Propager la confiance et la méfiance dans les algorithmes de recherche

Retour au sommaire de la lettre

Domaine :	Recherche	Référencement
Niveau :	Pour tous	Avancé

Quiconque s'intéresse au SEO et aux moteurs de recherche a déjà entendu parler de la notion de "TrustRank". Mais ce terme est trompeur et sa définition originale ne correspond pas à la façon dont Google envisage la confiance qu'il a dans un site. D'une façon générale, il existe de nombreux algorithmes permettant de propager des notions de confiance et, a contrario, de méfiance pour noter une page ou un site web et ainsi écarter le spam. En voici quelques exemples décrits dans cet article...

En réfléchissant au titre de l'article de ce mois-ci, nous nous sommes posés une question fondamentale : devions-nous y mettre le mot TrustRank, ou bien au contraire fallait-il l'éviter, pour que votre lecture ne commence pas sur un malentendu ? Comme vous le voyez, nous avons choisi de prendre le taureau par les cornes et d'en parler, sans le mettre dans le titre !

Dans cet article, nous allons donc évoquer les différents algorithmes qui ont été envisagés par les chercheurs dans le domaine de la recherche d'information pour propager la confiance, ou la non-confiance, sur le web. Si nous avons déjà évoqué le terme de TrustRank, c'est parce que cet algorithme, mis au point en 2004 par Zoltan Gyongyi, Hector Garcia-Molina et Jan Pedersen, est l'archétype des méthodes de propagation de la confiance. Son souci est qu'il a été longtemps confondu avec ce que Google appelait le trustrank, et qui n'avait pas de rapport (le vocable trustrank a d'ailleurs été abandonné par Google en 2009, qui lui a préféré le mot de personrank).

L'idée derrière la propagation de la confiance et de la non-confiance

L'idée de base est plutôt simple : le problème des méthodes qui calculent un score à partir des liens sur le web (comme le PageRank par exemple) est l'initialisation : si toutes les pages sont égales, les pages de spam finissent par gagner, car leur nombre est virtuellement infini. En revanche, si on sait repérer quelques pages qui sont de grande qualité et confiance, on va pouvoir s'en servir comme point de départ pour propager une information de qualité selon l'adage « les amis de mes amis sont mes amis ».

Inversement, si je repère quelques voyous, je vais partir du principe que les amis des malfaiteurs sont coupables par association. Ainsi une page de confiance qui fait un lien vers un tiers donnera un boost à ce tiers, et une page de faible qualité qui ferait un lien à un site tiers le plomberait, même légèrement.

Il n'y a donc pas de grand secret derrière ces méthodes : il s'agit ni plus ni moins que de mécaniser les heuristiques que nous appliquons dans notre vie : l'ami des gens honnêtes est honnête, celui des pourris l'est aussi.

Le mot a donc été lancé : « mécaniser », et c'est là que la complexité apparaît.

Le calcul de confiance : point de vue local ou global ?

Il existe deux grands courants de pensée en matière de propagation de confiance : celui du calcul local et celui du calcul global. Dans le premier cas, l'information de confiance est propre aux nœuds du graphe de relation (les sites web dans notre cas), ce qui signifie qu'il va y avoir plusieurs valeurs de confiance pour chaque site. Cela modélise le fait que

je peux avoir confiance en un site donné, tandis qu'une autre personne pourrait voir les choses différemment. Dans le second cas, l'information est globale : chaque site à un score de confiance, qui est utilisé par tout le monde.

Le lecteur l'aura deviné, la notion qui prévaut est celle du calcul global. Cela s'explique de deux façons : tout d'abord le calcul global est plus facile du point de vue de la complexité algorithmique et donc du point de vue de la puissance de calcul. Mais ensuite, il y a une raison plus « philosophique ». En effet, les moteurs de recherche sont principalement normatifs : ils « savent » ce qui est bien pour tout le monde, sans nuance ni réelle personnalisation, et la confiance est pour eux un absolu.

Confiance ou non-confiance ?

Techniquement, propager de la confiance ou de la non-confiance, ou de l'indifférence, ou n'importe quelle autre quantité, c'est la même chose. Mais il faut se poser la question de ce que cela change intrinsèquement de faire l'un plutôt que l'autre. En pratique, propager de la confiance est plus efficace que propager de la non-confiance. Si on voulait résumer à l'emporte-pièce : la discrimination positive est plus efficace que la discrimination négative.

