Méthodes de tri des résultats des moteurs de recherche

Jean-Pierre Lardy

Méthodes de tri des résultats des moteurs de recherche

LARDY Jean-Pierre
UNIVERSITE CLAUDE BERNARD - LYON I
Email: **lardy@.univ-lyon1.fr**
http://urfist.univ-lyon1.fr
http://www.adbs.fr/adbs/viepro/sinfoint/lardy/risi.htm

mots-clé : SEARCH ENGINE RANKING, LINK POPULARITY, RELEVANCE RANKING

La publication sur le web continue à croître sans relâche[1] et le milliard de pages HTML statiques vient d'être dépassé selon les estimations d'une étude de Inktomi[2]. Si les annuaires, produits manuellement, présentent l'avantage de classer les sites par thèmes, les moteurs, produits automatiquement, sont les outils les plus utiles pour fouiller le web.
La masse d'informations disponibles conduit malheureusement à des résultats pléthoriques la plupart du temps, ce qui déroute les utilisateurs. Aussi les moteurs de recherche ont développé des méthodes de tri automatiques des résultats[3]. Cela leur permet aussi de se distinguer les uns des autres. Dans la pratique aucune méthode de tri n'est parfaite mais cette variété offre à l'utilisateur la possibilité de traquer l'information de différentes manières et augmente donc ses chances d'améliorer ses recherches.
Le but du classement est d'afficher dans les 10 à 20 premières réponses les documents répondant le mieux à la question. Si on ne trouve pas ce que l'on cherche dans les toutes premières pages de résultats, il faut reformuler la question. Pour cela il faut comprendre les mécanismes sous-jacents pour en tirer vraiment profit.

On peut considérer trois grandes méthodes de tri.

Le tri par pertinence

Cette méthode repose sur des travaux de recherche déjà anciens de Robertson et Sparckjones[4], mis en pratique dans le logiciel d'indexation WAIS à la fin des années 80.
Les résultats d'une requête sont affichés selon un ordre déterminé par le calcul d'un score pour chaque réponse. La pertinence est basée sur les cinq facteurs suivants appliqués aux termes de la question :

1. **Le poids d'un mot dans un document** est déterminé par sa place dans le document : il est maximum pour le titre et le début du texte; à l'intérieur il est plus important si le mot est en majuscule.
2. **La densité** est basée sur la fréquence d'occurrence dans un document par rapport à la taille du document. Si deux documents contiennent le même nombre d'occurrences, le document le plus petit sera favorisé.
3. **Le poids d'un mot dans la base** est basé sur la fréquence d'occurrence pour toute la base de données. Les mots peu fréquents dans le corpus sont favorisés. Les mots vides sont soit éliminés, soit sous-évalués.
4. **La correspondance d'expression** est basée sur la similarité entre l'expression de la question et l'expression correspondante dans un document. Un document contenant une expression identique à celle de la question reçoit le poids le plus élevé.
5. **La relation de proximité** est basée sur la proximité des termes de la question entre eux dans le document. Les termes proches sont favorisés.

Cette technique a montré son efficacité dans le cadre des bases de données WAIS assez homogènes et peu volumineuses.
Elle a été reprise dans les moteurs de recherche apparus à partir de 1994 et basés sur les techniques d'exploration du web par les robots[5]. Cependant l'algorithme exact n'est jamais connu[6] car il est considéré comme secret industriel et quelquefois protégé par un brevet (cas d'Excite).
Les documents HTML peuvent contenir dans l'entête des informations concernant le contenu du document. Ces méta-données correspondent aux balises TITLE, META keywords et META description. Une étude a montré qu'elles étaient malheureusement peu utilisées. Certains moteurs de recherche en tiennent compte dans leur calcul.
Cependant le tri par pertinence présente l'inconvénient d'être facile à détourner par des auteurs désireux de placer leurs pages en tête de liste : pour cela il suffit de répéter les mots importants soit dans l'entête, soit dans le texte en utilisant des techniques de spamming (écrire le texte en blanc sur fond blanc par exemple) pour modifier à son avantage le classement. Les moteurs ont réagi en détectant ses techniques.

Cette méthode est utilisée par AltaVista, Ecila, Excite, FAST, HoBot, Inktomi, Lokace, Voila. Le résultat dépend beaucoup de la question et l'on choisira, chaque fois que cela est possible, des termes précis et non ambigus.

Tri par popularité

Les limites du tri par pertinence ont conduit à rechercher d'autres méthodes reposant sur des principes tout à fait différents et indépendants du contenu des documents. Connues sous le nom de tri par popularité, on distingue :

La méthode basée sur la co-citation

Lancé en 1998 par deux étudiants de l'Université de Stanford, Google classe les documents grâce à la combinaison de plusieurs facteurs dont le principal PageRank[7]. Ce dernier utilise le nombre de liens pointant sur les pages. L'article de Page et Sergey[8] en donne une description.
Plusieurs moteurs de recherche offrent cette fonctionnalité. Avec AltaVista il faut entrer :

+link:www.site.com –host:www.site.com

Cela permet à n'importe quel auteur de pages de découvrir les liens pointant sur son œuvre.

Google évalue l'importance d'une page par les liens qu'elle reçoit mais analyse en plus la page qui contient le lien. Les liens des pages "importantes" pèsent plus lourdement et aident à découvrir d'autres pages "importantes". Ainsi le tri est indépendant du contenu et évite les dérapages de la méthode précédente, le choix des liens étant laissé à la libre décision des millions d'auteurs de pages HTML. Il faut cependant noter que cette technique défavorise les pages récentes et donc inconnues.

