Filtrages et synthèses de masse sur internet

Etat de l'art et prospective

Alain Lelu

Le succès de l'internet est dû, pour beaucoup, à ce qu'il permet tout à la fois la communication en temps réel, en temps différé, et l'édition des écrits, du son et de l'image, c'est-à-dire leur capitalisation et leur mise à disposition, édition désormais à la portée de (presque) tous. L'exploitation de ces flux et de ces stocks d'information très informelle a attiré en premier lieu l'attention des services de renseignements, en particulier de la NSA (National Security Agency), l'agence de renseignement militaire américaine. Le financement d'équipes de recherche par cet organisme, la révélation du réseau d'écoutes planétaire Echelon [CON], et de la puissance informatique concentrée à Fort Meade pour le traitement de ces flux (5 Teraflops, équivalente à celle de 1 000 PC récents en parallèle), ont montré que beaucoup de techniques de filtrage d'informations, dont certaines auraient pu relever il y a peu de la science-fiction, sont d'ores et déjà mises en œuvre. Nous passerons en revue ces techniques, dont l'utilisation constitue un enjeu considérable pour les institutions civiles, les entreprises et les particuliers, que ce soit pour chercher des informations sur un sujet précis, pour préciser ce qui est connu, ou pour dégager ce que l'on ne connaît pas, ou mal, sur des thèmes généraux. Le passage du filtrage de corpus textuels volumineux par mots-clés (*word spotting*) au filtrage par thèmes (*topic spotting*), améliore le recueil d'informations bien ciblées. Les progrès de l'extraction automatique de thèmes, cette fois en vue d'en tirer

LCN, volume 3, n° 1-2002, pages 171 à 196

des synthèses interprétables, est gros de conséquences culturelles, sociales, politiques que nous entrevoyons à peine, mais qui accompagneront de façon incontournable, pour le meilleur et pour le pire, le monde « branché » qui se dessine sous nos yeux.

Quelques ordres de grandeur

Victime de son succès, l'internet affiche un taux de croissance exponentiel, qui s'exprime, sous toutes ses formes (nombre de serveurs, de pages, d'internautes, de messages, etc.), en nombre de mois pour un doublement – typiquement, entre 10 et 18 mois. Même si quelques signes d'accalmie de cette croissance se font sentir dans les pays occidentaux, le potentiel que représentent les pays en développement ne laisse présager aucun répit, et il nous faudra désormais vivre avec des chiffres astronomiques[1] dont l'étendue nous échappe.

Web indexé, web indexable et web invisible

En matière de décompte des pages web, domaine où toute précision est inutile tant les délais entre rédaction, publication et lecture rendent obsolète toute estimation, contentons nous d'affirmer qu'en fin 2001 plusieurs milliards de pages sont en ligne, constituant le web « indexable », à la portée des moteurs de recherche. Pris dans leur ensemble, ceux-ci n'en indexent réellement « que » de l'ordre d'un milliard (Google permettrait d'en atteindre 3 milliards).

Au-delà de ce web potentiellement visible, beaucoup de sites donnent accès à des systèmes d'information en ligne, payants ou gratuits, où la notion de page n'a plus de sens puisqu'il s'agit d'information créée

1. Ne faudrait-il pas parler plutôt de chiffres « biologiques » ? Les milliards de pages web ne sont-elles pas à mettre en rapport avec les 100 milliards de neurones dans notre cerveau, les 3 milliards de paires de bases dans notre patrimoine génétique, le million de milliards de protéines différentes chez les espèces vivantes sur terre, etc. ? Pour la première fois dans l'histoire, nous sommes confrontés à l'interaction avec des milliards d'objets différents à portée de main – toute page web est à une vingtaine de clics de souris de n'importe quelle autre –, d'appréhension subjective sans rapport avec les milliards d'unités monétaires répétitives que nous avons côtoyé jusqu'ici. N'ayons aucun souci à nous faire quant à l'usage des puissances de calcul et des capacités de stockage faramineuses que laissent présager pour bientôt la loi du doublement tous les 18 mois, observée jusqu'ici en informatique avec une régularité de métronome : ils trouveront leur emploi dans la croissance elle aussi exponentielle des ressources qu'ils exploiteront.

dynamiquement et mise en page de façon instantanée : il s'agit d'une bonne partie du « web invisible », dont peu s'aventurent à estimer la taille, mais où là aussi le nombre des enregistrements documentaires doit se compter en milliards. D'après la société Bright Planet, de l'ordre de 100 000 bases de données sont accessibles sur le web, dont 95 % gratuitement, représentant l'équivalent de 300 milliards de pages composées dynamiquement – les 60 plus grosses bases documentaires, comme Medline aux Etats-Unis [MED] ou Pascal en France [PAS], en fournissent près d'un tiers.

Les flux et les stocks : messageries et forums

A-t-on atteint le milliard de messages e-mail émis quotidiennement dans le monde ? Si ce n'est fait, c'est une question de mois, d'un an tout au plus. La messagerie électronique est en train de rattraper l'ordre de grandeur des flux téléphoniques, et dans de nombreux pays en développement son faible coût et sa fiabilité par rapport aux services postaux lui assurent une croissance soutenue – n'évoquons même pas son insensibilité aux attaques biologiques, rappelée à notre attention par l'actualité récente !

Une catégorie particulière de messages présente un grand intérêt du fait même de leur accumulation : il s'agit des contributions envoyées aux milliers de forums de discussion UseNet, qui couvrent un très large éventail de sujets, des divers aspects de l'informatique à ceux de la culture, en passant par la religion et la politique, images des préoccupations spontanées des internautes depuis une vingtaine d'années. Le fait que ces messages forment la trame de fils qui se répondent les uns aux autres, et sont disponibles dans des archives en ligne exhaustives [GOOa], en fait un capital culturel précieux de 700 millions de messages, différent de celui accumulé dans les bibliothèques du monde, car formé par sédimentation d'avis et réponses au style souvent télégraphique et familier, mais ce capital est pourvu d'un intérêt propre pour ses contributeurs comme pour ses observateurs.

Exploiter les informations structurées explicites

Le filtrage par les « en-têtes »

Les informations circulant sur l'internet sont constituées de parties non structurées (texte libre, images ou sons, programmes compilés) enveloppées dans de l'information structurée : adresses de départ et de destination, dates, volume, type physique… Exploiter ces données d'« en-têtes » est le B-A-BA de l'organisation et de la recherche d'information, ce que font tous les

logiciels de messagerie et de participation à des forums (*news groups*) en nous proposant des listes de messages triées par date, par auteur, par destinataire, etc.

Les services de renseignement ne manquent pas d'exploiter cette ressource et, compte tenu de l'énormité des flux à surveiller, continuent à se baser principalement sur la bonne vieille méthode d'écoute des adresses suspectes : Echelon semble intercepter la totalité des messages transitant par satellites[2], internet et téléphone confondus, et la lecture ou l'écoute directe, humaine, des communications suspectes est la façon la plus fiable de recueillir de l'information (à noter que certains services secrets de ce réseau dominé par les Etats-Unis n'ont pas les moyens de traiter les communications téléphoniques, trop gourmandes en temps d'écoute ! [HAG 96]). Le tri à réaliser pour rester dans des limites budgétaires supportables semble au plus de 1 pour 1 million de messages. D'où l'intérêt d'autres méthodes de filtrage, plus automatisées et plus adaptées à un sondage exhaustif de flux.

