Search engines for scholarly research: a brief update for clinicians

D. John Doyle MD PhD FRCPC

NO longer imprisoned in libraries, medical information can now be readily obtained online, with Web-based journal articles, professional updates, practice guidelines, evidence-based care guides, and much more. However, this vast selection of resources may present a challenge for some clinicians, who may be overwhelmed by the available options. Complicating this has been an increasing popularity of new forms of publishing.[1–4] While traditional vehicles for the dissemination of scholarly information like books and paper journals remain popular, academics are also contributing articles to electronic journals, as well as self-publishing articles on the Web and publishing multimedia material like instructional videos in CD-ROM and DVD format.

The availability of academic material in such a wide variety of formats, however, has led to new challenges, especially for busy clinicians trying to keep abreast of how the latest software developments may influence their scholarly and continuing medical education needs. In this context it is useful to consider how search engines function, and how academic search engines and other software tools compare with general search engines when searching the medical literature.

Search engines

Search engines have become ubiquitous to Internet users. While the Google search engine (www.google.com) is the best known and most commonly used, there are a great many others from which to choose, for example Alta Vista (www.altavista.com), Lycos (www.lycos.com) and AOL Search (search.aol.com), to name a few. Despite this large selection, however, a great many Web users simply use Google to the exclusion of all other search engines. This is largely because of Google's exceptional speed and accuracy, its excellent user interface, and because of its massive database (over 1 billion pages indexed).

"Meta" search engines combine and integrate the results of several search engines into one engine. For instance, MetaCrawler (http://www.metacrawler.com) searches Google, Yahoo Search, MSN Search, Ask.com, About, MIVA, LookSmart and other services to obtain results. To search for patents, users can visit the United States Patent and Trademark Office at www.uspto.gov/patft/, the Google patent search engine at www.google.com/patents, or FreePatentsOnline at www.freepatentsonline.com.

Search engines employ software ("spiders" or "crawlers") to automatically scan Web documents and construct databases as to what is to be found where. When one enters a query using a particular search engine, the request is checked against a database, and the best-matching results are returned.

Unless a Web document's author specifies the keywords for the document (for instance, by using "meta tags"), it is the job of the search engine to decide on them, isolating and indexing those words that appear to be noteworthy. Some search engines index every word on every page to be sure that nothing has been missed; others are more selective, using a set of rules as to what words are generally important. In this latter case, the search engine may give special importance to words in the title of a page, or words used near the beginning of a document, or words that are repeated numerous times throughout.

Many search engines index Web documents by examining the meta tags located in the HTML code at the beginning of a document. This allows the author to specifically identify a number of keywords for indexing purposes. However, because this feature is frequently misused to obtain falsely high rankings (a process sometimes known as "spamdexing") many search engines such as Google do not rely on meta tags. The heart of Google's search algorithm is PageRank™, a system for ranking pages developed

CAN J ANESTH 2007 / 54: 5 / pp 336–341

From the Department of General Anesthesiology, Cleveland Clinic, Cleveland, Ohio, USA.
Address correspondence to: Dr. D. John Doyle, Department of General Anesthesiology – E31, Cleveland Clinic, 9500 Euclid Avenue E31, Cleveland, Ohio, USA 44195. E-mail: doylej@ccf.org
Support was provided solely from institutional and/or departmental sources. The author has no financial relationship with any products, devices or drugs mentioned in this report.

by Google founders Larry Page and Sergey Brin. In essence, PageRank relies on a candidate page's link relations with other pages (amongst other things) as an indicator of an individual page's value.

Not all of the Web is accessible to search engines. For instance, if a Web page has never been linked to any other page, search engine spiders will not find it. Thus, the only way a newly constructed page that that has never been linked can appear in a search engine is for its address to be sent by an individual (such as the page's owner) to the various search engine authorities. All search engines offer easy ways to do this. That part of the Web that has not been accessed by search engine spiders is called the "Invisible Web"; it is estimated to be larger than the visible portion of the Web.

