Les moteurs sont déjà entrés dans la 4ème dimension : le temps

Retour au sommaire de la lettre

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

La dimension temporelle est devenu un critère incontournable de pertinence pour les moteurs de recherche. La fraîcheur de l'index est également un point très important pour la qualité des résultats renvoyés. De quelle manière les moteurs prennent-ils en compte cette notion de temporalité ? Donnent-ils plus d'importance aux pages anciennes ou aux pages récentes ? Comment déterminent-ils les pics d'actualité ? Quels critères sont pris en compte pour calculer l'"âge" d'une page ? Cet article tente de répondre à toutes ces questions...

Voila déjà dix ans, plusieurs chercheurs (en particulier Kumar (1999), Cho et Garcia-Molina (2000) et Kleinberg (2000), mais la réflexion est plus ancienne car on retrouve des articles de bibliométrie parlant de ce thème dès 1955 (Garfield)) spécialistes du domaine de l' « information retrieval » (extraction d'information, la « science des moteurs de recherche ») avaient remarqué que la prise en compte de la dimension temporelle était indispensable pour construire un algorithme performant pour un moteur de recherche.

Pourtant un moteur comme Google a très longtemps négligé la collecte d'information sur l'évolution de ce qu'ils appellent les signaux, c'est-à-dire les critères utilisés dans l'algorithme. Tout en accordant dès l'origine une grande attention à d'autres critères liés au temps comme la fraîcheur de l'index et l'âge des pages.

L'un des premiers indices spectaculaires de l'existence d'une prise en compte de critères d'évolution temporelle dans l'algorithme de Google est malgré tout assez ancien : au cours du printemps 2004 des observateurs ont noté un phénomène étrange affectant de nouveaux sites et les empêchant d'apparaître en tête des résultats (ce phénomène a semble-t'il été observé dès début 2004 mais « théorisé » un peu plus tard). Il fut baptisé « effet sandbox » par Barry Schwartz de Seroundtable. Depuis, les référenceurs ont tendance à appeler « sandbox » tout et n'importe quoi, mais la plupart des phénomènes assimilés à la sandbox présentent tous des analogies troublantes avec ce que l'on peut produire par la technique d'analyse temporelle des liens (TLA) dont nous parlerons plus tard.

Depuis lors, les indices d'une prise en compte de multiples critères temporels dans l'algorithme se multiplient, mais sans que cela soit forcément remarqué et discuté dans les forums et les blogs.

Dans cet article, nous allons donc essayer de faire le point sur les différentes techniques qu'un moteur peut utiliser pour tenir compte de l'évolution temporelle des pages, des liens et du web dans son algorithme, tout en s'attachant à indiquer les limites de ces méthodes.

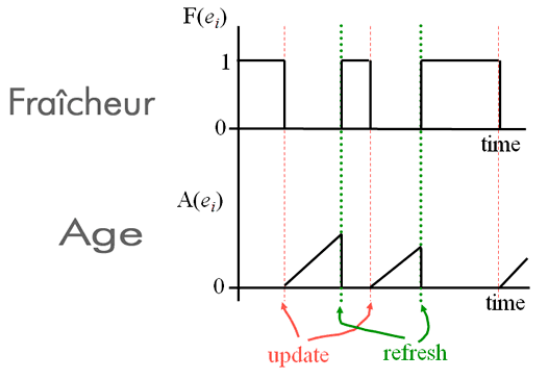
La problématique de l'âge et de la fraîcheur

Le premier élément temporel que l'on doit prendre un compte pour réaliser un moteur performant, c'est la « fraîcheur » de l'information restituée. Cette notion de fraîcheur n'a rien à voir avec l'âge d'une page : on mesure le caractère récent ou non de l'extraction d'information (sa date de dernière mise à jour), que la page soit ancienne ou qu'elle vienne d'apparaître sur le web.

L'âge d'une page est donc une mesure du temps qui s'est écoulé depuis son apparition. Dans le contexte d'un outil de recherche, il faut définir deux âges : l'âge de la page web proprement dite, qui mesure le délai depuis son apparition sur le web, et l'âge de la page dans l'index, qui mesure le temps écoulé depuis son indexation. Plus le crawl est efficace, plus l'index est « frais », et plus les deux âges vont être proches.

