Typologie et mode de fonctionnement des outils de recherche d'information sur internet en Biologie/Médecine

Christophe Boudry

Dossier technique

Typologie et mode de fonctionnement des outils de recherche d'information sur internet en Biologie/Médecine

Christophe Boudry

Adresse
C. Boudry
Unité Régionale de Formation à l'Information Scientifique et Technique de Paris /Ecole nationale des Chartes
17 rue des Bernardins
75005 Paris
e-mail: boudry@ccr.jussieu.fr

Les outils nécessaire au repérage des informations pertinentes dans le domaine de la biologie ainsi que leur mode de fonctionnement sont souvent trop mal connus des utilisateurs du réseau internet. L'objet de cet article est de détailler ce fonctionnement, afin de permettre aux biologistes de mieux les utiliser et ainsi, d'augmenter leur potentialité à localiser des informations pertinentes sur internet.
De façon assez schématique et en se basant sur leur mode de fonctionnement, il est possible de distinguer 3 catégories d'outils de recherche d'information sur internet: les moteurs ou robots de recherche, les méta-moteurs et les annuaires *(Tableau I)*.

Moteurs de recherche (ou robots)

Il existe à l'heure actuelle plusieurs milliers de moteurs de recherche sur internet [1]. La principale caractéristique de ces outils réside dans leur mode de fonctionnement totalement automatisé. Ils sont principalement constitués de 3 parties: le "spider" (encore appelé "crawler" ou robot), l'index ou base de données et le logiciel/interface d'interrogation. L'objectif du spider est de repérer les pages web par suivi récursif des liens présents dans les pages *(figure 1.1)*. Il assure ainsi la lecture des données des pages HTML et le repérage des liens vers d'autres pages afin de constituer l'index. L'index est en effet le lieu de stockage et d'indexation des pages web visitées et repérées par le spider. L'entrée de pages web dans l'index est également rendue possible par la visite du spider de pages, soumises volontairement par leurs créateurs, au moteur de recherche *(figure 1.2)*. Ce référencement peut être gratuit ou payant, avec dans le deuxième cas l'assurance que la page web apparaisse dans les premières réponses à la suite de la saisie d'un mot clé donné et/ou que sa fréquence de mise a jour dans l'index soit élevée. En tout état de cause, l'économie marchande semble être amenée dans l'avenir à prendre une place de plus en plus importante dans ces processus de référencement [2]. L'indexation du code source HTML des pages web est réalisée de façon totalement automatisée *(figure 1.3)*, de manière non descriptive, quasiment en texte intégral (seules sont exclues certaines parties comme par exemple les commentaires figurant dans le code source HTML). Certaines données, comme l'adresse, le titre de la page, ou les méta-données (description et mots clés associés par leurs auteurs aux pages web) sont parfois indexées dans des champs spécifiques. Afin d'actualiser le contenu de ces pages web dans l'index, le spider se rend à intervalle défini sur les pages web déjà indexées, pour mettre à jour leur contenu dans l'index. De par son caractère automatisée, l'indexation réalisée est très