PubMed

PubMed (www.pubmed.gov) is a popular search engine service of the U.S. National Library of Medicine (NLM) that provides free access to citations from the biomedical literature, as well as consumer health information, research tools and access to a variety of molecular biology resources. The service covers over 4,800 journals published in more than 70 countries, covering a time span from 1966 to the present.

PubMed offers two general search strategies: searching by medical subject headings (MeSH) *vs* text word. In a study of the otolaryngology-head and neck surgery literature by Chang *et al.*[5] the authors found that the "MeSH search provided a more efficient search than the text word search" and recommended that users begin their search using MeSH terms. However, they also recommended that once a key article was identified, that users should use the "related articles" feature to get more citations.

A very helpful synopsis to using PubMed's many search features is available in PDF format at http://nnlm.gov/training/resources/pmtri.pdf. A version called PubMed for Handhelds is available at http://pubmedhh.nlm.nih.gov/nlm/. A recent review article on the use of PubMed by Vincent *et al.*[6] will be of value for users wanting to master PubMed's advanced features.

Users unhappy with PubMed may wish to try Hubmed (www.hubmed.org), "an alternative interface to the PubMed medical literature database". Special features of HubMed include "date- or relevance-ranked search results; Web feeds for regular updates of published literature matching any search; clustering and graphical display of related articles; expansion of query terms; direct export of citation metadata in many formats; linking of keywords to external sources of information; manual categorization (tagging) and storage of interesting articles".

Google Scholar

Released in late 2005, Google Scholar (scholar.google.com) is a Web search engine that indexes the full-text of much of the academic literature, covering a rich variety of disciplines. For physicians, Google Scholar provides a much more focused academic searching tool than the parent Google search engine. However, articles published by Elsevier, the world's largest scientific publisher, are not included in its collection; these articles may be accessed either from Elsevier's freely available Scirus service (scirus.com) or their subscription-based Scopus service (scopus.com). Note also that articles found by Google Scholar are not necessarily available for free download, especially for recently published material.

Google Scholar offers a number of helpful advanced features. For example, one can specify keywords which must appear in both the article and the publication name or restrict search results within one or more of seven general areas of research. For users of bibliographic management systems, Google Scholar supports the importation of citations in RefWorks, RefMan, EndNote, and BibTeX formats. One especially welcome feature of Google Scholar is also available on all the Google family sites: spelling correction. For example, if you search for "glycopyrolate" it will ask "Did you mean: glycopyrrolate" while still displaying a selection of found citations, all containing the spelling error. Google Scholar also has a "cited by" feature that lists citations of articles that have cited the article being viewed, as well as a "related articles" feature that provides a list of similar articles.

Windows Live Academic

Windows Live Academic (academic.live.com) is a part of the Windows Live Search group of tools (www.live.com). Although similar to Google Scholar, it is not nearly as rich in search options. Still, it has many useful features, such as a "slider" to adjust the amount of detail displayed; the ability to group and sort results by author, journal, conference and date; and author "live links" that allow one to automatically search for articles from a particular author by simply clicking on the hyperlink of the author's name. Of interest, rather than searching the Internet for scholarly material, Windows Live Academic search results come directly from "trusted sources", such as publishers of academic journals.

HighWire Press

Another alternative to PubMed that may appeal to users is HighWire Press, a division of the Stanford University Libraries. HighWire Press now hosts over

1,000 journals online, including the Canadian Journal of Anesthesia. It may be visited at highwire.stanford.edu. In a study comparing PubMed and HighWire Press[7] the authors found that "using HighWire Press resulted in a higher likelihood of retrieving the desired article and higher number of search results than the same search on PubMed" while PubMed "was faster than HighWire Press in delivering search results".