Une autre stratégie tentée par le FBI consiste à rendre obligatoire l'installation, chez chaque prestataire de messagerie aux Etats-Unis, du logiciel Carnivore [EPI] permettant de mettre sur écoute certains comptes, sur décision judiciaire – idée qui a provoqué une forte émotion aux Etats-Unis devant les abus possibles et le manque de contrôle effectif du dispositif.

Exploiter les liens entre pages web

De quoi la Toile est-elle tissée ?

Le vocable « la Toile » (*The Web*) provient des liens dont sont pourvues les pages, liens qui les rattachent à d'autres pages ou à d'autres points d'ancrage dans les mêmes pages. L'ensemble de ces liens forme un

2. Les satellites de télécommunication constituent une cible idéale pour un réseau tel qu'Echelon, car ils concentrent tout le trafic international, *a priori* plus « sensible », et peuvent être écoutés librement et discrètement à partir d'un nombre limité de stations d'interception. L'évolution actuelle vers un renforcement du trafic crypté d'une part (70 % des communications internationales de la péninsule arabique le sont), sur fibre optique sous-marine d'autre part, complique sérieusement cette tâche, sans aller jusqu'à la rendre impossible... On comprend que l'existence de *backdoors* dans certains logiciels, révélée par deux affaires récentes, l'une concernant le système d'exploitation Windows, l'autre un modem professionnel français, ainsi que la limitation de la sécurité des clés de cryptage pour les logiciels vendus hors des USA, soit devenues un enjeu aussi stratégique que secret – au fait, que deviennent les ennuis judiciaires de Microsoft ?

enchevêtrement, un feutre plus qu'un tissu régulier, comparable aux liens de citation qu'on trouve dans les articles scientifiques et qui traduisent la reconnaissance par l'auteur d'un certain nombre de travaux antérieurs comme références appropriées – que ce soit pour les citer comme étapes de son cheminement ou pour les critiquer.

Dès lors que l'informatique permet de collationner « quelque part », par exemple dans un serveur documentaire de *preprints* ou dans un moteur de recherche web, tous les liens issus d'un grand nombre d'unités documentaires (textes et descriptions documentaires, pages web…), il devient possible d'enrichir le point de vue du surfeur à la recherche d'informations :

– sur le moteur Google [GOOb], tout ensemble de pages obtenu à partir d'une requête classique par mot(s) présente en tête de liste les pages vers lesquelles pointent le plus de liens, donc les plus « populaires » ou reconnues[3] ;

– le projet Clever du laboratoire IBM Almaden Research Center [CHA 98] parvient, en partant d'un tel ensemble de pages élargi aux pages amont liées et aval pointées, à repérer les pages les plus reconnues (*authorities*) et les meilleures pages carrefour (*hubs*), celles qui pointent vers le plus de pages reconnues. Pour réaliser ce projet, on attribue itérativement une note de reconnaissance aux premières, et une note de centralité aux deuxièmes : la note de centralité n'est autre que la somme des notes de reconnaissance des pages vers lesquelles pointe une page donnée, et réciproquement pour la note de reconnaissance. Le processus semble se mordre la queue, mais on démontre mathématiquement qu'en partant de notes arbitraires (par exemple toutes égales) le processus converge au bout de quelques itérations vers deux ensembles stables de pages : les mieux reconnues (*authorities*), et les plus « rayonnantes » (*hubs*). D'où le nom CLEVER, *Computing Large Eigen Values for Enhanced Retrieval*, où *eigenvalue* (en français *valeur propre*) désigne techniquement ce vers quoi ce processus se stabilise, quand on part d'un tableau (matrice, en termes techniques) de liens entre pages référentes et pages référées ;

3. Malgré le danger des phénomènes auto-entretenus de « rentes de notoriété » disqualifiant définitivement les sources obscures et sans-grade, force est de reconnaître que ce mécanisme est bien utile quand on ne recherche pas spécialement l'originalité des points de vue, et surtout qu'on veut éviter de suivre des centaines de liens à pertinence problématique… Google semble aujourd'hui le moteur le plus consulté et à la plus vaste couverture (grâce à sa prise en compte des liens, il permet d'atteindre beaucoup plus de pages que de pages indexées).

– le serveur de *preprints* et publications, auto-archivées CiteSeer [CIT] donne accès, pour chaque article, à de nombreux liens directs ou calculés : articles cités, mais aussi liens calculés à partir des articles citants, des articles partageant les mêmes citations (« cocitations »), ou à partir des mots ou phrases communs dans les résumés (cf. plus bas)…

Ces possibilités d'exploiter de façon « panoptique » les liens hypertextuels nous montrent qu'il est faux de considérer le World Wide web comme un support d'édition libre, soustrait à tout phénomène de validation par les pairs : n'importe qui peut publier à peu près n'importe quoi, mais peu de gens y accèderont si cette œuvre n'est pas référée dans les pages émanant au moins d'un micro-milieu partageant un micro-point de vue sur un micro-sujet ! « Se faire référer », et pas seulement par les moteurs web, est un enjeu essentiel pour qui publie un site ; la marche initiale à franchir est moins élevée que sur support papier, mais la logique éditoriale de la circulation de la reconnaissance entre pairs y joue tout autant.

Le web est propice à l'éclosion et au renforcement d'innombrables communautés d'intérêts, passionnés de numismatique comme militants rassemblés par un événement politique. Il paraît évident que les services de renseignements et de police utilisent les liens pour cerner les contours de « cliques » qui peuvent se référencer mutuellement sur des sujets sensibles, tout en employant un langage codé (*hackers*, terrorisme biologique ou nucléaire, nazis, pédophilie…).

Les liens hypertextes forment le tissu direct et explicite de la Toile ; mais bien d'autres types de liens peuvent être mis à jour pour qui dispose de la capacité de traiter le contenu d'un grand nombre de textes : ce sont les liens calculés, ou déduits, implicites.

Calculer des liens à partir du contenu même des informations sur l'internet

Pour peu qu'on se trouve en mesure de définir un indice numérique de ressemblance entre deux documents, une génération automatique de liens hypertextuels s'en déduit : il suffit de calculer systématiquement et patiemment (pas de problème pour un ordinateur…) les valeurs de cet indice entre un document donné et tous les autres, et de décider d'un seuil de valeur au-dessus duquel on considérera que des liens existent. En général on présente à l'utilisateur la liste, par ordre de valeur de lien décroissante, des documents liés, et celui-ci décide de les explorer, ou pas, en progressant plus ou moins dans cette liste.

