Linkkit : une nouvelle technologie 100% française

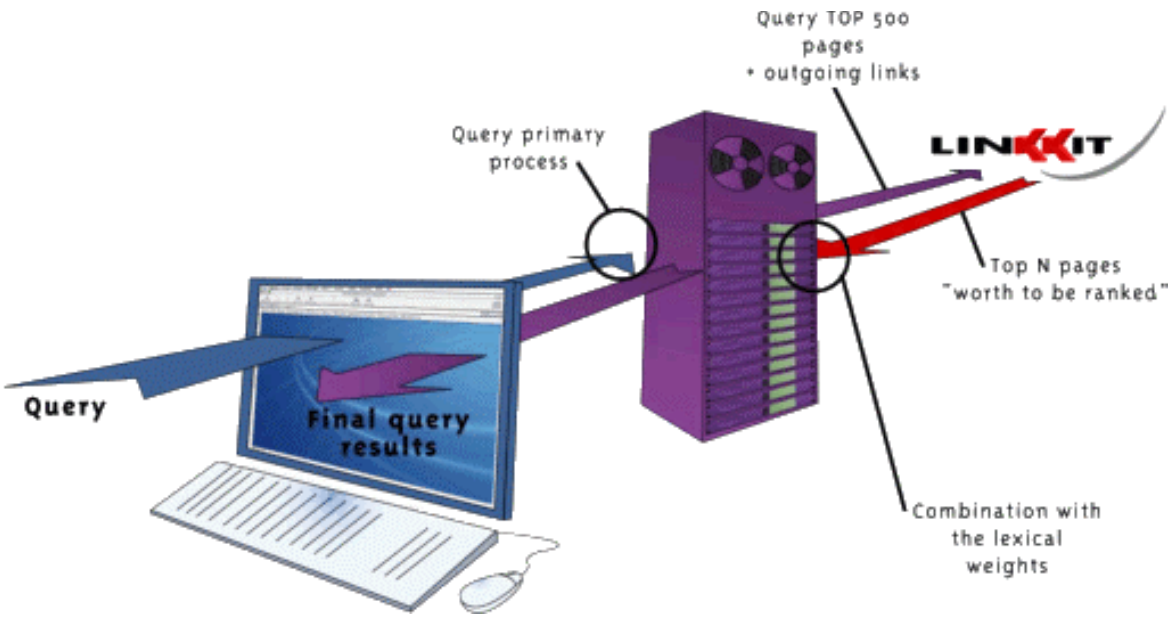
Retour au sommaire de la lettre

Le monde des moteurs de recherche est un monde en perpétuel changement, et ce ce qui en fait son charme et son côté passionnant ! Dans ce cadre, une société française, Linkkit, basée à La Seyne Sur Mer (près de Toulon), vient de développer une nouvelle technologie permettant, selon ses concepteurs, d'augmenter la pertinence d'un moteur de recherche.

Notons bien, en préliminaire de cet article, que la solution qui va être décrite dans ces pages ne constitue pas une solution "moteur de recherche" en elle-même. Elle se greffe, s'adapte à un moteur existant dans le but d'augmenter la pertinence des résultats déjà fournis par l'outil de recherche. Ainsi, la technologie conçue par Linkkit pourrait être utilisée (voire rachetée, car il s'agit là de l'objectif de ses concepteurs) par une société comme Fast, AltaVista (ou Overture :-)), Google ou encore Inktomi (Yahoo!).

Une analyse "communautaire" des liens du Web

Comment fonctionne la technologie développée par Linkkit ? C'est finalement assez simple... Le système peut être expliqué en plusieurs étapes. Imaginons que La technologie soit utilisée par un moteur, que nous baptiserons du nom imaginaire de "GooVista" :-)



1. L'utilisateur entre sa requête sur le formulaire de GooVista et lance la recherche.

2. Comme pour une requête tout à fait "classique", GooVista effectue une recherche dans son index et trouve les pages les plus pertinentes selon son propre algorithme de pertinence.