Other applications: toolbars and searching for images

Individuals preparing illustrated lectures (for instance, using PowerPoint) often make use of "image" search engines to obtain downloadable graphic images that can be pasted into a presentation. Examples of image search engines include Yahoo (http://images.search.yahoo.com), Google Image Search (http://images.google.com) and Alta Vista Image Search (http://www.altavista.com/image).

Users should remember, however, that while these images may be freely downloadable, unless they are in the public domain or have an appropriate Creative Commons license, all images remain the intellectual property of their lawful owners and their use in commercial publications usually requires permission.[8]

Finally, search toolbars allow one to search and navigate the Web without first going directly to a search engine. Google Toolbar (http://toobar.google.com) is one example; in addition to providing Web searches, it offers a "pop-up" blocker, has the capability to fill out Web forms semi-automatically, offers a spelling checker for Web mail, and can translate Web pages into English.

In conclusion, many electronic portals exist to access the medical literature online. While general search engines may be effective, upgrades to long-established search engines such as PubMed, and new academic search engines such as Google Scholar, HighWire Press and Windows Live Academic offer an array of user-friendly and advanced features that will benefit physicians accessing the medical literature online.

Moteurs de recherche pour la recherche universitaire : brève mise à jour pour les cliniciens

L'information médicale n'est plus enfermée dans les bibliothèques et est maintenant aisément accessible en ligne par le biais d'articles de revues électroniques, de mises à jour professionnelles, de guides de pratique, de guides de soins basés sur des données probantes, et de nombreuses autres sources. Toutefois, cette ample sélection de ressources peut constituer un défi pour certains cliniciens, qui se sentiraient engloutis sous les options disponibles. Pour compliquer le tout, de nouvelles formes de publication connaissent une popularité grandissante.[1-4] Bien que les véhicules traditionnels servant à la diffusion d'informations scientifiques, tels que les livres et les journaux sur papier, demeurent populaires, les universitaires écrivent également des articles pour des journaux électroniques et publient eux-mêmes des articles sur le Web ainsi que du matériel multimédia comme des vidéos didactiques en format CD-ROM et DVD.

La disponibilité de matériel scientifique dans une si grande diversité de formats induit toutefois de nouveaux défis. C'est particulièrement le cas pour les cliniciens très occupés qui tentent de se tenir au courant de la manière dont les derniers développements informatiques peuvent influencer leurs besoins d'érudition et de formation médicale continue. Dans un tel contexte, il est utile de réfléchir au fonctionnement des moteurs de recherche ainsi que d'examiner la manière dont les moteurs de recherche scientifiques et autres outils informatiques se comparent aux moteurs de recherche généraux lors d'une recherche dans la littérature médicale.

Les moteurs de recherche

Les moteurs de recherche sont devenus omniprésents pour les utilisateurs d'Internet. Le moteur de recherche Google (www.google.com) est le plus connu et le plus utilisé, mais il en existe de nombreux autres à disposition, par exemple Alta Vista (www.altavista.com), Lycos (www.lycos.com), et AOL Search (search.aol.com), pour ne citer que ceux-ci. Malgré cette ample sélection, de nombreux utilisateurs Web n'utilisent toutefois que Google, à l'exclusion de tous les autres moteurs de recherche. Ceci est en grande partie dû à

la vitesse et à la précision exceptionnelles de Google, à son excellente interface utilisateur et à cause de son énorme base de données (plus d'un milliard de pages répertoriées).

Les 'méta'-moteurs de recherche combinent et intègrent les résultats de plusieurs moteurs de recherche dans un seul moteur. Par exemple, MetaCrawler (http://www.metacrawler.com) cherche dans Google, Yahoo Search, MSN Search, Ask.com, About, MIVA, LookSmart et d'autres services pour obtenir des résultats. Pour rechercher des brevets, les utilisateurs peuvent visiter le United States Patent and Trademark Office (Bureau des brevets et marques déposées des États-Unis) sous www.uspto.gov/patft/, le moteur de recherche de brevets de Google sous www.google.com/patents, ou FreePatentsOnline sous www.freepatentsonline.com.