imparfaite et présente de nombreuses limites: pas d'indexation du contenu associé aux pages web telles les images et les autres fichiers liés aux pages web, pas d'indexation des pages orphelines ou en accès reservé ou bien encore ayant un contenu dynamique provenant d'une base de données *(figure 1.4).* Il en résulte que la quantité de données accessibles via les moteurs de recherche est faible comparée à la totalité des données présentes sur internet (d'où la création de l'expression "web invisible"). Si les moteurs généralistes rivalisent actuellement pour inclure le plus grand nombre de pages possibles dans leur index (plus de 1 milliard pour google selon [3]), certains moteurs, pour proposer un contenu plus spécifique, conditionnent l'entrée des pages web dans leur index, à leur appartenance à un champs d'activité donnée (c'est le cas de www.scirus.com et www.search4engine.com pour les sciences, de www.bioview.com pour la biologie).
Le logiciel et l'interface de recherche permettent à l'utilisateur de saisir sa requête en utilisant un ou plusieurs termes, qu'il pense être représentatif de sa recherche d'information, tout en respectant une certaine syntaxe *(figure 1.5)*. Certains moteurs, tels www.search4science.com proposent aux utilisateurs de restreindre ou d'étendre leurs recherches, soit en saisissant directement des termes complémentaires à ceux de la requête, soit en choisissant des termes, qui se rapportent à la requête suggérés par l'interface. D'autres moteurs offrent l'opportunité de restreindre la recherche à des champs spécifiques. Par exemple sur www.altavista.com, l'utilisateur peut restreindre sa recherche aux pages dont l'adresse contient un terme précis (la saisie de "host:univ" restreindra la recherche aux pages dont l'adresse contient le terme "univ" et qui par conséquent, sont des pages web universitaires) ou aux titres des pages web (par la saisie de "title:terme_recherché"). Il faut noter l'existence d'une extrême variabilité des syntaxes d'interrogation entre moteurs (problème qui concerne également les annuaires) *(Tableau II)* obligant les usagers à un effort d'adaptation important, dès que l'utilisation d'un nouveau moteur s'avère nécessaire [4]. Les termes de la requête sont alors recherchés dans l'index *(figure 1.6)* et une liste de pages web est proposée à l'utilisateur, selon un ordre de pertinence donné *(figure 1.7)*, pages sur lesquelles il peut se rendre via un lien hypertexte *(figure 1.8).* L'ordre de pertinence peut être par exemple basé sur le calcul de la fréquence d'apparition des termes de la requête dans une pages web et/ou considérer sa popularité par le biais du nombre de pages ayant un lien vers elle. L'ordre de pertinence des réponses pourrait bien être à l'avenir de plus en plus influencé par la présentation systématique, dans les premiers résultats, de pages provenant d'un référencement payant. Les réponses peuvent être également proposées regroupées par répertoires thématiques (c'est le cas de www.northernligth.com) ou par mots clés qui se rapportent aux termes de la requête (c'est le cas de www.exalead.com), elles peuvent encore être présentées sous forme de carte graphique (c'est le cas de www.kartoo.com).

Méta-moteurs

Le principe des méta-moteurs est de permettre l'interrogation simultanée de plusieurs index de moteurs de recherche différents. La saisie de la requête *(figure 2.1)* s'effectue à travers une interface unique qui peut être accessible via un site web (c'est le cas de www.metacrawler.com) ou via un logiciel qu'il est nécessaire d'installer sur un poste client (c'est le cas de Copernic qu'il est possible de télécharger à l'adresse www.copernic.com). La requête est alors soumise aux différents moteurs de recherche interrogés *(figure 2.2)* dont le nombre varie de quelques uns à plusieurs dizaines selon les méta-moteurs considérés. Les réponses, provenant de ces différents moteurs subissent le plus souvent une élimination des doublons et sont présentées par ordre de pertinence à l'usager *(figure 2.3)*, qui peut se rendre sur chaque page proposée, via un lien hypertexte *(figure 2.4)*. Le principal intérêt de ces outils est d'augmenter la taille de l'index interrogé. Cependant, si la plupart des méta-moteurs assurent une traduction de la requête pour l'adapter à la syntaxe de chacun des moteurs

interrogés, l'utilisation de requêtes complexes génère le plus souvent des réponses très bruitées.

Annuaires

L'objectif principal de ce type d'outils n'est pas de chercher à indexer un maximum de pages web, mais de privilégier la qualité des pages présentes dans leur index. L'entrée dans l'index est en effet supervisée par des indexeurs qui sélectionnent des pages web à partir de pages trouvées sur internet *(figure 3.1)* ou soumises par leurs créateurs *(figure 3.2)*, selon des critères tels la qualité et l'intérêt de leur contenu et/ou leur appartenance à une discipline ou une communauté donnée. La sélection de certaines pages web, notamment des pages commerciales, est parfois payante. L'indexeur créé alors une fiche décrivant le contenu de chaque page sélectionnée, lui affecte une catégorie d'appartenance et la stocke dans l'index *(figure 3.3)*. Les pages web sont classées en fonction de leur contenu dans des catégories et sous catégories pré-définies. L'archétype de ce type de classification est la classification Dewey, principalement utilisée dans les bibliothèques, mais également par certains annuaires de recherche sur internet (c'est le cas de bubl.ac.uk/link). Le nombre de pages web indexées varie d'un annuaire à l'autre, de quelques milliers pour les annuaires spécifiques à plusieurs millions pour certains annuaires généralistes. Pour effectuer sa recherche, l'usager a deux possibilités *(figure 3.4)*, il peut saisir une requête en utilisant un ou plusieurs termes qui sont alors recherchés dans les fiches descriptives présentes dans l'index *(figure 3.5)*, la liste des fiches descriptives correspondant à sa requête lui est alors proposée *(figure 3.6)*. Il peut également se déplacer en "furetant" dans les catégories ("browsing" pour les anglo-saxons) via des liens hypertextes à la recherche de pages pouvant répondre à son questionnement. Dans les deux cas, l'accès au contenu des pages web est obtenu via un lien hypertexte *(figure 3.7)*.

