aléatoire y soit sera grande, et réciproquement.

D'un point de vue mathématique, cette probabilité est identifiée en calculant la distribution stationnaire de la chaine de Markov associée au graphe du web, mais nous ne rentrerons pas dans des détails trop scientifiques. Il faut simplement savoir que cela permet de trouver la formule de la figure 1, qui calcule itérativement le PageRank de chaque page.

Que nous dit cette formule ? Qu'à l'origine, on donne à chaque page un PageRank égal à (1/N), N étant le nombre total de pages sur le web, puis qu'ensuite les règles de transmission contribuent à hauteur d'une proportion c au PageRank de la page et que le facteur (1/N) contribue à hauteur d'une proportion (1-c). c est appellé *Damping Factor* et vaut 85% dans l'article initial de Brin et Page [1]. Cela signifie donc deux choses qu'il faut retenir : le PageRank vaut de 0 à 1 (exclus l'un et l'autre) et la somme de tous les PageRank vaut 1.

On voit que finalement, le PageRank est une quantité très intuitive, basé sur une modélisation du comportement des utilisateurs du Web. Le problème qui s'est posé par la suite est celui de la fidélité de ce modèle au comportement réel des internautes. Est-ce qu'on peut vraiment croire que la probabilité de suivre un lien va uniquement dépendre du nombre de liens sortants depuis la page que l'on visite ? La réponse est bien évidemment

négative, et le bon sens veut qu'*a minima* cette probabilité dépende du contenu de la page source et de la page cible. En effet, si un internaute est sur une page qui parle de cuisine espagnole, il y a plus de chance qu'il suive un lien qui l'emmène vers une page parlant de chorizo que vers une page qui aborde le problème du matériau de brasure en plomberie.

Un surfeur aléatoire qui a de la suite dans les idées

Pour faire évoluer ses algorithmes et fournir des résultats de meilleure qualité, les moteurs sont donc tentés d'utiliser un modèle plus réaliste que celui du surfeur aléatoire « basique ». En embarquant des informations thématiques dans le modèle, on obtient un meilleur réalisme. De nombreux chercheurs ont travaillé sur le sujet, le lecteur intéressé pourra lire l'article [3] qui présente une approche inefficace mais qui est historiquement la première, et surtout l'article [4] de Taher Haveliwala, qui est l'inventeur du PageRank thématique. Ceci étant, la présentation que nous faisons ici est largement tiré des travaux de Brian Davison et son équipe [5], ces derniers ayant mieux formalisé la notion.

Le surfeur aléatoire sensible à la thématique détient une richesse de comportement plus importante que le surfeur aléatoire classique. Il a trois possibilités comportementales :

1. Il peut se téléporter sur une page tirée au hasard, et dans ce cas il va se concentrer sur une des thématiques contenues dans cette page.
2. Il peut suivre un lien vers une nouvelle page, en restant concentré sur la même thématique.
3. Il peut suivre un lien vers une nouvelle page, et opérer un « switch » thématique en se concentrant sur une des thématiques de la page cible, qui n'était pas celle sur laquelle il était concentré auparavant.

On voit immédiatement une chose importante : chaque page possède des scores thématiques : on sait quelles sont les thématiques portées par la page, et dans quelle mesure elles sont importantes ou pas au sein du contenu de la page. Par exemple, une page qui aborde le sujet des voitures et de l'immobilier se verra attribuée une ventilation sur chacune de ces thématiques : elle sera (par exemple) à 30% dans la thématique voiture et à 70% dans la thématique immobilier.

