

Glossaire du chercheur internet

Projet glossaire collaboratif - Document compilé par Françoise QUAIRE pour l'ADBS – 2001

Dernière mise à jour : dimanche 20 mai 2001

AGENT INTELLIGENT

Logiciel visant à faciliter la recherche et la gestion de l'information sur l'Internet.

ALGORITHME DE PERTINENCE

Relevancy algorithm : Méthode utilisée par un moteur de recherche ou un répertoire pour relier les mots-clés d'une requête avec le contenu de chaque page, de telle sorte que les pages Web trouvées correspondent bien au sujet de la requête. Chaque outil de recherche est susceptible d'utiliser un algorithme différent et de le changer ou de l'améliorer. Voir aussi : critères de tri.

ANNUAIRE

Voir Répertoire

APPLET JAVA

Programme en langage Java, téléchargeable et exécutable sur n'importe quel type de processeur. Les principaux navigateurs sont capables de lire et d'exécuter les applets Java. Il est possible que la présence de ce programme stoppe l'indexation d'une page par le robot d'un moteur de recherche.

ARAIGNEE (traduction littérale de Spider)

C'est la partie du moteur de recherche qui "surfe" sur le net, enregistre les URLs, classe les mots-clés et le texte de chaque page qu'il trouve.

ASPIRATEUR DE SITE

Outil permettant de copier un site Web à distance pour le relire ensuite en local, depuis son propre disque dur

.BRUIT

Réponse non pertinente fournie lors d'une recherche d'information.

CADRE

Voir Frame

CGI

Acronyme de Common Gateway Interface, logiciel qui facilite la communication entre un serveur Web et des programmes fonctionnant hors de ce serveur ; par exemple, des programmes qui traitent des formulaires interactifs ou qui recherchent des informations dans des bases de données sur le serveur, suite à la requête d'un utilisateur.

CONTENU DYNAMIQUE

Il s'agit de pages Web avec des informations qui changent ou sont changées automatiquement en fonction d'une base de données ou d'éléments provenant de l'utilisateur. (ex. suffixe .asp, .cfm, .cgi ou .shtml dans l'URL). Voir aussi Page dynamique

CRITERE DE TRI

Façon automatique de sélectionner les résultats retournés par le moteur de recherche, afin de présenter en début de liste ceux qui correspondent le mieux à la requête. On distingue généralement le tri par pertinence du tri par popularité. Voir aussi : algorithme de pertinence.

CRYPTAGE

Moyen de rendre secrète la communication informatique grâce à des logiciels d'encodage de données. Seul le possesseur de la clé de décodage peut interpréter le message. Aussi appelé "chiffrement".

DENSITE DES MOTS CLES

Une des propriétés qui permet d'indiquer l'importance de certains mots dans le texte d'une page Web. Certains outils de recherche utilisent cette propriété pour le tri. La formule de calcul = nombre d'occurrences du terme demandé / nombre de termes de la page en question, une fois éliminés les mots vides.

DIRECTORY

Voir Répertoire

EN-TETE heading tags

Ce sont les commandes qui se trouvent en tête des pages html. Certains moteurs de recherche donnent plus d'importance et de poids au texte qui s'y trouve.

EQUATION DE RECHERCHE

Formulation d'une question sous forme mots clés reliés par des termes logiques (ou opérateurs). Voir aussi opérateurs booléens.

FRAME

Synonymes : trame, cadre . Il s'agit d'une technique de programmation en html pour présenter deux ou plusieurs documents séparés dans une fenêtre du navigateur, les documents agissant l'un sur l'autre à travers différentes sous-fenêtres. Les frames posent souvent de gros problème d'indexation aux moteurs de recherche.

JAVA

Langage multiplateforme, créé par Sun, capable de s'exécuter à l'intérieur d'une page Web. Souvent utilisé pour créer des applets.

JAVASCRIPT

Langage simple interprété qui permet d'exécuter des petites tâches au sein des pages html.

KEYWORD

Voir Mot-clé

LIEN A L'ARRIVEE

Un lien hypertexte vers une page particulière venant de quelque part et apportant du trafic à cette page. Les liens à l'arrivée sont souvent un instrument de mesure pour connaître la popularité d'une page.

LIEN MORT

Un lien qui ne mène plus à une page ou à un site, soit parce que le serveur est en panne, soit parce que la page a été déplacée ou bien n'existe plus. La plupart des outils de recherche ont des techniques pour oter de telles pages de leur liste automatiquement. Mais l'Internet continuant à croître quotidiennement, il leur devient de plus en plus difficile de contrôler régulièrement toutes ces pages.