Si les différences entre moteurs et annuaires persistent quand a leur mode de fonctionnement, ces 2 types d'outils peuvent être désormais interrogés fréquemment à partir de la même interface. Par exemple, l'annuaire Yahoo! (www.yahoo.com) utilise les résultats du moteur de recherche Google (www.google.com) en cas de recherche infructueuse dans son index, et à l'inverse Google donne la possibilité à ses utilisateurs d'effectuer une recherche dans l'annuaire Open Directory Project (www.dmoz.org).
Si le monde des annuaires et des méta-moteurs semble maintenant relativement stable du point de vue des technologies déployées, celui des moteurs est en revanche encore en pleine ébullition. Il ne se passe en effet pas un mois sans que de nouveaux moteurs de recherche, proposant des technologies et mode de fonctionnement innovants, apparaissent. Souhaitons simplement que ces évolutions aillent dans le sens d'une plus grande facilité d'utilisation pour les usagers, de l'augmentation des performances de ces outils afin de localiser des informations toujours plus pertinentes dans le domaine de la biologie sans que l'économie marchande ne vienne (trop) fausser la donne.

Références

1. http://www.searchengineguide.com/. Page consultée le 5 octobre 2001.
2. Andrieu O. L'avenir du référencement sur les outils de recherche. Bases/Netsources 2001 ; 33 : 10 - 12.
3. http://www.searchenginewatch.com/reports/sizes.html. Page consultée le 4 octobre 2001.
4. Dong X, Su L. Search engines on the world wide web and information retrieval on the internet: a review and evaluation. Online & CD ROM Review 1997 ; 21 : 67 - 81.

Remerciements: L'auteur remercie L. Salamatian pour sa participation à l'étude comparative des syntaxes d'interrogation des différents outils de recherche présentés et P. Herlin pour ses conseils et la relecture du manuscrit.

