

personnes atteintes du syndrome de Kartagener. Ces personnes sont plus souvent sujettes aux infections respiratoires, car les cils qui repoussent normalement les micro-organismes hors des voies respiratoires sont absents et, de plus, ils ont souvent un situs

inversus. Les organes des souris porteuses d'une mutation d'une protéine qui compose les cils s'orientent aléatoirement.

En 1998, Nobutaka Hirokawa et ses collègues de l'Université de Tokyo ont observé que les cils des cellules de

nœud de souris tournent, dans le liquide qui entoure l'embryon, en sens inverse des aiguilles d'une montre. Ce mouvement crée un courant qui rassemble vers la gauche du nœud les facteurs importants tels que l'acide rétinoïque et les protéines Nodal et Lefty. Cette accumulation de liquide et de protéines sur le côté gauche correspond peut-être au signal qui déclenche la rupture de la symétrie embryonnaire. La façon dont nos organes internes se développent est donc commandée par une caractéristique architecturale des cellules du nœud : le sens de rotation des cils.

Ce processus dépend probablement de l'asymétrie des molécules qui dirigent le mouvement ciliaire. Cependant, pour la moitié des individus totalement dépourvus de cils au niveau du nœud, les organes s'organisent normalement : les cils ne sont pas nécessaires au développement des organes, mais interviennent plutôt dans leur orientation et dans leur positionnement.

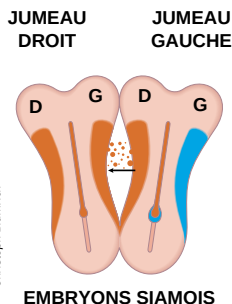
La compréhension de l'orientation gauche-droite au cours de l'embryogenèse a récemment progressé grâce à la découverte des gènes qui agissent de manière asymétrique chez l'embryon précoce. La nature exacte de l'initiation de l'asymétrie dans l'embryon reste inconnue, mais l'identification récente de protéines impliquées dans les événements suivants facilitera la découverte des protéines impliquées dans d'autres aspects du développement. La mise au point de nouvelles méthodes de diagnostic prénatal pour des pathologies liées à ces protéines sera alors possible.

Les siamois

Chez certains siamois, les organes se développent à une place anormale, qui dépend de l'endroit par où ils sont reliés. Pour la moitié des siamois soudés par le côté (siamois bicéphales), telles Abigail et Brittany Hensel (voir la photographie), l'un des membres de la paire, souvent celui qui est à droite, a son cœur à droite et non à gauche. Les siamois bicéphales se développent à partir d'un ovule unique fécondé, qui ne se divise pas complètement : les deux embryons qui se forment sont alors souvent parallèles (voir la figure). Selon l'équipe de Clifford Tabin et ses collègues de Harvard, cette proximité permet aux facteurs sécrétés, qui s'expriment dans un nœud de l'un des jumeaux, d'avoir une influence sur l'autre. La protéine nom-



Steve Weverka, Impact Visuals



mée Activine β B produite du côté droit du jumeau de gauche (*en rouge*), réprime probablement la production de la protéine Sonic Hedge-

hog du côté gauche du jumeau de droite. Celui-ci a donc autant de chances d'avoir un cœur à droite qu'à gauche.



Juan M. Hurle, University of Cantabria

6. LA VITAMINE A RÈGLE le développement du cœur (*en rouge*) de jeunes embryons de poulet. En présence d'une concentration normale d'acide rétinoïque (une forme de la vitamine A), le cœur s'incurve correctement (*à gauche*). Un excès d'acide rétinoïque provoque une déviation inverse (*à droite*). Les embryologistes pensent que l'acide rétinoïque favorise la régulation des gènes, tel le gène Nodal, qui gouvernent l'asymétrie gauche-droite.

Juan Carlos Ispisúa BELMONTE est professeur de biologie du développement à l'Université de Californie, à San Diego.

William B. WOOD, *Left-Right Asymmetry in Animal Development*, in *Annual Review of Cell and Developmental Biology*, vol. 13, pp. 53-82, 1997.

Michael LEVIN et Mark MERCOLA, *The Compulsion of Chirality : Toward an Understanding of Left-Right Asymmetry*, in *Genes and Development*, vol. 12, pp. 763-769, 15 mars 1998.