Pourquoi ? Pour beaucoup de raisons, mais la première est sociale : la confiance se propage de manière plus lointaine que la méfiance. On recommande facilement l'ami d'un ami d'un ami, mais on évitera de dire du mal d'une personne qu'on ne connaît que par plusieurs rebonds. Plus globalement, les gens sont mesurés dans la critique, et moins dans le compliment (ceci varie peut être selon les pays) et donc l'information de non-confiance est plus difficile à obtenir. Mais une autre raison est bassement pragmatique : sur le Web, on considère les liens comme porteurs de l'information, et généralement c'est pour indiquer la confiance qu'on fait un lien. En l'absence d'information sur le lien pour indiquer si on parle d'un site en bien ou en mal, les algorithmes ont du mal à prendre en compte simultanément confiance et non-confiance.

Sur ce dernier point, le SEO avisé pourra évoquer le nofollow, qui est un tag qui modère notre propos, en neutralisant la confiance associée à certains liens.

Retour au TrustRank

Et si vous avez bien suivi, vous avez bien sûr compris qu'il ne s'agit pas ici de la notion proposée par Google, mais bien de l'algorithme mis en place par Gyongyi, Garcia-Molina et Pedersen : le TrustRank est un score associé à chaque page web et qui est relativement similaire au pagerank dans l'idée. Certaines pages ont un score de TrustRank initial qui est propagé de page en page grâce aux liens. L'idée est de repérer grâce à ce mécanisme les pages qui ne sont clairement pas du spam car elles bénéficient de nombreux liens depuis des pages de confiance. Le TrustRank est construit en deux étapes : une étape de sélection du noyau de pages web qui vont servir de référence pour l'initialisation de l'algorithme, et une étape de propagation du score à partir du noyau initial.

Cette première étape est cruciale puisque les pages choisies pour être dans le noyau se verront attribuer un score (maximum) de 1, qui sera ensuite propagé via les liens. Inutile de dire qu'un lien depuis une page avec un score de 1 permet d'avoir un très haut score également.

Si on n'a pas entendu parler du TrustRank comme « super algo qui dégomme le spam », c'est à cause des problèmes liés à cette étape de sélection. La seule bonne méthode est coûteuse puisqu'il s'agit de faire noter les pages par des êtres humains (faire du *quality rating* donc, ce qui est complexe sur des gros volumes de pages). Si on souhaite en revanche réaliser la sélection de manière automatique, on tombe toujours dans le même travers : un spammeur doué réussira toujours à placer des pages dans le noyau.

En effet, parmi les méthodes de sélection automatique, on trouve le tirage de pages au hasard (dans ce cas, une page du noyau sur cinq est du spam) ou le choix des pages à fort PageRank (mauvaise idée, car le métier des spammeurs est de créer des pages à fort PageRank).

La seule idée qui semble efficace est celle de l'utilisation du PageRank inverse. Le PageRank inverse correspond au PageRank standard, mais calculé sur le graphe du web au sein duquel on inverse le sens de tous les liens. En faisant cela, on va minimiser la sélection des pages qui obtiennent leur PageRank grâce à des liens de faible valeur mais en très grand nombre (pour éviter les *mass linkers*). Ceci étant, un bon référenceur réussira toujours à mettre en place quelques pages compatibles avec ce critère du PageRank inverse, et donc à placer quelques pages dans le noyau.

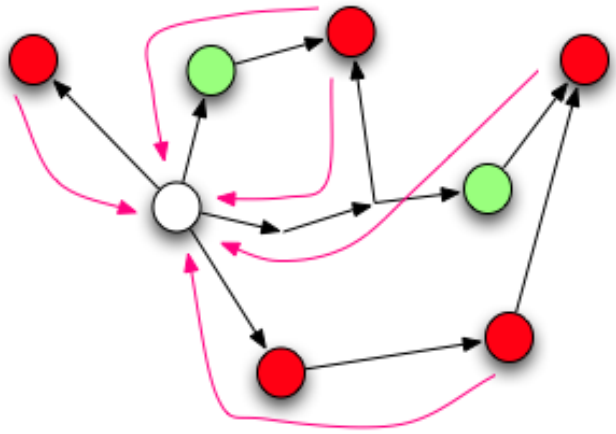
Bref, on voit que la sélection pose problème. Mais même la méthode de transmission de la confiance est à discuter. En effet, il n'existe pas de nombreuses manières de procéder. On peut transmettre toute sa confiance : « si tu es de confiance et que tu me dis que Robert est de confiance, alors je lui confie ma maison, ma vie, mes économies ». On peut, de manière plus réaliste, faire une transmission amortie : « j'ai plus confiance en Robert qui est recommandé par un ami, qu'en Jean qui est recommandé par la sœur du cousin de ce même ami ». Enfin, on peut faire une transmission « splittée » : « quand Robert me recommande Victoria seulement, j'ai plus confiance en elle que si Robert m'avait recommandé Andrea et Victoria ». Là aussi, on peut imaginer un souci : aucune de ses transmissions n'est vraiment en phase avec ce que l'on comprend intuitivement de la notion de confiance.