Liens calculés à partir du vocabulaire partagé par chaque paire de pages

Compter le nombre de mots communs à deux textes est une façon aisée de les rapprocher globalement : s'il n'y en a aucun, ces textes n'ont clairement rien à voir entre eux, s'il y en a beaucoup, alors ils parlent de sujets voisins. A partir de cette idée simple, voire simpliste, un bon nombre d'indices de ressemblance entre textes ont été proposés pour compenser les inévitables inconvénients de cette idée de base. Le « cosinus TF-IDF » de Salton en est le plus répandu, et répond à deux exigences :

– donner moins d'importance aux mots fréquents, communs à la majorité des documents, et qui n'apportent quasiment aucune discrimination entre ceux-ci : on donne à chaque mot dans un document d'autant plus de poids qu'il y sera fréquent ; et au contraire on lui donnera d'autant moins de poids qu'on le trouvera dans le plus grand nombre de documents. D'où le nom de cette pondération : *Term Frequency, Inverse Document Frequency* ;

– ne pas dépendre de la taille de chaque texte : l'indicateur de Salton est normalisé, entre 0 % et 100 %, pour ne pas favoriser les grands textes au détriment des petits.

Un tel indicateur demande des comptages et des calculs numériques qui le lient au modèle vectoriel en documentation, dont Gerald Salton fut le héraut depuis les années 1960 [BAE 99].

Un certain nombre de moteurs de recherche documentaire sur le web incluent cette fonction de recherche de documents similaires (*similarity ranking*), parfois aussi dénommée *relevance feed-back* (retour de pertinence) dans la mesure où l'utilisateur, confronté à une liste de documents issus d'une requête précédente, en choisit un qu'il considère comme pertinent, et dont les proches risquent fort de l'être également (cf. fonction *related articles* dans [MED]). En d'autres termes, on fait suivre une requête booléenne, nécessairement limitée, par une « requête-document » beaucoup plus riche en mots potentiellement pertinents.

Plus radicalement, dès lors qu'on est libéré des liens directs, explicites, entre documents, rien n'empêche de considérer un ensemble de textes indexés (c'est-à-dire décrits par des mots) comme un réseau de mots : chaque mot étant caractérisé par ses fréquences dans tous les documents où il est présent, le même type d'indicateur de ressemblance vu plus haut peut être défini, mais cette fois entre deux mots, créant de cette façon des « couronnes » de mots proches d'un mot donné. Cette opération, possible dans quelques systèmes documentaires (elle s'appelle parfois *Zoom*), a été implantée dans le passé sur les moteurs Excite et AltaVista, et reste en démonstration sur [LSI] ; elle a le mérite de suggérer à l'utilisateur des mots

pertinents présents dans le corpus qui ne lui venaient pas à l'esprit – ce problème du décalage entre vocabulaire de l'utilisateur et vocabulaire du corpus est reconnu comme une des difficultés récurrentes de la recherche d'information[4].

A noter que la qualité, la pertinence, de ces liens calculés dépend en ligne directe de la qualité de l'indexation : si celle-ci se contente d'élever chaque chaîne de caractères à la dignité de mot-clé (indexation *full text*, en texte intégral), beaucoup de « bruit » viendra polluer les exploitations ultérieures [LSI]. La qualité de celles-ci sera améliorée par une analyse morpho-syntaxique du corpus permettant de dégager les formes normalisées des mots, ou *lemmes* (verbes rapportés à l'infinitif, adjectifs au masculin singulier…), et surtout de dégager les termes composés, véritables atomes sémantiques, les plus proches d'une indexation idéale par d'authentiques concepts. Pour un exemple de moteur web utilisant un traitement linguistique, voir [NOM].

Autres types de liens calculés

Caractériser un texte par un profil de fréquences de mots n'est pas la seule façon de le décrire : plusieurs travaux [DAM 95], [LEL 98] ont montré qu'il était aussi possible de le caractériser par un profil de fréquences de n-grammes[5], ce qui rend cette technique indépendante de la langue et du type de codage des caractères, donc particulièrement appropriée pour les langues asiatiques, sans séparateurs de mots. La qualité des représentations synthétiques obtenues dépend de la qualité des filtrages statistiques réalisés en amont, question toujours ouverte [HAL 01] qui seule permettra de rivaliser avec la qualité obtenue à partir de termes lemmatisés et filtrés – inutile de préciser que dans ce cas, les traitements linguistiques deviennent inutiles ! On dégage alors par des « surlignages flous » des termes simples et

4. Le logiciel Neuronav, décrit dans [LEL 01] et [AUB 01] généralise ces opérations vectorielles [LEL 02] : de façon itérative, un paquet de documents pertinent est caractérisé par ses mots les plus typiques, puis une sélection parmi les mots proches d'un ou plusieurs mots intéressants élargit le cercle des mots pertinents, qui appellent alors davantage de documents à retenir, etc., jusqu'à ce que l'utilisateur ait le sentiment d'avoir « fait le tour du problème », sans pour autant avoir consulté des milliers de pages.

5. Les n-grammes sont les suites de caractères de longueur fixe n (par exemple, 2, 3,…) obtenues en promenant sur le texte une fenêtre de n caractères. Les premiers bigrammes de la présente note sont : *Le, es, s_, _n,* … On collationne tous les n-grammes différents dans tous les textes, et chaque texte est caractérisé par le profil des fréquences de ses n-grammes.

composés candidats au statut de mots d'index, caractéristiques de chaque texte, qui en présentent, en quelque sorte, un résumé thématique.

Les pages multimédias peuvent aussi être à l'origine d'autres liens calculés : par exemple, ceux qui sont déduits sur la base d'indicateurs numériques de couleurs, de textures ou de formes sur les images, dont on pouvait voir la préfiguration dans la fonction *similar images* d'AltaVista/Recherche d'images (cf. démo. [IKO]). Les mélodies, et bien d'autres caractéristiques du son (les empreintes vocales individuelles ?), pourraient en être les équivalents pour les informations sonores.

Enfin la norme XML, quand elle parviendra à supplanter HTML pour un nombre significatif de pages, pourra permettre d'enrichir l'éventail des possibilités : à partir du typage des liens, on dégagera autant de synthèses que de type de liens, et on élargira, au moyen de balises sémantiques caractérisant les mots, les possibilités de dégager des liens. Notons toutefois qu'il ne suffit pas de mettre techniquement à disposition ces possibilités de typage et de balisage pour qu'elles soient utilisées : seule une minorité de pages HTML profite de la possibilité actuelle qu'on a de renseigner les champs MétaTags d'information structurée (auteurs, mots-clés...). Mais ceci est un autre débat.

Filtrer l'information

Le schéma dominant de la recherche d'information sur l'internet est celui de la requête, c'est-à-dire d'une question que l'on pose, qu'un système automatique « comprend » et à laquelle il fournit une réponse. Le schème alternatif du butinage (*browsing*) au gré des liens entre pages, bien que séduisant au départ et homologue à la démarche d'associations d'idées prônée par le pionnier de l'hypertexte Vannevar Bush, donne lieu à en pratique à des séances de dérive étourdissantes, vertigineuses, dont on ressort l'esprit broyé, malaxé, avec un sentiment de difficulté à capitaliser, à faire un bilan ; l'internaute de base a du mal à passer maître dans le maniement des fragiles traces collectées par les navigateurs, et dans la gestion des listes de favoris, toujours à la merci d'un reparamétrage intempestif, d'une réinstallation de logiciels ou d'un changement de configuration !