Ann F. RAMSDELL et H. Joseph YOST, *Molecular Mechanisms of Vertebrate Left-Right Development*, in *Trends in Genetics*, vol. 14, n° 11, pp. 459-465, novembre 1998.

Recherche intelligente sur l'Internet

LES MEMBRES DU PROJET CLEVER

La quantité d'informations en ligne sur le réseau Internet a tant augmenté que les outils actuels de recherche manquent d'efficacité. On améliore les recherches quand on analyse comment les pages du réseau mondial sont interconnectées.

Aujourd'hui, le réseau Internet donne accès à plusieurs centaines de millions de pages et s'enrichit d'un million de pages par jour. Plus d'un milliard de connexions entre les pages, les hyperliens, structurent cet impressionnant volume d'informations. Pour la première fois dans l'histoire, des millions de personnes ont accès, de leur domicile ou de leur bureau, aux productions intellectuelles d'une partie importante de la population de la planète.

Toutefois, en raison de son développement rapide et anarchique, Internet est un réseau d'informations sans organisation ni structure, de sorte que la recherche efficace d'information y est difficile. Le réseau est devenu un bric-à-brac gigantesque : les pages sont écrites par des personnes aux formations, aux intérêts ou aux motivations variés. Certaines pages ne contiennent que quelques caractères ; d'autres plusieurs centaines de milliers. Selon les cas, les informations sont justes, fausses, précises, floues ; la propagande avoisine les documents officiels et les inepties.

Les moteurs de recherche actuels cherchent des mots-clés présents dans les pages, mais ils identifient ainsi des dizaines de milliers de pages, dont la plupart sont inutiles. Comment localiser rapidement l'information voulue, et seulement elle, tout en étant sûr qu'elle soit authentique et fiable ?

Nous avons mis au point un nouveau type de moteur de recherche qui exploite la plus importante des ressources du réseau : ses innombrables hyperliens. Grâce à l'analyse de ces interconnexions, notre système localise

automatiquement deux types de pages : celles qui font autorité et celles qui servent de pivot. Les premières sont supposées être les meilleures sources d'information sur un sujet particulier ; les secondes sont des groupes de liens qui pointent vers les premières. Notre méthodologie permettra aux utilisateurs de localiser rapidement et efficacement la plupart des informations qu'ils désirent.

Les défis des moteurs de recherche

Les moteurs de recherche sont des programmes, accessibles aux utilisateurs du réseau, qui consultent une gigantesque mémoire où est enregistré pour chaque mot-clé présent dans les pages d'Internet, une liste des pages contenant ce mot. Un groupe de telles listes est nommé index. Ainsi, quand un utilisateur cherche des informations sur la revue *Pour la Science*, le moteur de recherche lui communique la liste de toutes les pages qui contiennent ce groupe de mots.

