

de 140 mots environ. Diverses améliorations pondèrent l'importance des mots selon qu'ils apparaissent dans les titres, au début des chapitres ou en caractères plus gros que le corps du texte.

De nombreux sites abusent de ces règles, afin que leurs pages soient bien classées. C'est pour cette raison que le titre de certaines pages est étrangement rédigé, tel «billets avion soldés, billets avion soldés, billets avion soldés». D'autres sites incluent plusieurs fois dans leurs pages des phrases judicieusement choisies, dans

des polices et des couleurs invisibles à l'œil. De telles pratiques, qui inondent les moteurs de recherche avec des informations non pertinentes (le *spamming*), compliquent la tâche des gestionnaires de ces moteurs (le mot anglais *spam* vient d'un épisode de la série télévisée *Monty Python's Flying Circus*, où le mot *Spam*, une marque de jambonneau, couvre les dialogues).

Même sans cette difficulté, le bien-fondé des méthodes de recherche textuelles est douteuse : des pages qui intéresseraient beaucoup un utilisateur

ne contiennent pas le mot qui a été transmis au moteur de recherche, tandis que d'autres pages qui contiennent le mot spécifié sont sans intérêt. Cette seconde difficulté provient de l'abondance de synonymes et de mots polysémiques (des mots qui ont plusieurs sens) : une demande d'information sur les «automobiles» laissera de côté une kyrielle de pages qui évoquent les «voitures». En raison des polysémies, une simple requête sur «jaguar» donnera des milliers de pages concernant, entre autres, l'automobile de ce nom, le félin et une équipe de football américain.

On améliore les recherches en utilisant les relations sémantiques entre les mots. Certains moteurs de recherche utilisent de telles relations, nommées «réseaux sémantiques», identifiées par les linguistes. À la réception de la requête «automobile», un moteur de recherche couplé à un réseau sémantique détermine d'abord qu'«automobile» est un mot équivalent à «voiture», puis il donne la liste des pages qui contiennent l'un des deux mots. Toutefois, cette méthode est à double tranchant : elle résout le problème des synonymes, mais aggrave celui de la polysémie.

De surcroît, l'établissement des réseaux sémantiques est une tâche redoutable, car, surtout sur le réseau Internet, multiculturel, les néologismes fleurissent. Des mots sont apparus, tels que «FAQ» (Foire Aux Questions), «zines» (diminutif de fanzines, de petits journaux diffusés sur le réseau) ou «internaute» ; d'autres mots ont de nouvelles acceptions, tels «surfer» ou «naviguer».

Recherche avec les hyperliens

Le projet *Clever* de la Société IBM voulait résoudre ces difficultés. Très tôt, nous avons compris que l'indexation ou la récupération de pages perdait beaucoup à ne se fonder que sur le texte : plus d'un milliard de liens ne révèlent-ils pas les relations entre les pages. Comment tirer parti de ces informations ?

Lorsqu'ils recherchent «Harvard», par exemple, beaucoup d'internautes désirent en fait des informations sur les équipes sportives qui s'y trouvent. Toutefois, plus d'un million de pages contiennent ce mot, et la page d'accueil de l'Université Harvard n'est pas celle qui l'utilise le plus souvent, de sorte qu'elle ne sera pas la première choisie, ni même jugée importante à l'aide des fonctions de

La «cinétique chimique» sur l'Internet

Tous les moteurs de recherche ne donnent pas de résultats équivalents. Nous avons comparé des moteurs de recherche en analysant les réponses à la requête «cinétique chimique» (l'étude de la vitesse de réactions chimiques).

Plusieurs objectifs se dégagent pour cette recherche : le moyen d'établir un contact entre enseignants pour une mise en commun d'expériences pédagogiques, la recherche des enseignements de troisième cycle dispensés dans cette matière et la recherche de documentation.

Nous avons comparé les résultats donnés par une sélection de moteurs francophones (*Ecila*, *Eureka*, *Excites*, *Francité*, *Lokace*, *Yahoo! France*), à laquelle nous avons ajouté deux moteurs anglophones (*AltaVista* et *GoTo*) et deux métamoteurs, des moteurs qui recherchent dans plusieurs moteurs (*MetaCrawler*, *Savvy Search* et *Trouvez*).