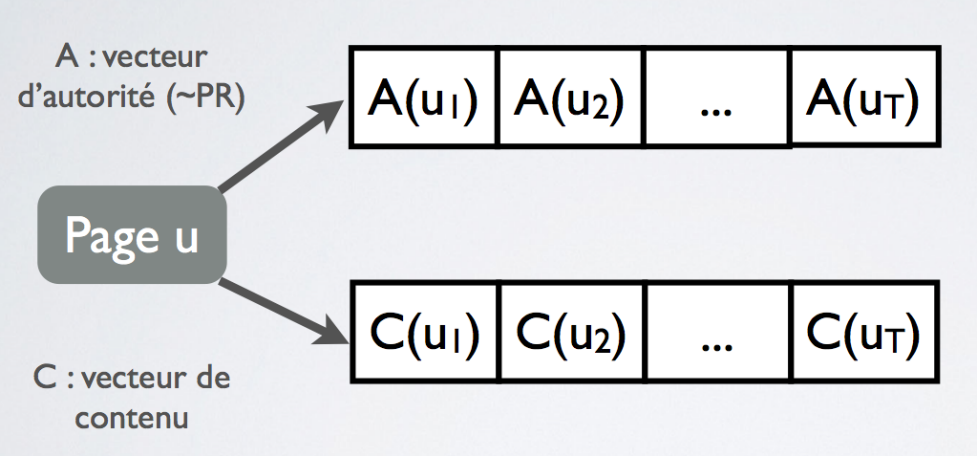

Fig.2. Attribution de 2 vecteurs à une page web.

D'un point de vue plus formel, chaque page va se voir associée deux vecteurs (figure 2).
Le vecteur A est le vecteur d'autorité, c'est celui qui correspond aux PageRank thématiques. Vous remarquerez ici l'usage du pluriel : le PageRank thématique est en fait un PageRank par thématique, la somme de tous les PageRank thématiques d'une page étant égale au PageRank usuel de la page. Il est très important de s'en rendre compte : le PageRank thématique n'est ni plus ni moins qu'une ventilation du PageRank standard par thématique.

Où sont les thématiques ? Ici les thèmes sont notés 1, 2, etc. u1 signifie « le thème 1 dans la page u ». Donc A(u1) est l'autorité de la page u sur le thème 1. On a un vecteur de taille 100 si le moteur utilise 100 thématiques différentes pour son algorithme de classement. Il est impossible de savoir combien de thématiques un moteur moderne considère, ni ce qu'est exactement une thématique. Pour un être humain, la notion de thématique est assez claire, mais pour un algorithme cela peut tout simplement être un regroupement de pages ayant des profils de TF-IDF similaires (voir nos articles précédents sur ce sujet dans la lettre R&R). C'est pour cette raison qu'il est parfois difficile de saisir la logique existant derrière les résultats fournis par le moteur.

Le deuxième vecteur est celui d'adéquation au contenu. Il est dans la figure noté C. C(u1) est le score d'adéquation de la page u pour le thème u1. Si par exemple le thème 1 est le sport, et que la page parle à 30% de sport, C(u1) vaudra 0,3. La somme de toutes les valeurs du vecteur C vaut 1.

Pour calculer les valeurs du vecteur A, on a un mécanisme de transmission itérée similaire au PageRank usuel, quoi que légèrement plus complexe, puisque le surfeur aléatoire a maintenant trois actions possibles au lieu de deux.

La figure 3 donne la formule exacte. On voit que la structure du calcul est inchangée avec d'un coté une transmission par les liens et de l'autre une « téléportation », mais un nouveau paramètre est apparu : alpha, qui est la probabilité que le surfeur a de changer de thématique (c'est ce qui lui permet de déterminer si il fait l'action 2 ou l'action 3 parmi celles qu'il peut faire). Cette probabilité est sans doute une donnée d'usage globale pour les moteurs actuels, mais cela pourrait devenir un paramètre de personnalisation des résultats très facilement.