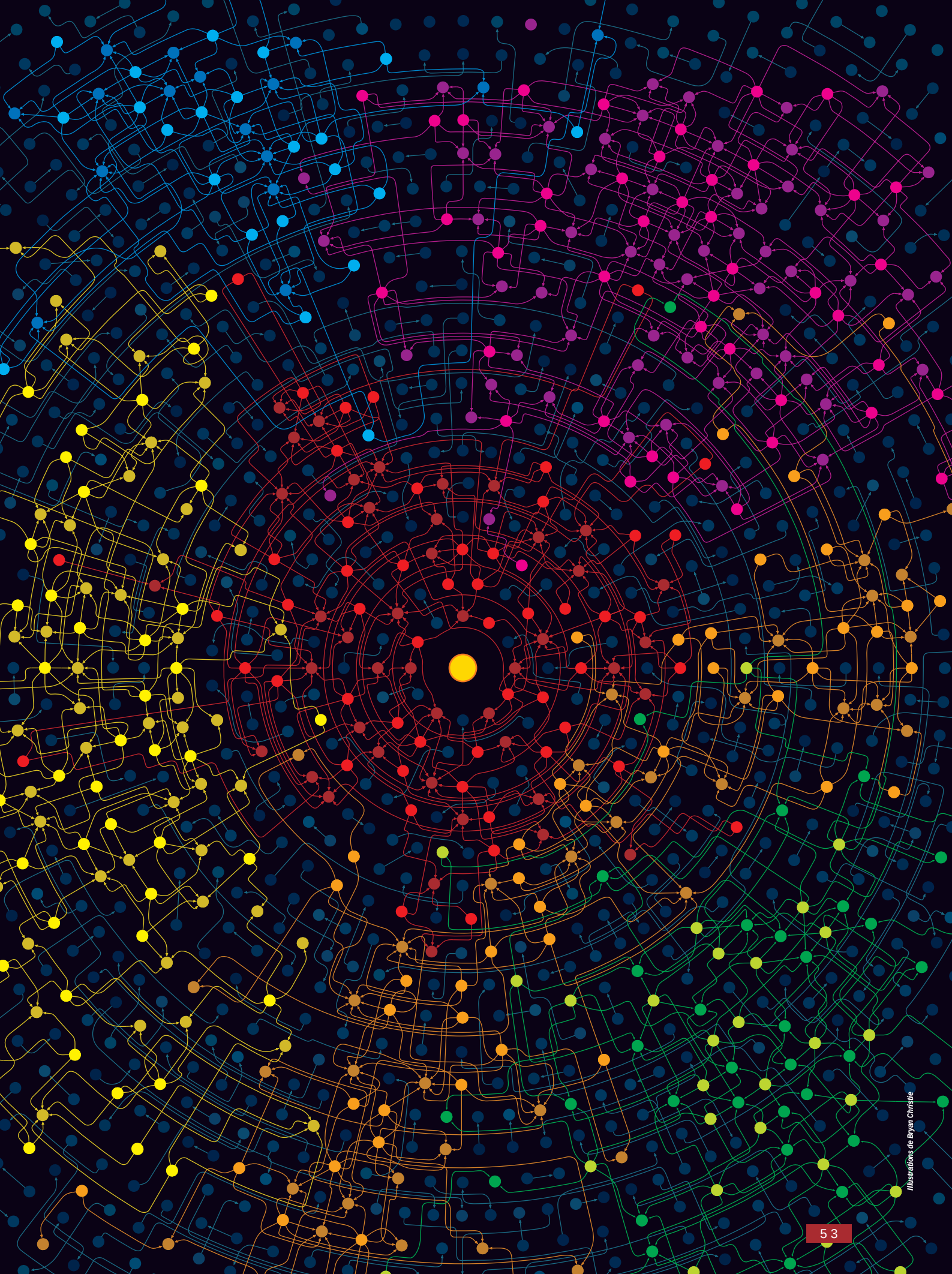
La création et la maintenance de ces index sont des tâches très lourdes (voir *La recherche d'informations*, par Clifford Lynch, *Pour la Science*, mai 1997) et c'est toujours un problème redoutable que de déterminer quelle information retourner en réponse à la question d'un utilisateur. Supposons que quelqu'un

cherche des informations sur la compagnie aérienne népalaise *Nepal Airways*. Sur la centaine de pages qui contiennent aujourd'hui cette expression, comment décider quelles sont les 20 meilleures ? Cette question est difficile, parce qu'on ne sait pas mesurer ce que « meilleure » veut dire : la réponse dépend du demandeur.

Des moteurs de recherche comme *AltaVista*, *HotBot*, *Lycos*, *Infoseek* et *Excite* utilisent des règles empiriques pour ordonner les réponses. Ces « recettes », nommées fonctions de classement, sont appliquées non seulement aux requêtes précises (*Nepal Airways*), mais aussi à des requêtes plus générales, comme « avion », un mot qui apparaît dans 150 000 pages, ou son équivalent en anglais, *plane*, qui apparaît dans plus d'un million de pages. Comment un moteur de recherche retiendra-t-il seulement 20 pages ?

On obtient un classement très simple en comptant le nombre de fois où le mot demandé apparaît dans les pages. On peut aussi placer en premier les pages où le mot requis apparaît le plus tôt. Toutefois, de telles approches ne donnent pas toujours de bons résultats. Par exemple, la recherche du mot *contribuable* avec de telles règles de classement trouverait que *Polyptote*, extrait des *Exercices de style* de Raymond Queneau, répond très bien à la requête, puisqu'il contient 22 occurrences de ce mot dans un texte

1. LES COMMUNAUTÉS DU CYBERESPACE (montrées ici par des couleurs) cohabitent sur le réseau. L'exploration de la structure du réseau a révélé des communautés sur des sujets aussi variés que les marées noires sur la côte Ouest du Japon, les pompiers d'Australie ou les ressources pour les Turcs vivant aux États-Unis. Le réseau comprend des centaines de milliers de petites communautés de ce type.