Il existe également deux façons de définir la « fraîcheur ». La première part d'une notion binaire : une page dans l'index est « à jour », ou elle ne l'est pas. Une page dans l'index d'un moteur est considérée comme « à jour » si la version contenue dans l'index est la même que celle que l'on trouve sur le web. Cela signifie donc qu'elle n'a pas été mise à jour depuis. La « fraîcheur » de l'index est alors définie comme le taux de pages « à jour » rapporté à la taille de l'index (on devrait d'ailleurs plutôt parler de taux de pages « à jour » plutôt que de fraîcheur, ce qui éviterait les ambiguïtés, mais les habitudes sont prises...). Une deuxième définition consiste à mesurer le délai qui s'est écoulé depuis le dernier « crawl » de la page. Plus ce délai est court, plus la page est considérée comme « fraîche ».

Pourquoi deux ratios pour mesurer la fraîcheur ? Parce que dans certains cas particuliers, ces mesures échouent à déterminer la qualité de l'index. Beaucoup de pages du web par exemple ne changent jamais. Les moteurs de recherche en tiennent compte, et ne vont les crawler que de loin en loin. Dans ce cas, la mesure la plus intuitive de la « fraîcheur », c'est-à-dire la deuxième, aura tendance à indiquer une mauvaise qualité de l'index alors que des pages non crawlées depuis des lustres sont néanmoins parfaitement « à jour ». A l'inverse, si l'on prend par exemple des pages contenant un fil d'actualité, la page indexée ne sera en général pas « à jour ». Il est important de savoir par contre si l'on dispose d'une version très récente de la page.



L'illustration ci-dessus représente un graphique symbolisant l'évolution des mesures de l'âge et de la fraîcheur dans l'index d'un moteur. Lorsque la page vient d'être crawlée, son « âge » dans l'index retombe à zéro, et sa fraîcheur reste à 1 tant que la page est considérée comme à jour. La fraîcheur tombe à zéro si la page de l'index n'est plus à jour (on est ici dans la conception : fraîcheur = taux de pages à jour). L'idée pour un moteur parfait est de déclencher un « recrawl » dès qu'une page est considérée comme obsolète.

Quelles sont les performances des moteurs en matière de fraîcheur de l'index ?

Dirk Lewandowski de l'université des Sciences Appliquées de Hambourg a mené une étude sur trois ans (entre 2005 et 2008) sur les performances des trois principaux moteurs de recherche sur le critère de la « fraîcheur » de l'index. Son étude publiée en 2008 dans le JIS (*Journal of Information Science*) révèle une très grande variabilité des performances entre les moteurs et aussi dans le temps !

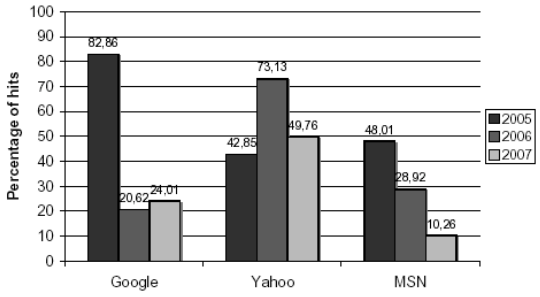


Fig. 1 ***Percent of pages that are up to date 2005–2007***

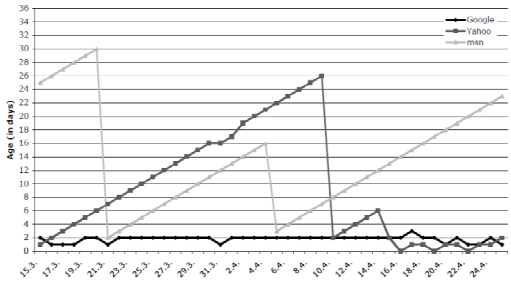


Fig. 9 Update pattern for German Wikipedia main page

La courbe de gauche montre qu'en 2005, le taux de pages à jour sur Google est très supérieur à celui de ses concurrents sur l'échantillon de pages étudié. Ce leadership est perdu en 2006, pour être retrouvé en 2007, mais avec des performances bien moindres. Il semble en fait qu'aucun des moteurs en 2007 n'avait résolu complètement le problème de la mise à jour de leur index. Il faut dire que le problème devient de plus en plus complexe au fil du temps avec l'évolution du web. L'exploitation des sitemaps xml fait ainsi partie des solutions trouvées par les moteurs pour améliorer la situation.

