questions de communication

Questions de communication

14 | 2008
Moteurs de recherche. Usages et enjeux

Attentes *versus* réalité

Les fonctionnalités nécessaires aux moteurs pour la recherche Web en 2008

Expectations versus Reality. Search Engine Features needed for Web Research in 2008

Judit Bar-Ilan

OpenEdition Journals

Édition électronique
URL : http://journals.openedition.org/questionsdecommunication/719
DOI : 10.4000/questionsdecommunication.719
ISSN : 2259-8901

Éditeur
Presses universitaires de Lorraine

Édition imprimée
Date de publication : 1 décembre 2008
Pagination : 49-74
ISBN : 978-2-86480-981-4
ISSN : 1633-5961

Référence électronique
Judit Bar-Ilan, « Attentes *versus* réalité », *Questions de communication* [En ligne], 14 | 2008, mis en ligne le 23 septembre 2015, consulté le 01 mai 2019. URL : http://journals.openedition.org/questionsdecommunication/719 ; DOI : 10.4000/questionsdecommunication.719

questions de communication, 2008, 14, 49-74

> DOSSIER

JUDIT BAR-ILAN
Bar-Ilan Université, Ramat Gan, Israël
barilaj@mail.biu.ac.il

ATTENTES *VERSUS* RÉALITÉ. LES FONCTIONNALITÉS NÉCESSAIRES AUX MOTEURS POUR LA RECHERCHE WEB EN 2008

Résumé. — Les chercheurs qui analysent le Web s'appuient sur des données qui sont souvent collectées à l'aide des moteurs de recherche. Dans une précédente contribution (Bar-Ilan, 2005), l'auteur a proposé une liste d'objectifs pour le moteur de recherche idéal en expliquant le besoin de fonctionnalités spécifiques pour ce type d'activité. Ici, il revisite cette liste et examine si les principaux moteurs de recherche actuels peuvent répondre, au moins partiellement, aux exigences de l'outil de recherche idéal. Les principaux outils de recherche sont commerciaux et destinés à l'utilisateur « moyen » et non au chercheur scientifique qui analyse le Web, ils ne peuvent donc pas satisfaire toutes les demandes.

Mots clés. — Webométrie, moteurs de recherche, fonctionnalités, besoins des chercheurs.

Il est communément admis que le Web est une source majeure d'information et de communication, à la fois dans la vie professionnelle et dans la vie quotidienne. Les principaux outils d'accès à l'information sur le Web sont les moteurs de recherche (voir par exemple Fallows, 2005 ou Dutton, Helpser, 2007). Actuellement, le moteur de recherche le plus utilisé dans le monde occidental est de loin Google, suivi de Yahoo et Live Search (Sullivan, 2007). En France, selon Nielsen/Netratings (2008), les compagnies les plus prisées sont, dans l'ordre : Google Inc., Microsoft, France Telecom, Iliad, PagesJaunes et Yahoo. Cependant, toutes ne sont pas spécialisées dans la recherche d'information en ligne, nous en concluons donc que Google, Live Search (propriété de Microsoft) et Yahoo sont les moteurs de recherche les plus populaires en France.

Dans cette contribution, nous étudions l'usage des moteurs pour la recherche scientifique sur le Web, c'est-à-dire celle qui analyse les données disponibles en ligne. On emploie parfois une autre terminologie : la webométrie, définie par Lennart Björneborn et Peter Ingwersen (2004 : 1217) comme « l'étude quantitative de la construction et de l'utilisation des ressources informationnelles, des structures et technologies du Web, inspirée de la bibliométrie et de l'infométrie » ou encore la webologie, définie par Alireza Noruzi (2004) comme « l'étude du Web, de sa structure, son organisation, sa topologie, ses fonctions, ses caractéristiques, ses interconnexions et son développement ». Le terme webologie est plus général et, dans les milieux académiques, le terme webométrie est plus répandu. Pour ne pas limiter notre approche, nous utilisons ici l'expression plus générale de « recherche Web ». On peut noter que l'expression « science du Web », dont le but « est de comprendre à la fois le développement du Web et de créer des approches qui permettent l'émergence de nouveaux modèles puissants et plus fertiles » (Berners-Lee *et al.*, 2006 : 269), est aussi pertinente de ce point de vue.

Parfois, les chercheurs peuvent se passer d'interroger les moteurs commerciaux pour la collecte des données nécessaire à l'analyse du Web. Le logiciel SocSciBot[1], développé par Mike Thelwall, offre cette possibilité : il est disponible librement pour quiconque dispose des ressources nécessaires au fonctionnement du robot de collecte et au stockage de ses résultats. D'autres (par exemple Spink, Jansen, 2004) ont la chance de pouvoir accéder aux historiques d'interrogation des moteurs de recherche commerciaux. Quelques études du Web ont été menées par des chercheurs qui travaillent pour les moteurs de recherche (par exemple Broder *et al.,* 2000 ou Fetterly *et al.,* 2004), ce qui leur a permis d'explorer le Web ou d'accéder aux données collectées par le moteur.

[1] http://socscibot.wlv.ac.uk/

Cependant, le plus souvent, les chercheurs qui étudient le Web recourent aux outils de recherche d'information librement disponibles : les moteurs de recherche ou les archives collectées dans le cadre des projets de conservation du Web (par exemple, l'archive d'internet, http://www.archive.org). Actuellement, les projets d'archivage du Web ont soit un accès limité (pour des raisons de droits d'auteurs), soit ils ne fournissent que des outils restreints de recherche plein texte, voire aucun ; ils sont alors d'une utilité très limitée pour la recherche scientifique en matière de Web. Les meilleurs outils libres d'accès sont donc les principaux moteurs de recherche. D'autres moteurs dotés de fonctionnalités spécifiques pour la recherche Web pourraient exister mais, selon nous, la couverture de l'outil de recherche – la taille de l'index – compte beaucoup ; c'est pourquoi nous limitons notre analyse des outils de recherche aux seuls moteurs de recherche généralistes.

Dans un premier temps, nous établirons la liste des fonctionnalités souhaitées, dont l'importance pour la recherche scientifique sur le Web a déjà été expliquée dans le détail (Bar-Ilan, 2005). Dans un deuxième temps, nous examinerons si les trois principaux moteurs de recherche (Google, Live Search et Yahoo) remplissent ces conditions avant de développer les conséquences de notre étude. Les moteurs de recherche changent constamment, aussi tenons-nous à souligner que nos remarques sur ces outils sont fondées sur les résultats des interrogations des moteurs effectuées en avril 2008. Pour appuyer les résultats, nous avons sauvegardé et documenté chaque exemple présenté ici, et nous pourrons fournir au lecteur intéressé une copie de nos exemples et toute la documentation sur lesquels repose cet article. Dans certains cas, nous avons ajouté ou modifié certains exemples suite aux remarques des relecteurs, c'est pourquoi certains cas datent de juin 2008 (quand les exemples ne datent pas d'avril 2008, nous l'indiquons clairement).

Tableau 1 : Attentes. Les fonctionnalités du moteur de recherche idéal.