En l'absence de repères solides pour répondre aux questions – Ou suis-je ? Saurais-je plus tard revenir à tel endroit où je me trouvais tout à l'heure ? Vers où me diriger ? – la moins mauvaise métaphore reste peut-être aujourd'hui celle du dialogue avec la machine, du couple requête/réponse

au moyen de mots répertoriés par la machine et signifiants pour l'homme. On se souvient qu'en posant telle requête sur tel moteur on retrouvera sans coup férir ce qui nous avait intéressé il y a quelque temps.

Le filtrage est la généralisation de ce schéma à un flux de données : une fois une requête mise au point, on l'applique à intervalle régulier au nouvel état d'un stock d'informations, ou à un flux de messages. Les documentalistes appellent cela depuis longtemps « diffusion sélective d'information », et les informaticiens l'ont remis au goût du jour sous le vocable d'agents un peu abusivement qualifiés d'intelligents.

Filtrage par localisation de mots (word-spotting)

Le but du filtrage est de recueillir un maximum de documents sémantiquement homogènes, répondant à une problématique claire du point de vue de la compréhension humaine. Or le langage est truffé de pièges sémantiques – polysémies, homonymies, synonymies – qui ne sont pas des exceptions, mais la règle dans toutes les langues. C'est en effet le contexte sémantique qui est l'élément de base, commun et fédérateur, dans la communication humaine, et qui permet de résoudre sans problème ces difficultés, qui n'en sont que pour l'ordinateur, voué par construction à ne traiter que des codes, ou chaînes de caractères, définis de façon univoque, et indifférent à la notion de contexte. En informatique, le chemin est long de la forme graphique d'un mot à son sens ; il est jalonné par plusieurs générations de techniques, utilisées pour commencer à prendre en compte ces questions.

Techniques basées sur les chaînes de caractères

Historiquement, l'assimilation abusive 1 mot = 1 chaîne de caractères[6] a été la première façon et la plus simple de prendre en compte l'intégralité du texte sur un ordinateur. Elle le reste aujourd'hui dans l'écrasante majorité des moteurs de recherche web ; parmi ceux-ci, seuls ceux qui sont dotés de capacités de reconnaissance de la langue des pages sont en mesure d'éliminer les mots grammaticaux comme *le* ou *the*, au moyen d'un antidictionnaire. Grâce à des opérateurs de troncature permettant d'ignorer les terminaisons, il est possible de composer des requêtes booléennes et

6. Cette assimilation n'est possible que pour les langues à séparateur entre mots, comme l'anglais et les langues latines. Dans les langues sans séparateurs, comme les langues asiatiques, seul le contexte permet à un humain d'effectuer cette séparation. On notera que la majorité des moteurs de recherche en langue chinoise sont basés sur une indexation des textes par les bi- et tri-grammes (cf. note précédente).

d'adjacence (comme : *base* NEAR donnée**) qui reportent sur l'auteur de la requête le poids de la prise en considération des phénomènes syntaxiques, et des compromis bruit documentaire/silence qui en résulteront.

Il semble que le réseau Echelon ait fait un usage intensif de ces techniques pour traiter les flux de messages interceptés. D'abord un signe : les ordinateurs situés dans chaque station d'écoute satellite sont appelés *dictionnaires* ; ensuite on sait que des matériels informatiques spécialisés dans la reconnaissance à la volée de chaînes de caractères ont été acquis par la NSA (machines Fast Data Finder, de Paracel Inc.), quand les capacités de traitement des processeurs universels n'étaient pas encore suffisantes. En octobre 1999, pour la journée militante de brouillage « Jam Echelon Day » ont été mises en circulation sur l'internet des listes de mots « sensibles » (*trigger words, spook words*), qui laissent parfois dubitatif [LIN] : à côté de sigles d'organismes gouvernementaux ou militaires, de mots propres au monde du renseignement, on trouve des termes comme *Java, zip, fish, 3, Elvis, Scully* ou *Roswell*… Dépouiller tous les messages contenant un ou même une combinaison limitée de ces termes paraît une tâche surhumaine, et surtout naïve. Peut-être ces listes, si elles ne sont pas un pur canular, sont-elles utiles pour surveiller certains milieux *hackers* ou militants s'intéressant explicitement aux questions de sécurité informatique, cryptographie ou renseignement. Il paraît évident que les noms de code que sont peut-être *3, Elvis* ou *Scully* ne peuvent être utilisés sans une équation de recherche longue et complexe établissant en gros le contexte dans lequel ces mots sont employés ! Pour de véritables délinquants soupçonnant une interception, le contexte est établi par les seules adresses de départ et d'arrivée, et un vocabulaire codé anodin échappe à ce type de filtrage (*le jour* pour *le 11 Septembre*…). Mais Echelon est-il fait pour repérer des délinquants, ou pour surveiller la routine des activités politiques et gouvernementales, des projets industriels « normaux », dans un monde fortement ancré dans une culture de la compétition ?

Traitements linguistiques

La raison pour laquelle le repérage « bête » de chaînes de caractères reste prédominant est que la majorité des pages web et des messages électroniques ne sont pas de vrais textes, conformes, même de loin, aux canons de la langue, sans parler de la multiplicité des langues autres que l'anglais, de plus en plus présentes sur ce média. Par contre les dépêches d'agence, les articles de journaux, les résumés bibliographiques – qu'on trouve aussi sur l'internet –, s'en rapprochent davantage, et il est alors possible d'en faire l'analyse morpho-syntaxique pour étiqueter

grammaticalement chaque mot, le ramener à sa forme normalisée (lemme), et repérer sur une base statistique ou linguistique les termes composés. Certains moteurs de recherche documentaire, dédiés à la veille en entreprise (veille stratégique, technique, concurrentielle, d'image) et quelques rares moteurs web [NOM], [PER] font appel à de telles techniques, à des degrés très disparates d'élaboration et de fiabilité. Elles accroissent sans conteste la qualité et la précision de l'indexation, mais peuvent aussi engendrer un bruit de nature différente de celui de l'indexation *full text*. Là aussi le compromis est difficile entre 1) la tolérance aux fautes d'orthographe et de syntaxe, aux néologismes, 2) la fiabilité de l'analyse syntaxique (beaucoup de cas ne se résolvent que par le contexte sémantique, généralement hors de portée), et 3) la puissance de calcul nécessaire.

Le filtrage est alors plus simple à définir et de meilleure qualité que sur de l'indexation en texte intégral, mais ici aussi le mot n'est pas le concept, et bien des « faux négatifs » passent entre les mailles, alors que beaucoup de « faux positifs » sont recueillis. Des campagnes d'évaluation de systèmes de recherche d'information, comme TREC aux Etats-Unis [TRE] et Amaryllis en France [AMA], sont organisées pour comparer des systèmes commerciaux ou expérimentaux à partir de corpus et requêtes de filtrage communes, dont les résultats sont synthétisés par des courbes « rappel *vs* précision »[7].

Traitements linguistiques + sémantiques

Depuis toujours les chercheurs en intelligence artificielle symbolique caressent le rêve de traduire le langage naturel en une langue « idéale » où les mots auraient un sens univoque et parviendraient à décrire les concepts et relations entre concepts sous-jacents à tout texte. En somme, de passer des données à la connaissance.