Les obstacles à la détermination de l'âge d'une page ou d'un lien

Pour déterminer l'âge d'une page sur le web, l'idée la plus simple est d'utiliser les informations de date de création et de dernière modification communiquées par le serveur web dans l'en-tête http. Le problème c'est que les serveurs web ne renvoient pas toujours cette information, et que parfois ils renvoient dans le champ « *last modified* »... la date de consultation de la page.

Plusieurs études indiquent que le nombre de pages pour lesquelles on dispose d'une date de dernière modification fiable est inférieure à 50% (et même moins de 40% dans l'étude de Einat Carmel et alter en 2004). La date de création de la page est encore plus fantaisiste : elle ne correspond que rarement à la date de première mise en ligne de la page.

Ceci explique pourquoi Google conseille aux webmasters de bien configurer leurs serveurs web pour renvoyer des dates correctes. Mais ces conseils prodigués depuis des années n'ont pas amélioré la situation. N'hésitez cependant pas à bien vérifier la configuration de votre serveur à ce niveau si vous désirez connaître une meilleure indexation de vos documents !

Un autre problème difficile à résoudre résulte de la complexité des changements qui interviennent sur les pages. Les pages ont souvent un comportement composite, avec une partie de ses composants qui restent fixes et d'autres qui évoluent, dans des proportions et à des rythmes variés. Un site d'actualité verra de nouvelles actus chasser les anciennes rapidement, chaque nouvelle brève étant importante à indexer. Mais le volume de texte de la page reste constant. A l'inverse, une page éditoriale accompagnée de commentaires verra son volume de texte augmenter, mais chaque commentaire ajouté n'a pas individuellement une grande valeur par rapport au reste de la page. Ces différences de comportement d'actualisation et d'importance des informations contenues dans les zones modifiées imposent la mise en place de stratégies sophistiquées de recrawl.

Pour illustrer ce point, voici ci-dessous des graphes issus d'un article paru en 2008 à propos d'un travail de recherche mené pour Yahoo et signé par Christopher Olston et Sandeep Pandey ("*Recrawl Scheduling Based on Information Longevity*").