Les moteurs de recherche utilisent des logiciels (robots nommés « spiders » ou « crawlers ») afin de parcourir automatiquement des documents Web et de construire des bases de données quant à ce qui peut être trouvé à quel endroit. Lorsqu'un utilisateur saisit une demande en se servant d'un moteur de recherche en particulier, la demande est vérifiée par rapport à une base de données, et les résultats les plus pertinents sont présentés.

Sauf si l'auteur du document Web spécifie les mots clés du document (en utilisant par exemple des 'balises Méta'), c'est au moteur de recherche de décider quels sont ces mots clés en isolant et répertoriant les mots qui semblent dignes d'intérêt. Certains moteurs de recherche répertorient chaque mot de chaque page afin de s'assurer que rien ne manque ; d'autres sont plus sélectifs et se servent d'un ensemble de règles afin de déterminer quels mots sont en général importants. Dans ce cas, le moteur de recherche peut accorder une importance spéciale aux mots du titre d'une page, ou aux mots utilisés vers le début du document, ou aux mots souvent répétés dans le document.

De nombreux moteurs de recherche répertorient les documents Web en examinant les balises Méta situées dans le code HTML au début d'un document. Ceci permet à l'auteur d'identifier spécifiquement un nombre de mots clés dans un objectif de classement. Toutefois, parce que cette fonction est souvent détournée afin d'obtenir un classement élevé abusif (une procédure parfois nommée 'spamdexing', ou référencement abusif), nombre de moteurs de recherche tels que Google ne s'appuient pas sur les balises Méta. Le cœur de l'algorithme de recherche de Google est PageRank™, un système de classement des pages développé par les fondateurs de Google Larry Page et Sergey Brin. PageRank s'appuie, entre autres, sur les relations de liens d'une page candidate à d'autres pages comme indicateur de la valeur d'une page individuelle.

Le Web n'est pas complètement accessible aux moteurs de recherche. Par exemple, si une page Web n'a jamais été reliée à aucune autre page, les robots des moteurs de recherche ne la trouveront pas. Dès lors, le seul moyen pour une page nouvellement construite, qui n'a jamais été reliée à aucune autre, d'apparaître dans un moteur de recherche est si son adresse est envoyée par une personne (tel que le propriétaire de la page) aux différentes administrations des moteurs de recherche. Tous les moteurs de recherche offrent des moyens faciles d'effectuer cette démarche. La partie du Web qui n'a pas été touchée par les robots des moteurs de recherche est appelée la 'Toile Invisible' ; on estime qu'elle est plus vaste que la partie visible du Web.

PubMed

PubMed (www.pubmed.gov) est un service de moteur de recherche populaire de la U.S. National Library of Medicine (NLM) (Bibliothèque nationale de médecine des États-Unis) qui fournit un accès gratuit à des citations de la littérature biomédicale ainsi que des informations sur la santé aux consommateurs, des outils de recherche et un accès à une diversité de ressources sur la biologie moléculaire. Le service couvre plus de 4 800 journaux et revues publiés dans plus de 70 pays, allant de 1966 à nos jours.

PubMed met à disposition deux stratégies de recherche générales : la recherche par sujet (MeSH) ou par mot du texte. Dans une étude de Chang *et coll.*[5] portant sur la littérature concernant la chirurgie otorhinolaryngologique, de la tête et du cou, les auteurs ont découvert que la « recherche MeSH a permis une recherche plus efficace que la recherche par mot du texte » et ont conseillé aux utilisateurs de commencer leur recherche en se servant des termes MeSH. Néanmoins, ils ont également recommandé aux utilisateurs de se servir de la fonction 'articles apparentés' une fois l'article principal identifié, afin d'obtenir davantage de citations.