De nombreuses propositions de « langues-pivots » sémantiques ont été avancées (y compris… le latin, que sa rigueur syntaxique et logique, aux antipodes de l'anglais, autant que son statut de véritable langage humain ne classent pas parmi les plus farfelues !), mais ils se sont heurtés au rôle décisif du contexte sémantique dans toute langue, malgré l'existence d'une syntaxe qui paraissait assez facilement formalisable. D'où une pléthore de travaux, en particulier dans le domaine médical, pour établir des langages-pivots de spécialité et des traducteurs automatique dans les deux sens, parfois

7. Rappel = nombre de documents pertinents retrouvés/nombre de documents retrouvés,
Précision = nombre de documents pertinents retrouvés/nombre de documents pertinents.

opératoires dans quelques micro-domaines. Des spécialités aux noms prometteurs sont apparues au fil des ans : traitement automatique du langage naturel (TALN), bases de connaissances, ontologies, apprentissage symbolique (*Machine Learning*), *Message Understanding* (compréhension de messages), *Knowledge Management and Extraction* (gestion et extraction de connaissances), entre autres. Ces derniers courants cités, issus au départ de la culture scientifique de l'IA symbolique, prônent une réconciliation et une intégration avec leur rivale de toujours, l'informatique « molle » des calculs numériques et de la statistique, revivifiée par les modèles neuronaux, les algorithmes génétiques et la vie artificielle ; informatique molle car rien n'y est tout blanc ou tout noir, et les structures hiérarchiques chères à l'IA symbolique peuvent y être considérées comme des commodités de programmation plus que des traductions de la réalité profonde des choses [LEL 95], [ROB 02].

Parmi les réalisations, celle qui semble à notre connaissance la plus opérationnelle à grande échelle est due à une équipe française, celle de Christian Krumeich, à l'origine chez Thomson, qui a mis au point pour le renseignement militaire français (DGSE et DRM) le logiciel Taïga (traitement automatique de l'information géopolitique d'actualité), conçu à l'origine pour dépouiller les bases documentaires russes au moment de la Perestroïka [KRU 94]. Quelques dizaines de postes de travail sont en fonctionnement, pour dépouiller des sources structurées, homogènes et rédigées en toutes langues, après un travail considérable de formalisation des connaissances pour alimenter son moteur sémantique, que peu d'entreprises ou d'institutions peuvent se permettre de financer, dans des domaines bien définis comme la construction aéronautique ou la prospection pétrolière. Cette technique semble peu compatible avec le filtrage à grande échelle de messages informels, comme le réalise Echelon.

Filtrage par localisation de contextes sémantiques prédéfinis **(topic-spotting)**

Dès lors que le sens d'un mot dépend de son contexte, l'objectif de localisation de chaque atome de sens derrière chaque mot paraît peu réaliste dans l'état de l'art actuel. Il paraît plus raisonnable, et de toute façon satisfaisant pour une large palette d'applications, de se rabattre sur l'objectif, tout compte fait moins ambitieux, de repérage de *contextes* sémantiques, c'est-à-dire d'ensemble de textes qui en gros parlent de la même chose, du même sujet ; ce sujet peut impliquer un grand nombre de mots qui chacun peuvent intervenir dans d'autres contextes, mais que seul le contexte de ce sujet précis met en relation, fait intervenir ensemble. C'est ainsi qu'on peut

« piéger » un contexte sémantique, et c'est plus facile qu'il n'y paraît : il suffit de rassembler un nombre suffisant de textes en rapport, de *notre* point de vue humain, avec ce sujet, et qui en présentent un échantillon significatif des expressions possibles. A partir de là, plusieurs approches sont réalisables, selon la façon dont on représente les textes, dont on les abstrait.

On peut les représenter par des profils de mots. A notre connaissance, les applications de cette approche concernent le problème du classement (en anglais : *classification*) de documents textuels, c'est-à-dire de l'attribution automatique à tout nouveau document répertorié d'une catégorie prise dans une liste prédéfinie, par exemple l'attribution à une page web d'une rubrique dans l'arborescence d'un guide d'information en ligne.

De nombreuses méthodes statistiques, souvent formalisées en tant que modèles neuronaux, sont disponibles pour résoudre ce premier problème dit de « discrimination », linéaire ou non (on parle aussi d'*apprentissage supervisé*). C'est ainsi qu'à partir d'une base initiale de pages web classées manuellement dans les catégories de son guide en ligne, le moteur de recherche Voilà [VOI] fait appel à une technique issue de France Télécom Développement (ex-CNET) pour assurer le classement du flux des nouvelles pages répertoriées, plutôt que de confier cette tâche, sous le contrôle indispensable et coûteux, de son personnel documentaliste, aux éditeurs de sites désirant se faire répertorier comme le font la plupart des autres guides en ligne.

On peut représenter les textes par leurs profils de n-grammes (cf. note plus haut). Nous avons connaissance d'une seule application de ce type, mais elle est importante puisqu'elle semble mise en œuvre au sein du réseau Echelon. Elle consiste à faire définir par un analyste le sujet qui l'intéresse sous forme d'un ensemble de textes, puis à comparer au profil global de n-grammes de cet ensemble les profils de tous les textes balayés, pour ne retenir que ceux dont l'indicateur de similarité est supérieur à un certain seuil. Cette technique a été publiée en 1995 [DAM 95] et un brevet a été pris par la NSA. Elle présente l'avantage considérable de ne dépendre ni de la langue, ni de l'écriture, et d'être robuste face aux fautes diverses, coquilles et autres résidus de débalisation des textes : après constitution dans une langue donnée d'une base d'exemples par un expert, les textes japonais ou chinois sont filtrés aussi efficacement que les textes anglais.

Enfin, bien que l'on sorte *stricto sensu* du domaine de l'internet actuel, nous décrirons une technique liée au monde du renseignement et située à la limite de l'état de l'art, mais lourde de conséquences : il s'agit du repérage automatique du sujet des conversations téléphoniques.

Les auteurs du logiciel Semantic Forest, financé et utilisé par la NSA, ont publié leur approche [SCH 97] et pris un brevet. Le principe en est simple : chaque mot de la langue peut être décrit par l ' « arbre » de ses significations, plus ou moins touffu selon le nombre de concepts associés, chacun doté d'un poids en fonction de sa fréquence et d'autres considérations. Dans l'exemple connu « L'astronome a épousé une étoile », *astronome* est défini par le concept d'*homme* avant d'être lié à celui d'*astre* (notes respectives : disons 1 et 0.5), *épouser* est lié à parts égales à *homme* et *femme* (1 et 1), *étoile* peut-être indifféremment une *femme* ou un *astre* (1 et 1). La somme des notes pour chaque concept avantage *homme* (2) et *femme* (2) au détriment d'*astre* (1,5), établissant ainsi un contexte de mariage charnel plutôt que métaphorique… A plus grande échelle, on comprend que les transcriptions automatiques de conversations téléphoniques, longues suites de mots incertains et le plus souvent erronés, puissent donner, après percolation à travers les ramifications d'un dictionnaire sémantique une indication sur leur teneur, un peu comme nous devinons petit à petit de quoi il s'agit quand nous entendons deux étrangers parler entre eux dans une langue que nous dominons mal.