1	couverture
2	fiabilité et exactitude
3	transparence, renseignements, clarté de la documentation
4	actualisation
5	temps de réponse, accessibilité
6	objectivité – pas d'influence publicitaire et pas d'influence sur l'environnement
7	tous les résultats recensés peuvent être retrouvés
8	résultats conservés en mémoire cache
9	classement, différentes options de tri
10	recherche de qualité dans les langues autres que l'anglais

11	interface de programmation d'application (API) disponible
12	requêtes booléennes intégrales, diversité des opérateurs d'interrogation
13	techniques avancées de recherche pour l'analyse des liens
14	diversité d'opérateurs de recherche avancée
15	fonctions supplémentaires : possibilité de recherche ou non sur le radical, troncature gauche/droite, jokers, sensibilité ou non à la casse des caractères, correcteur orthographique, possibilité d'exclure ou non les sites hors service
16	assistance durant la recherche : rétroaction de pertinence, recherche de pages similaires ou apparentées, personnalisation
17	possibilité de combiner toutes les fonctions dans une même requête (et nombre de termes de recherche illimité), possibilité de construire des sous-ensembles à partir de résultats précédents (emboîtement de requêtes)
18	possibilité d'adapter les résultats affichés
19	indexation de tout le document
20	possibilités de recherche non-textuelle

Réalité-fonctionnalités des moteurs de recherche actuels

Couverture

On sait bien que les moteurs de recherche ne couvrent pas – et ne peuvent couvrir – tout le Web (Bharat, Broder, 1998 ; Lawrence, Giles, 1998, 1999 ; Gulli, Signorini, 2005). On mesure l'étendue de la couverture d'un moteur de recherche par le nombre de pages qu'il indexe. Dans ce domaine, plus grand est l'index, meilleur n'est pas forcément le moteur. Toutefois, dans le cas où les moteurs de recherche sont utilisés pour collecter des données du Web, la couverture la plus large possible est un critère essentiel. Une des premières études fondée sur l'analyse des résultats des moteurs de recherche est celle de Ronald Rousseau (1997) sur les sitations[2], et ses conclusions sont fortement influencées par la couverture des moteurs utilisés. On peut essayer d'évaluer la limite inférieure du nombre de pages indexées en soumettant une requête sur un mot fréquent, par exemple sur l'article *a* en anglais. Le 2 avril 2008, pour cette requête, Google affichait plus de quatorze milliards de résultats, Live Search plus de sept milliards et demi et Yahoo presque 31 milliards (tableau 2).

[2] R. Rousseau utilise le terme de *sitation* proposé par G. Mc Kiernan (1996) dans le sens de citation de site web, par analogie avec les citations de textes, tout en soulignant que le sens des deux mots est sensiblement différent.

La couverture non uniforme des ressources en ligne par les moteurs pose un autre problème. Herbert Snyder et Howard Rosenbaum (1999) ont montré que le taux de couverture des principaux domaines de l'internet par les différents moteurs de recherche n'est pas le même. Dans une étude de la couverture des grands domaines nationaux, Mike Thelwall (2000) a constaté qu'elle était extrêmement inégale. Les résultats obtenus au moment où nous écrivons cet article montrent que la couverture des moteurs de recherche est encore inégale (tableau 2). En soumettant différentes requêtes, nous avons obtenu des ratios différents concernant la taille des index des moteurs de recherche[3].

Tableau 2 : Taille relative des moteurs de recherche.

2 avril 2008	***a*** en anglais	***le*** en français	и en Russe	site:.com
Google	14 010 000 000	42 800 000	24 200 000	13 160 000 000
Live Search	7 540 000 000	378 000 000	213 000 000	7 190 000 000
Yahoo	30 900 000 000	1 380 000 000	1 920 000 000	9 241 889 263
	Rapports de taille des moteurs			
Google	1,85	0,11	0,11	1,83
Live Search	1	1	1	1
Yahoo	4	3,65	9,01	1,29

Ces résultats semblent suggérer que l'index de Yahoo est le plus grand. Cependant, il est bien connu que le nombre de résultats indiqués par les moteurs de recherche n'est pas précis et on ne sait pas clairement si les différents moteurs de recherche interprètent la requête de la même manière (c'est-à-dire considèrent *a* comme un mot en tant que tel).

Fiabilité et exactitude

Lorsque l'on conduit une recherche scientifique, la fiabilité et l'exactitude des outils de collecte de données sont de la plus haute importance. Les principaux moteurs de recherche s'adressent au grand public pour qui ces caractéristiques sont moins importantes, du fait qu'il se concentre seulement sur les dix ou vingt premiers résultats.

[3] Sauf lorsque nous le précisons, les résultats de Google proviennent de google.com, ceux de Yahoo de search.yahoo.com et ceux de Live Search de son interface en anglais en choisissant la région États-Unis (http://www.live.com/?scope=web&mkt=en-US). Dans nos requêtes, nous n'avons pas limité les résultats à un langage donné, à l'exception de quelques exemples explicitement indiquées.

Ronald Rousseau (1998-1999) a relevé des fluctuations quotidiennes dans le nombre des résultats de recherche d'AltaVista ; ces variations ont été comparées au nombre de résultats rapportés par Northern Light, qui croissaient de manière continue. Judit Bar-Ilan (2000) a observé des variations quotidiennes importantes dans les résultats de recherche de Hotbot par rapport à Snap, alors que ces deux outils de recherche reposaient sur la technologie Inktomi. L'auteure a proposé un ensemble de mesures pour évaluer la stabilité d'un moteur de recherche dans le temps (Bar Ilan, 2002a).

Souvent, le nombre de réponses indiqué change au fur et à mesure que l'on explore plus en profondeur la liste de résultats. Par exemple, nous avons recherché sur Google « *digifeed* » le 3 avril 2008. Au début, cette requête donnait 686 résultats, puis nous avons voulu consulter l'ensemble complet des résultats (en incluant ceux qui avaient été omis initialement). Cette fois, le nombre de résultats indiqués passait à 693, mais au moment où nous consultions la dernière page, ce nombre a diminué à 510. Dans une certaine mesure, Google apporte un début de réponse à la question de ces modifications du nombre de résultats indiqués lorsqu'on n'inclut pas les pages ignorées par le moteur[4], mais pas dans le cas où l'on clique sur le lien « relancer la recherche en incluant les pages ignorées » en bas de la page de résultats. Live Search donnait au début 620 résultats mais, quand nous avons atteint les dernières pages de résultats, ce nombre s'est réduit à 198. Mike Thelwall (2008) a également observé ce comportement en étudiant un grand nombre de requêtes. Yahoo a rapporté et affiché le nombre de résultats le plus stable pour cette requête spécifique : au début, il indiquait 740 résultats et 713 ont été réellement retrouvés.

Nous avions montré (Bar-Ilan, 2005) que Google était un peu faible en « mathématique de moteur de recherche ». La situation ne s'est pas beaucoup améliorée en 2008 (tableau 3) :

Tableau 3 : Nombre de résultats obtenus avec Google en avril 2008.

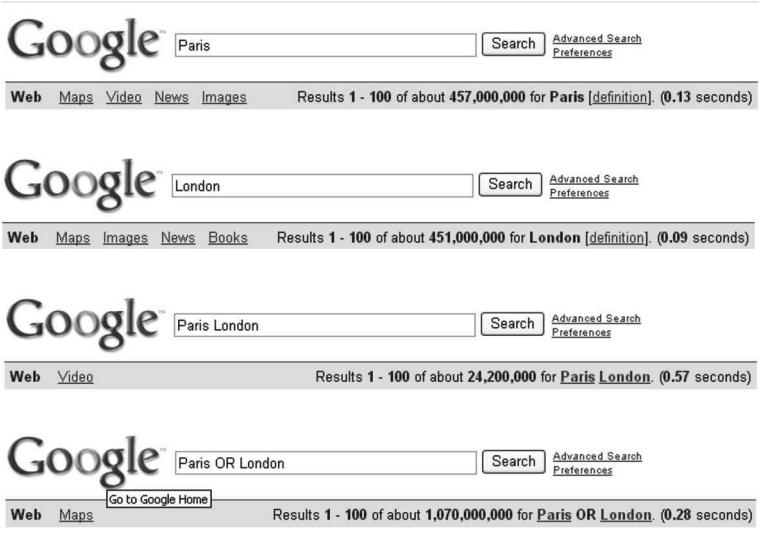
Requête	Nombre de résultats
Paris	457 millions
London	451 millions
Paris London (interprété comme Paris ET London)	24,2 millions
Paris OR London (OR signifie OU)	1 070 millions

[4] Google Inc., *Conseils sur la recherche avancée : disparition des résultats* http://www.google.fr/support/bin/answer.py?answer=499&topic=13913 (consulté le 28 juin 2008).

Selon la théorie ensembliste :
$|A \cup B| = |A| + |B| - |A \cap B|$

Le nombre de résultats pour « Paris OR London » devrait être 457+451-24,2=883,8 millions et non 1 070 millions de résultats.