3. C'est ici qu'intervient une étape supplémentaire : Au lieu de fournir directement ses résultats à l'internaute, GooVista les redirige vers le module "Linkkit". Il fournit, par exemple, les 500 premiers liens qu'il a trouvés avec, pour chacun d'entre eux, les liens sortants issusde chacune des pages trouvées. Le "lot de pages" de pages fourni peut alors être qualifié de "communautaire" ou "contextuel" puisqu'a priori, tous ces documents parlent de la même chose, du même thème. Enfin, en tout cas, si le moteur GooVista n'est pas le plus mauvais moteur du Web :-)

4. Linkkit reçoit ces données et analyse alors les connexions entre elles en s'aidant de la liste des liens sortants fournis par le moteur. Linkkit applique donc ses algorithmes propriétaires (protégés par brevet) en temps réel à un espace communautaire de pages traitant d'un même sujet pour produire le résultat.

Les algorithmes mis en œuvre par Linkkit appartiennent à la famille de l'"analyse relationnelle contextuelle". Ils ont été transposés à partir des indicateurs de "densité" et de "centralité" utilisés en sociologie dans l'analyse des réseaux sociaux. Transposés au monde du web, ces algorithmes ont été adaptés pour intégrer l'existence de "spam relationnel", décrit plus loin dans cet article.

Linkkit renvoie alors le "Top 500" initial au moteur, mais reclassé en fonction des interconnexions des documents : les pages considérées comme étant les plus pertinentes seront celles qui seront le plus souvent "pointées" par les pages du lot fourni. Le travail effectué ne présente donc aucune composante lexicale (prise en compte dans la première étape par le moteur lui-même) et est uniquement basé sur l'interconnectivité des documents à l'intérieur d'un lot de pages pour réorganiser un premier classement effectué par un moteur existant.

Le but, l'objectif de Linkkit est de trouver les 10 meilleurs résultats pour une requête, les plus pertinents, qui sont parfois disséminés dans le "Top 500" du moteur, et de les remonter au niveau du "Top 10".

5. Dernière étape : Linkkit fournit le "Top 500" reclassé au moteur qui affiche les nouveaux résultats à l'internaute, soit de façon brute soit en les retravaillant éventuellement une deuxième fois grâce à ses propres citères de pertinence.

Il est important de comprendre que les algorithmes mis en œuvre par Linkkit sont "contextuels" et non pas "absolus". Lorsque Google calcule son indicateur PageRank, il le définit chaque mois sur la base du positionnement de la page dans l'environnement des 3 milliards des autres pages de son index. Le PageRank reflète alors la "popularité absolue" d'une page sur le Web. Ainsi, une page telle que www.Yahoo.com aura-t-elle de grandes chances d'avoir un bon PageRank.

Au contraire, l'algorithme de Linkkit est "contextuel" et se définit sur la base d'une communauté de pages pour laquelle, pour une requête telle que "Linux", la page www.linux.org aura sans doute une légitimité supérieure à la page www.Yahoo.com. A la différence de la popularité absolue d'une page, Linkkit définit la pertinence comme étant **la légitimité de la page dans son contexte de référence**. La légitimité d'une page au sens de Linkkit ne saurait néanmoins se réduire à une simple "popularité" (calcul du nombre de liens entrants sur une page donnée) à l'intérieur des 500 pages retenues comme base de travail pour les algorithmes. En effet, Linkkit a vérifié que l'application d'une "popularité simple" à l'intérieur de l'ensemble initial des 500 pages ne conduisait pas à des résultats satisfaisants en termes de pertinence, en raison de la propension (légitime) de certains webmasters, qui souhaitent améliorer le classement de leurs sites par les moteurs en augmentant artificiellement le nombre de liens entre les divers sites qu'ils gèrent avec qui ils peuvent avoir une entente de réciprocité au niveau d'échange de liens hypertextes. Ces pratiques, appelées "**SPAM relationnel"**, doivent être détectées par les moteurs de recherche au risque de biaiser significativement leurs résultats.

Selon ses concepteurs, les algorithmes de Linkkit repèrent en dynamique le SPAM relationnel et l'atténuent voire l'éliminent purement et simplement.