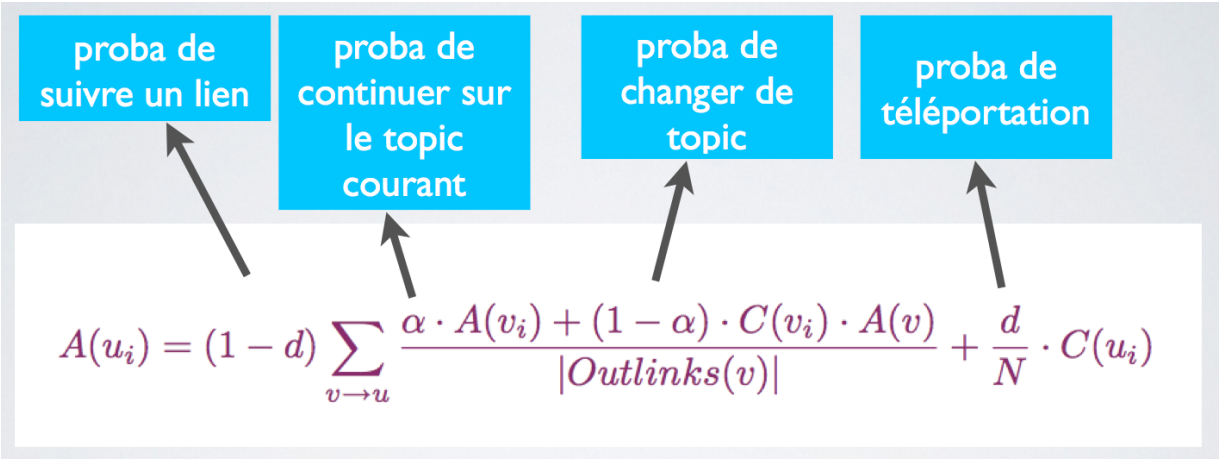

Fig.3. Formule du PageRank thématique

Mettre en action le PageRank thématique au niveau du moteur

A ce stade de l'explication sur le PageRank thématique, il ne semble pas y avoir de rapport entre les requêtes et les deux vecteurs de chaque page. Mais c'est pourtant le cas. Voici comment un moteur utilise toute cette machinerie, ce qui nous permettra d'éclaircir toutes ces notions au travers d'exemples :

1. Quand une requête est tapée par un utilisateur, le moteur va créer un vecteur d'adéquation au contenu pour cette requête. Par exemple, la requête « Jaguar » va être considérée comme étant à 30% dans la thématique « animal » et à 70% dans la thématique « voiture ».

2. Le moteur prend alors dans le vecteur A le PageRank de chaque page pour les thématiques liées à la requête, pondéré par le score d'adéquation de la page à ces thématiques (l'information est dans le vecteur C). Le score de chaque page pour la requête est la somme de ces PageRank pondérés. La formule est présentée dans la figure ci-dessous :

$$Score_q(u) = \sum_k A(u_k) \cdot C(q_k)$$

3. Les pages sont alors classées par score. On a donc un classement potentiellement différent pour chaque requête.

La question qui se pose naturellement maintenant est de savoir si les résultats obtenus par le moteur avec le PageRank thématique sont de meilleure qualité que ceux qu'il obtenait avec le PageRank standard.

Brian Davison et son équipe ont mené des expériences pour mettre en évidence les différences de qualité des deux algorithmes. Pour cela ils ont choisi plusieurs requêtes, certaines compétitives, d'autres moins, et ils ont appliqué les deux