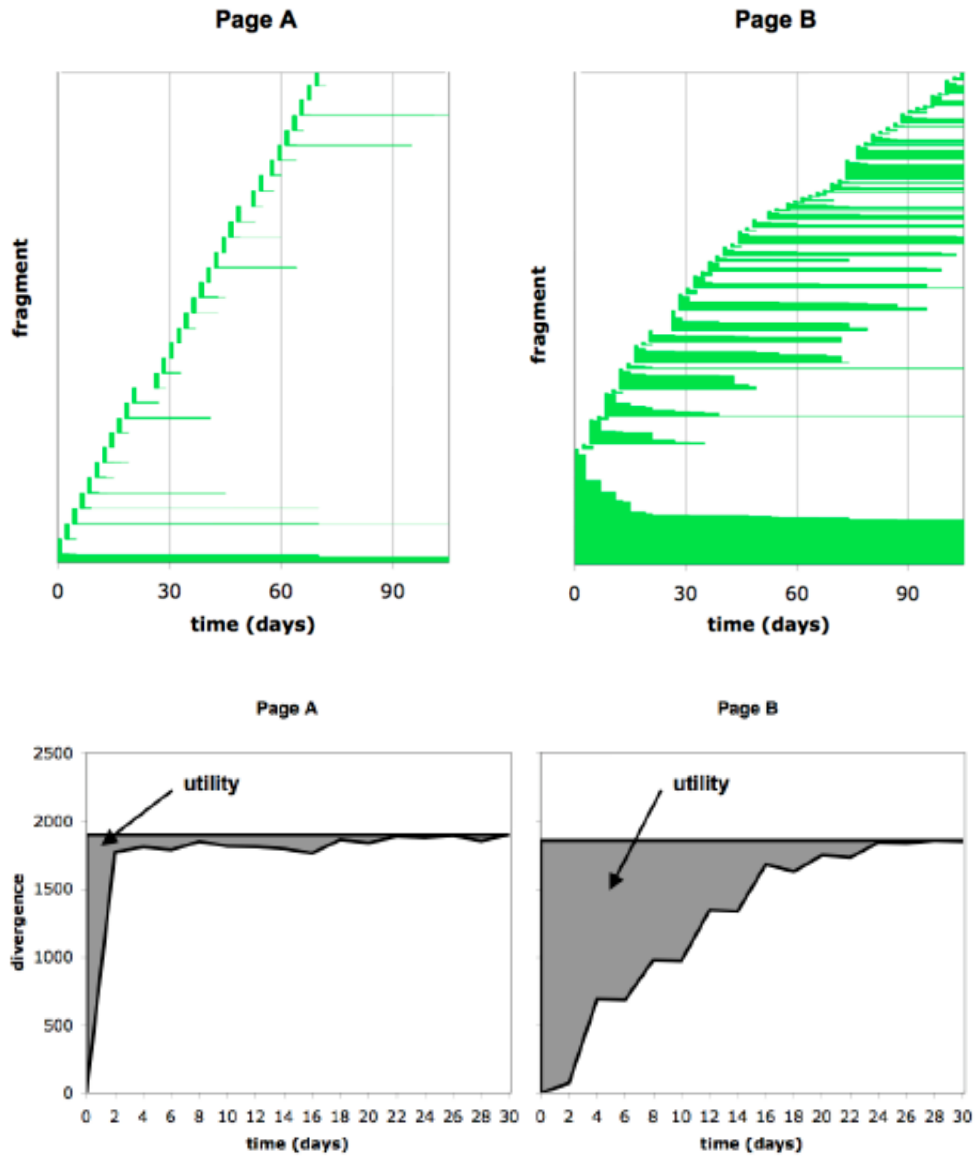


Figure 2: Two divergence graphs, with utility.

Le graphe de gauche montre l'évolution de deux pages au fil du temps. La page A est une page de type « actualités » dans laquelle un grand nombre de zones changent au fil du temps, un élément chassant l'autre, chaque élément ayant une faible durée de vie. La page de droite est une page d'un site de recettes de cuisine : une partie de la page ne bouge pas, le reste de la page est constitué de commentaires qui ont une certaine durée de vie. Le graphe de droite visualise la divergence au fil du temps entre la page stockée dans l'index du moteur et la page en ligne sur le site. Pour la page A, un nouveau crawl s'impose au bout deux jours, tandis que pour la page B, la divergence est plus progressive.

Quelles pages favoriser dans l'algorithme ? Les pages anciennes ou les pages récentes ?

Les premières versions des algorithmes des moteurs avaient une tendance à accorder plus de poids à une page « ancienne » qu'à une page récente. Cela peut avoir du sens : une page ancienne, au sein d'un site en ligne depuis longtemps, peut être considérée comme plus fiable. D'abord, elle est moins suspecte d'être une page créée pour un objectif de spam. Ensuite, au fil du temps, de telles pages reçoivent de plus en plus de liens ce qui fait qu'un algorithme

comme le Pagerank donne à la longue une prime aux pages anciennes qui sont considérées donc comme plus importantes. Parfois à raison, parfois à tort.

Le problème, c'est que lorsqu'un internaute cherche une information liée à l'actualité, il y a de fortes chances que les pages pertinentes sur cette requête soient « anciennes ». Il faut donc ajouter dans l'algorithme un système qui permet de favoriser à bon escient des pages récentes (une « prime de fraîcheur » en quelque sorte).

Mais par ailleurs, au fil des ans, avec l'évolution de l'internet, il faut aussi compter avec l'obsolescence progressive des informations contenues dans les pages. Le problème qui était anecdotique en 1999 lors des balbutiements de Google est devenu réellement sérieux depuis. Il faut donc élaborer des solutions pour détecter les pages obsolètes, et pour éviter de leur donner une importance qu'elles ont perdu !