Ce système n'est pas sans rappeler le brevet obtenu par Google, intitulé "Ranking search results by reranking the results based on local inter-connectivity" (N° US 6,526,440), et qui a été décrit dans la lettre R&R du mois dernier (mars 2003)... Christophe Vaucher, l'un des créateurs de Linkkit, qui travaille sur ce projet depuis début 2000, estime cependant que sa technologie n'est pas identique à celle décrite dans le brevet de Google : "le brevet de Google protége la combinaison linéaire des scores de ranking des modes global et local (respectivement "popularité globale" d'une page, indépendante de la requête (PageRank) et "popularité contextuelle"). Linkkit ne réutilise pas le score global pour son classement final, et ne tombe donc pas sous le coup du brevet de Google, offrant une porte de sortie à ses moteurs partenaires pour aller vers le traitement contextuel (là ou réside le plus grand potentiel d'amélioration de la pertinence à court et moyen terme, selon tous les moteurs) sans être obligé de négocier une licence avec Google."

La technologie Linkkit est le fruit et le développement d'une thèse menée par Eric Boutin, l'un des créateurs de la société avec Christophe Vaucher, sur ce thème. Un premier brevet a été déposé à l'INPI en novembre 2000 et plusieurs autres sont en cours de validation. Le projet a reçu plusieurs fois le soutien de l'Anvar (100 KF en 2000 et 80 000 euros en 2001/2002). La technologie est aujourd'hui opérationnelle et a nécessité l'équivalent de 10 années-homme de développement à ce jour.

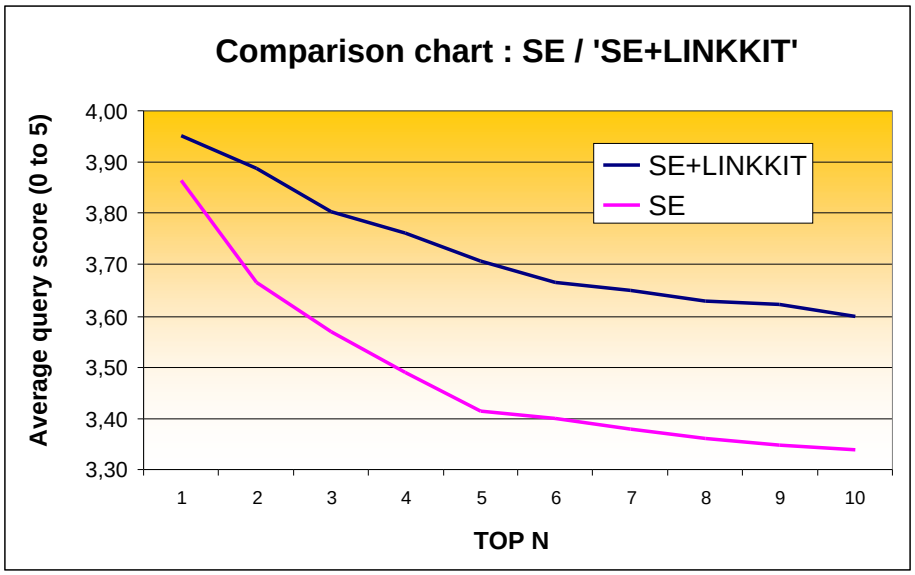
Une technologie consommatrice de ressources ?

On peut cependant se poser la question des ressources consommées par Linkkit, puisque son utilisation force le moteur à mettre en place une phase, une étape supplémentaire dans sa procédure de "ranking". Selon Christophe Vaucher, "alors que, par exemple, Google utilise un parc de plus de 15.000 PC pour son moteur, le surcoût matériel qu'il connaîtrait, s'il choisissait d'utiliser la technologie Linkkit, ne serait que de quelques dizaines de PC, de l'ordre de 1 pour mille environ".

La technologie a déjà été testée par un des grands acteurs mondiaux du monde de la recherche d'information (un accord de confidentialité interdit de révéler son nom).