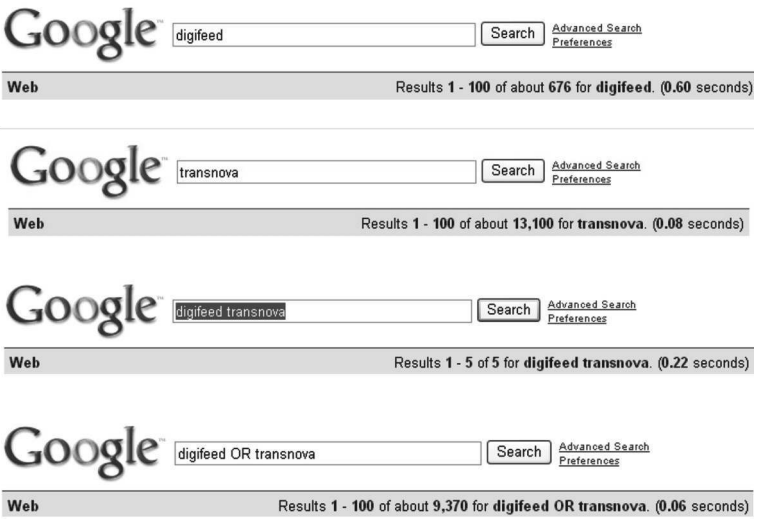
Figure 1 : La logique booléenne de Google.



Une explication possible de ces résultats équivoques pourrait être la suivante : le nombre total de résultats indiqué étant nettement plus élevé que le nombre de documents affichés, il s'agit seulement d'approximations grossières[5]. Ici, nous reproduisons un exemple déjà traité (Bar-Ilan, 2005) où le nombre de résultats rapporté est « faible ». Pour les requêtes « *digifeed* » et « *transnova* », Google donnait respectivement 676 et 13 100 résultats le 4 avril 2008. La requête avec l'opérateur ET retournait seulement 5 résultats mais, pour la requête OU, on obtenait 9 370 résultats, ce qui est bien inférieur au nombre de résultats pour seul le mot *transnova*. Il y a donc encore là quelque chose qui ne pose problème (figure 2).

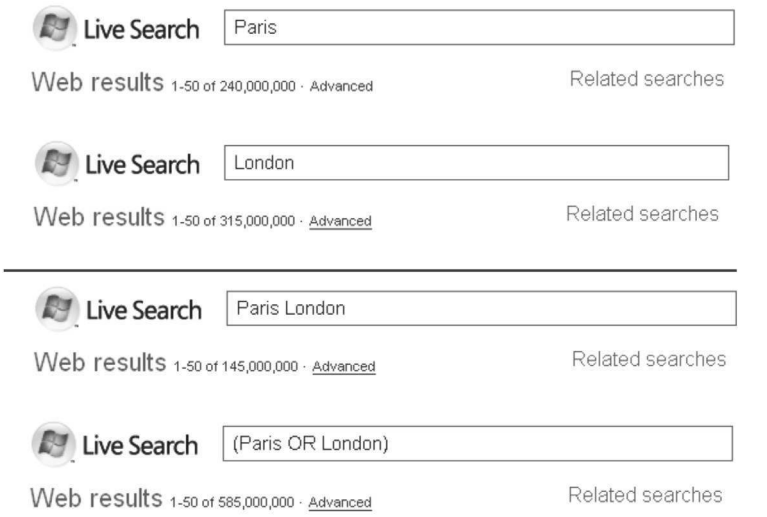
[5] Voir aussi sur ce point les analyses de J. Veronis (2005a ; 2005b ; 2005c).

Figure 2 : Requêtes *digifeed* et *transnova*.



L'inclusion et l'exclusion posent également problème à Live Search. Considérons à nouveau l'exemple de Paris-London : Paris donne 240 millions de résultats, London 315 millions résultats, Paris ET London 145 millions résultats, enfin Paris OU London 585 millions résultats (figure 3), alors que la requête Paris OU London devrait produire seulement 410 millions résultats.

Figure 3 : La logique booléenne de Live Search.



Nous n'avons pas rencontré ce type de problèmes sur Yahoo.

Transparence, renseignements, clarté de la documentation

Nous avons vu *supra* que les moteurs de recherche ne sont pas toujours parfaits. Parfois, il est impossible d'atteindre cette perfection (par exemple, il est impossible de couvrir tout le Web) ; mais, parfois aussi, les moteurs de recherche sont à l'origine de ces problèmes et ils les connaissent. Malheureusement, ils ne les signalent pas toujours. Par exemple, nous avons montré (Bar-Ilan, 2002b) que Google ne donne pas le nombre réel de pages de liens vers un site donné qui figurent dans son index. Quelques temps après, Google a reconnu cette politique sans fournir aucune explication[6]. Les moteurs de recherche ne rendent pas compte des raisons pour lesquelles ils ne retrouvent pas des documents qu'ils indexent sur certaines requêtes pour lesquelles ces documents devraient absolument apparaître (Mettro, Nieuwenhuysen, 2001).

Considérons par exemple la requête link:http://ques2com.ciril.fr sur Google. Celle-ci devrait lister les pages Web ayant un lien vers la page d'accueil de la revue *Questions de communication*. La requête a donné 53 résultats le 3 avril 2008 (figure 4). La même requête sur Yahoo donnait 1 510 résultats. La plupart des résultats des deux moteurs de recherche proviennent du blog irrealTV (http://irrealTV.blogspot.com) dont la barre latérale contient un lien vers la page d'accueil du site de la revue. Yahoo indexe environ 300 pages de ce blog, alors que Google en indexe 1620 (figure 5), mais présente seulement 48 de ces pages en réponse à la requête sur le lien. Pourquoi ? Nous n'avons pas de réponse à cette question. En outre, Yahoo présente au moins 90 autres pages qui comportent des liens vers la page d'accueil du site de la revue, alors que Google en montre seulement 5. Nous avons vérifié l'une de ces pages (http://www.aston.ac.uk/lss/staff/grundmannr.jsp): elle est indexée par Google et la copie en cache de Google contient un lien vers la page d'accueil de journal. Pourquoi cette page est-elle exclue des résultats ? Une fois de plus, nous n'avons pas de réponse. Notons que l'on peut vérifier si une certaine adresse URL est indexée par les moteurs de recherche en soumettant une requête formée de cet URL.

[6] Searchenginewatch Forum, 2004, *Google say not reporting all backlinks*, http://forums.searchenginewatch.com/showthread.php?t=2423 (consulté le 3 avril 2008).

Détail intéressant, Live Search n'affiche qu'un seul lien vers la page d'accueil de la revue. Ce moteur n'indexe qu'une partie du blog irrealTV (31 pages) et indexe également http://www.aston.ac.uk/lss/staff/grundmannr.jsp, mais aucune de ces pages n'est montrée en résultat pour la requête sur le lien. Un examen plus approfondi de l'unique résultat suggère que Live Search ne traite pas les requêtes portant sur des liens, bien que ce soit l'une des options proposées en « recherche avancée ». C'est comme si le moteur cherchait la chaine de caractère « link:http://ques2com.ciril.fr » dans les pages Web, au lieu de rechercher des liens hypertextes (figure 6).

À noter que les requêtes sur Yahoo portant sur les liens sont transférées à sa page Site Explore (http://siteexplorer.search.yahoo.com/) qui propose des filtres assez utiles pour l'analyse de liens (par exemple, exclure les liens du site auquel appartient la page, voir les liens pointant vers une page spécifique ou vers tout le sous dossier) et permet également de sauvegarder les résultats sous forme d'un fichier au format tsv (séparation par des tabulations), ce qui simplifie le processus de collecte de données.

Figure 4 : Les résultats de Google pour la requête sur le lien.

Figure 5 : Pages répertoriées d'irrealtv.blogspot.com.