L'une des méthodes envisageable pour détecter cette obsolescence est de tester si une page a cessé de recevoir régulièrement de nouveaux liens, ou des liens depuis des pages actualisées. Cette approche s'appelle l'analyse temporelle des liens.

L'analyse temporelle des liens

En 2004, un article publié par des chercheurs du laboratoire IBM d'Haifa (*Trend detection through Temporal Link Analysis*, par Amitay, Carmel, Herscovici, Lempel, Soffer, laboratoire de recherche IBM d'Haifa Israël : http://einat.webir.org/JASIS_2004_Temporal_Links_Analysis.pdf) a fait un peu de bruit au sein des spécialistes des outils de recherche. Après avoir remarqué que les algorithmes des moteurs (en particulier le Pagerank de Google - même s'ils notent par ailleurs que l'idée de créer un algorithme tenant compte de ce critère était bien présente dans les premiers articles de Larry Page sur le pagerank) ne prenaient pas en compte la temporalité, et pour améliorer le système, ils ont proposé une nouvelle approche qu'ils ont baptisé TLA (*Temporal Link Analysis*). Le principe de l'analyse temporelle des liens (ATL en français) consiste tout bonnement à exploiter les informations suivantes :
- Les dates de création des pages ;
- La date de dernière modification ;
- La date de disparition d'une page (date à partir de laquelle le serveur à renvoyé un code 404) ;
- La date d'apparition des liens sur les pages ;
- Et la date de disparition des liens sur les pages ;

A propos des codes 404 et 410

La nécessité pour les moteurs de bien comprendre la signification d'un code 404 explique pourquoi Google a récemment annoncé qu'il voulait traiter les codes 410 différemment. Voici la signification théorique des codes 404 et 410 :
404 : Not found => le serveur indique que la page désignée par l'url ne correspond pas à une ressource connue.
410 : Gone => le serveur indique que la page désignée par l'url n'existe plus (mais indique aussi qu'elle a existé et que l'url est connue).
Jusqu'ici, Google considérait de la même façon les deux types de code renvoyé. Ce n'est plus le cas, et Google recommande dorénavant l'utilisation du code 410 pour les pages qui ont disparu, pour indiquer le caractère « permanent » de cette disparition. Voir : http://www.google.com/support/forum/p/Webmasters/thread?tid=1bc5206b5e0fac47&hl=en#all

L'article n'explique pas de manière explicite comment « exploiter » les données issues de l'ATL, mais donne plusieurs pistes. La première est de détecter certains schémas de croissance anormale des liens pointant vers une page, afin de détecter l'existence d'une stratégie de « link spam ». La seconde est d'identifier les pages obsolètes. La troisième consiste à introduire dans le pagerank une modification tenant compte du critère de temporalité.

$$t(x,y) = w_{t1} \cdot \frac{f(y)}{\sum_{(x,z)\in E} f(z)} + w_{t2} \cdot \frac{f(x,y)}{\sum_{(x,z)\in E} f(x,z)} + w_{t3} \cdot \frac{avg\{f(v,y \mid (v,y) \in E\}}{\sum_{(x,w)\in E} avg\{f(v,w \mid (v,w) \in E\}}$$

Voici, ci-dessus, un exemple de formule de pagerank modifiée à partir d'une ATL. L'Analyse Temporelle des Liens a depuis 2004 engendré une littérature scientifique abondante et de très nombreuses études, expériences et applications. Cette variante a été présentée par Klaus Berberich de l'institut Max Planck...