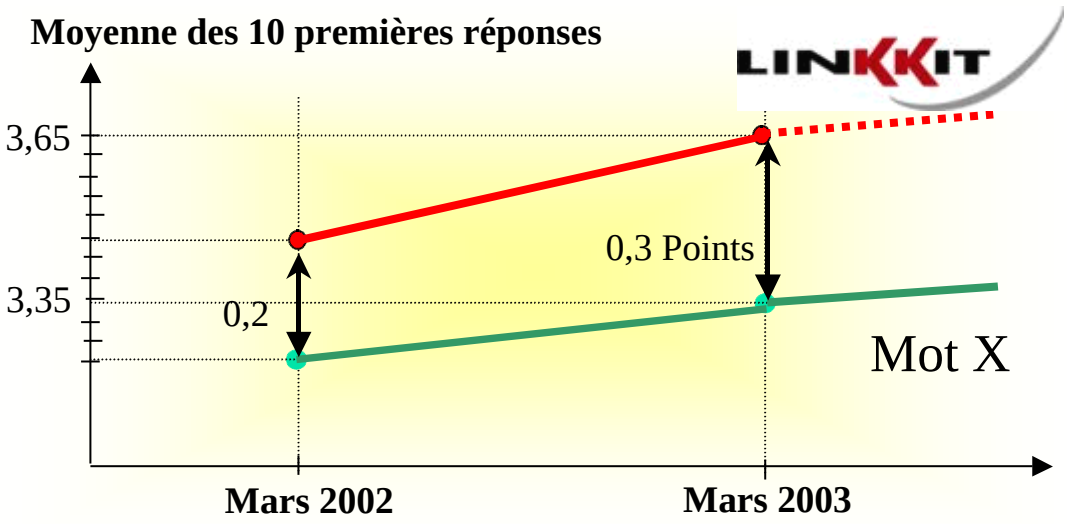
D'autre part, la société a mis en place un système de comparaison des résultats fournis par sa solution par rapport aux mêmes résultats fournis par les moteurs de recherche actuels, selon une méthodologie qui vous sera présentée dans les pages qui suivent.

Linkkit a notamment été testé en mars 2003 sur 350 requêtes (amenant la notation de 5 000 pages) par rapport au moteur majeur donc parlions quelques lignes plus haut et qui s'est prêté aux tests de la société. Ce moteur de recherche a donc "greffé" Linkkit sur ses algorithmes et a comparé la pertinence des résultats fournis. Voici les enseignements de ce comparatif sur les 10 premiers liens fournis ("SE" = Search Engine) :



En abscisse, on retrouve les valeurs moyennes comparatives des TOP N (N premiers résultats du moteur et du "moteur+Linkkit", pour N variant de 1 à 10.
Interprétons les valeurs du Top 10. La moyenne des 10 premières notes du moteur est d'environ 3,35 sur une échelle qui va de 0 à 5. Cette valeur correspond à la moyenne arithmétique simple des 10 premières pages renvoyées par le moteur, chaque page faisant l'objet d'une note de 0 à 5. On s'aperçoit que là ou le moteur obtient une moyenne de 3,35, Linkkit obtient une moyenne de 3,65, soit 3 dixièmes de mieux. Cette différence peut sembler mineure. Elle représente en fait, selon Linkkit, un écart très significatif, supérieur à l'écart mesuré aujourd'hui, selon la métrique employée, entre Google et ses principaux concurrents.

D'autre part, le test a été répété pendant un an, de mars 2002 à mars 2003, pour tester l'évolution de la notation au fil des mois :



Alors que le moteur X a progressé en pertinence de 0,2 points en un an, les résultats fournis par Linkkit ont, eux, progressé de 0,3 points sur la même période.

Selon les tests effectués, la technologie proposée par Linkkit semble donc apporter un réel plus au moteur de recherche qui l'utilise. Bien sûr, tous ces résultats demandent maintenant à être évalués par un organisme indépendant. La société serait d'ailleurs en train d'évaluer le coût d'un "benchmark" officiel par plusieurs laboratoires spécialisés pour obtenir un comparatif de ses performances totalement indépendant. Son but est de transférer sa technologie à un moteur de recherche qui désirerait augmenter significativement la pertinence de ses résultats, tout en utilisant des algorithmes qui ne tomberaient pas sous le joug de brevets déjà déposés par les moteurs concurrents.

Arrivera-t-elle à ses fins ? c'est certainement ce que nous verrons d'ici à la fin de l'année...

Comment évaluer la pertinence d'un moteur ?

