Un point sur les brevets : Le Page Rank de Google

Retour au sommaire de la lettre

Nous en avons parlé le mois dernier dans cette lettre, Google détient un certain nombre de brevets auprès de l'USPTO, organisme gérant les brevets aux Etats-Unis. **Larry Page**, co-fondateur de Google, est propriétaire (pour The Board of Trustees of the Leland Stanford Junior University) du brevet dénommé "**Method for node ranking in a linked database**" (numéro 6,285,999), qui décrit le principe du PageRank (http://patft.uspto.gov/netacgi/nph-Parser?Sect1=PTO2&Sect2=HITOFF&p=1&u=/netahtml/search-bool.html&r=1&f=G&l=50&co1=AND&d=ptxt&s1='Page+Lawrence'.INZZ.&OS=IN/"Page+Lawrence"&RS=IN/"Page+Lawrence"). On peut d'ailleurs s'étonner que ce brevet n'appartienne pas à Google mais à l'université de Stanford. Si Larry Page s'en allait de Google, qu'adviendrait-il de ses algorithmes de pertinence, fortement basés sur ce brevet ? Le mot "Google" n'est pas énoncé une seule fois dans le texte de ce brevet...

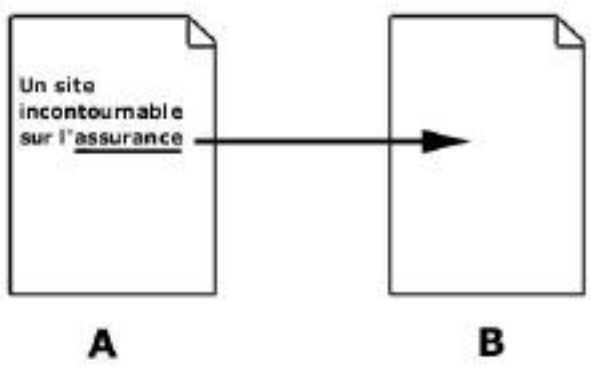
Nous nous sommes penché de façon approfondie sur les explications fournies sur le site de l'USPTO au sujet de ce brevet, afin de mieux comprendre les mécanismes de classement de pertinence de Google.

Voici ce qu'il y est expliqué : l'invention présentée dans le document se rapporte à l'analyse de l'interconnectivité de documents dans des bases de données comparables au Web. Plus particulièrement, elle se rapporte à la façon de donner des classements, des notes, à des "noeuds" de ces bases de données.

Le document relate tout d'abord un bref historique de la façon dont les algorithmes des moteurs de recherche ont été imaginés depuis le début du web, selon certains critères de pertinence :
- Nombre d'occurrence du mot demandé.
- Date de dernière modification du document.
- Proximité des termes demandés entre eux dans le document.
- Etc.

La première conclusion est que ces seules méthodologies ne sont pas assez précises pour fournir des résultats très pertinents. De plus, elles sont potentiellement fortement sujettes au spam.

Le projet HyperLink Search Engine (qui se trouvait à l'adresse http://rankdex.gari.com/, mais celle-ci ne répond plus) est cité par Larry Page comme l'un des premiers outils de recherche à avoir utilisé l'analyse des liens entrants d'une page pour identifier du contenu pertinent. Cet outil utilisait le texte du lien pointant vers le document pour caractériser la pertinence de ce dernier. Exemple :



Si un document A a mis en place un lien vers un document B avec le texte indiqué (le lien est proposé sur le mot "assurance"), le document B sera bien classé sur le mot clé contenu dans le texte du lien du document A (donc, ici, "assurance").

Cette idée d'associer la pertinence d'un document au texte des liens pointant vers lui avait été implémentée dans un premier temps sur l'outil de recherche World Wide Web Worm (http://www.inf.utfsm.cl/~vparada/html/wwww.html), un très ancien (à l'échelle de l'Internet) moteur. Le but était de se servir non pas du contenu de la page en question pour la classer, mais plutôt de celui des pages pointant vers elle. Ingénieux...