Si cette technique paraît bien au point et sans mystère, la plus grande difficulté[8] réside en amont dans le processus de transcription automatique, aux deux niveaux de la reconnaissance des phonèmes et de leur regroupement en mots : à partir de quel taux d'erreur de reconnaissance Semantic Forest perd-il toute fiabilité, pour du son téléphonique et de la parole continue multilocuteurs ? Un mot correctement reconnu sur 3 ? sur 10 ? sur 100 ? Cependant on ne peut pas nier que le problème soit circonscrit, et l'on peut dire que le *topic spotting* sur les conversations téléphoniques ne dépend plus que d'une multitude de mises au point et micro-progrès à tous les niveaux de la chaîne, qui se sont déjà produits ou se produiront nécessairement un jour.

Le succès des approches supervisées de détection de thèmes que nous venons de passer en revue repose sur la qualité du travail de catégorisation opéré par les analystes qui les initialisent, par regroupement manuel de documents autour d'un thème ou mise à jour d'un dictionnaire sémantique. Mais ces opérateurs peuvent 1) faire une mauvaise analyse, c'est-à-dire vouloir retrouver dans les données des concepts vagues ou à la mode qui n'y sont pas, 2) faire une bonne analyse, mais disposer d'un ensemble de textes

8. Une autre difficulté de taille est d'établir – et surtout de mettre à jour – un dictionnaire sémantique complet inventoriant les sens métaphoriques et argotiques, les détournement de mots courants dans les vocabulaires techniques, les régionalismes… vaste programme ! Tâche infinie !

insuffisant qualitativement ou quantitativement pour en rendre compte. D'où l'intérêt pour qui ne s'intéresse pas à des thématiques trop ponctuelles, d'une approche *non supervisée*, où un processus automatique repère les groupements « objectifs » de documents, les zones de forte densité de l'espace des données[9].

Faire émerger l'essentiel, sans idées préconçues, à partir d'un enchevêtrement de liens

Quand on s'intéresse aux liens *explicites* entre pages web, et qu'on s'aperçoit que la structure des liens sur un site de taille moyenne devient vite inextricable, la première idée qui vient à l'esprit pour en offrir une vue d'ensemble est de dessiner pour l'utilisateur, sur le plan de l'écran, un graphe représentant les pages (« nœuds ») par des ronds ou des rectangles, et les liens par des arcs entre ces nœuds. Très vite se posent des problèmes de présentation optimale, comme disposer les nœuds de façon à ce que les arcs évitent au maximum de se croiser, ou disposer les libellés des nœuds en évitant au maximum qu'ils se recouvrent. Mais dès que le nombre de nœuds dépasse quelques dizaines, malgré les trésors de « design » dépensés jusqu'ici (nœuds et libellés déplaçables, utilisation de la couleur, de la perspective…), le graphe devient illisible et d'autres techniques de visualisation doivent prendre la relève, comme la possibilité de centrer une vue *fish-eye* sur un nœud particulier quand on clique dessus – les relations de voisinage de ce nœud apparaissent alors clairement, tandis que les relations entre voisins de plus en plus éloignés se trouvent « tassées » à la périphérie du graphe [WAL].

Représenter de façon accessible des graphes immenses comme ceux qu'on peut définir sur le web est un domaine de recherche en soi, mais se heurte toujours aux limites physiques de l'écran : une solution pour en sortir est de passer à un niveau supérieur de synthèse et d'abstraction, c'est-à-dire de grouper les nœuds par paquets, ou classes, de rôles semblables ; les nœuds d'une classe pointant vers (ou étant pointés par) *grosso modo* le même ensemble de voisins. En termes plus formels, ceci revient à classer – sans superviseur, c'est-à-dire sans idée préconçue – les éléments dont les relations sont décrites par un tableau carré croisant ces éléments avec eux-mêmes ; la

9. Le but étant d'extraire du sens, c'est-à-dire des thèmes humainement « parlants », cette démarche ne dispense pas d'un travail de va-et-vient entre l'interprétation des regroupements obtenus, la lecture des plus typiques des textes regroupés, et le reparamétrage de l'analyse, que ce soit pour en ajuster la finesse (le « grossissement ») ou pour préciser les contours du corpus analysé.

force de chaque relation est alors traduite par un nombre à l'intersection d'une ligne et d'une colonne, par exemple un « 1 » si telle page pointe vers telle autre, un « zéro » sinon.

Peu de travaux sur de telles représentations de faisceaux de liens explicites sur le web semblent avoir débouché de façon opérationnelle ; par contre rien n'empêche de les envisager quand le tableau carré évoqué ci-dessus traduit des liens *calculés* : par exemple, d'un tableau de cooccurrences de mots dans un ensemble de pages on peut tirer des classes de mots homogènes, traduisant les grands thèmes dont il est question dans l'ensemble des pages. La méthode des mots associés [COU 96], [NOY 96] qui fonde le logiciel Sampler, ou a fondé la fonction AltaVista/Refine, part d'un tel principe ; elle place les classes de mots les unes par rapport aux autres sur un graphe global, lisible et aéré, et les mots eux-mêmes forment des graphes locaux pour chaque classe ; l'épaisseur des traits entre classes ou mots traduit la force du lien d'apparition dans les mêmes pages qui les unit.

Un bon nombre d'autres méthodes traitent de tels tableaux carrés de cooccurrence de mots. Mais l'information qu'ils véhiculent peut être extraite directement du tableau rectangulaire brut (documents × mots) répertoriant la présence ou la fréquence de chaque mot dans chaque document, qui a servi à le construire. Ce tableau sert de matière première à deux grandes familles de méthodes, dites directes, de synthèse d'informations, ainsi qu'à une famille hybride.

Méthodes factorielles

Un ensemble de textes décrits par des fréquences de mots (ou de n-grammes) peut être considéré techniquement, même si c'est inaccessible à notre intuition, comme un nuage de points, représentant chacun un texte, dans un espace comportant autant de dimensions que de mots (ou de n-grammes) différents répertoriés. Le principe de l'analyse factorielle consiste à réduire ce nombre de dimensions en « photographiant » le nuage de façon optimale, c'est-à-dire en perdant le moins d'information possible quant à la forme générale du nuage, en respectant au mieux les distances originelles entre points. Pour prendre une analogie triviale, une girafe, « objet » défini dans 3 dimensions, sera mieux caractérisée sur une photo, c'est-à-dire en deux dimensions, si on la saisit de profil plutôt qu'en vue de dessus ou de devant, et pour réduire la surface de cette photo, en découpant celle-ci parallèlement à « l'axe principal » de l'animal, c'est-à-dire son cou…

Une première façon d'exploiter cette réduction de dimensions est de représenter le nuage de points dans ses deux ou trois dimensions

« synthétiques » les plus marquantes : c'est de cette façon qu'on utilise le plus souvent l'analyse factorielle des correspondances (AFC), variante devenue en trois décennies une méthode de routine pour le dépouillement d'enquêtes, parmi d'autres, dans beaucoup de pays. Grâce à des logiciels comme SPAD/T, Hyperbase, Le Sphinx [JAD] il est devenu possible d'appliquer cette technique à la représentation synthétique de corpus textuels, en particulier ceux issus d'internet, dans un objectif d'analyse de discours ou de réponses à des questions ouvertes, voire d'étude littéraire [VIP 97]. Les cartes factorielles ainsi obtenues font coexister les points-textes avec les points-mots (analyser les textes revient à analyser les mots), ce qui permet généralement d'interpréter, de donner une signification aux axes obtenus, pour obtenir une vue d'ensemble des logiques à l'œuvre dans le corpus.