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

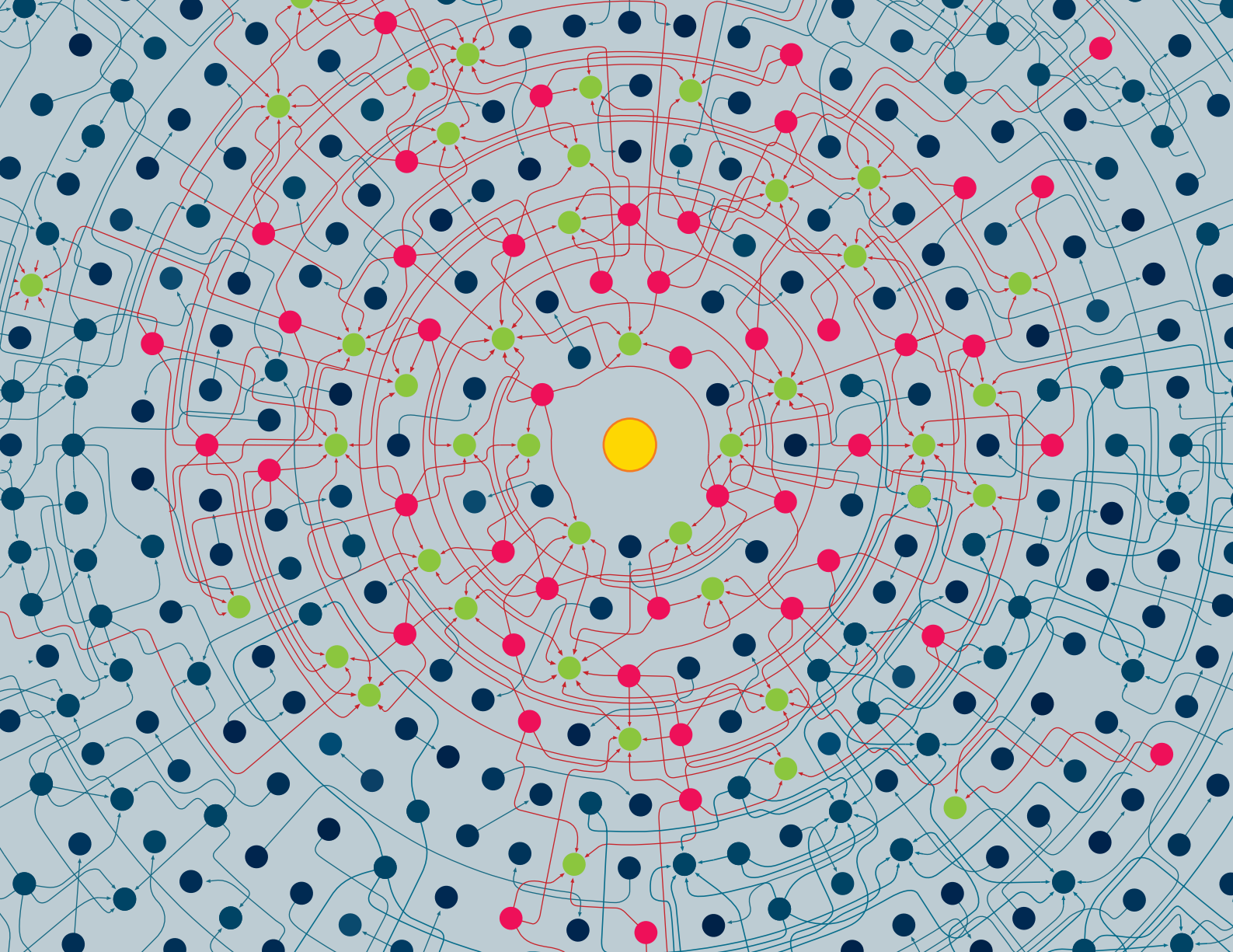
de pages traitant de cinétique chimique jugées pertinentes).