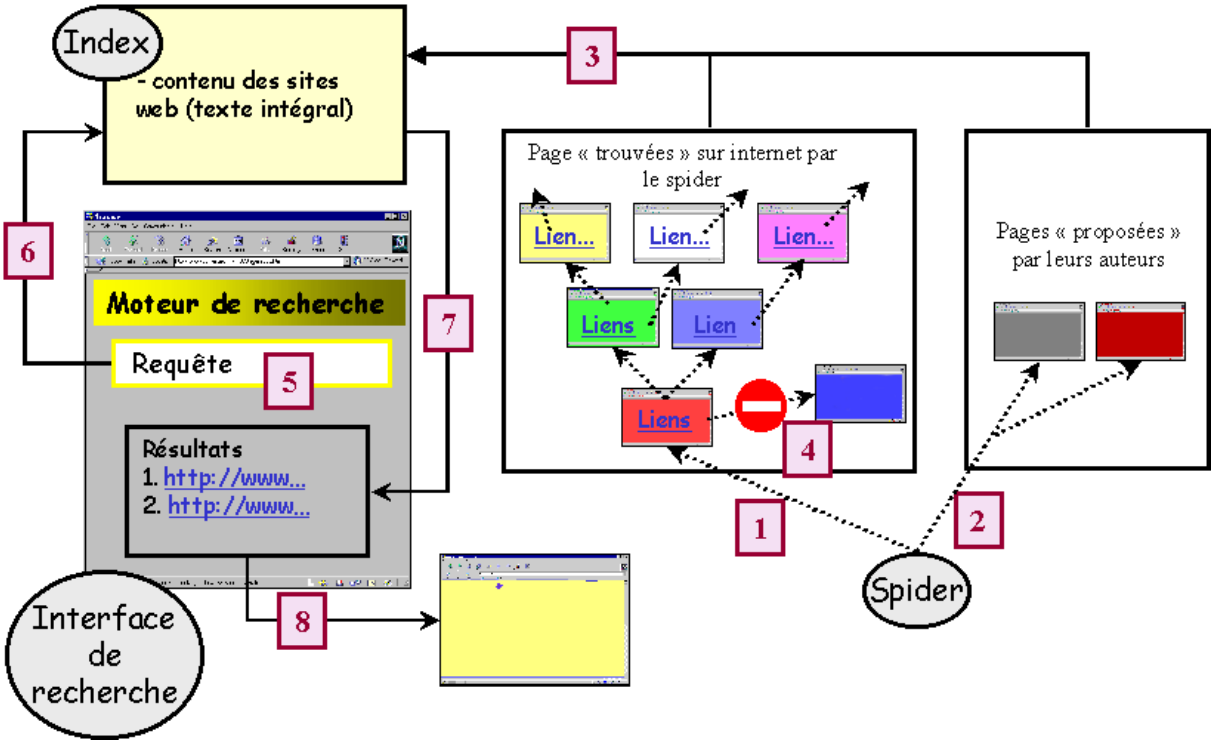

Figure 1. **Mode de fonctionnement des moteurs de recherche**

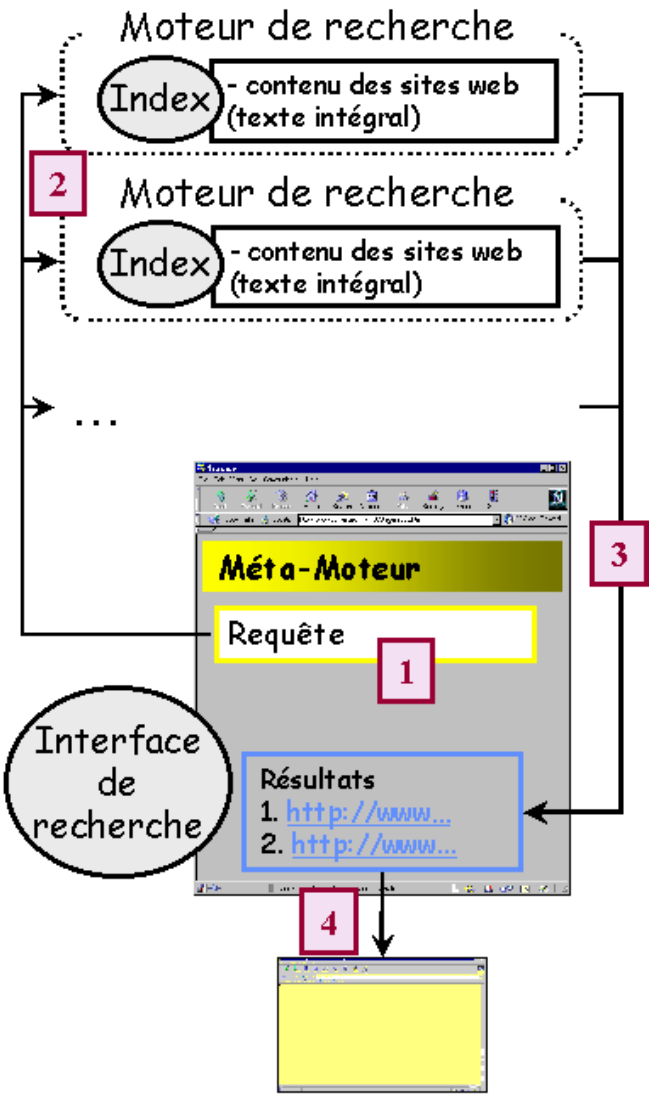

Figure 2. **Mode de fonctionnement des méta-moteurs.**

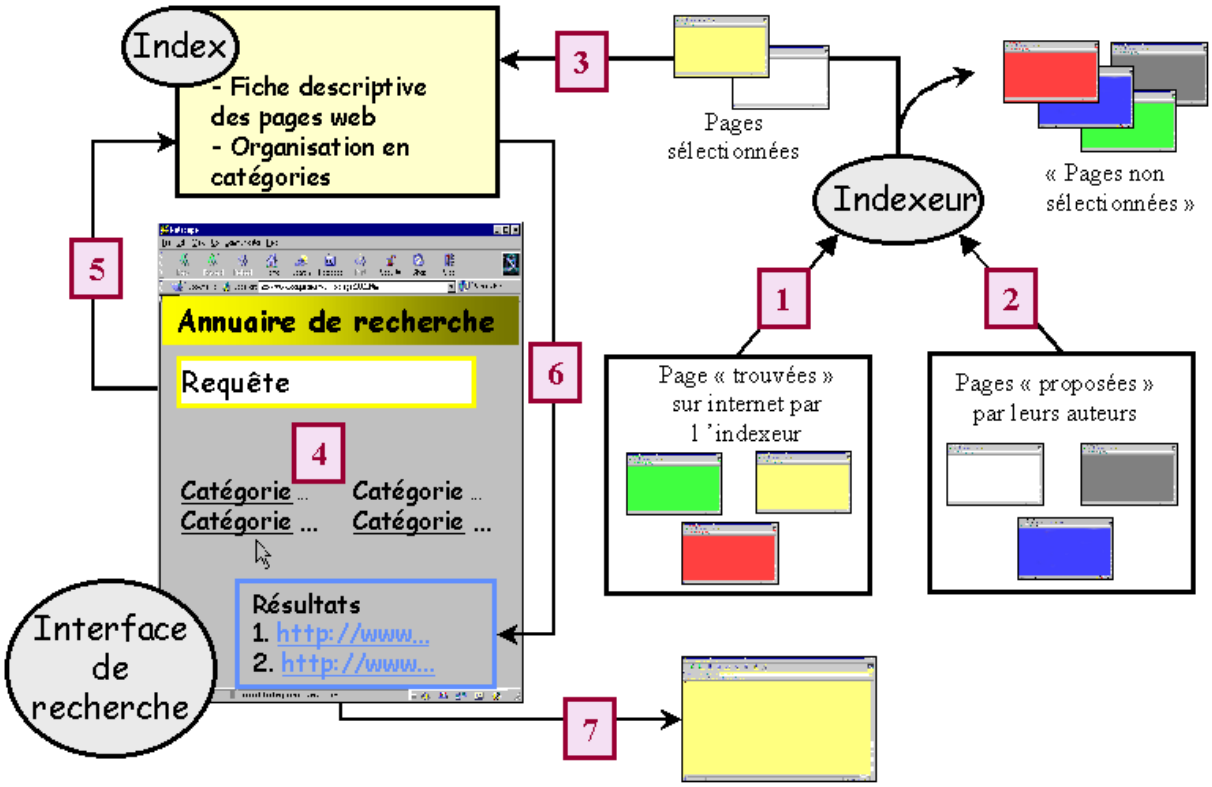

Figure 3. **Mode de fonctionnement des annuaires**

	Moteurs	Meta-moteurs	Annuaires
Généralistes	• www.google.com • www.altavista.com • www.northernlight.com • www.exalead.com • www.kartoo.com • …	• www.copernic.com • www.ixquick.com • www.metacrawler.com • www.dogpile.com • …	• www.yahoo.com • www.dmoz.org • www-english.lycos.com • www.about.com • …
Spécifiques au domaine scientifique	• www.scirus.com • www.search4science.com		• vlib.org • infomine.ucr.edu/Main.html • sunsite.berkeley.edu/InternetIndex/ • …
Spécifique au domaine de la biologie	• www.bioview.com		• www.pasteur.fr/recherche/BNB/bnb-fr.html • bubl.ac.uk/link • bioresearch.ac.uk • www.biofinder.org/index_e.html • www.library.ucsf.edu/biosites/ • biology.miningco.com/education/biology • www.sciencekomm.at • …

Tableau I. Principaux outils de recherche disponibles sur internet pour la recherche d'information en biologie. Les outils sont dits généralistes s'ils couvrent tous les domaines de connaissance, spécifiques s'ils couvrent uniquement le champ scientifique ou une discipline donnée. Des panoramas plus complets des outils généralistes et spécifiques au domaine "scientifique" et au domaine de la biologie peuvent être consultés respectivement aux adresses suivantes: (www.abondance.com) et (www.ccr.jussieu.fr/urfist/biolo/bioguide2/frame.htm).