META-MOTEUR ou META-CHERCHEUR

Meta search engine, parallel search engines

Un outil qui, pour une même requête, interroge en parallèle (simultanément) plusieurs moteurs de recherche et/ou répertoires et compile les résultats avant de les présenter. Parfois qualifiés d'agents semi-intelligents, de métachercheurs...etc. Différentes orthographe et plusieurs définitions plus ou moins exactes se rencontrent : métamoteurs, méta-moteurs, metamoteurs, meta-moteurs, méta-chercheurs, multimoteurs,... etc. On distingue généralement 3 générations de méta-moteurs :

Première génération : ils rassemblent sur une même interface un certain nombre de moteurs et outils de recherche. Ils ne permettent pas une interrogation simultanée mais offre néanmoins un gain de temps.

Seconde génération : ces méta-moteurs interrogent simultanément plusieurs outils de recherche. Mais ils affichent les résultats moteurs par moteurs sans éliminer les doublons ni procéder à l'analyse de pertinence. L'utilisateur doit évaluer les résultats en se connectant site par site.

Troisième génération : ces méta-moteurs, les plus sophistiqués, sélectionnent les sites dans différents moteurs, dédoublonnent les réponses et affichent les résultats selon des critères de pertinence ou par type de document.

META TAG

Construction placée dans l'en-tête html des pages Web, fournissant des informations qui ne sont pas visibles par les navigateurs. Les plus courants des meta-tags sont KEYWORDS et DESCRIPTION. MOT CLE ou key-word : Mot ou groupe de mot, éventuellement dans une forme lexicographique normalisée, choisi dans le titre ou le texte d'un document, caractérisé par le contenu et permettant la recherche de ce document.

MOTEUR DE RECHERCHE

Search Engine et chercheur au Québec.

Programme qui indexe le contenu de différentes ressources Internet, et plus particulièrement de sites Web, et qui permet à l'internaute de rechercher de l'information selon différents paramètres, en se servant de mots clés, et d'avoir accès à l'information ainsi trouvée.

Mode de fonctionnement : des robots logiciels (appelés crawlers ou spiders) scrutent le Web, vont de page en page et sauvegardent alors le contenu texte des pages rencontrées, constituant ainsi un index, c'est-à-dire une collection plus ou moins grande de pages Web. Le robot logiciel repasse selon des délais plus ou moins fréquents sur les pages qu'il a indexées au préalable, pour en sauvegarder une version plus récente. On dit alors qu'il rafraîchit sa base (ou son index).

Lorsque l'internaute saisit un mot clé dans le formulaire proposé, le moteur va en rechercher les occurrences dans son index, i.e. dans le contenu texte des pages Web sauvegardées au préalable. Une fois le "lot" de pages contenant le terme demandé identifié, le moteur classe les pages par ordre de pertinence, selon un ordre et un algorithme spécifique. (voir algorithme de pertinence et aussi critère de tri).

A tort, le terme moteur de recherche est souvent utilisé tant pour un répertoire que pour un vrai moteur.

Voir aussi : outil de recherche.

MOTEUR THEMATIQUE

Synonyme : moteur spécialisé : Il procède par catégorisation automatique de pages, généralement à partir de catégories prédéfinies et de mots-clés préétablis.

OPERATEUR BOOLEEN

Pour effectuer une recherche par mots clés, on couple souvent une suite de mots grâce à des opérateurs booléens. Venant du nom de George Boole (mathématicien britannique) ces opérateurs permettent d'élargir ou de restreindre la recherche en imposant certains mots et en en excluant d'autres. Il existe plusieurs opérateurs booléens :

le ET [AND] (la recherche se fait obligatoirement sur les deux mots saisis);

le OU [OR] (la recherche se fait soit sur l'un des mots saisis, soit sur l'autre, soit sur les deux);

le SAUF [WITHOUT] (qui exclue le terme en question de la recherche).

OUTIL DE RECHERCHE

Terme générique pour tout service de recherche d'information sur le Web, combinant bien souvent désormais les procédés d'un répertoire et ceux d'un moteur de recherche, plus parfois de moteurs spécialisés.

P TO P

Point to Point, Peer to Peer ou encore People to People. Type de connexion qui met en communication deux interlocuteurs, et seulement deux. Des applications comme Napster ou Gnutella emploient cette technique et permettent à des utilisateurs de permuter, d'échanger des fichiers directement entre eux, sans passer par un serveur d'hébergement.

PAGE DYNAMIQUE

Page HTML dont le contenu n'est pas situé dans un fichier enregistré sur le serveur mais générées "à la volée" par une application informatique à partir d'un modèle de document HTML en accédant à des informations situées dans une (ou des) base(s) de données. Les techniques utilisées sont variables, CGI langage de script, API propriétaires permettant de créer un lien entre la base de données et le serveur HTTP.

PAGE STATIQUE

Page HTML dont le contenu est situé dans un fichier figé, enregistré sur le serveur Web.