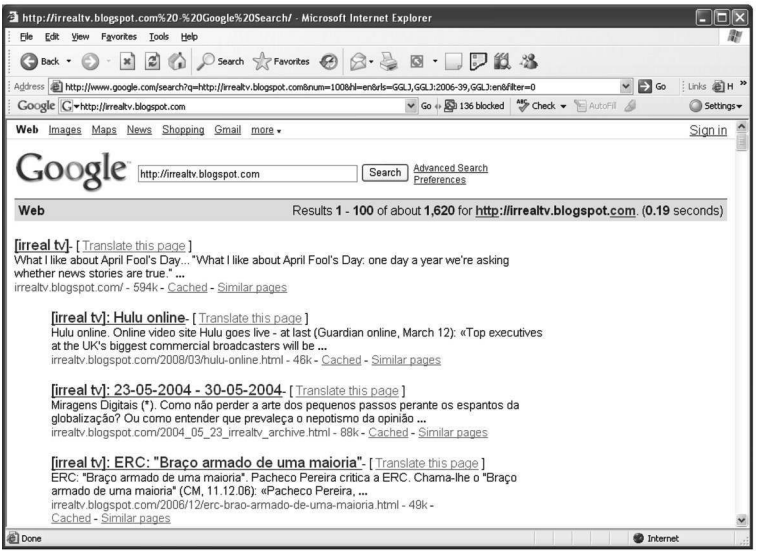
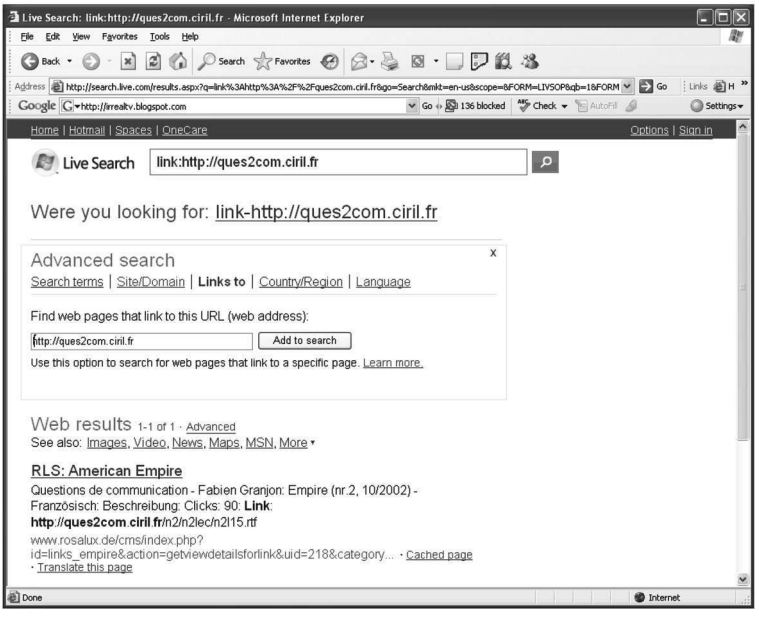


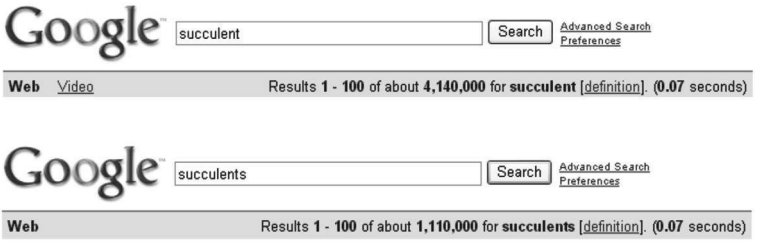
Figure 6 : Recherche de lien sur LiveSearch.



La documentation est parfois incomplète. Par exemple, il n'y a aucun détail sur les algorithmes de classement que les moteurs de recherche appliquent. On peut le comprendre du fait de la concurrence entre les moteurs de recherche et de la lutte constante contre les spammeurs qui essayent d'augmenter de manière artificielle leur classement dans les résultats.

Par ailleurs, l'algorithme de recherche sur le radical des mots que Google applique est peu clair : il ajoute d'autres mots à ceux de la requête d'origine de l'utilisateur[7]. Mais la recherche sur le radical est-elle appliquée systématiquement ? Comme on peut le voir dans la figure 7, ce n'est toujours pas le cas.

Figure 7 : Succulent *vs* succulents.



Live Search surligne également la forme du mot au pluriel sur une recherche sur le singulier (et le singulier pour une recherche sur la forme plurielle), mais nous n'avons trouvé aucune mention sur ce point dans les pages d'aide. Ce moteur n'applique pas non plus systématiquement la recherche sur le radical : en effet, pour le mot *succulent* nous avons obtenu 3 870 000 résultats alors que le mot *succulents* donne 1 010 000 résultats. Quant à Yahoo, on ne sait pas clairement s'il applique une recherche sur le radical : ce n'est pas mentionné dans ses pages d'aide et ce traitement de la requête n'est pas apparent lorsqu'on examine les résultats d'une recherche. Pour la requête *succulent*, nous avons obtenu 12 millions de résultats, contre 4,54 millions de résultats pour « *succulents* », puis 0,454 million de résultats pour « *succulent succulents* » et 15,7 millions de résultats pour « *succulent* OR *succulents* ».

Actualisation

Bien sûr, quand nous analysons les données du Web, nous nous intéressons aux pages les plus récentes, mais les moteurs exécutent leur recherche sur les pages qu'ils ont trouvées lors de leur dernière collecte. Les moteurs de recherche ne peuvent pas mettre à jour leurs index dès que le contenu d'un page Web est modifié. La fréquence des mises à jour dépend de la politique qu'ils mettent en œuvre. Prenons, par exemple, l'entrée de Wikipedia « 2008 Tibetan unrest »[8]. Cette page a été indexée par Google pour la dernière fois le 27 mars 2008 (c'est la date de la

[7] Google Inc., 2008, *The basics of Google search*, http://www.google.com/intl/en/help/basics.html (consulté le 1er avril 2008).

[8] En français, « Troubles au Tibet en 2008 » (http://en.wikipedia.org/wiki/2008_Tibetan_unrest).

version en cache) et, depuis le 5 avril 2008, la page a été modifiée 312 fois selon l'historique des modifications de la page. Live Search a indexé cette page pour la dernière fois le 22 mars 2008, manquant ainsi 701 modifications de cette entrée jusqu'au 5 avril 2008. Yahoo ne fournit pas de date pour la version en cache des pages qu'il indexe mais, en nous basant sur la liste des événements figurant sur la page en cache, elle semble être assez récente.

Temps de réponse, accessibilité

Chacun des trois moteurs de recherche a d'excellents temps de réponse et fournit un service 24 heures sur 24 et 7 jours sur 7.

Objectivité – pas d'influence publicitaire, pas d'influence sur l'environnement

Les moteurs de recherche sont les principaux outils qui nous permettent d'accéder à l'information en ligne. La plupart des utilisateurs consultent seulement les tout premiers résultats (voir par exemple Spink, Jansen, 2004), ces outils ont donc une forte influence sur la sélection de l'information en ligne qu'ils consultent. Lucas Introna et Helen Nissenbaum (2000) analysent cette question de la politique des moteurs de recherche. La communauté des utilisateurs peut parfois influencer l'ordre dans lequel les moteurs de recherche présentent les résultats, en mettant en œuvre des bombes Google (voir par exemple Bar-Ilan, 2007). Bien sûr, il n'existe pas de solution simple pour atteindre l'idéal d'objectivité. Les requêtes des utilisateurs sont brèves et les ensembles de résultats très grands, les algorithmes de classement des moteurs influencent donc de manière importante quelle information l'utilisateur verra réellement.

Tous les résultats recensés peuvent être consultés

Il s'agit d'une fonctionnalité très importante pour l'étude du Web. Comme nous l'avons vu, il n'est pas possible de se fier au nombre total de résultats indiqué par les moteurs. De plus, selon les objectifs des investigations menées sur le Web, il peut être nécessaire d'analyser les pages retrouvées pour une requête donnée. Actuellement, tous les moteurs de recherche évoqués dans cet article limitent à 1 000 le nombre de résultats qu'ils veulent bien montrer pour une requête donnée. Il est possible de contourner partiellement ce problème en employant différentes techniques de partitionnement, par exemple

en incluant ou en excluant des termes dans la requête, ou encore en limitant la recherche à des sites ou des formats de fichier spécifiques[9].

Résultats en cache

Google, Live Search et Yahoo donnent accès à une version en cache des pages, ce qui est très utile à la fois pour essayer de comprendre pourquoi le moteur a retrouvé tel résultat pour une requête donnée et pouvoir consulter les pages devenues inaccessibles. Google et Live Search indiquent aussi la date à laquelle la page a été mise en cache, ce qui est une fonctionnalité supplémentaire bien utile. Pour les versions plus anciennes des pages Web, on peut tenter de consulter les archives de l'internet (http://www.archive.org) ou les archives nationales du Web[10].