L'inconvénient principal de la méthode découle de cette représentation : dès que l'on a plusieurs milliers de documents et de mots, ce qui arrive très vite sur l'internet, les cartes deviennent illisibles ; d'autre part il n'est pas certain que deux ou trois dimensions suffisent pour résumer l'essentiel d'un corpus très fourni.

Une autre façon d'opérer est de ne pas chercher à interpréter les nouveaux axes de coordonnées obtenus, et de se contenter d'utiliser la propriété de réduction du nombre de dimensions, afin de se situer dans un espace sémantiquement significatif et dégagé de la part de bruit et d'arbitraire inhérente au langage naturel : c'est ce que réalise la méthode Latent Semantic Analysis [DEE 90], dont une variante a été utilisée au départ par le moteur Excite pour présenter à l'utilisateur 1) la similarité entre pages web dans un espace de dimensions réduites, empiriquement établi cependant à quelques centaines de dimensions, 2) la similarité entre mots dans le même espace. De cette façon, deux documents qui n'ont à la limite aucun mot commun peuvent être constatés comme proches s'ils parlent du même sujet. Des applications psycho-pédagogiques sont en ligne sur le site [LSA], en particulier la notation automatique de rédactions…

Méthodes de classification automatique

A ne pas confondre avec les méthodes de classement mentionnées plus haut, qui sont des méthodes supervisées : les méthodes de classification automatiques dégagent d'elles-mêmes, sans qu'on ait besoin de leur injecter de connaissances sur les « bonnes » classes à détecter, des ensembles de documents homogènes, en maximisant l'homogénéité interne des classes et l'hétérogénéité des classes entre elles, du point de vue du vocabulaire employé.

Plusieurs familles de méthodes sont utilisées : les méthodes hiérarchiques, ascendantes ou descendantes, comme celle qui est à l'œuvre dans le logiciel d'analyse de données textuelles Alceste [REI 00], les méthodes dites à centres mobiles, comme celle utilisée pour des applications scientométriques dans [NOY], où une carte d'ensemble de type factorielle montre la disposition des classes entre elles, plutôt que celle des mots et des documents, trop nombreux.

Le modèle neuronal dit carte auto-organisatrice (*Self-Organizing Map*) de T. Kohonen [SOM] permet de réaliser simultanément une classification des documents et un placement des classes sur une grille représentant un pavage régulier de « neurones » incarnant les classes. Pour une application à la navigation dans une base bibliographique sur l'internet, voir le site [SIM].

Méthodes mixtes

Dans notre propre ligne de recherche, nous avons développé au CNRS/INIST puis aux universités de Paris 8 et Franche Comté des méthodes intermédiaires entre l'analyse factorielle et la classification automatique, qu'on peut décrire comme des méthodes de classification floue et recouvrante, où chaque thème regroupe des documents homogènes dotés chacun d'une valeur de « typicité », de centralité dans ce thème, ainsi que de mots caractéristiques de ce thème, eux aussi avec une valeur de centralité. Cette représentation est propice à la traduction d'effets de contexte, car un mot peut apparaître comme central dans plusieurs contextes, prenant des sens légèrement ou profondément différents. Un document traitant de plusieurs sujets apparaîtra lui aussi comme important dans plusieurs thèmes.

On peut également qualifier ces méthodes de factorielles, dans la mesure où les centralités d'un document ou d'un mot définissent ses coordonnées dans un repère qui n'est plus orthogonal, comme dans les analyses factorielles classiques, mais oblique, où chaque axe factoriel correspond à un thème *interprétable*, même s'il vient au cinquantième rang par ordre d'importance (il est rare de pouvoir interpréter les axes d'analyse factorielle classique au-delà du quatrième ou cinquième). Pour plus de précision sur les algorithmes, voir [LEL 94], sur les logiciels NeuroNav et CartoWeb : [LEL 01], [CAR].

Des synthèses pour quoi faire ?

Pour qui a les moyens de « visualiser les masses de liens », des réseaux sociaux deviennent lisibles sur le Net ; à la différence des réseaux de la socialité habituelle, la proximité physique, géographique peut y être remplacée – ou traduite – par celle directement exprimée par les liens hypertextes, mais aussi par la (ou les) proximité(s) déduite(s) par ces observatoires du virtuel que sont les moteurs de recherche, qui se dotent petit à petit d'» instruments d'optique » de plus en plus puissants : Google donne les meilleures pages de référence sur un sujet donné, Clever ne va pas tarder à nous donner en plus les meilleures pages-passerelles, Kartoo [KAR] nous rend évidents, par paquets réduits d'une quinzaine de pages, les liens d'association de vocabulaire au sein de chaque paquet, liens qu'AltaVista/Refine a échoué commercialement, et peut-être ergonomiquement[10], à nous rendre évidents à l'échelle de milliers de pages-réponses. Mais des dizaines de prototypes en gestation dans les laboratoires, comme [ZAM 99], [NEU], [PAC], sont en train de définir les contours des « optiques » de la prochaine génération, celles qui nous permettront de combiner les bons filtres, le bon tissu à observer (liens explicites ou implicites ? de quelle nature ?…), et le bon grossissement, tout paramètre que nous apprendrons à adapter itérativement à nos objectifs de recherche d'information.

Une autre voie possible pour l'exploitation de ce tissu de liens nous est suggérée par le développement de la scientométrie, ou étude de la vie des courants scientifiques et techniques à partir des fiches bibliographiques stockées dans les grandes bases documentaires mondiales (articles scientifiques, brevets…) : depuis les 2 ou 3 décennies que ces bases existent sur support électronique, elles constituent la matière première de l'observation du mouvement vivant des sciences et techniques par des sociologues [POL 96] et des chargés d'études auprès des institutions où se décident (ou s'infléchissent) les politiques scientifiques (ministères, grands organismes de recherche…). Pour les sociologues du Centre de sociologie de l'innovation [CAL 89] les articles scientifiques et les brevets représentent bien plus que des traces de l'activité scientifique et technique : ils en sont les objets-acteurs principaux – les citations et le vocabulaire employé, les « mots-bannières », créent la dynamique d'alliance et d'exclusion entre « acteurs-réseaux » humains et non humains. C'est dans et pour ce milieu que sont apparues en premier les méthodes de synthèse d'informations, de

10. Les mots n'étaient que les chaînes de caractères brutes répertoriées par AltaVista, et l'accès aux documents typiques de chaque *cluster* de mots était difficile.

représentations cartographiques, appliquées à la même époque (début des années 1980) aux Etats-Unis aux liens de cocitation entre articles [GAR 98], et en France puis aux Pays-Bas aux liens calculés à partir du vocabulaire d'indexation [COU 96], [NOY 98].