Pour complémenter la présentation de sa technologie, Linkkit nous a fourni sa méthodologie de comparaison de la pertinence d'un outil de recherche. Elle nous a semblé intéressante, aussi, il nous a paru pertinent de vous la présenter, avec leur accord. **Laissons la parole à Linkkit, qui présente donc sa méthodologie dans les pages suivantes :**

Remarques préliminaires

L'évaluation de la pertinence des pages retournées par un outil de recherche (annuaire, moteur, meta-moteur) est sans doute l'aspect le plus subjectif de la technologie des moteurs de recherche, tout en représentant un passage obligé pour faire évoluer une technologie vers de "meilleurs" résultats. Or "meilleurs" résultats ne peut vouloir dire que "plus grande satisfaction" des internautes devant les résultats, puisque ce sont les internautes et eux seuls, qui "font" et "défont" les moteurs, en venant "gonfler" ou non la masse des recherches sur tel ou tel site, le trafic sur un site donné étant le principal moyen de rentabilisation du moteur.

Bien qu'il existe un grand nombre de critères possibles pour l'évaluation des résultats des moteurs, nous avons choisi de nous écarter volontairement (un peu et parfois beaucoup) des systèmes de notation tels que ceux d'eTesting Labs (ayant testé un panel de moteur à la demande de Google en Sept 2000), ZD Labs (ayant testé un panel de moteurs à la demande d'Altavista en Mai 2000), qui ne se positionnent pas vraiment en matière de satisfaction de l'internaute. En outre, les critères choisis dans ces études, financées par un des moteurs évalués, sont nécessairement un peu biaisés à l'avantage du commanditaire, ce qui transparaît dans les conclusions de ces études.

Nous estimons que le critère de "pertinence" des résultats est le critère majeur de satisfaction de la très grande majorité des internautes, très loin devant tous les autres, et nous avons en

conséquence choisi de centrer notre démarche comparative uniquement sur la "pertinence" des résultats au sens de l'internaute.

En conséquence, dans les tests que nous avons conduits, nous avons choisi :
- De n'évaluer que la différence de pertinence entre divers moteurs et notre technologie.
- De ne retenir aucun autre critère pouvant venir artificiellement biaiser les résultats : la présentation, la rapidité de réponse, les fonctions spéciales, etc. n'entrent pas en ligne de compte dans nos tests.
- D'avoir une interprétation de la "pertinence" uniquement au sens de la satisfaction des internautes.

Profil des internautes visés par LINKKIT

On peut distinguer 2 principales catégories d'internautes :
- L'internaute "grand public", qui va rechercher des informations dans le domaine de la vie courante. Il ne saisit en moyenne que 1 à 2 mots clés par requête. Ses principaux centres d'intérêts sont :
    * Le multimédia (photos, musiques, films, acteurs, hommes/femmes publics, etc.).
    * Sports.
    * Finance.
    * Infos générales ET achat en ligne sur les sujets suivants : jardinage, voiture, biens de consommation, culture (livres, concerts, etc.), loisirs, voyages, etc.
    * Politique, informations.
    * Sexe.
- L'internaute "professionnel", qui va rechercher des informations dans le domaine professionnel. Il saisit en moyenne plus de 2 mots clés par requête. Ses principaux centres d'intérêts vont se situer dans les domaines suivants :
    * Science et technologie.
    * Tous secteurs professionnels.

On pourrait également proposer une segmentation un peu différente :
- L'internaute "expérimenté", qui est un habitué des techniques informatiques et qui va mieux anticiper ce qu'est capable de lui renvoyer une base de données. Il saisira en moyenne plus de 2 mots clés par requête, de manière à la rendre très précise, et procédera plus facilement par itérations successives. Le moteur lui renverra alors en moyenne beaucoup moins de résultats, mais ceux-ci seront dans l'ensemble plus pertinents. Nous estimons cette population d'internautes à moins de 10% de la population totale accédant aux moteurs. Cette proportion devrait encore diminuer avec la progression de l'accès "grand public " aux moteurs.
- L'internaute "M. Toulemonde", qui n'est pas un habitué des techniques informatiques. Il saisira en moyenne moins de 2 mots clés par requête. Le moteur lui renverra alors beaucoup de résultats, pour lesquels il ne sera pas évident pour le moteur d'effectuer le tri du TOP 10 le plus pertinent.