La méthode basée sur la mesure d'audience

La société DirectHit a été fondée en avril 98 et propose de trier les pages en fonction du nombre de visites qu'elles reçoivent. DirectHit analyse le comportement d'un internaute dans l'utilisation d'un moteur de recherche : sur la page d'accueil, il saisit un ou plusieurs mots de recherche dans un formulaire, consulte la page de résultats classés par ordre de pertinence, choisit l'un d'entre eux, va sur le site correspondant pour le consulter. Si la page ne lui

convient pas, il revient sur la page de résultats du moteur, choisit un autre lien, etc. jusqu'à ce qu'il ait trouvé un document pertinent. DirectHit enregistre ce comportement pour tenter de trouver les pages les plus "populaires" sur un moteur de recherche et ainsi améliorer leur classement. Il fonctionne, en règle générale, en tâche de fond sur un moteur existant. A chaque consultation d'un utilisateur, DirectHit note sur quel lien celui-ci a cliqué et quel était le rang de ce lien. Il mesure le temps passer sur une page avant que l'utilisateur ne revienne aux résultats. S'il ne revient pas, il en "déduit" que le site proposé était pertinent. Il sera alors mieux classé dans les résultats suivants, lors d'une interrogation sur le même mot-clé. Ainsi les interrogations et la façon d'interroger et de naviguer des internautes vont enrichir la base données de DirectHit.
Cette méthode comme la précédente pénalise les pages récentes mais évite le spamming.

DirectHit peut être interrogé directement sur son site mais alimente aussi les résultats de nombreux outils de recherche comme HotBot, LookSmart et des sites Web comme celui de ZDNet.

L'annuaire Snap utilise une technique appelée "Global brain", classant les sites selon leur popularité auprès des internautes, afin de les inclure dans ses algorithmes de pertinence.

Tri par calcul dynamique de catégories

NorthernLight, lancé en Août 1997, propose le classement des documents trouvés dans des dossiers (clustering) constitués automatiquement en fonction des réponses. Un dossier peut lui-même être constitué de sous-dossiers. Quatre types existent :
Subject (e.g., hypertension, baseball, camping, expert systems, desserts)
Type (e.g., press releases, product reviews, resumes, recipes)
Source (e.g. commercial Web sites, personal pages, magazines, encyclopedias, databases)
Language (e.g., English, German, French, Spanish)

Dans chaque dossier final, les réponses sont triées par pertinence.

Autres recherche

Chez IBM John Kleinberger[9] a développé l'algorithme HITS (Hypertext-Induced Topic Search) pour identifier les sources d'autorité parmi les pages HTML en fonction du nombre de liens qu'elles possèdent. Le moteur de recherche CLEVER[10] (Clientside Eigenvector Enhanced Retrieval) utilise cette technique pour caractériser les pages.

Conclusion

On constate donc que les méthodes de classement des résultants se sont diversifiées ce qui complique la tache des auteurs[11] de pages mais augmente les chances de trouver de bons résultats dans la jungle que représente actuellement le web.
En fait le mieux serait sans doute de combiner ces différentes méthodes : c'est ce que vient d'annoncer NorthernLight[12].

Moteurs cités
AltaVista www.altavista.fr, DirectHit www.directhit.com, Ecila www.ecila.com, Excite www.excite.com, FAST www.alltheweb.com, Google www.google.com, HoBot www.hotbot.com, Inktomi www.inktomi.com, Lokace www.lokace.com, NorthernLight www.northernlight.com, Snap www.snap.com, Voila www.voila.fr

---

Bibliographie

[1] - Lawrence S., Giles L. “Accessibility and Distribution of Information on the Web”, Nature 400(6740): 107-109, 1999
http://wwwmetrics.com/ [dernière visite le 29/03/2000]
http://www.notess.com/search/stats/nature99.shtml [dernière visite le 29/03/2000] (Summary and analysis by Greg R. Notess)

[2] - “Web Surpasses One Billion Documents“, 18 Janvier 2000
http://www.inktomi.com/new/press/billion.html [dernière visite le 29/03/2000]

[3] - Courtois M.P., Berry M.W. “Results Ranking in Web Search Engines”, ONLINE, mai 1999
http://www.onlineinc.com/onlinemag/OL1999/courtois5.html

[4] - Robertson S. E., Sparckjones K. “Relevance weighting of search terms”, Journal of the American society for Information Science, 27 (3): 129-146, 1976

[5] - Koster M. “The Web robots FAQ“
http://info.webcrawler.com/mak/projects/robots/faq.html [visité le 29/03/2000]

[6] - Hassani Md, Lardy J. P. “Techniques de tri des moteurs de recherche“
DEA SIC Université LYON I, 1998

[7] - Page L. “PageRank: Bringing Order to the Web”
http://www-pcd.stanford.edu/~page/papers/pagerank/ [visité le 29/03/2000]

[8] - Brin S., Page L. “The Anatomy of a Large-Scale Hypertextual Web Search Engine”
http://www7.scu.edu.au/programme/fullpapers/1921/com1921.htm [visité le 29/03/2000]

[9] - Kleinberg J. “Authoritative sources in a hyperlinked environment” Proc. 9th ACM-SIAM Symposium on Discrete Algorithms, 1998

[10] - The CLEVER project
http://www.almaden.ibm.com/cs/k53/clever.html [visité le 29/03/2000]

[11] - Search Engines - Submission Tips, Help and Use.
http://www.sofer.com/research/engines.html [visité le 29/03/2000]

[12] - Feldman S. “Northern Light Adds Link Popularity to Relevance Ranking Factors”, 1999
http://www.infotoday.com/newsbreaks/nb1101-3.htm