Un résumé très utile quant à l'utilisation des nombreuses fonctions de recherche de PubMed est disponible en format PDF sur http://nnlm.gov/training/resources/pmtri.pdf. Une version nommée PubMed for Handhelds (PubMed pour appareils de poche) est disponible sur http://pubmedhh.nlm.nih.gov/nlm/. Un compte rendu récent sur l'utilisation de PubMed par Vincent *et coll.*[6] s'avérera profitable aux utilisateurs souhaitant maîtriser les options avancées de PubMed.

Les utilisateurs insatisfaits de PubMed pourraient tester HubMed (www.hubmed.org), « une interface de rechange à la base de données de littérature médicale de PubMed. » Parmi les fonctions spéciales de HubMed citons « des résultats de recherche classés par date ou par pertinence, des apports Web pour des mises à jour régulières de la littérature publiée pour correspondre à n'importe quelle recherche, le regroupement et l'affichage graphique d'articles apparentés, l'expansion des termes de recherche, l'exportation directe des méta-données de citation dans de nombreux formats, la connexion de mots clés à des sources d'information externes, la catégorisation manuelle (tagging, ou balisage), et la conservation d'articles intéressants ».

Google Scholar

Apparu à la fin de 2005, Google Scholar (scholar.google.com) est un moteur de recherche Web qui répertorie le texte entier d'une importante partie de la littérature scientifique, couvrant une variété importante de disciplines. Google Scholar fournit aux médecins un outil de recherche scientifique bien plus pointu que le moteur de recherche parent Google. Cependant les articles publiés par Elsevier, le plus important éditeur scientifique au monde, ne sont pas inclus dans cette collection ; on peut accéder à ces articles soit par le service gratuit d'Elsevier Scirus (scirus.com), ou par son service par abonnement Scopus (www.scopus.com). Il est à noter que les articles trouvés par Google Scholar, plus particulièrement les publications récentes, ne sont pas forcément disponibles pour le téléchargement gratuit.

Google Scholar offre de nombreuses options avancées utiles. Par exemple, on peut spécifier les mots clés devant apparaître dans l'article ainsi que dans le nom de la publication, ou limiter les résultats de recherche dans un ou plusieurs domaines généraux de recherche, au nombre de sept. Pour les utilisateurs se servant de systèmes de gestion bibliographique, Google Scholar permet l'importation de citations dans les formats RefWorks, RefMan, EndNote et BibTeX. Une option très intéressante de Google Scholar, également disponible sur tous les sites de la famille Google, est le correcteur d'orthographe. Par exemple, si une recherche de « glycopyrolate » est lancée, le moteur va proposer « Essayez avec cette orthographe : glycopyrrolate » tout en affichant une sélection des résultats obtenus contenant tous la faute d'orthographe. Google Scholar offre également une option « cité par » qui énumère les citations d'articles ayant fait référence à l'article visionné, ainsi qu'une option « articles apparentés » qui fournit une liste d'articles similaires.

Windows Live Academic

Windows Live Academic (academic.live.com) fait partie du groupe d'outils de Windows Live Search (www.live.com). Bien que similaire à Google Scholar, il n'offre pas autant d'options de recherche. Mais il possède tout de même plusieurs options utiles, comme par exemple un 'glisseur' (slider) qui permet d'ajuster la quantité de détails affichée ; la possibilité de regrouper et de classer les résultats par auteur, journal, conférence et date, et des 'liens en direct' aux auteurs qui permettent à l'utilisateur de chercher automatiquement les articles d'un auteur en particulier en cliquant simplement sur le lien hypertexte du nom de l'auteur. Plutôt que de fouiller l'Internet à la recherche de matériel didactique, les résultats de recherche de Windows Live Academic proviennent directement de 'sources sûres', telles que d'éditeurs de revues scientifiques, ce qui peut être intéressant.