Le cœur de cible Linkkit correspond à cette dernière catégorie, qui se confond essentiellement avec les internautes grand public, ainsi que les utilisateurs professionnels non spécialisés ou formés en informatique/bureautique. Pour tous ces utilisateurs, il est souhaitable de disposer d'une technologie donnant des résultats satisfaisants, même pour des requêtes composées de 1 ou 2 mots clés. En effet, le nombre de mots clés saisis en moyenne ne dépasse pas 2, alors que les réponses renvoyées par la base de données du moteur peuvent être très nombreuses (plusieurs milliers à plusieurs dizaines de milliers de pages par requête).
C'est sur ces requêtes que va se concentrer l'action des annonceurs, principale source de revenu à terme des moteurs. D'ou la cible Linkkit.

Profil requêtes des internautes – représentativité de l'échantillon de test

Construction des requêtes

Les internautes effectuant une recherche sur le web saisissent à 95% des mots clés plutôt que des phrases. Par ailleurs, pour les 5% restants, les phrases sont interprétées dans leur grande majorité par les moteurs par des mots clés, après suppression des "mots vides".
Nous ne considérerons dans notre étude principalement que des requêtes saisies à partir d'un ou plusieurs mots clés.

Langue des requêtes

Nous nous devons, dans le cadre d'une évaluation se voulant représentative des requêtes sur l'ensemble des moteurs, de respecter la proportion des langues présentes dans les requêtes, soit près de 75 à 80% de requêtes anglophones. En première approximation, et pour des raisons pratiques évidentes, nous avons considéré que le Web francophone était représentatif, pour l'essentiel, des pages dans d'autres langues que l'anglais, et avons essentiellement évalué des requêtes francophones, pour la part dans nos tests des 20 à 25% de requêtes qui ne sont pas de langue anglophone.

Interprétation de la notion de "pertinence" selon Linkkit

Nous pensons que la pertinence d'une page, pour un internaute, correspond à la distance qui le sépare de l'information de référence pour la requête

Qu'attend en effet l'internaute "M. Toulemonde" des résultats de sa requête ?

- Idéalement, de trouver toute l'information recherchée dans une seule et même page.
- Il acceptera néanmoins d'avoir à consulter 2 ou 3 autres pages, si elles sont accessibles :
    * Soit dans les positions immédiatement suivantes du moteur.
    * Soit au travers d'un lien bien visible et accessible sur la page en cours, même si la page derrière le lien ne figure pas dans les premiers résultats du moteur (notion de "bon hub").

Ensuite, pour les pages accédées, les critères suivant sont à retenir, dans une ordre décroissant d'importance :
* **Centralité** de la page par rapport au sujet de la requête : une page ne traitant que partiellement du sujet de la requête, dans un paragraphe, par exemple, n'est que partiellement intéressante, sauf si elle possède un lien bien visible vers une page très intéressante.
* **Exhaustivité** du contenu de la page par rapport au sujet de la requête : une page comprenant l'information ou les liens vers tous les sous-domaines de la requête (ex pour la requête "Nicole Kidman" : bibliographie, listes et descriptions de ses films, photos, fan club, liens pour acheter des K7 vidéos ou télécharger des films, etc.) aura une meilleure note qu'une page ne couvrant qu'un des sous-domaines (ex : uniquement les photos pour Nicole Kidman). Si la requête est "nicole kidman pictures", à l'inverse, les sites très exhaustifs sur Nicole Kidman n'auront pas la note maximale.
* **Qualité** du contenu de la page : les informations émanent-elles d'une source fiable, autorisée ?
* **Fraîcheur** de l'information contenue dans la page : date de dernière mise à jour. Ce paramètre devra compter un peu plus si l'information sur le sujet de la requête est susceptible d'évoluer rapidement avec le temps. Dans le cas contraire, il sera minoré.
* **Présentation** : présentation générale, facilité d'accès à l'information, esthétique : Une page riche en contenu, mais trop dense sera bien moins notée qu'une page bien claire et légère, renvoyant à l'information voulue par des liens bien visibles.

Quantification de la notion de "pertinence" selon LINKKIT

Pour des raisons d'efficacité de notation, nous raisonnons essentiellement sur les 2 principaux critères, qui sont :
- la centralité (sur 2 points) : est-ce que la page est centrée ou non sur le sujet ?
- l'exhaustivité (sur 3 points)

Nous obtenons alors une note sur 5.