La recherche s'est effectuée uniquement par mot clé, car cette méthode est commune à tous les moteurs. Plusieurs requêtes ont été saisies avec des contraintes de plus en plus fortes. Nous avons d'abord fait la requête *cinétique*, puis *+cinétique +chimique* (qui recherche toutes les pages où les deux termes sont présents) et enfin *cinétique chimique*, qui renvoie les pages où l'expression «cinétique chimique» est présente.

Nous avons classé les moteurs selon deux critères : la quantité de pages indexées et la qualité (renvois

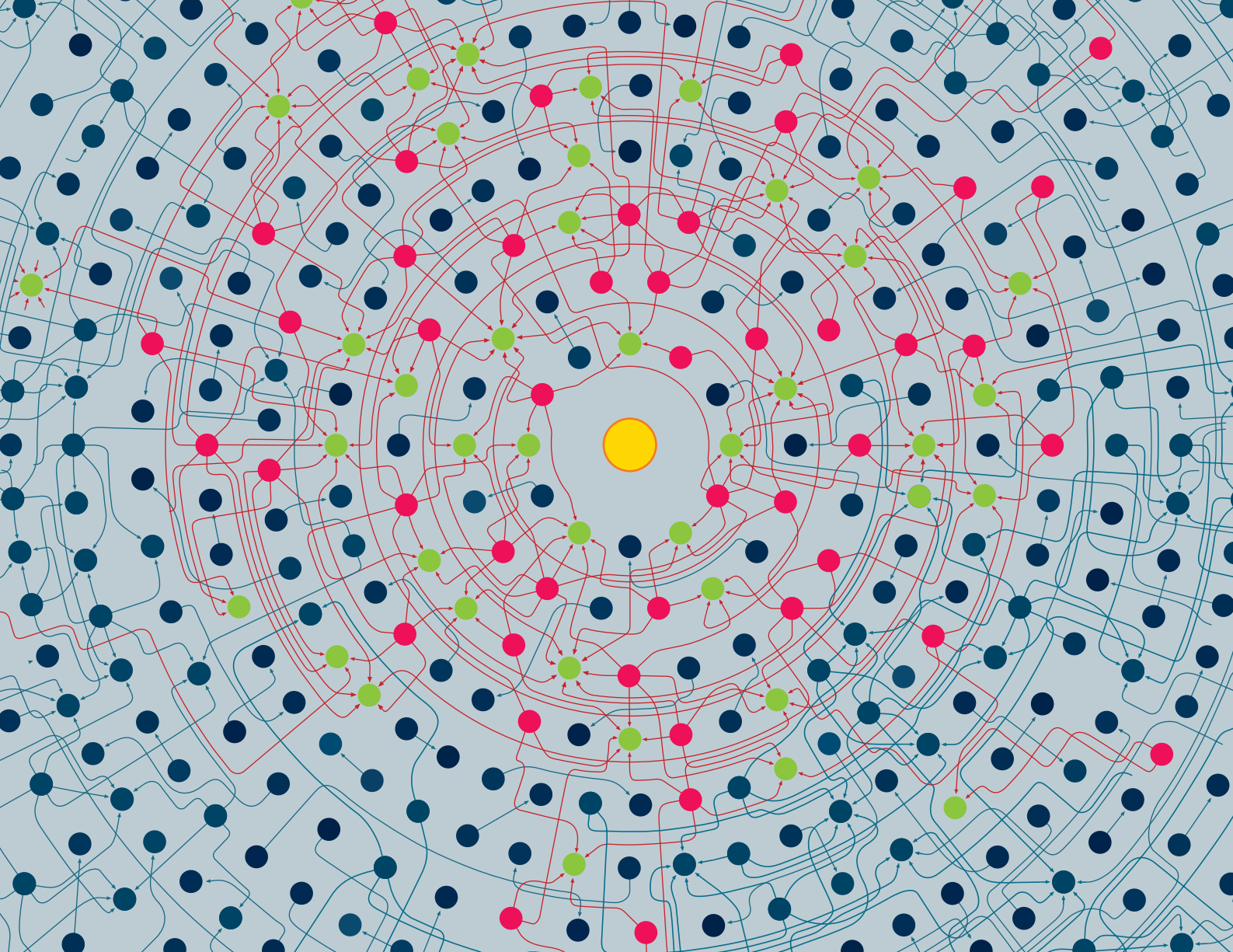
de pages traitant de cinétique chimique jugées pertinentes).