	Bioview	*Scirus*	*Search4science*	*Altavista*	*Google*	*Copernic*	*Infomine*	*Open Directory Project*
Opérateur par défaut (1)	ET	ET	ET	ET	ET	ET	ET	ET
Opérateur ET (2)	NON	OUI (choisir "All the words")	NON	OUI (AND)	NON	OUI (choisir "Recherche de tous les mots" ou saisir AND)	OUI (AND)	OUI (AND)
Opérateur OU (3)	NON	OUI (choisir "Any of the words")	NON	OUI (OR)	OUI (OR)	OUI (choisir "Recherche de n'importe quel mot ou saisir OR)	OUI (NOT)	OUI (OR)
Opérateur SAUF (4)	NON	OUI (choisir ANDNOT)	NON	OUI (AND NOT)	NON	OUI (NOT)	NON	OUI (ANDNOT)
Opérateur arithmétique + (5)	NON	NON	OUI (+)	OUI (+)	OUI (+)	NON	NON	OUI (+)
Opérateur arithmétique - (6)	NON	NON	NON	OUI (-)	OUI (-)	NON	NON	OUI (-)
Recherche de phrase (7)	NON	OUI (choisir "Exact phrase")	NON	OUI (" ")	OUI (" ")	OUI (choisir "Recher de l'expression exacte")	OUI (" ")	OUI (" ")
Troncature (8)	NON	NON	NON	OUI (*)	NON	OUI (*)	OUI	OUI (*)

Tableau II. Disponibilité des principaux opérateurs de recherche et syntaxe d'interrogation (entre parenthèses) des moteurs de recherche Bioview, Scirus, Search4science, Altavista et Google du méta-moteur Copernic et des annuaires Open Directory Project et Infomine. Pour connaître la syntaxe exacte d'autres outils, il est possible de consulter www.abondance.com ou www.calvin.edu/library/searreso/internet/searengi.stm pour les outils de recherche généralistes, www.ccr.jussieu.fr/urfist/biolo/frame.htm pour les outils de recherche dans les domaines scientifiques et biologiques.

(1) L'opérateur par défaut est l'opérateur qui est utilisé "par défaut" lorsque l'utilisateur saisit deux termes à la suite sans aucune autre précision (ex: terme1 terme2). Les opérateurs booléens ET, OU et SAUF permettent de créér des requêtes complexes. (2) La saisie de "terme1 ET terme2" permet de localiser des pages web où apparaissent simultanément les 2 termes recherchés. (3) La saisie de "terme1 OR terme2" permet de localiser des pages web où apparaissent au moins l'un des deux termes recherchés. (4) La saisie de "terme1 NOT terme2" permet de localiser des pages web où apparaît le mot "terme1" sans que n'apparaissent le mot "terme2". (5,6) Les opérateurs arithmétiques + et - permettent de localiser des pages web ou apparaît un terme (syntaxe: +terme) ou d'exclure les pages web ou apparaît un mot (syntaxe :-terme). Si ces opérateurs arithmétiques permettent d'obtenir des résultats équivalents à l'utilisation de l'opérateur booléen ET ("terme1 ET terme2" équivaut à "+terme1 +terme2") ou de l'opérateur booléen SAUF ("terme1 SAUF terme2" équivaut à "+terme1 -terme2"), il n'est en aucune façon possible d'obtenir l'équivalent de l'opérateur booléen OU. (7) La recherche de phrases permet de localiser des pages web contenant une expression donnée. La syntaxe la plus courante consiste à inclure l'expression entre guillemets. (8) La troncature, dont la syntaxe la plus courante est "*" permet la recherche de variants orthographiques. Par exemple, la saisie de apoptos* permet de localiser des pages ou figure par exemple apoptosis ou apoptose.