Les autres critères temporels

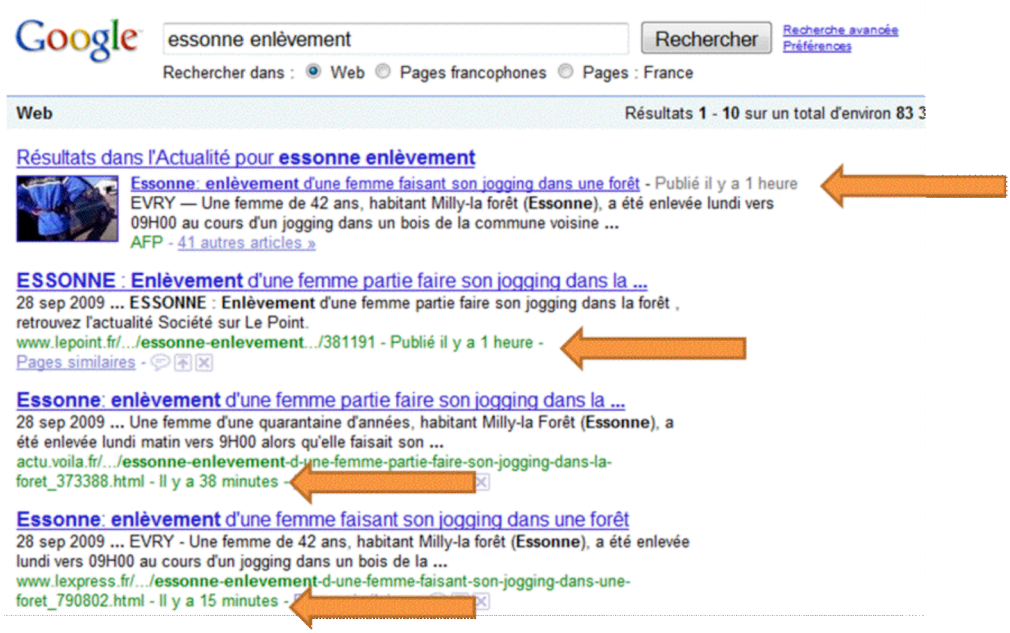
L'ATL ne permet pas d'exploiter tous les signaux temporels exploitables sur le web. Un brevet publié par Google en 2005 en a révélé bien d'autres (*Information Retrieval Based on Historical Data* : http://www.seomoz.org/article/google-historical-data-patent). Le schéma ci-dessous résume l'essentiel des critères évoqués dans ce brevet :

Un exemple d'analyse temporelle des flux de requêtes : les requêtes QDF

Dans un article du New York Times (*Google Keeps Tweaking Its Search Engine*, NY Times edition du 3 juin 2007 : http://www.nytimes.com/2007/06/03/business/yourmoney/03google.html) rédigé à partir de conversations avec Amit Singhal (le « maître » de l'algorithme chez Google) et d'Uri Manber, il a été fait mention pour la première fois de l'existence des requêtes QDF (*Query Deserves Freshness*). Si on en croit les informations « lâchées » par différents interlocuteurs à Mountain View, il semble que les dirigeants de Google se soient émus de l'absence de résultats « frais » sur la requête « tsunami » au lendemain de la catastrophe de décembre 2004. Plus récemment, le fait que l'on ne trouve pas d'informations financières fraîches sur Google au moment de leur entrée à Bourse à achevé de convaincre d'introduire un changement dans l'algorithme.

Une requête « QDF » est une requête qui d'un seul coup vient à être tapée en un laps de temps très court par un grand nombre d'internautes. Ce comportement permet de déceler qu'un évènement survient qui provoque un grand nombre de recherches d'information sur ce sujet. Quand une telle explosion de la fréquence de frappe d'une requête donnée est décelée,

elle déclenche la mise en service d'un algorithme de classement des résultats alternatif, qui fait la part belle aux documents « frais » et « récents » et aux sources d'actualités.



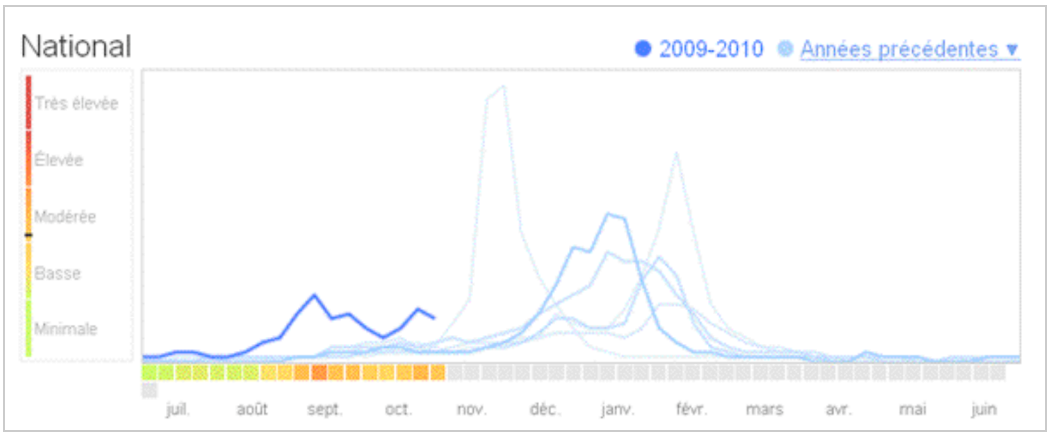
On voit ci-dessus un exemple du comportement actuel de Google sur une requête portant sur un sujet d'actualité. La page de résultat sur « enlèvement essonne » se remplit de pages créées quelques minutes auparavant à partir d'une dépêche de l'AFP publiée une heure auparavant.