POPULARITE

Synonyme : notoriété. Mesure le nombre et la qualité des liens pointant vers une page particulière. Plusieurs moteurs de recherche utilisent de plus en plus ce procédé dans le processus de tri.

PORTAIL

Terme générique pour désigner un site qui sert de point d'entrée sur l'Internet pour un nombre important d'utilisateurs. Un site portail offre une multitude de services différents depuis la page d'accueil.

POSITIONNEMENT ou Ranking

Processus de classement des sites, des pages Web dans un moteur de recherche ou un répertoire afin que les sites les plus pertinents apparaissent en premier sur la page résultat lors d'une requête.

REGROUPEMENT ou Cluster

Affichage d'une seule adresse pour chaque site Web sur la page des résultats d'un outil de recherche. Cette méthode permet d'éviter qu'un petit nombre de sites occupe toutes les premières positions de résultats et en facilite la lecture pour l'utilisateur.

REPertoire Synonymes : Catalogue, index thématique, liste thématique. Synonyme communément utilisé à tort : annuaire (un répertoire n'est pas annuel !).

Liste de sites Web classés dans des catégories thématiques. Le classement est effectué par des personnes physiques en fonction de diverses informations soit fournies au moment du référencement, soit déduites après la visite du site par les indexeurs du répertoire. Les catégories sont créées et gérées humainement. L'unité de classement est le site. La valeur ajoutée d'un répertoire tient en la qualité de son système de classification et à l'insertion éventuelle de commentaires et descriptions enrichies pour chaque site référencé. La plupart des répertoires proposent une recherche par mot-clé sur le titre des sites, les mots de la description, et les catégories concernées. On peut distinguer plusieurs types de répertoires.

REPERTOIRE GENERALISTE

Répertoire ayant vocation à indexer tous les sites et qui n'effectuent une censure que sur la base de principes prédéfinis (par exemple, des sites manifestement illégaux ou dont le référencement cherche à induire l'internaute en erreur). Exemple : Yahoo, Nomade

REPERTOIRE SPECIALISE

Répertoire dont les sites répertoriés relèvent tous d'un domaine ou d'un secteur particulier (le vin, le tourisme, le sport, l'agriculture, etc.). Un répertoire spécialisé peut, par exemple, ne prendre en compte que les entreprises d'un secteur, ou les produits d'un domaine. Ne pas confondre avec un moteur thématique. Exemple : Qualisteam.com dans le domaine bancaire et financier.

REPERTOIRE SELECTIF

Répertoire dont les gestionnaires privilégient les sites de meilleure qualité et excluent les sites qu'ils n'estiment pas suffisamment intéressants. Exemple : bonWeb.com

REPERTOIRE CONTRIBUTIF

Synonymes : répertoire ouvert, (open directory). Répertoire dont l'enrichissement est effectué par différentes équipes d'internautes. Ces répertoires confient la responsabilité d'une ou plusieurs catégories soit à des internautes experts reconnus dans leur domaine et rémunérés pour leur prestation (exemple : About.com), soit à des internautes bénévoles dont la compétence dans le domaine couvert par cette catégorie a été vérifiée. Ces internautes reçoivent alors les demandes de référencement de leur catégorie, décident ou non de référencer les sites et, le cas échéant, rédigent eux-même la description du site (exemple : dmoz - Open Directory Project).

REPERTOIRES D'OUTILS DE RECHERCHE

Synonymes : listes de listes, répertoires de répertoires - Répertoires spécialisés dans le référencement de répertoires et d'outils de recherche (moteurs de recherche, méta-moteurs, etc.). Exemple : 7alpha, Beaucoup

REQUETE

Synonymes : query, terme recherché. Mot, expression ou groupe de mots employés pour interroger un outil de recherche afin de localiser des pages sur le sujet recherché.

ROBOT

Programmes de navigation qui suivent les liens hypertextes des pages Web mais qui ne sont pas directement sous contrôle humain. Exemples : les spiders ou araignées des moteurs de recherche.

ROBOTS.TXT

Fichier texte déposé dans le répertoire principal d'un site Web pour interdire l'accès aux robots de certaines pages ou sous-répertoires du site.

SILENCE

Désigne l'ensemble des documents pertinents non retrouvés lors d'une recherche.

SITE FEDERATEUR

Synonymes : site de référence, portail spécialisé Site spécialisé sur un thème précis (ex. le vin) proposant plusieurs types de ressources. Par exemple : répertoire spécialisé, liens vers des répertoires ou pages de liens spécialisées, articles en texte intégral ou bibliographie en ligne, actualités du secteur, événements du secteur, accès à base(s) de données, etc. Le fin du fin d'un site fédérateur consiste à créer une communauté autour de lui (via forums, newsletters, etc) pour devenir le point de référence du domaine.