Les autres critères (Qualité du contenu de la page, Fraîcheur de l'information contenue dans la page, Présentation) feront alors l'objet d'un malus éventuel, de 0,2 à 1 point pour chaque critère, et pour chaque page.

Remarques importantes :

Ce système de notation n'est pas rigide, et reste essentiellement dépendant de la requête : en fonction du sujet de la requête, chaque critère pourra se voir modulé de + ou – 0,5 points, voire plus.

En outre, une page permettant d'accéder à une deuxième page par un lien bien en évidence va obtenir la note qu'aurait obtenue la 2ème page si elle avait été renvoyée en direct, moins 0,2 à 1 points, suivant la qualité et la visibilité du lien qui permet d'y accéder.

Toutes les pages sont notées en aveugle.

Pour une requête donnée, les pages issues des divers TOP 10 moteurs et du TOP 10 de Linkkit sont "mélangées" et triées par ordre alphabétique, par exemple sur un tableau Excel, avant d'être notée : l'examinateur ne connaît pas la provenance des pages lorsqu'il leur donne une note, ce qui garantit l'indépendance de la notation.

Gestion des erreurs 40X

Le taux moyen d'erreurs 40X (404, etc.) pour les pages renvoyées par les moteurs varie en fonction du moteur autour de la valeur moyenne de 7 à 9%, selon nos propres statistiques.
Notre "process" élimine actuellement les pages 40X de la base de travail de nos algorithmes, ce qui ne sera pas le cas dans la configuration réelle d'utilisation de notre technologie, couplée avec la base de données du moteur client. Pour ne pas désavantager le moteur, nous avons choisi dans cette phase de test de ne pas prendre en compte les erreurs 40X ni du côté du moteur, ni de notre côté : ces pages sont purement supprimées des résultats, et devraient faire l'objet selon nous d'un critère séparé, lié à la fraîcheur de rafraîchissement de la base de données du moteur.

Google et les autres

Selon la grande majorité des analystes, la progression rapide de Google jusqu'au 55% de parts de marché qu'il occupe en avril 2003, notamment en France, est principalement dûe à l'écart de pertinence de ses résultats avec ceux de ses concurrents.
Nous avons mesuré cet écart, selon notre système de notation, sur une première base significative de 192 requêtes. Nous sommes parvenus aux résultats suivants : Google devance son suivant immédiat, le moteur C de 0.2 pts , les moteurs A et D de 0.33 pts, selon la métrique Linkkit.
Ceci est à comparer aux 0.3 points en moyenne que fait gagner la technologie Linkkit à chaque moteur.

Etalon de pertinence suivant notre système de notation

0.3 point d'écart entre deux notes sur 5 ne représente pas une variation importante en soi, a priori. Il faut pourtant croire que cette variation est très significative, puisque 0.3 point d'écart entre 2 moteurs, suivant notre système de notation, est précisément supérieur à l'écart moyen (le gap) de pertinence entre Google et ses suivants, que nous mesurons actuellement à une valeur de 0.25 dans notre métrique.

La notion de valeur moyenne de l'écart d'un point, apparemment anodin, masque en fait une autre réalité, celle de la régularité des résultats des moteurs.

En effet, pour plus de la moitié des requêtes, disons les 2/3, les moteurs en général pourront produire des résultats plutôt semblables : ils vont donc obtenir des notes très proches des autres moteurs, même des meilleurs. 0.3 d'écart en moyenne entre 2 moteurs signifie alors dans ce contexte, que pour une requête sur 3, l'écart va être de l'ordre de 1 points, ce qui est tout simplement considérable : en lieu et place de 4 pages de notes 4,5/5 sur les 10 pages figurent 4 pages de notes 2,5/5 (4 pages d'information très moyenne remplacent 4 pages de référence).

Les coordonnées de la société Linkkit :

LINKKIT
ZA des Playes – Av de Rome – Valparc 2 – 83500 La Seyne
Tél : 04 94 10 20 50
Contact : Christophe Vaucher (contact@linkkit.com)
Site Web : http://www.linkkit.com