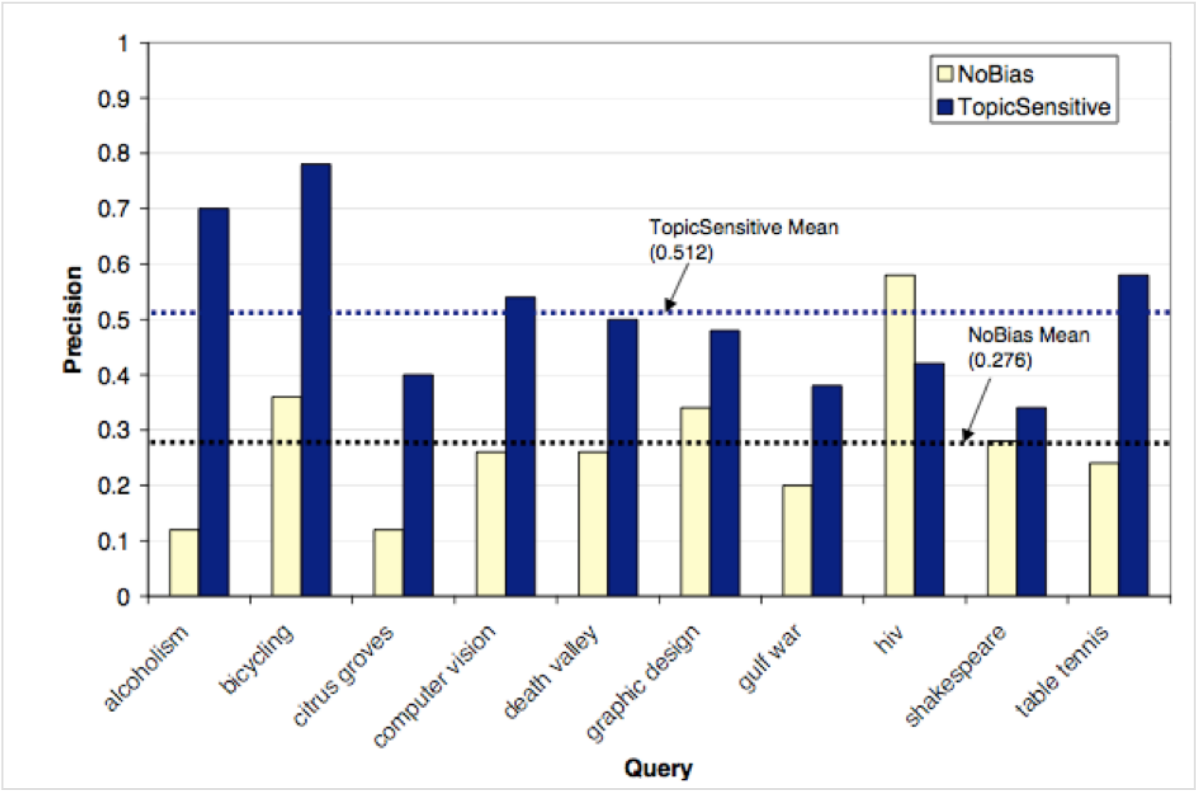

Fig.4. Résultats de l'expérience de Brian Davison

types de PageRank pour construire des SERP, et on ensuite demandé à des volontaires de noter ces SERP. Nous avons déjà évoqué dans un précédent article de la lettre R&R la problématique de l'évaluation des SERP, il suffit juste ici de dire que la métrique utilisé par Brian Davison et son équipe est la précision à 10 (p@10).

La figure 4 (page précédente) présente les résultats de leur expérience.

Les barres jaunes (NoBias) correspondent au PageRank usuel, tandis que celles en bleue (TopicSensitive) correspondent au PageRank thématique. Premier point important, la qualité globale perçue selon l'algorithme utilisé : le PageRank standard obtient une précision moyenne de 0,276, tandis que le PageRank thématique obtient 0,512. C'est une amélioration de la qualité perçue de plus de 85% ! Le PageRank thématique semble donc réellement améliorer les SERP, et on comprend sans hésiter pourquoi un moteur peut vouloir mettre en place un tel mécanisme.

Si on regarde en détail ces résultats, on voit que sur certaines requêtes comme, par exemple, « alcoholism », on a une amélioration de près de 700%. On voit d'ailleurs qu'il existe toujours une amélioration substentielle, sauf dans deux cas. Le premier est la requête « Shakespeare » sur laquelle l'amélioration n'est pas significative. L'explication est simple : sur cette requête non compétitive, les liens sont tous thématisés et donc le passage au PageRank thématique ne change rien aux scores des pages de la thématique. Le deuxième cas délicat est sur la requête « HIV ». Sur celle-ci, on trouve même une dégradation. Ce phénomène est très connu et est inhérent aux pages institutionnelles et aux requêtes associées (c'est le cas de la requête « HIV »). Les pages institutionnelles se lient entre elles, et sont donc très puissantes en terme de PageRank usuel, mais elles ne sont pas nécessairement dans la même thématique et donc le passage au PageRank thématique dégrade leurs scores.

Au final, on voit toutefois que le PageRank thématique est réellement un plus pour un moteur de recherche, et pour ses utilisateurs. Mais quel est l'impact de cet algorithme pour le référencement de sites web ?