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

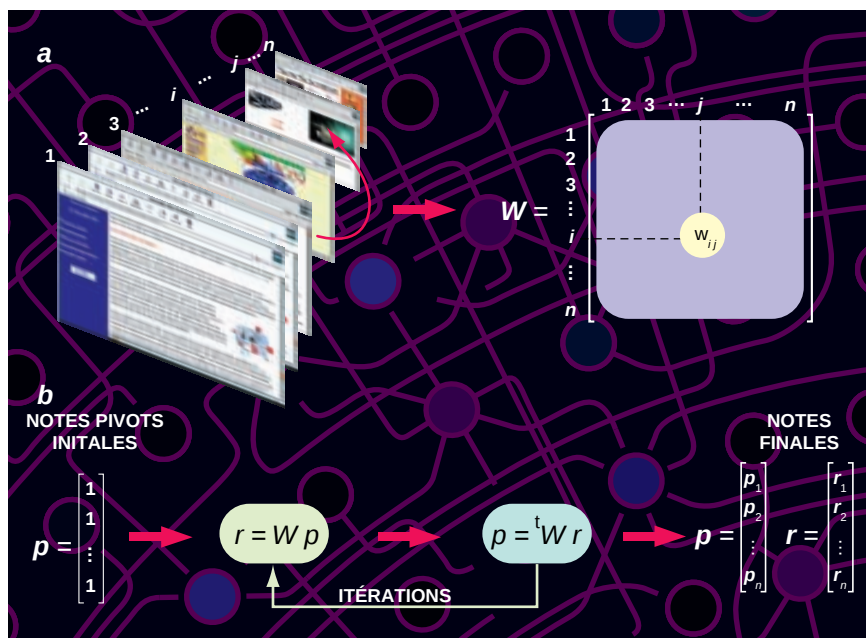
En plus de ces sites de référence, le réseau contient beaucoup de sites d'un autre type, des «pivots», qui sont connectés à ces endroits prestigieux et qui répercutent leur influence. Les sites pivots sont variés : ce sont aussi bien les listes professionnelles réalisées par les sites commerciaux que les sites personnels où l'on trouve des pages telles que «Mes sites préférés». Nous supposons qu'une page de référence est citée par beaucoup de bons pivots, et qu'un pivot utile est un site qui pointe vers beaucoup de sites de référence.

Comment obtenir une méthode d'identification des sites de référence et des sites pivots à partir de définitions aussi circulaires? Nous examinons d'abord un certain nombre de pages pouvant *a priori* répondre convenablement à une requête particulière, et nous attribuons à chacune d'elles une estimation de la fonction référence ou de la fonction pivot. À partir de ces évaluations initiales, nous exécutons ensuite un algorithme itératif en deux étapes.

La première utilise les estimations des sites de référence pour améliorer les estimations des sites pivots : nous repérons les meilleurs sites de références selon cette première estimation, puis nous analysons quelles pages pointent vers eux et jugeons ces sites comme étant de bons pivots. Lors de la deuxième opération, nous utilisons ces informations actualisées sur les pivots pour affiner les évaluations des sites de référence : nous regardons vers où pointent les meilleurs pivots et considérons ces sites comme de bonnes références. La répétition de ces corrections affine les résultats.

Nous avons programmé cet algorithme dans *Clever*, un moteur de recherche prototype. Pour chaque requête sur un sujet, par exemple l'hépatite B, *Clever* obtient d'abord une liste de 200 pages en utilisant un moteur de recherche classique comme *AltaVista*. Le système ajoute alors à ces 200 pages toutes celles qui y sont connectées par un hyperlien entrant ou sortant. Dans notre expérimentation, nous recueillons ainsi 2 000 à 5 000 pages, que nous nommons l'ensemble de base.

Chacune de ces pages reçoit deux notes initiales, une en tant que référence et une autre en tant que pivot. Puis *Clever* affine ces valeurs : pour chacune des pages, il calcule une nouvelle note de référence en additionnant les notes



3. COMMENT FAIRE MATHÉMATIQUEMENT LA RECHERCHE des pages de référence et des pages pivots? À partir d'un ensemble de n pages, on construit la matrice W dont chaque élément représente l'importance d'un lien d'une page à une autre (a). Par exemple, si la page i a beaucoup de liens vers la page j , l'élément w_{ij} de la matrice est élevé. Pour commencer le calcul, on pose toutes les notes pivots des pages égales à un (b), puis on calcule les notes de référence et les nouvelles notes pivots suivant un processus itératif : on obtient les notes de référence (r) en multipliant les notes pivots par la matrice W et on obtient les nouvelles notes pivots (p) en multipliant les notes de référence trouvées à l'itération précédente par la matrice transposée de W (${}^t W$, symétrique de W par rapport à la diagonale). Les calculs se terminent lorsque les notes sont à peu près stables. Ces notes peuvent alors être utilisées pour la détermination des meilleurs sites de référence et des meilleurs pivots.