Un calcul basé sur la récursivité

Le brevet déposé par Larry Page reprend l'idée de l'analyse des liens vers un document. Dans un premier temps, et de façon basique, il définit, pour un document A, un "taux de citation" r(A) égal au nombre N de pages ayant placé un lien vers lui :

$r(A) = N$

Mais, dans un univers hétérogène comme le Web, cette définition simpliste n'est pas satisfaisante. Il est nécessaire d'aller plus loin et de ne pas noter la simple quantité des liens, mais également leur "qualité". C'est donc ici qu'entre en lice la notion de récursivité du calcul du PageRank : le PageRank d'une page dépend non seulement du nombre de liens pointant vers elle, mais également du PageRank des documents qui les contiennent.

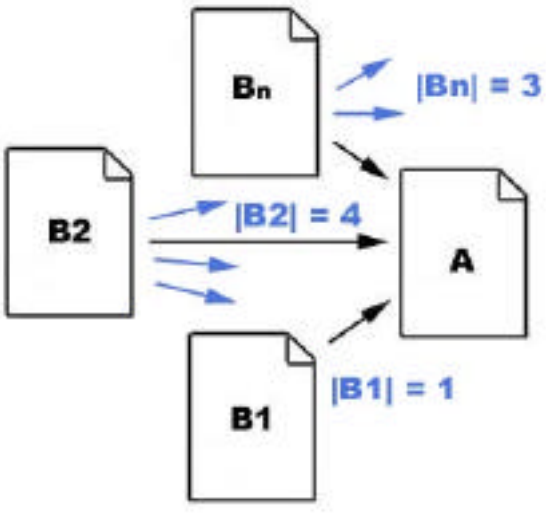
Voici la phrase détaillant le système du PageRank, dans le brevet de Larry Page (rappelons que le terme de PageRank vient du nom même de son inventeur et non pas du mot "page") : "The method comprises the steps of obtaining a plurality of documents, at least some of the documents being linked documents, at least some of the documents being linking documents, and at least some of the documents being both linked documents and linking documents, each of the linked documents being pointed to by a link in one or more of the linking documents; assigning a score to each of the linked documents based on scores of the one or more linking documents; and processing the linked documents according to their scores."

Bref, après avoir pris une aspirine, nous pouvons continuer...

En prenant en compte le "poids" respectif, l'importance de chaque lien pointant sur une page, l'équation donnant un double aspect quantitatif et qualitatif au "taux de citation" devient :

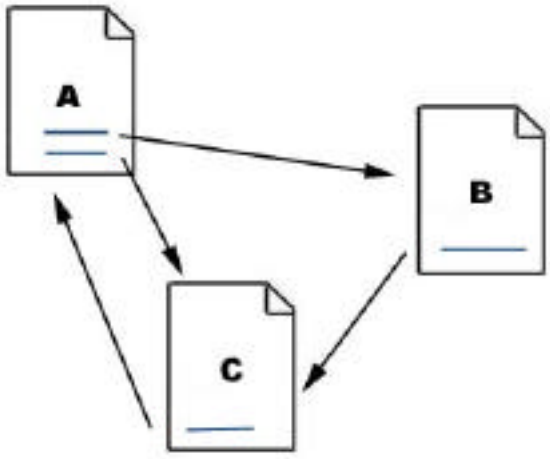
$$r(A) = \frac{x}{N} + (1-x)\left(\frac{r(B1)}{|B1|} + \ldots + \frac{r(Bn)}{|Bn|}\right)$$

où B1, ... Bn représentent les n pages pointant vers A, r(Bn) le "taux de citation" de la page Bn, |B1|...|Bn| indiquant le nombre de liens sortant de B1... Bn. "x" est une constante dont la valeur est comprise entre 0 et 1 (proche de 0,1 le plus souvent) et N représente le nombre total de pages du Web. Ces valeurs sont reprises sur le diagramme ci-dessous :



Le but de cette équation est de prendre en compte la "qualité" des liens pointant vers la page A. Plus le "taux de citation" des pages Bx sera fort, plus celui de la page A augmentera... Ainsi, un document ayant un seul lien depuis une page très "populaire" pourra être mieux classé qu'un autre document pointé par 10 documents très peu "populaires".

Illustrons cette formule par un calcul simple :



Le document A contient 2 liens, vers B et vers C.
Le document B contient 1 lien, vers C.
Le document C contient 1 lien, vers A.

A a un seul lien pointant vers lui (de C).
C a un seul lien pointant vers lui (de B).