Classement, différentes options de tri

Le moteur MSN, version précédente de Live Search, permettait aux utilisateurs de modifier dans une certaine mesure l'ordre d'affichage des résultats de la recherche en paramétrant l'importance relative des critères de classement (Bar-Ilan, 2005). Cette option a été supprimée et, à présent, aucun des principaux moteurs de recherche ne permet aux utilisateurs de modifier l'ordre d'affichage des résultats de sa recherche, même s'ils le font de manière implicite en s'appuyant sur le comportement spécifique de recherche de l'utilisateur. La personnalisation des résultats par les moteurs de recherche est un sujet de recherche controversé (voir Levene, 2006 : chapitre 6.4 ; Jeh, Widom, 2003 ; Liu *et al.*, 2004).

Nous pouvons montrer qu'une personnalisation implicite existe, selon le pays d'où la requête est émise : l'ordre des résultats de recherche est différent quand on compare, par exemple, les résultats de Google.com à ceux de Google.fr (France) ou de Google.de (Allemagne) – voir les figures 8, 9 et 10.

[9] Pour des explications sur ces techniques, le lecteur peut se référer à Thelwall (2008) ou à Bar-Ilan, Peritz (à paraître).

[10] Voir les pages du Consortium international pour la préservation de l'internet (IIPC) qui propose des liens vers certains projets nationaux : http://www.netpreserve.org/

Figure 8 : Résultats de recherche pour « Louvre » sur Google.com.

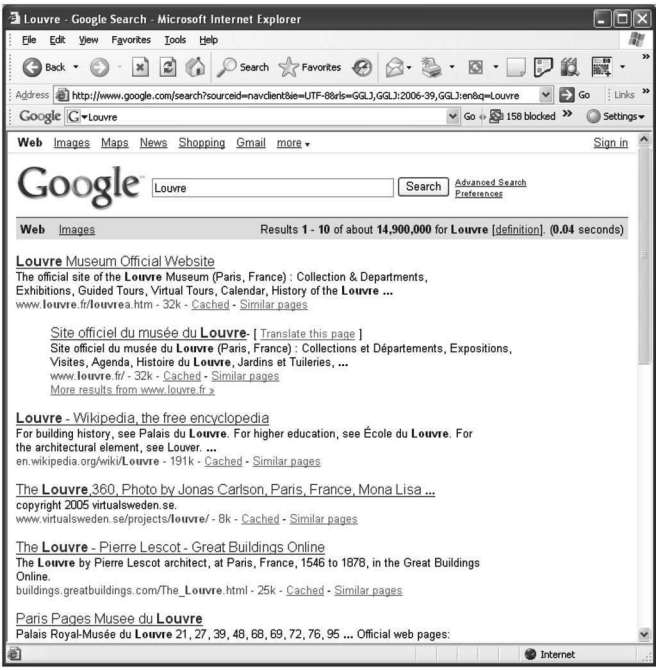


Figure 9 : Résultats de recherche pour « Louvre » sur Google.fr.

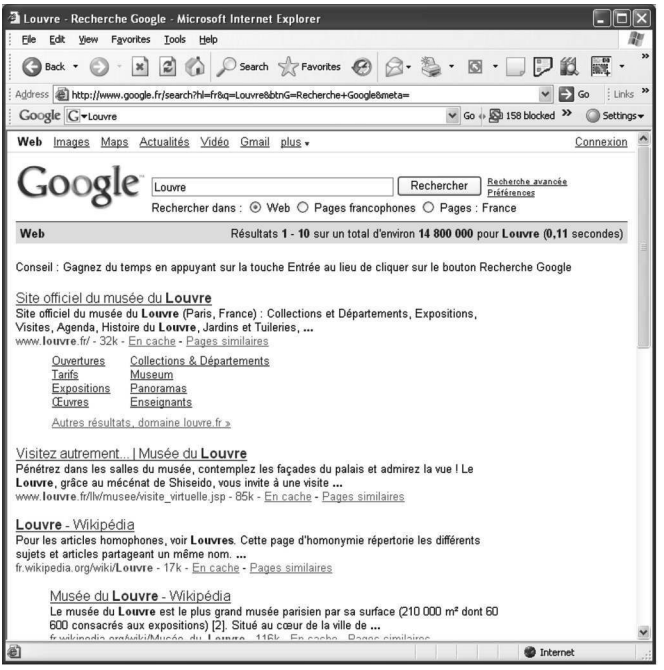
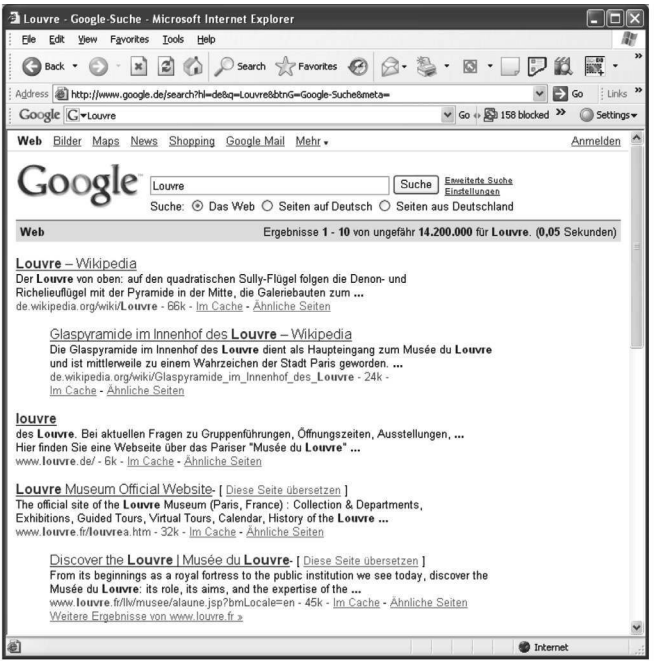


Figure 10 : Résultats de la recherche « Louvre » sur Google.de



On peut noter que le nombre total de résultats indiqué est pratiquement identique dans ces trois cas. Toutefois,les premiers résultats affichés sont différents, même si l'on ne limite pas la recherche à un langage donné.

Recherche de qualité dans d'autres langues que l'anglais

L'anglais est une langue dont la morphologie est simple et la plupart des recherches scientifiques menées en recherche d'information (*information retrieval*) reposent sur des corpus en langue anglaise. Par exemple, si nous prenons le cas de la langue française, plusieurs questions se posent : 1. Comment le moteur traite-t-il les accents ? ; 2. Comment le moteur traite-t-il les articles définis ou les pronoms possessifs dans le cas où ce ne sont pas des mots autonomes mais une partie du mot ? Il est important de comprendre la manière dont les moteurs gèrent les différentes langues, parce que cela influence la qualité des résultats obtenus pour une requête sur un terme donné (Lazarinis *et al.*, 2007; Bar-Ilan, Gutman, 2005). Pour répondre à ces questions, nous avons soumis des requêtes à chacun des moteurs en variant et en combinant des termes orthographiés de quatre manières différentes : « électricité », « l'électricité », « d'électricité » et « electricite ». Nous avons aussi testé l'ajout du signe + devant un terme

de la requête, cette syntaxe étant censée forcer le moteur à respecter l'orthographe exacte du terme[11], c'est-à-dire la présence des accents. Les résultats sont plutôt décevants. Windows Live et Yahoo ne semblent pas sensibles à la casse des caractères. Google semble montrer une certaine sensibilité à la casse mais plutôt inégale, même si les requêtes « électricité » et « +électricité » produisent un nombre total de résultats identiques, de même que « l'électricité » et « +l'électricité » ; nous avons obtenu une différence très importante quant au nombre total de résultats pour « électricité l'électricité » et pour « +électricité l'électricité ». Nous devons toutefois tenir compte du fait que le nombre total de résultats indiqué par les moteurs de recherche est souvent peu fiable.

Tableau 4 : Nombre total de résultats indiqué par les moteurs pour chaque requête (28 juin 2008).