des pivots qui pointent vers elle, une note pivot en additionnant les notes de référence des pages vers lesquelles elle pointe. Ainsi la note de référence d'une page augmente quand beaucoup de pivots bien notés pointent vers cette page et une page qui pointe vers des sites de référence bien notés a sa propre note de pivot qui augmente en proportion. *Clever* répète ces calculs jusqu'à ce que les notes se stabilisent. Puis il classe les sites selon les notes de référence et de pivot. Remarquons que cet algorithme n'interdit pas qu'une page soit bien notée dans les deux catégories, ce qui arrive parfois.

Pour mieux comprendre les opérations effectuées par l'algorithme utilisé, représentons le réseau Internet par un immense réseau de sites reliés d'une manière apparemment aléatoire. Pour un certain ensemble de pages contenant un mot donné, *Clever* cible les structures de liens les plus denses entre ces pages.

La sommation itérative des notes de référence et de pivot peut être analysée dans le cadre de l'algèbre linéaire : l'itération effectuée correspond à la multiplication répétée d'un vecteur (une colonne de nombres représentant les

notes de référence ou les notes pivots) par une matrice, c'est-à-dire un tableau à deux dimensions, où chaque élément représente un lien d'une page à une autre dans l'ensemble de base. Après quelques itérations, les composantes des vecteurs pivots et des vecteurs de référence deviennent stationnaires et révèlent quelles pages sont les meilleurs pivots ou les meilleures références. Ces colonnes de nombres stabilisés sont des «vecteurs propres».

Le processus itératif converge rapidement vers un ensemble de notes à peu près stable. Dans le cas d'un ensemble de base de 3 000 pages, le calcul converge en cinq itérations seulement et, ce qui est plus intéressant, le résultat est indépendant du choix des notes initiales : la méthode fonctionne même si toutes les notes initiales sont égales à un, de sorte que les notes finales sont intrinsèques aux pages retenues (voir la figure 3).

L'algorithme de *Clever* a l'avantage de regrouper automatiquement les sites en catégories. Par exemple, une recherche sur l'avortement produit deux types de sites, ceux qui sont pour et ceux qui sont contre, parce que les

pages d'une catégorie sont plus souvent liées entre elles qu'à celles de l'autre communauté.

Plus généralement, *Clever* révèle la structure sous-jacente du réseau. Bien que ce dernier ait crû de façon fulgurante et anarchique, il possède une structure intrinsèque – bien qu'imparfaite – fondée sur les liaisons entre pages.

Le lien avec l'analyse de citations

Par son principe, *Clever* ressemble à l'analyse de citations, l'étude de la manière dont les articles scientifiques se citent : la principale mesure de l'importance des articles est le «facteur d'impact», qui juge les publications d'après le nombre de références que les autres articles en font. Mis au point par Eugene Garfield, un spécialiste de l'information scientifique, le facteur d'impact préside à l'«Index des citations scientifiques» (*Science Citation Index*).

Sur le réseau, le facteur d'impact d'une page pourrait être le nombre de liens qui pointent vers elle. Toutefois, cette approche simpliste favoriserait, même pour des requêtes sur des sujets spécifiques, les pages généralistes très connues, telle la page d'accueil du *Monde*.

Dans le domaine de l'analyse des citations, les chercheurs ont tenté d'améliorer la mesure de E. Garfield, qui compte de la même façon toutes les références. Ne serait-on pas avisé de donner plus de poids aux citations d'un journal plus important ? Le problème, avec cette approche, est qu'on donne une définition circulaire du mot «importance», comme dans la définition des pivots et des références.

Dès 1976, Gabriel Pinsky et Francis Narin, du Laboratoire de recherches du CHI, dans le New Jersey, ont éliminé cet obstacle en mettant au point une méthode itérative de calcul qui aboutit à un ensemble stable de notes, qu'ils nomment poids d'influence. Toutefois, ils ne distinguent pas référence et pivot.

Cette différence révèle une particularité entre les pages Internet et la littérature scientifique classique : dans le cyberspace, des sites de référence en compétition, comme *Netscape* et *Microsoft*, ne reconnaissent pas leur existence mutuelle et ne peuvent donc être reliés que par un ensemble de pivots tiers ; inversement, les revues scientifiques rivaux pratiquent des

citations croisées qui rendent le rôle des pivots moins crucial.