Nous avons ici évoqué le TrustRank, mais il faut savoir qu'il existe de nombreux travaux similaires. Gyonguy et al. ont par exemple mis en place le concept de « masse de spam » qui est très proche de celui de TrustRank. Dans leur algorithme, chaque page se voit associer une masse de spam relative : il s'agit de la fraction de son PageRank qui provient de pages de spam. Plus cette fraction est forte, plus la page en question sera suspecte.

Face au TrustRank : l'anti-trustrank

A une époque, la notion d'anti-TrustRank circulait chez les SEO américains (avec la rumeur – vraie ou fausse – que le PR0 était totalement dû à un score d'anti-trustrank très haut). Il s'agit de l'inverse du TrustRank. L'idée de la méthode est simple : autant un webmaster n'a pas le contrôle sur les liens entrants, autant il fait ce qu'il veut sur les liens sortants. En conséquence, on va « punir » éventuellement - en rétro-propageant une non-confiance - les sites qui font des liens vers des spammeurs, de façon plus ou moins massive.

Techniquement, il s'agit d'identifier des pages de spam (ce qui est plus facile à faire que d'identifier des pages de confiance), puis on va parcourir aléatoirement les liens en sens inverse. Si, en remontant depuis un pool assez gros de pages de spam, on tombe souvent sur une même page, c'est qu'elle est source de beaucoup de chemins qui amènent au spam, et donc qu'il s'agit a priori d'une page de mauvaise qualité.



Sur l'image ci-dessus, on voit que la page correspondant au nœud blanc est à l'origine de nombreux chemins qui vont vers le spam (les pages rouges) et de peu de chemins vers des pages légitimes (les pages vertes). Sa masse relative est donc haute, et elle a toutes les chances d'être pénalisée par l'algorithme de l'anti-trustrank.

Là aussi, un grand nombres de méthodes proches ont été mises en place par les chercheurs. On peut mentionner les méthodes qui utilisent un noyau de pages de spam, mais qui, au lieu de faire de la propagation, vont réaliser des marches aléatoires autour du spam, et vont croiser cette information de proximité avec une analyse du contenu (par exemple quantifier le taux de duplication, avec l'idée que les spammeurs font beaucoup de réutilisation de contenu). Le principe est proche, mais la méthode est un peu plus subtile, et donne des résultats plus fiable, mais trouvant moins de spam (moins de faux positifs, mais aussi une efficacité réduite).

Conclusion

Pour les chercheurs, la conclusion concernant les méthodes de propagation est pour l'instant plutôt décevante. La plupart des méthodes sont entachées de nombreux défauts : elles sont difficiles à initialiser correctement, elles sont loin de l'idée qu'on veut réellement mettre en place, et enfin soit elles sont efficaces mais se trompent souvent, soit elles sont inefficaces mais font peu d'erreur. Bref, pour l'instant, voici au moins une chose qui ne doit pas empêcher les SEO de dormir !

Sources

Z. Gyöngyi, H. Garcia-Molina, J. Pedersen. Combating Web Spam with TrustRank. 30th International Conference on Very Large Data Bases (VLDB), Toronto, Ontario, Canada, 2004.
http://infolab.stanford.edu/~zoltan/publications/gyongyi2004combating.pdf

C.-N. Ziegler and G. Lausen, "Propagation models for trust and distrust in social networks," Information Systems Frontiers, vol. 7, no. 4–5, pp. 337–358, December 2005.
http://www2.informatik.uni-freiburg.de/~cziegler/papers/ISF-05-CR.pdf

Z. P. Gyongyi, P. Berkhin, H. Garcia-Molina, and J. Pedersen, "Link spam detection based on mass estimation," in Proceedings of the 32nd International Conference on Very Large Databases (VLDB), pp. 439–450, 2006.
http://infolab.stanford.edu/~zoltan/publications/gyongyi2006link.pdf

Gyöngyi, Pavel Berkhin, Hector Garcia-Molina, Jan Pedersen. Link Spam Detection Based on Mass Estimation. 32nd International Conference on Very Large Data Bases (VLDB),

Seoul, Korea, 2006.
http://infolab.stanford.edu/~zoltan/publications/gyongyi2006link.pdf

B.Wu,V.Goel,andD.B.Davison,"Propagatingtrustanddistrusttodemote Web spam," in Workshop on Models of Trust for the Web (MTW), May 2006.
http://www.cse.lehigh.edu/~brian/pubs/2006/MTW/propagating-trust.pdf

B. Wu and K. Chellapilla, "Extracting link spam using biased random walks from spam seed sets," in Proceedings of the 3rd International Workshop on Adversarial Information Retrieval on the Web (AIRWeb), pp. 37–44, New York, NY, USA: ACM Press, 2007.

***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/