Requête	Google.fr avec l'option « pages francophones »	Live Search en sélectionnant la région France et la langue « français (France) »	Yahoo France (fr.yahoo.com), avec l'option « en français »
électricité	10 300 000	7 030 000	28 200 000
+électricité	10 300 000	7 030 000	28 200 000
l'électricité	1 790 000	2 090 000	6 970 000
+l'électricité	1 790 000	2 090 000	6 970 000
d'électricité	2 490 000	2 200 000	6 830 000
+d'électricité	2 530 000	2 200 000	6 830 000
electricite	11 700 000	7 030 000	28 200 000
+electricite	1 790 000	7 030 000	28 200 000
électricité -l'électricité	1 330 000	4 920 000	21 200 000
+électricité -l'électricité	1 180 000	4 920 000	21 200 000
électricité l'électricité	11 700 000	2 090 000	7 190 000
+électricité +l'électricité	1 920 000	2 090 000	7 190 000
+électricité l'électricité	2 230 000	2 090 000	7 190 000
l'électricité -électricité	40 000	0	4

Mise à disposition d'une API

Actuellement les trois moteurs de recherche fournissent une interface de programmation d'application (API), ce qui permet aux développeurs de créer des applications et aux chercheurs scientifiques qui analysent le Web d'adapter les résultats de recherche à leurs besoins. Cependant, il faut noter que les résultats des recherches exécutées *via* les API ne sont en général pas identiques aux résultats obtenus grâce aux interfaces

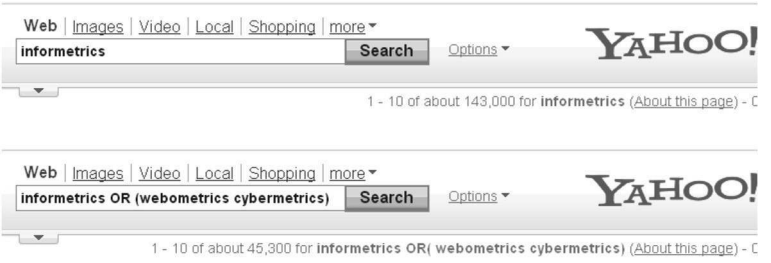
[11] Nous remercions le relecteur de cette contribution qui nous a fait cette suggestion.

de recherche ordinaires[12]. La mise à disposition de ces API améliore la recherche scientifique sur le Web parce que cela permet aux chercheurs, entre autres résultats, de récupérer automatiquement de grandes quantités de données, comme c'est le cas par exemple pour l'application *LexiURL Searcher*[13] de Mike Thelwall.

Requêtes booléennes intégrales, diversité des opérateurs

Les fonctions booléennes permettent aux chercheurs scientifiques d'affiner leurs requêtes et de dépasser – au moins en partie – la limite de 1000 résultats par requête (Thelwall, 2008 ; Bar-Ilan, Peritz, à paraître). Le grand public recourt rarement aux requêtes booléennes, les moteurs n'investissent donc pas d'efforts en ce sens. Google ne fournit pas toutes les fonctionnalités de la recherche booléenne, c'est-à-dire la possibilité de formuler des requêtes booléennes complexes et d'utiliser les parenthèses. Yahoo permettait de formuler de vraies requêtes booléennes, mais il semble que ce ne soit plus le cas, puisque le nombre de résultats pour la requête « informetrics OR (*webometrics cybermetrics*) » est inférieur à celui obtenu pour la requête sur le seul terme « informetrics » (figure 11) :

Figure 11 : Essai d'une requête booléenne sur Yahoo.



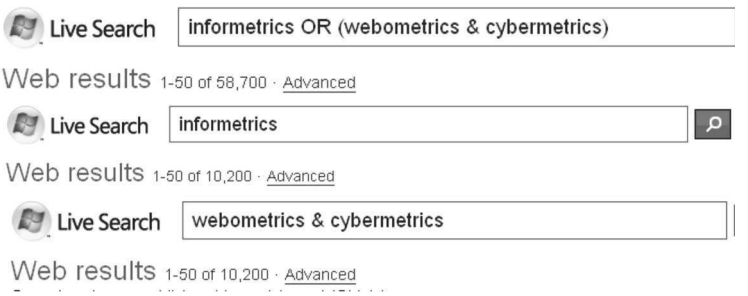
L'interprétation de Live Search pose également problème (figure 12), puisque le nombre de résultats pour la requête « *informetrics* OR (*webometrics* & *cybermetrics*) » ne devrait pas dépasser la somme des résultats obtenus pour « *informetrics* » et pour « (*webometrics* &

[12] Le site http://bsd119.ib.hu-berlin.de/~ft/index_e.html, maintenu par Mayr et Tosques, montre des différences dans les résultats obtenus avec les interfaces habituelles et les API de Google et Yahoo.

[13] http://lexiurlsearcher.blogspot.com/

cybermetrics) ». Actuellement, Live Search est le seul moteur qui prétende dans ses pages d'aide permettre la formulation de requêtes booléennes complexes[14].

Figure 12 : Essai d'une requête booléenne sur Live Search.



L'opérateur NOT (ou le signe - devant le terme) est utile en tant qu'opérateur autonome, parce qu'il permet d'essayer d'estimer la taille de la base de données du moteur en soumettant des requêtes de la forme $+x$ et $-x$, où x est le terme recherché. Actuellement, aucun des trois moteurs ne traite de telles requêtes[15].

Recherche avancée pour l'analyse de liens

Comme nous l'avons déjà mentionné, Yahoo fournit des fonctions utiles pour l'analyse de liens et de sites *via* Siteexplorer (http://siteexplorer.search.yahoo.com/). Google permet seulement de rechercher des liens vers des pages Web spécifiques, et non sur des sites entiers ou des sous-parties de sites ; en outre, il ne mentionne pas tous les liens, comme nous l'avons montré dans la sous-section 3. Sur ce plan, Google présente de sérieuses insuffisances pour la recherche scientifique puisqu'il n'est pas possible d'utiliser le moteur pour une analyse sérieuse des liens. S'il n'est pas possible de chercher les liens vers un site ou de combiner avec d'autres opérateurs les requêtes portant sur les liens, c'est peut-être parce que cet opérateur est rarement utilisé par le grand public ; seuls les webmestres essaient de savoir quelles sont les pages qui pointent vers leurs sites. Google ne se sent donc pas obligé de permettre de telles requêtes. Quelques études utilisent la fonction link: avec d'autres opérateurs (par exemple Larson, 1996 ; Thellwall, 2002 ; Vaughan, 2004 ;

[14] Windows Live, 2008, *Advanced search options*, http://help.live.com/help.aspx?project=wl_searchv1&market=en US&querytype=keyword&query=hcraesbew&domain=search.live.com:80 (consulté le 1er avr. 2008).

[15] Une explication possible suggérée par un des relecteurs de l'article : un tel traitement pourrait être trop coûteux en calcul pour les moteurs.

Bar-Ilan, 2004). Nous avons vu (figure 6) que Live Search ne permet pas non plus, à l'heure actuelle, de faire des recherches sur des liens, même si cette option apparaît dans son formulaire de recherche avancée.

Diversité des opérateurs de recherche avancée

Pour pouvoir formuler de manière plus précise des requêtes, nous avons besoin d'un certain nombre de moyens de restreindre les recherches, notamment sur les dates, les domaines, les langages, la zone géographique, les formats de fichiers, l'emplacement dans le fichier (par exemple, le titre, l'URL ou une ancre). Les opérateurs avancés qui existent actuellement permettent d'affiner une recherche sur une période particulière, un domaine ou une partie du document[16]. Google les nomme « opérateurs de recherche avancée », Yahoo emploie en anglais la terminologie *search meta terms* (méta-termes de recherche) et « mots clés spéciaux » en français, alors que Live Search décrit des moyens de « focaliser votre recherche » (« *focus your search* ») ce qui est traduit dans ses pages en français par « mots clés de recherche avancée ». Ces opérateurs peuvent s'avérer très utiles, par exemple lorsqu'on veut trouver combien d'URL du site *x* ont un lien vers un site ou une page donnée (voir Larson 1996 ; Thelwall, 2002 ; Vaughan, 2004 ; Bar-Ilan, 2004).

Par le passé, les moteurs permettaient aussi des recherches par dates : il était possible, par exemple, de limiter la recherche aux pages créées ou mises à jour dans une période particulière. Les résultats étaient assez problématiques du fait que, souvent, les serveurs web ne donnent pas les dates correctes. Actuellement Yahoo et Google permettent seulement de limiter les recherches aux pages mises à jour dans les derniers mois (pour plus de détails, voir les pages de recherche avancée de ces moteurs). Si l'on passe par l'interface avancée d'Alta Vista (http://www.altavista.com/web/adv), il est encore possible de limiter les recherches sur certaines dates. Alta Vista affiche plus ou moins les mêmes résultats que Yahoo, vu que Yahoo a racheté Alta Vista.