HighWire Press

HighWire Press, une filiale des bibliothèques de l'université de Stanford, constitue une autre alternative attrayante à PubMed. HighWire Press héberge actuellement plus de 1000 journaux en ligne, y compris le Journal canadien d'anesthésie. Le site peut être trouvé sur highwire.stanford.edu. Dans une étude comparant PubMed et HighWire Press,[7] les auteurs ont découvert que « l'utilisation de HighWire Press résultait en une probabilité plus élevée de retrouver l'article désiré ainsi qu'un nombre plus élevé de résultats de recherche que la même recherche sur PubMed », alors que PubMed « était plus rapide que HighWire Press pour la livraison des résultats de recherche ».

Autres applications : barres d'outils et recherche d'images

Les personnes préparant des conférences illustrées (par exemple en se servant de PowerPoint) ont souvent recours à des moteurs de recherche 'd'images' afin d'obtenir des images graphiques téléchargeables qui peuvent être insérées dans une présentation. Quelques exemples de moteurs de recherche d'images : Yahoo (http://images.search.yahoo.com), Google Image Search (http://images.google.com) et Alta Vista Image Search (http://www.altavista.com/image).

Il faut cependant que les utilisateurs soient conscients que ces images, même si elles sont disponibles gratuitement pour le téléchargement, et à moins qu'elles ne fassent partie du domaine public ou possèdent une autorisation Creative Commons appropriée, demeurent la propriété intellectuelle de leurs propriétaires légitimes et que leur utilisation dans des publications commerciales requiert généralement une autorisation.[8]

En dernier lieu, les barres d'outils de recherche permettent de fouiller et de naviguer sur le Web sans passer d'abord par un moteur de recherche. La barre d'outils Google (http://toolbar.google.com) en est un exemple ; en plus de fournir des recherches sur le Web, cette barre d'outils fournit un bloqueur de fenêtres intrusives (pop-up), peut utiliser la saisie semi-automatique pour remplir des formulaires en ligne, fournit un correcteur d'orthographe pour le courrier électronique et peut traduire des pages Web vers l'anglais.

Pour conclure, de nombreux portails électroniques existent pour accéder à la littérature médicale en ligne. Bien que des moteurs de recherche généraux puissent être efficaces, les mises à niveau de moteurs de recherche bien établis tels que PubMed ou les nouveaux moteurs de recherche scientifiques tels que Google Scholar, HighWire Press et Windows Live Academic offre un éventail d'options avancées faciles à utiliser qui profiteront aux médecins se servant de la littérature médicale en ligne.

References

1 *Davidson LA*. The end of print: digitization and its consequence--revolutionary changes in scholarly and social communication and in scientific research. Int J Toxicol 2005; 24: 25–34.

2 *Frank M*. Access to the scientific literature--a difficult balance. N Engl J Med 2006; 354: 1552–5.

3 *Sperr EV Jr*. Libraries and the future of scholarly communication. Mol Cancer 2006; Nov 7 5: 58. Available from URL; (http://www.molecular-cancer.com/content/5/1/58).

4 *Albert KM*. Open access: implications for scholarly publishing and medical libraries. J Med Libr Assoc 2006; 94: 253–62.

5 *Chang AA, Heskett KM, Davidson TM*. Searching the literature using medical subject headings versus text word with PubMed. Laryngoscope 2006; 116: 336–40.

6 *Vincent B, Vincent M, Ferreira CG*. Making PubMed searching simple: learning to retrieve medical literature through interactive problem solving. Oncologist 2006; 11: 243–51. Available from URL; (http://theoncologist.alphamedpress.org/cgi/content/full/11/3/243).

7 *Vanhecke TE, Barnes MA, Zimmerman J, Shoichet S*. PubMed vs. HighWire Press: a head-to-head comparison of two medical literature search engines. Comput Biol Med 2006 Dec 19; [Epub ahead of print].

8 *Farrell A*. Getting the picture: medical images on the Web. Med Ref Serv Q 2006; 25: 47–54.