

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