Une autre restriction concerne le nombre de termes que l'on peut insérer dans une requête. Auparavant, Google limitait à 10 le nombre de mots de la requête, mais est passé à 32 il y a quelques temps. Cela restreint toujours les possibilités car des requêtes plus longues permettent d'outrepasser les limites imposées sur le nombre de résultats affichés.

[16] Les opérateurs de recherche avancée que Google propose actuellement sont énumérés sur la page http://www.google.com/help/operators.html et pour Yahoo, sur la page http://help.yahoo.com/l/us/yahoo/search/basics/basics-04.html.

Yahoo semble limiter le nombre de mots dans une requête à 50 et Live Search à 19 mais cela n'est pas annoncé dans les pages d'aide de ces moteurs de recherche.

Fonctions supplémentaires

Un certain nombre de fonctions supplémentaires sont utiles : pouvoir rechercher ou non sur le radical d'un mot, utiliser la troncature gauche/droite, des jokers[17], régler la sensibilité à la casse des caractères, disposer d'un correcteur d'orthographe, choisir d'exclure ou non les sites hors service. Chacun des trois moteurs de recherche propose une forme de correction orthographique. Tous ne distinguent pas les majuscules et les minuscules. Il n'est pas possible de choisir d'activer ou non la recherche de radical. Pour l'heure, ils ne permettent pas l'utilisation de caractères jokers ou tout autre opérateur de troncature. Du point de vue du chercheur scientifique qui analyse le Web, il vaudrait mieux que le moteur le laisse décider de recourir ou non à la recherche sur le radical, de distinguer majuscules et minuscules ou de recourir à la troncature. Actuellement, seul Live Search permet de choisir d'exclure ou non les sites hors service ainsi que Yahoo, à condition de passer par l'interface d'Alta Vista (www.altavista.com).

Assistance durant la recherche

Il est utile de disposer de fonctionnalités de rétroaction de pertinence, de pouvoir rechercher des pages similaires ou liées. Quant à la personnalisation, c'est un sujet brûlant et controversé[18] (voir Murray, Teevan, 2007). D'un côté, elle est supposée améliorer les recherches des utilisateurs mais, d'un autre, elle soulève des inquiétudes sur les données privées puisque, pour fournir des résultats personnalisés, le moteur doit détenir des informations – implicites ou explicites – sur l'utilisateur.

Possibilité de combiner toutes les fonctions dans une même requête

Les moteurs devraient permettre de combiner en une seule requête toutes les fonctions disponibles et ne pas limiter le nombre de termes

---

[17] Un caractère joker ou caractère générique permet de remplacer n'importe quelle lettre dans un mot. Par exemple, si le caractère joker est ?, « ver?e » retrouvera aussi bien « verse » que « verte ».

[18] Lire à ce sujet H. G. Hotchkiss, « The pros & cons of personalization », *Search Engine Land,* 9 mars 2007, http://searchengineland.com/070309-081324.php (consulté le 28 juin 2008).

spécifiés dans l'interrogation. Il serait également souhaitable de pouvoir construire des sous-ensembles à partir des résultats précédents, par emboîtement de requêtes. Tous les moteurs de recherche limitent le nombre de termes que l'on peut utiliser dans une requête. Ce peut être un point faible pour les chercheurs qui analysent le Web quand ils conçoivent des requêtes précises. Google ne permet pas les recherches booléennes complexes (notamment l'utilisation de parenthèses) ni de combiner l'opérateur *link :* avec un autre terme dans la requête, ce qui est un limite importante pour les recherches scientifiques sur le Web.

Possibilité d'adapter les résultats affichés

Les moteurs répondent à certains besoins dans ce domaine. Live Search et Yahoo (si l'on passe par l'interface d'AltaVista) permettent d'exclure ou non les sites hors service. L'utilisateur peut paramétrer le nombre de résultats affichés par page pour chacun des trois moteurs. Mais aucun d'eux n'utilise des techniques de regroupement des résultats (*clustering*) et les utilisateurs ne peuvent pas modifier le format d'affichage pour obtenir des résultats spécifiques, sauf sur des points mineurs (par exemple masquer ou non le lien « plus de pages sur ce site » sur Yahoo). Tous les moteurs permettent aux utilisateurs d'activer un filtre pour les sites réservés aux adultes.

Indexation de tout le document

On ne sait pas exactement s'il existe une limite de taille pour l'indexation des documents. Ces limites ont existé par le passé. Si l'on prend l'exemple du livre numérique *Amusements in Mathematics* d'Henry Ernest Dudeney, sur le site du Projet Gutenberg[19], Google l'indexe ; cependant, une recherche sur une expression exacte qui apparaît vers la page 235 n'affiche aucun résultat. Cela semble indiquer que Google n'indexe les documents longs que jusqu'à une certain point. Live Search indexe le livre jusqu'à la page 219 environ et Yahoo ne peut retrouver des expressions qui apparaissent dans ce livre après la page 55. Le fait de ne pas indexer le document entier a un effet négatif sur le rappel[20].

[19] Dudeney H. E, 1917, *Amusements in Mathematics*, version numérique sur le site du projet Gutenberg à l'adresse: http://www.gutenberg.org/files/16713/16713-h/16713-h.ht

[20] Pour une définition du rappel (Levene, 2006 : 25).

Recherches non textuelles

Actuellement, les trois moteurs permettent la recherche d'images, en se fondant principalement sur des descriptions textuelles. Il y a encore beaucoup de progrès à faire dans le domaine de la recherche d'information multimédia. Les fonctions de recherche d'images qui pourraient présenter un intérêt pour les chercheurs qui analysent le Web ne sont pas traitées dans cet article.

Conclusion

Les moteurs commerciaux actuels sont encore loin de l'outil idéal de recherche en ligne dont rêvent les chercheurs qui analysent le Web. Ce dont nous avons besoin est un outil puissant, fiable et flexible pour servir la communauté scientifique. L'ensemble des fonctions idéales est, de façon inévitable, quelque peu subjective et fondée sur mes préférences personnelles, mais les recherches conduites et publiées par d'autres confirment le besoin de beaucoup de ces fonctionnalités. Bien sûr, la fiabilité est indispensable pour les recherches académiques. Avec des temps de réponse longs, la collecte de donnée à grande échelle devient impossible et c'est donc un besoin de base. L'étendue de la couverture du moteur est de la plus grande importance pour les analyses quantitatives. Les mises à jour fréquentes de la base de données du moteur sont nécessaires notamment pour les études sur l'évolution du Web. Le fait de pouvoir retrouver tous les résultats de recherche indiqués est crucial pour étudier le contenu des pages Web sur un sujet, dans un domaine ou sur un site donné. Les hyperliens sont l'un des mécanismes de base du Web et présentent un grand intérêt pour en comprendre la structure. Il est assez facile d'analyser les liens sortant d'un site ou d'une page donnée ; mais, si le chercheur n'a pas de robot de collecte, il doit compter sur les outils fournis par les moteurs de recherche pour accéder aux liens qui pointent vers un site ou vers une page. Les fonctionnalités linguistiques des moteurs pour les langues autres que l'anglais prennent de l'importance avec l'augmentation de la proportion de pages non anglophones en ligne. Les autres fonctionnalités que nous avons listées répondent aux principaux objectifs des chercheurs scientifiques et facilitent la collecte des données, ou en fournissent des moyens détournés dans le cas où les moteurs de recherche n'offrent pas certaines fonctions de base.

Certaines de ces fonctionnalités idéales sont difficiles, voire impossibles, à atteindre – par exemple une couverture complète du Web ou une mise à jour instantanée –, mais sur certains points, les moteurs pourraient s'améliorer. Par exemple, ils pourraient fournir l'ensemble complet des résultats pour une requête. Ils pourraient améliorer leurs estimations du

nombre total de résultats pour une requête, permettre des requêtes booléennes plus complètes ou des investigations sur les liens. Les moteurs de recherche ont un but commercial et leur principal objectif est de servir les intérêts des utilisateurs communs, mais on pourrait espérer qu'ils trouvent le moyen de satisfaire aussi les besoins des chercheurs scientifiques.