Le cocon sémantique est légitime

Il est important de dire que la notion même de cocon sémantique a d'abord été mise en avant par Laurent Bourrelly (voir son article [6]) pour proposer une méthodologie permettant de faire du linking interne sémantique au sein d'un site web. Il est d'ailleurs sans doute celui a qui proposé ce nom particulièrement évocateur. Nous n'allons pas ici expliquer en détail cette méthodologie pour laquelle Laurent propose même une formation (voir [7]) ou qui a déjà été expliquée en détail par Christian Méline (voir [8]), mais nous pouvons donner ici quelques informations supplémentaires à ce sujet.

Tout d'abord, toute la « machinerie » du cocon sémantique est légitime, et c'est la notion même de PageRank sémantique qui justifie la méthodologie. En faisant prioritairement des liens entre pages d'une même thématique, on maximise la transmission de popularité. Par ailleurs, si on veut passer d'une thématique à une autre *via* un lien (ou plutôt, si on veut faire des liens entre pages de thématiques différentes), alors il faut réaliser un glissement sémantique progressif, ce sera

plus productif qu'un lien abrupt entre deux pages sans aucun rapport.

Autre point important : la construction du cocon sémantique est largement présenté pour l'optimisation interne d'un site, mais elle est tout aussi valable pour les opérations de netlinking externe. Allons même plus loin : la logique de cocon sémantique facilite l'acquisition de liens, car en faisant légèrement glisser les thématiques, on crée du potentiel de liens plus important. L'exemple typique et un peu convenu, c'est celui du e-commerçant qui vend un produit « dur », par exemple du câble électrique. Comment trouver des liens vers une page de vente de câbles, tout en préservant l'adéquation thématique ? Avec une logique de cocon, c'est possible : il suffit d'entourer la page de vente de page en léger glissement, comme par exemple une page sur la fabrication des câbles, une autre sur les métiers de la fabrication, etc. Ces pages trouveront plus facilement des liens (par exemple, la page sur les métiers peut obtenir des liens depuis des sites web de lycées).

Enfin, mais là on sort de la technique et du référencement, la logique de cocon permet généralement d'améliorer le service au visiteur du site. En pensant la création du contenu, on crée naturellement des contenus utiles, ce qui est une bonne chose pour les visiteurs.

Conclusion

La plupart des référenceurs n'en ont pas conscience, mais la prise en compte de la transmission thématique pour le netlinking est réellement un levier très important. L'expérience que nous avons pu avoir suite à nos formations est d'ailleurs que c'est un levier que les référenceurs mettent prioritairement en action une fois qu'ils le connaissent. Nous vous encourageons à faire de même !

Références

[1] Page, L., Brin, S., Motwani, R., & Winograd, T. (1999). *The PageRank citation ranking: bringing order to the Web.* http://ilpubs.stanford.edu:8090/422/1/1999-66.pdf

[2] Largillier, T., & Peyronnet, S. (2014). *Algorithmique du web: autour du PageRank.* Informatique Mathématique Une photographie en 2014. Presses Universitaires de Perpignan.

[3] Richardson, M., & Domingos, P. (2001). *The Intelligent surfer: Probabilistic Combination of Link and Content Information in PageRank.* In NIPS (pp. 1441-1448). http://alchemy.cs.washington.edu/papers/pdfs/richardson-domingos02a.pdf

[4] Haveliwala, T. H. (2003). *Topic-sensitive PageRank: A context-sensitive ranking algorithm for web search.* Knowledge and Data Engineering, IEEE Transactions on, 15(4), 784-796. http://ilpubs.stanford.edu:8090/750/1/2003-29.pdf

[5] Nie, L., Davison, B. D., & Qi, X. (2006, August). *Topical link analysis for web search. In Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 91-98). ACM.

[6] *Le cocon sémantique, l'arme fatale du SEO ?* Laurent Bourrelly. http://www.laurentbourrelly.com/blog/1631.php

[7] http://www.laurentbourrelly.com/formations/cocon-semantique/

[8] http://www.referencement-naturel-white-hat.fr/cocons-semantiques/

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/*