Donc : r(A) = r(C)

B a un seul lien pointant vers lui, mais ce lien émane de A qui, lui-même, contient deux liens sortants.

Donc : r(B) = r(A)/2

C a un lien depuis B (lien sortant unique de B) et un lien depuis A (1 lien sur les 2 que contient A).

Donc : r(C) = r(A)/2 + r(B)/1 = r(A)/2 + r(B)

On obtient donc rapidement les valeurs :
r(A) = 0,4
r(B) = 0,2
r(C) = 0,4

Dans notre cas, reprenons l'équation de départ :

$$r(A) = \frac{x}{N} + (1-x)\left(\frac{r(B1)}{|B1|} + \ldots + \frac{r(Bn)}{|Bn|}\right)$$

Dans cet exemple, N=3. Prenons, par défaut, une variable x=0,5. Donc : x/n = 0,5/3 (=1/6) et 1-x=0,5 (=1/2). On obtient les résultats suivants :

r(A) = 1/6 +r(C)/2

De même :

r(B) = 1/6 +r(A)/4

et :

r(C) = 1/6 +r(A)/4 + r(B)/2

D'où les résultats :

r(A) = 0,358
r(B) = 0,256
r(C) = 0,385

Etc. C'est ainsi qu'est calculé le PageRank par Google. Larry Page indique simplement que, lors du calcul, seules deux itérations (ou trois, rarement plus) sont effectuées pour calculer le PageRank et que la constante N est égale au nombre de pages contenant au moins un lien hypertexte. Les pages sans lien ne sont donc pas prises en compte dans le calcul, ce qui semble logique.

Le brevet indique également que cette méthode peut être complétée par d'autres facteurs, comme :
- Le fait de donner un poids plus fort aux liens externes plutôt qu'aux liens émanant du site lui-même (liens internes) qui sont cependant pris en compte.
- L'"importance" des sites pointant vers la page (institution, auteur "connu", localisation géographique).
- Le lien provenant d'une page d'acceuil peut être surévalué par rapport à un lien provenant d'une page interne.
- La situation du lien à l'intérieur de la page : un lien en haut de document peut avoir plus de poids qu'un lien en fin de page.
- La mise en exergue du lien : la taille de la police de caractère, en gras, etc. peuvent également avoir plus de poids.
- La date de dernière modification des pages contenant les liens : plus celle-ci est récente, plus le lien a d'importance.
- Le texte du lien (voir début de l'article) peut avoir une forte importance.
- La présentation du brevet indique que ce dernier peut également être utilisé en corrélation avec l'utilisation statistique des pages qu'en font les internautes. Une page très souvent lue (i.e. cliquée dans les pages de résultats ?) pourra ainsi avoir une importance plus grande...

Conclusion

Que devons-nous retenir de ce brevet ? Qu'il est très important de bien "choisir" les pages qui vont pointer vers votre site (si tant est que vous puissiez les choisir :-)) :
- Plutôt des pages d'accueil ou des pages disposant d'un bon PageRank (utilisez la Googlebar pour cela, afin de visualiser de façon instantanée le PageRank des pages affichées sur votre navigateur).
- Plutôt des pages disposant de peu de liens sortants.
- Les liens vers vos pages devront si possible se trouver en haut des documents pointant vers vous, et toute mise en exergue (gras, grande taille de caractères, etc.) sera un "plus".
- Enfin, soignez, lorsque cela est possible, les textes des liens qui pointent vers vos documents : plus ils contiendront de mots clés relatifs à votre activité et mieux ce sera.

Malheureusement, il est très complexe de maîtriser la façon dont les autres sites web vont créer des liens vers vos pages. Alors, pourquoi ne pas leur proposer une petite "trousse à outils" ou des exemples de liens, sous forme de "templates", sur une page ad'hoc ? Les webmasters désirant mettre en place un lien vers votre site auront ainsi à leur disposition toutes les infos nécessaires pour le réaliser. Libre à eux de suivre vos directives ou non, bien sûr. Faites attention également à ne pas en faire trop : vous pourriez vite tomber dans le spam, ce que n'aime pas du tout (avec raison...) Google... Tout est question de bon sens, un webmaster averti en vaut toujours deux...