Dans un article antérieur (Bar-Ilan, 2005), nous avions promis de revisiter périodiquement les fonctions proposées par les moteurs de recherche. Le présent article met à jour nos résultats pour la première fois. Comme dans le précédent, nous recommandons de construire un ensemble de tests selon des règles méthodologiques claires et planifiées, qui détaillent la façon d'exécuter et de documenter ces tests pour évaluer la performance des moteurs de manière périodique[21]. Nous espérons que la communauté des chercheurs s'emparera du challenge qui consiste à instaurer de telles règles méthodologiques et que ces règles motiveront les moteurs de recherche à mieux répondre aux besoins de la recherche scientifique.

Références

Bar-Ilan J., 2000, « Evaluating the stability of the search tools Hotbot and Snap : a case study », *Online information review*, 24(6), pp. 439-449.

— 2002a, « Methods for Measuring Search Engine Performance over Time », *Journal of the American Society for Information Science and Technology*, 54(3), pp. 308-319.

— 2002b, « How Much Information Search Engines Disclose on the Links to a Web Page ? - A Longitudinal Case Study of the "Cybermetrics" Home Page », *Journal of Information Science*, 28(6) pp. 455-466.

— 2005, « Expectations versus reality – Search engine features needed for Web research at mid 2005 », *Cybermetrics*, 9(1), paper 2 [en ligne] http://www.cindoc.csic.es/cybermetrics/articles/v9i1p2.html (consulté le 1er avr. 2008).

— 2007, « Google bombing from a time perspective », *Journal of Computer-Mediated Communication*, 12(3), article 8, http://jcmc.indiana.edu/vol12/issue3/bar-ilan.html (consulté le 1er avr. 2008).

Bar-Ilan J., Gutman T., 2005, « How do search engines respond to some non-English queries ? », *Journal of Information Science*, 31(1), pp. 13-28.

[21] Nous avons intégralement sauvegardé toutes les recherches effectuées pour cet article, et nous avons bien noté la date à laquelle les recherches ont été effectuées. Cela est nécessaire, puisqu'il sera sans doute impossible de reproduire les résultats exacts que nous avons obtenus pour nos requêtes. Nous sommes prêts à fournir au lecteur intéressé les données brutes de cette étude.

Bar-Ilan J., Peritz B. C., « The lifespan of "informetrics" "on the Web : An eight year study (1998-2006) », *Scientometrics*, à paraître.

Bharat K., Broder A., 1998, « A technique for measuring the relative size and overlap of public Web search engines », *in* : *Proceedings of the 7th International World Wide Web Conference*, http://www.ra.ethz.ch/CDstore/www7/1937/com1937.htm (consulté le 1er avr. 2008).

Björneborn L., Ingwersen P., 2004, « Toward a basic framework for webometrics », *Journal of the American Society for Information Science and Technology*, 55(14), pp. 1216-1227.

Broder A., Kumar R., Maghoul F., Raghavan P., Rajagopalan S., Stata R., Tomlins A., Wiener J., 2000, « Graph structure in the Web », *Computer Networks*, 33(1-6), pp. 309-320.

Dutton W. H., Helsper E. J., 2007, *The Internet in Britain : 2007*, Oxford Internet Institute, University of Oxford, Grande Bretagne, http://www.oii.ox.ac.uk/research/oxis/OxIS2007_Report.pdf (consulté le 1er avr. 2008).

Fallows, D., 2005, *Search engine users. Pew Internet & American Life Project*, http://www.pewinternet.org/pdfs/PIP_Searchengine_users.pdf (consulté le 1er avr. 2008).

Fetterly D., Manasse M., Najork M., Wiener J., 2004, « A large-scale study of the evolution of Web pages », *Software : Practice and Experience*, 34(2), pp. 213-237.

Gulli A., Signorini A., 2005, « The indexable Web is more than 11.5 billion pages », pp. 902-903, *in : Special Interest Tracks and Posters of the 14th International Conference on World Wide Web*, http://www.cs.uiowa.edu/~asignori/web-size/size-indexable-web.pdf (consulté le 1er avr. 2008).

Introna L. D., Nissenbaum H., 2000, « Shaping the Web : Why the politics of search engines matters », *The Information Society*, 16, pp. 169-180.

Jeh G., Widom J., 2003, « Scaling personalized Web search », pp. 271-279, *in : Proceedings of the 12th International World Wide Web Conference*, Budapest, Hongrie.

Larson R. R., 1996, « Bibliometrics of the World Wide Web : An exploratory analysis of the intellectual structure of Cyberspace », pp. 71-78, *in : Proceedings of ASIS96*, http://sherlock.berkeley.edu/asis96/asis96.html (consulté le 28 juin 2008).

Lawrence S., Giles C. L., 1998, « Searching the World Wide Web », *Science*, 280 (5360), pp. 98-100.

— 1999, « Accessibility of information on the Web », *Nature*, 400, pp. 107-109.

Lazarinis F., Ferro J. V., Tait J., 2007, « Improving non-English Web searching » (iNEWS07), *SIGIR Forum*, 41(2), pp. 72-76.

Levene M., 2006, *An introduction to search engines and navigation*, Harlow, England, Addison-Wesley.

Liu F., Yu C., Meng W., 2004, « Personalized Web search for improving retrieval effectiveness », *IEEE Transactions on Knowledge and Data Engineering*, 16(1), pp. 28-40.

Mettrop W., Nieuwenhuysen P., 2001, « Internet search engines. Fluctuations in document accessibility », *Journal of Documentation*, 57(5), pp. 623-651.

Murray G. C., Teevan J., 2007, « Query log analysis : Social and technological challenges », *SIGIR Forum*, 41(2), pp. 112-120.

Nielsen/Netratings, 2008, *France: Top 10 parent companies: Month of February 2008. Home/Work panel*, http://www.nielsen-netratings.com/reports.jsp?section=pub_reports_intl&report=parent&period=monthly&panel_type=4&country=France (consulté le 1er avril 2008).

Noruzi A., 2004, « Introduction to Webology », *Webology*, 1(1), Article 1, http://www.webology.ir/2004/v1n1/a1.html (consulté le 28 juin 2008)

Rousseau R., 1997, « Sitations: An exploratory study », *Cybermetrics*, 1(1), http://www.cindoc.csic.es/cybermetrics/articles/v1i1p1.html (consulté le 28 juin 2008).

— 1998-99, « Daily time series of common single word searches on AltaVista and NorthernLight », *Cybermetrics*, 2/3(1), http://cybermetrics.cindoc.csic.es/articles/v2i1p2.html (consulté le 28 juin 2008).

Snyder H., Rosenbaum H., 1999, « Can search engines be used as tools for web-link analysis ? A critical view », *Journal of Documentation*, 55(4), pp. 375-384.

Spink A., Jansen B. J., 2004, *Web search: Public searching the Web*, London, Springer.

Sullivan D., 2007, « Comparing search popularity rating services: June to November 2007 », *Searchengineland*, http://searchengineland.com/071228-173523.php (consulté le 1er avr. 2008).

Thelwall M., 2000, « Web impact factors and search engine coverage », *Journal of Documentation*, 56(2), pp. 185-189.

— 2002, « An initial exploration of the link relationship between UK university Web sites », *ASLIB Proceedings*, 54(2), pp. 118-126.

— 2008, « Extracting accurate and complete results from search engines : Case study Windows Live », *Journal of the American Society for Information Science and Technology*, 59(1), pp. 38-50.

Veronis J., 2005a, « Web : Comptes bidons chez Google ? », *Technologies du langage*, http://aixtal.blogspot.com/2005/01/web-comptes-bidons-chez-google.html (consulté le 28 juin 2008).

— 2005b, « Le mystère des pages manquantes de Google résolu », *Technologies du langage*, http://aixtal.blogspot.com/2005/02/web-le-mystre-des-pages-manquantes-de.html (consulté le 28 juin 2008).

— 2005c, « Web : Google perd la boole », *Technologies du langage*, http://aixtal.blogspot.com/2005/01/web-google-perd-la-boole.html (consulté le 28 juin 2008).