D'autres équipes utilisent différemment les hyperliens, pour la recherche sur le réseau. À l'Université de Stanford, Serge Brin et Larry Page ont mis au point un moteur de recherche, nommé *Google*, qui utilise un classement fondé sur les poids d'influence de G. Pinsky et F. Narin. L'équipe de Stanford considère un internaute qui suit des liens en faisant un certain nombre de sauts au hasard, qui le mènent à certaines pages plus fréquemment qu'à d'autres. Ainsi *Google* trouve un seul type de pages universellement importantes, intuitivement celles qui sont fréquemment visitées lors d'une marche au hasard dans la structure des liens du réseau. Concrètement, pour chaque page, *Google* additionne les notes des sites qui pointent vers la page. C'est pourquoi, lors d'une requête spécifique, *Google* répond rapidement en trouvant toutes les pages qui contiennent le texte recherché et en les ordonnant selon leur rang prédéterminé.

Les moteurs *Google* et *Clever* diffèrent cependant par deux aspects importants. D'une part, *Google* élabore un ordre de classement initial et l'utilise quelle que soit la requête, alors que *Clever* construit un ensemble de base différent pour chaque terme recherché, puis fixe les priorités des pages dans le contexte de cette requête. Au total, *Google* répond plus rapidement. D'autre part, *Google* ne regarde que dans une seule direction, d'un lien vers un autre, alors que *Clever* regarde aussi dans l'autre direction, pour savoir quels sont les sites qui pointent vers une page de référence. *Clever* tire ainsi parti du phénomène sociologique qui fait que les humains construisent naturellement des contenus qui expriment leur expertise sur un sujet donné.

La recherche continue

Nous cherchons encore à améliorer *Clever*. Dans notre démarche, nous cherchons à associer texte et hyperliens. Pour ce faire, nous estimons que certains liens ont plus de poids que d'autres suivant la pertinence du texte qui se trouve en regard. Plus précisément, nous analysons le contenu des pages dans l'ensemble de base et nous cherchons le nombre d'occurrences et les positions relatives du terme recherché. Ces informations pondèrent ensuite certaines connexions entre

pages. Si, par exemple, le texte de la requête apparaît fréquemment à proximité d'un lien, alors la pondération correspondante est augmentée.

Nos premières expériences indiquent que ce raffinement améliore substantiellement la pertinence des résultats (une imperfection de *Clever* a parfois l'inconvénient d'élargir trop sa recherche, quand la requête est très précise : ainsi, si l'on cherche la Maison Ozenfant, de Le Corbusier, le système élargit sa recherche et retrouve des informations sur un sujet général comme l'architecture contemporaine). Nous expérimentons d'autres améliorations qui tiennent compte des divers styles de réalisation de pages Internet.

Nous avons aussi commencé à construire automatiquement des listes de ressources Internet, similaires aux annuaires de sites élaborés manuellement par les employés de sociétés comme *Yahoo!* ou *Infoseek*. Les premiers résultats sont aussi bons que ceux des annuaires. De surcroît, ce travail nous a montré que le réseau est plein de petites communautés très liées, avec des intérêts particuliers (les amateurs de sumo, les passionnés de yo-yo de la Creuse ou les amateurs de sauce ketchup au petit déjeuner). Nous recherchons comment découvrir automatiquement ces communautés cachées (voir la figure 1).

Le réseau Internet est aujourd'hui très différent de ce qu'il était il y a seulement cinq ans. Personne ne peut dire ce qu'il sera dans cinq ans. L'indexation du réseau deviendra-t-elle impossible ? Et comment la notion de recherche changera-t-elle alors ? Aujourd'hui, la seule chose dont nous soyons certains est que la croissance du réseau continuera à lancer des défis informatiques à ceux qui veulent naviguer dans cet océan d'informations.

Le projet *Clever* : Soumen Chakrabarti, Byron Dom, Ravi Kumar, Prabhakar Raghavan, Sridhar Rajgopalan et Andrew Tomkins travaillent au Centre de recherche Almaden d'IBM, à San Jose (Californie). Jon Kleinberg travaille au département d'informatique de l'Université Cornell. David Gibson termine sa thèse au département d'informatique de l'Université de Berkeley.

De nombreux liens pour cet article sont disponibles sur notre site Web : <http://www.pourlascience.com/numeros/pls-262/clever.htm>
