

La conception de nouveaux moteurs de recherche

En juillet 1998, le moteur *AltaVista* a récupéré 170 millions de pages (sur un total estimé à 350 à cette période), dont 125 millions ont été indexées, ce qui représente environ 800 gigaoctets de texte brut (un gigaoctet, ou milliards d'octets, correspond au contenu en texte de l'intégrale d'une grosse encyclopédie), soit, après traitement, à un index de près de 250 gigaoctets installé sur plus d'une vingtaine de machines de très haut de gamme (dix processeurs et dix gigaoctets de mémoire vive par machine). Ces machines sont interrogées 37 millions de fois par jour, en semaine, avec un temps de réponse moyen de 0,6 seconde. Ces quelques chiffres aident à apprécier la difficulté de la conception et de la mise en œuvre d'un moteur de recherche généraliste.

Afin d'améliorer la pertinence des documents retournés, les moteurs de recherche mettent en œuvre plusieurs solutions. L'une concerne les requêtes elles-mêmes, par exemple la correction orthographique ou la détection automatique des phrases. Ainsi, un moteur fournira de meilleures réponses si la requête est l'expression «effet de serre» plutôt que les trois mots «effet», «de» et «serre», car l'expression est plus discriminante que les trois mots pris séparément. Ce fait est bien connu des linguistes : le sens des textes est en grande partie contenu dans les groupes nominaux. Depuis l'été 1998, *AltaVista* reconnaît automatiquement les phrases, ce qui améliore sensiblement la qualité du moteur.

Cependant, le principal levier dont dispose l'architecte d'un moteur de recherche reste encore l'amélioration de l'algorithme d'évaluation de pertinence (le *ranking*). Sur le réseau, les algorithmes fondés sur la mise en correspondance des mots des requêtes et des mots contenus dans les documents trouvent rapidement leurs limites : documents volontairement biaisés par les «spammers», polysémie importante, nombre de documents trop élevé, duplication anarchique des documents, etc. Tout concourt à faire échouer les algorithmes classiques d'évaluation et de pertinence. Pour ces raisons de nouveaux algorithmes,

tels ceux utilisés par *Clever* ou *Goggle*, sont élaborés avec des premiers résultats prometteurs.

Un autre élément important de la conception d'un moteur de recherche réside en la prise en compte de la polysémie, problème particulièrement épineux sur l'Internet. Un acronyme comme «BSE» signifie, entre autres, *Bovine Spongiform Encephalopathy* («encéphalopathie spongiforme bovine»), *Breast Self Examination* («auto-examen des seins»), ou encore *Bombay Stock Exchange* («Bourse de Bombay»). Lorsqu'un utilisateur effectue la requête «BSE», il est donc essentiel que le moteur lui fournisse un moyen de préciser de manière simple l'objet réel de sa recherche. C'est ainsi qu'avec notre équipe de l'École des mines de Paris, nous avons mis au point la technique *Cow9*, utilisée par le moteur *AltaVista* depuis 1997.

Cette technique permet d'affiner les requêtes à l'aide d'une liste thématique construite automatiquement à partir des résultats des recherches. Les thèmes non pertinents peuvent être exclus d'un simple clic de souris, tandis qu'un zoom est possible sur les thèmes jugés les plus pertinents. Une approche similaire, permettant le classement des résultats des recherches dans des dossiers thématiques (dont la liste est établie manuellement), a plus récemment été déployée sur le moteur de recherche *NorthernLight*.

D'autres approches sont utilisées sur certains moteurs pour aider les utilisateurs dans leurs recherches. Citons par exemple la fonction *What's Related* de *Netscape*, qui propose des liens vers des pages au contenu proche d'une page donnée.

Des approches différentes de celles qui sont suivies par les moteurs améliorent aussi la recherche d'information pertinente, comme la recherche par nom de marques de *RealNames*, la reformulation de requêtes en questions aux réponses connues, voie suivie par *AskJeeves*, l'approche des anneaux (*rings* en anglais) qui consiste à relier entre eux par des liens hypertextes les sites aux contenus voisins (ce qui ne résout toutefois pas le problème de trouver un premier site situé

dans l'anneau), ou encore les méta-moteurs qui interrogent en parallèle plusieurs moteurs de recherche classiques et fusionnent ensuite de manière intelligente les résultats.

Sachant que le taux de couverture d'un moteur (c'est-à-dire la proportion de documents qui se trouve effectivement dans la base de données du moteur) est au maximum de 30 pour cent, l'utilisation de métamoteurs peut sembler intéressante. Malheureusement, la pratique montre que la valeur ajoutée de ces outils est bien faible et que, malgré son taux de couverture limité, un moteur classique estime mieux la pertinence des documents qu'un métamoteur.

Le monde des outils de navigation est en perpétuelle mutation, et les défis auxquels sont confrontés les architectes des moteurs de recherche sont légion. D'abord, comment espérer maintenir, à un coût raisonnable, un taux de couverture du Web suffisant quand le nombre de documents disponibles augmente de six pour cent par mois? Comment fournir des réponses pertinentes à un utilisateur qui fournit très peu d'indications sur ce qu'il recherche vraiment? Et comment amener, sans le décourager, l'utilisateur à en dire plus?

La prise en compte des liens hypertextes, comme avec *Clever*, améliore la pertinence des documents retournés, mais ces solutions sont loin d'être parfaites et on ne fera pas l'économie d'une réflexion approfondie sur les impacts socio-économiques et culturels de ces méthodes, notamment les impacts liés au risque de consolidation des positions dominantes et au risque de perte du «service universel» (c'est-à-dire la garantie de pouvoir se faire entendre sur le réseau) pour le citoyen et les minorités. Il faudra d'autres approches fondées sur la collaboration, comme *The Open Directory Project*, pour faire le pendant aux guides commerciaux, un peu à la manière dont le logiciel libre, tel le tandem GNU/Linux, est en train de s'imposer dans une industrie du logiciel dominée par quelques grands acteurs.

François BOURDONCLE
École des mines de Paris