SPAMDEXING

Création ou modification d'un document avec l'intention de tromper un répertoire ou un système de classement automatique. Toute technique visant à augmenter la position potentielle d'un site aux dépens de la qualité du corpus de l'outil de recherche peut également être considérée comme du spamdexing.

TECHNIQUES DE POSITIONNEMENT

Le fait de modifier sa page Web afin que les moteurs de recherche traitent la page comme la plus appropriée pour une requête spécifique, ou un ensemble de requêtes.

TRI PAR PERTINENCE

Méthode de classement automatique des résultats retournés par le moteur de recherche qui s'appuie sur le calcul d'un score pour chaque réponse. La pertinence est alors basée sur des facteurs comme :

le poids d'un mot déterminé par sa place dans le document;

la densité : fréquence d'occurrence dans un document par rapport à la taille du document;

le poids d'un mot dans la base et sa fréquence d'occurrence dans toute la base;

la correspondance d'expression : similarité entre l'expression de la requête et l'expression correspondante dans un document;

relation de proximité : proximité des termes de la requête entre eux et dans le document. Cette technique est apparue avec la première génération de moteurs de recherche (à partir de 1994) et présente l'inconvénient d'être facile à détourner par les référenceurs peu scrupuleux (Cf. Spamdexing). Elle est utilisée par des moteurs comme AltaVista, Excite, Inktomi, Voila...

TRI PAR POPULARITE

Méthode de classement automatique des résultats retournés par le moteur de recherche qui s'appuie soit sur le principe de citation (popularité) soit sur la mesure de l'audience.

Dans le premier cas, l'importance d'une page est évaluée en fonction des liens hypertexte qui pointent vers elle et en fonction de la nature du document qui la cite. Le tri est alors indépendant du contenu, mais les documents récents ou peu cités par les autres sont défavorisés. Google avec son système de PageRank en

est l'exemple le plus connu.

Dans le second cas, l'importance d'une page est fonction du nombre de visites reçues lors d'une requête sur un moteur de recherche. C'est l'analyse du comportement de l'internaute qui détermine la popularité d'une page par rapport à un mot-clé. Solution DirectHit utilisée par HotBot et LookSmart par exemple.

URL

Uniform Resource Locator : adresse d'un site Web.

VIRUS

Programme informatique conçu pour "infecter" ou détériorer le fonctionnement d'un logiciel ou d'un ordinateur sur lequel il est installé.

WEB (ou WWW ou Web)

"Toile d'araignée mondiale" : outil logiciel multimédia et hypertexte permettant d'effectuer des recherches de tous types sur le réseau, l'accès à l'information recherchée et sa visualisation. Cet interface graphique, accessible via Netscape ou Internet Explorer, a permis l'explosion actuelle d'Internet. Les utilisateurs peuvent y créer, y éditer ou y rechercher des documents. La taille du Web est en constante augmentation, dépassant allègrement le milliard de sites.

WEB INVISIBLE

Expression qui sous-entend "la part du Web invisible pour les moteurs de recherche" : l'ensemble des pages non localisables et/ou non indexables par ces outils. Le Web invisible correspond à plusieurs types de ressources :

Documents dans des formats différents du html (par exemple pdf, word, etc.);

Pages situées à l'intérieur d'une frame (cadre);

Pages dont les caractéristiques techniques rendent difficiles, sinon impossible l'indexation par les moteurs : javascripts modifiant le contenu, technologies propriétaires (par exemple flash, active X, java);

Pages qui n'ont fait l'objet ni d'un référencement direct , ni d'aucun lien d'une autre page;

Pages nécessitant une identification de la part de l'internaute;

Pages dont le contenu indique aux moteurs qu'ils ne doivent pas l'indexer;

Page produite à partir de bases de données ou d'applications, et dont l'URL comporte des paramètres non exploitables par la plupart des moteurs;

Page produite à partir de données saisies par l'utilisateur via un formulaire html. Exemple : les résultats de l'interrogation d'une base de données avec des critères de recherche entrés par l'utilisateur.

XML

Pour eXtensible Markup Language : langage de description et d'échange de documents structurés sur le Web. Il est le résultat de la coopération d'entreprises et de chercheurs partenaires du World Wide Web Consortium (W3C) dont l'objectif a été de définir un formalisme permettant de regrouper les concepts d'hypertextes, de bases de données, de formats d'échange et de publication.

Sources des définitions

Les définitions ci-dessus sont tirées des discussions de la liste de diffusion des formateurs Internet ADBS et des sites de référence suivants :

IDF.net

Un glossaire sur la technologie de recherche internet

<http://www.idf.net/mdr/glossaire.html>

Grand Dictionnaire Terminologique de la Langue Française

<http://www.granddictionnaire.com>

Jargon Français, Roland Trique

Le Jargon Français v. 3.2.51

<http://www.linux-france.org/prj/jargonf/>