L'analyse des tendances

Avec l'évolution du web il est devenu particulièrement intéressant de découvrir les tendances et les « buzz » qui naissent et meurent sur la toile. L'analyse de plusieurs types de signaux permet d'identifier ce qui à un moment donné constitue soit un élément nouveau, soit une information qui devient importante soit un changement de comportement.

Les requêtes tapées par les internautes sont une source particulièrement intéressante pour détecter ce qui fait l'actualité. Par exemple toute personne qui vit sont « quart d'heure de célébrité » (pour reprendre l'expression d'Andy Warhol) peut être identifiée parce que son nom est soudainement beaucoup plus tapé. Mais on peut également exploiter les données sur l'évolution des liens et des ancres, sur les billets publiés dans les blogs, ou l'évolution des liens entre personnes dans les réseaux sociaux.

L'une des applications les plus connues de l'exploitation des données temporelles et des flux de requêtes pour déterminer des tendances est Google Trends, et en particulier l'utilisation des données de Trends pour suivre l'évolution de la grippe.



Un enjeu : l'indexation de l'information en temps réel

Depuis quelques mois, une bataille de communication oppose Google, Bing, Yahoo et Twitter à propos de l'information « temps réel ». Twitter, mais aussi d'autres outils sociaux créent la possibilité d'avoir une information immédiate grâce aux internautes. Faites l'expérience : si vous voulez savoir si le RER A circule à nouveau, préférez Twitter au site de la RATP ou à l'AFP. Twitter a carressé un moment l'idée de créer seul un outil d'information temps réel, mais il semble depuis qu'ils préfèrent l'idée d'un partenariat avec un moteur majeur.

La temporalité : un élément à intégrer dans le référencement

Ce survol des techniques utilisées par les moteurs pour exploiter les informations temporelles met en lumière l'importance de ces critères dans l'algorithme. Il est donc indispensable de bien penser à l'effet de ces critères sur le référencement de ses sites.

En particulier, on s'attachera à réfléchir à la manière dont les pages web de vos sites naissent, meurent, changent, évoluent, pour savoir si ce comportement est cohérent, est lisible par les moteurs ou conforme à l'image que vous souhaitez donner de l'importance à ces pages.

Bibliographie et références

Site français de suivi de la grippe à partir de Google Trends :
http://www.google.org/flutrends/intl/fr/fr/

La temporalité : critère utile pour un moteur de recherche

Kumar, R., Raghavan, P., Rajagopalan, S., & Tomkins, A. (1999). *Trawling the Web for emerging cyber-communities*. Proceedings of the Eighth International World Wide Web Conference, Computer Networks & ISDN,
http://www.uzh.ch/home/mazzo/reports/www8conf/2166/pdf/pd1.pdf

Kraft, R., Hastor, E., & Stata, R. (2003, June). *TimeLinks: Exploring the evolving link structure of the Web. Proceedings of the Second Workshop on Algorithms and Models for the Web-Graph*, Budapest, Hungary.
http://cis.poly.edu/~qq_gan/papers/timelink.pdf

Kleinberg, J.M. (2002, July 23–26). *Bursty and hierarchical structure in streams. Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (KDD 2002)
http://www.cs.cornell.edu/home/kleinber/bhs.pdf