Pour l'étude quantitative, trois moteurs sortent du lot : ce sont *Carrefour*, *Nomade* et *AltaVista*. Ils renvoient 500 à 6 000 pages selon les contraintes imposées. D'autres moteurs donnent de bons résultats (entre 50 et 200 pages) et, enfin, d'autres sont insuffisants.

Pour l'étude qualitative, nous avons prédéfini comme pages pertinentes cinq pages d'universités ou d'écoles qui enseignent la cinétique chimique (un des objectifs précités) : l'Université catholique de Louvain (Belgique), l'École nationale supérieure des industries chimiques à Nancy, l'Université de Pau, l'Université Pierre et Marie Curie (Paris VI) et l'Université de Laval (Canada). Avec la saisie de «cinétique chimique», nous avons classé les différents moteurs en cinq groupes, du meilleur au moins bon, les meilleurs étant ceux qui contiennent les cinq pages d'accueil des institutions précitées parmi les 30 premières pages retournées. Les plus performants sont *Trouvez*, *SavvySearch*, *Metacrawler* et *GoTo*.

Les trois moteurs qui renvoyaient le plus grand nombre de pages ne sont pas les plus performants sur la qualité de l'information. On retrouve ainsi les préoccupations du projet *Clever* : la quantité n'est pas toujours liée à la qualité, d'où l'importance d'un moyen de retrouver une information pertinente.

J. BENAÏM et V. VIOSSAT
Université P. et M. Curie, Paris



2. L'INFORMATION EST STRUCTURÉE par les sites de référence et par les sites pivots. Les sites de référence (●) sont des sites qui contiennent de l'information et, en conséquence, sont référencés par de nombreuses pages. Sur le sujet des droits de l'homme,

par exemple, la page d'accueil d'*Amnesty International* est un tel site. Les pivots (●) sont des sites qui pointent vers de nombreux sites de référence ; ce sont par exemple des listes de liens ou les rubriques «Mes sites préférés» des pages personnelles.

classement classiques. Aucune caractéristique interne de cette page ne révèle son importance.

En effet, les concepteurs des sites Internet ont des intentions précises. Les grandes entreprises, par exemple, préfèrent souvent transmettre une image plutôt que de décrire leur activité réelle. Par exemple, la page d'accueil d'*IBM* ne contient pas le mot «ordinateur». Face à ce type de pages, les techniques de recherche classiques échouent.

Les concepteurs des moteurs de recherche sont tentés d'intervenir directement pour résoudre ces problèmes : pour certaines questions, l'humain n'est-il pas plus efficace que l'ordinateur ?

Plusieurs moteurs de recherche utilisent cette approche. Un annuaire de sites comme *Yahoo!* ne contient même que des pages sélectionnées

manuellement. Cependant, le nombre de requêtes possibles est virtuellement infini. Comment un nombre fini d'experts humains mettrait-il à jour toutes ces listes de réponses prédéfinies, alors que le nombre de pages croît au rythme d'un million par jour ?

Nous avons pris le problème par un autre bout : en utilisant les hyperliens, nous avons mis au point une technique automatique qui trouve les sites qui font autorité sur des sujets larges. Cette idée tire parti de la fonction des hyperliens : ce sont eux qui structurent les millions de pages en un tissu de connaissances. C'est grâce à eux que l'utilisateur navigue sur le réseau et découvre fortuitement des informations utiles, en passant par des pointeurs et par des recommandations de personnes qu'il n'a jamais rencontrées.

Cette approche est fondée sur l'hypothèse que chaque lien matérialise une parenté sémantique entre la page initiale et la page vers laquelle il pointe. Par exemple, un militant des droits de l'homme dont la page pointe vers la page d'accueil d'*Amnesty International* approuve clairement ce site.

Certains liens existent aussi pour de simples raisons de navigation («Cliquez ici pour retourner au menu principal»), de publicité («Les vacances de vos rêves ne sont qu'à un clic de souris») ou pour marquer un désaccord («Allez sur ce site pour lire ce que cet imbécile raconte»). Néanmoins, nous pensons que, dans l'ensemble, c'est-à-dire lorsque l'on considère beaucoup de liens, les pointeurs sont une marque d'adhésion et confèrent aux sites un statut de référence.

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