Sur la thématique de l'obsolescence

Bar-Yossef, Z, Broder, A. Z., Kumar, R. and Tomkins, A. *Sic Transit Gloria Telae: Towards an understanding of the web's decay*. Proceedings of the 13th International World Wide Web Conference, pages 328–337, New York, USA. ACM Press, 2004.
http://*portal.acm.org/ft_gateway.cfm?id=988716&type=pdf*
(ressource réservée aux abonnés de l'ACM)

Fraîcheur, âge et problème de la fréquence du recrawl

J. Cho and H. Garcia-Molina. *The Evolution of the Web and Implications for an Incremental Crawler*. Proc. VLDB, 2000.
http://oak.cs.ucla.edu/~cho/papers/cho-evol.pdf

J. Cho and H. Garcia-Molina. *Effective Page Refresh Policies for Web Crawlers*. ACM Transactions on Database Systems, 28(4), 2003.
http://oak.cs.ucla.edu/~cho/papers/cho-tods03.pdf

J. Cho and H. Garcia-Molina. *Estimating frequency of change. ACM Transactions on Internet Technology*, 3(3), 2003.
http://oak.cs.ucla.edu/~cho/papers/cho-freq.pdf

E. Coman, Z. Liu, and R. R. Weber. *Optimal robot scheduling for web search engines*. Journal of Scheduling, 1, 1998.
ftp://ftp.inria.fr/INRIA/publication/publi-pdf/RR/RR-3317.pdf

J. Edwards, K. S. McCurley, and J. A. Tomlin. *An Adaptive Model for Optimizing Performance of an Incremental Web Crawler.* Proc. WWW, 2001.
http://www.www10.org/cdrom/papers/pdf/p210.pdf

Analyse Temporelle des Liens

Trend Detection Through Temporal Link Analysis. Einat Amitay, David Carmel, Michael Herscovici, Ronny Lempel, and Aya Soffer / IBM Research Lab in Haifa (2004)
http://einat.webir.org/JASIS_2004_**Temporal_Links_Analysis**.pdf

Analyse de l'évolution des pages

D. Fetterly, M. Manasse, M. Najork, and J. L. Wiener. *A large-scale study of the evolution of web pages*. Proc. WWW, 2003.
http://research.microsoft.com/pubs/73808/p97-fetterly.pdf

Recrawl Scheduling Based on Information Longevity. Christopher Olston Yahoo! Research Sandeep Pandey Carnegie Mellon University
http://infolab.stanford.edu/~olston/publications/www08.pdf

A three-year study on the freshness of Web search engine databases - Dirk Lewandowski / Hamburg University of Applied Sciences, Hamburg, Germany
http://www.bui.haw-hamburg.de/fileadmin/user_upload/lewandowski/doc/freshness_web_search_engine_databases_JIS2008.pdf

Etude des tendances

Kumar, R., Novak, J., Raghavan, P., & Tomkins, A. (2003, May 20–24). *On the Bursty evolution of blogspace*. Proceedings of the 12th International World Wide Web Conference (WWW2003)
http://cui.unige.ch/tcs/cours/algoweb/2005/articles/p568-kumar.pdf

Popescul, A., Flake, G.W., Lawrence, S., Ungar, L.H., & Giles, C.L. (2000,May 22–24). *Clustering and identifying temporal trends in document databases*. Proceedings of IEEE Advances in Digital Libraries 2000
http://clgiles.ist.psu.edu/papers/ADL-2000-temporal-clusters-DLs.pdf

Brevets de Google sur l'exploitation des données historiques

Information retrieval based on historical data
http://appft1.uspto.gov/netacgi/nph-Parser?Sect1=PTO1&Sect2=HITOFF&d=PG01&p=1&u=%2Fnetahtml%2FPTO%2Fsrchnum.html&r=1&f=G&l=50&s1=%2220050071741%22.PGNR.&OS=DN/20050071741&RS=DN/20050071741

Systems and methods for determining document freshness
http://www.wipo.int/pctdb/en/wo.jsp?WO=2005033977

Philippe Yonnet, ***Directeur Technique @Position (http://www.aposition.com) et président de l'association SEO Camp (http://www.seo-camp.org/)***

Réagissez à cet article sur le blog des abonnés d'Abondance :
http://abonnes.abondance.com/blogpro/2009/11/les-moteurs-sont-deja-entres-dans-la.html