Si l'on extrapole à la Toile, la situation nouvelle dans laquelle il sera possible de repérer et d'explorer les zones de forte densité d'entrelacs de liens, liens tant explicites que calculés – service fourni *a priori* par les moteurs ou méta-moteurs de recherche -, sera largement inédite : en effet ces zones, aujourd'hui largement invisibles, marquent la condensation de processus collectifs, mis de façon volontaire et en connaissance de cause sur la vaste place publique que constitue le web par les individus sociaux que nous sommes. Les outils que nous avons brièvement mentionnés permettraient alors l'exploration et le parcours de cette « géographie virtuelle » et mouvante (pour une veille sur ce domaine, voir le site Cybergeography [CYB]).

Une réflexion voisine, bien que distincte par son objet, est menée actuellement par les milieux qui explorent le concept de « mémétique », ou métaphore de l'évolution génétique appliquée au domaine des idées [MEM]. Le concept de *Global Brain* issu de ce courant, propose la métaphore de l'entrelac des liens du web comparé à celui des neurones, liés dans la matière grise par les axones et leurs ramifications dendritiques. Ces pistes sont intéressantes, et sans doute riches de progrès possibles pour les sciences humaines – contribueront-elles à dégager des « observables » dont l'analyse puisse être susceptible de réfutation ? Elles débouchent parfois sur la généralisation, prématurée à notre avis, des conséquences de certains concepts considérés comme acquis, souvent dans une optique de critique sociale : ainsi [BOL 96] voit dans le *Global Brain* la naissance d'une pensée conformiste à l'échelle planétaire – alors que d'autres, y compris parmi les proches de la mémétique, y voient au contraire un émiettement en de multiples particularismes s'ignorant les uns des autres, et le renforcement d'une tribalisation générale.

Loin de ces spéculations à long terme, où les auteurs de science-fiction nous paraissent les mieux placés pour forcer le trait dans la description, apocalyptique ou non, de mondes possibles, nous nous contenterons de questions plus circonscrites. En effet, au-delà des opérations ponctuelles de *Web-mining*, la typologie des pages web [BOR 98], leur catégorisation automatique à partir de leurs liens ou de leurs contenus par des neurones artificiels sont des opérations à visée « panoramique » qui seront bientôt possibles dans la pratique quotidienne du web, ne nécessitant qu'un changement d'échelle par rapport aux prototypes des laboratoires. Sous

quelles formes se réalisera leur assimilation sociale ? Serviront-elles en premier lieu, comme d'autres avant elles, aux objectifs des services de renseignements, si tant est que ceux-ci se dotent de la culture permettant d'en tirer parti ? Ou bien, dans le prolongement de la scientométrie actuelle, fourniront-elles d'abord des techniques de « photographie du virtuel », matière première pour des analyses (enfin) réfutables, et bases d'une capitalisation d'acquis à la fois théoriques et empiriques dans les sciences humaines et sociales ? Apparaîtront-elles publiquement en priorité en tant qu'outils pour la consultation de collections homogènes et spécialisées sur le web, comme les revues électroniques ou les bases de *preprints*, ou encore les interfaces vers les bases de données du web invisible, comme [SIM] ? Ou bien leur utilisation par monsieur Tout-le-monde dans des moteurs de recherche, qui sont actuellement les mieux placés pour les proposer, sera-t-il au contraire l'événement déclencheur des autres usages, comme pourrait le préfigurer le succès de Kartoo [KAR] (cartographie à partir des mots à l'échelle restreinte d'une quinzaine de pages-réponses) ?

Autant de questions auxquelles nous ne nous hasarderons pas à répondre, échaudés par le constat de la lenteur de pénétration de ces techniques de *text-mining* (ou fouille de textes) dans les entreprises depuis une quinzaine d'années, malgré la généralisation de l'information sous forme électronique, malgré les discours incessants sur l'excès d'informations, sur la nécessité d'extraire et gérer la connaissance, sur la mémoire et les savoir-faire d'entreprise à préserver… Processus jalonné de multiples faillites ou difficultés chroniques des entreprises actuellement sur ce créneau de logiciels et services, processus incertain posant peut-être question de la place de la réflexion au-delà du court terme et des sciences sociales dans l'entreprise : la pente naturelle y est d'attendre des synthèses presse-bouton, des réponses aux questions posées dans la langue de bois à la mode à l'instant donné (bien décrite dans [VOL 00]), toutes choses antinomiques d'un processus d'aller-retour entre élaboration des données et des problématiques, paramétrages des analyses, et interprétation des résultats pour éclairer l'action.

On peut voir dans les perspectives de catégorisation et de cartographie du web une opportunité fantastique pour le développement des sciences sociales, qui y trouveront à la fois un gisement d'observations inespéré, autant que d'expérimentations participantes, sur les traces de la scientométrie, d'un impact sans doute considérable tant sur leurs théories que sur leurs pratiques.

On peut y voir aussi le danger de telles méthodes appliquées de façon perverse aux flux de messages personnels, pour l'instant textuels, plus tard

téléphoniques et visiophoniques, dans des buts de contrôle social. L'espionnage « ponctuel » d'aujourd'hui y serait remplacé par l'espionnage des âmes… Mais l'histoire n'est jamais écrite à l'avance : toute action suscite une réaction, immédiate ou latente ; d'ajustements en ajustements, de crises en crises, les choses vont leur cours, et nul ne peut sérieusement se targuer « d'avoir tout dit » à l'avance sur l'assimilation sociale d'une technologie de l'information et de la communication.

On peut y voir enfin le support d'une boucle d'autorégulation sociale d'un type nouveau, où les acteurs sociaux disposeraient en temps réel de cartographies de leur identité et de leurs relations, ce qui ne manquerait pas d'influer en retour sur cette identité et sur ces relations, et sur la dialectique transparence/opacité à l'œuvre dans toute communication humaine... Mais ceci relève pour l'heure, où une faible proportion de relations sociales est concernée par l'internet, malgré son développement, de la science (sociale)-fiction !

Références

[ADE 00] LELU A., Autour de NeuroWeb : acquis et interrogations sur la cartographie de l'information, séminaire ADEST du 22/05/2000.
www.upmf-grenoble.fr/adest/seminaires/lelu2000

[AMA] Campagnes d'évaluation AMARYLLIS et CLEF.
http ://amaryllis.inist.fr ; cf. aussi http ://clef.iei.pi.cnr.it

[AUB 01] AUBIN S., Présentation de Neuronav, séminaire ADEST du 13/11/2001.
www.upmf-grenoble.fr/adest/seminaires/diatopie.ppt

[BAE 99] BAEZA-YATES R., RIBEIRO-NETO B., *Modern Information Retrieval*, Reading, Ma., ACM Press et Addison-Wesley, 1999.

[BOL 96] BOLLEN J. & Heylighen F., « Algorithms for the self-organisation of distributed, multi-user networks. Possible application to the future World Wide Web », *Cybernetics and Systems '96*, R. Trappl ed., 1996, p. 911-916.
http ://pcp.vub.ac.be/papers/SelfOrganWWW.html

[BOR 98] BORZIC B., Un modèle de gestionnaire iteratif de flux informationnel sur internet, Thèse de doctorat, Information Scientifique et Technique, Paris, avril 1998, CNAM/CNRS.

[CAL 89] CALLON M., *La science et ses réseaux*, Paris, La Découverte, 1989.

[CAR] Démo. CartoWeb, www.diatopie.com/DemoCartoWeb.com

[CHA 98] CHAKRABARTI S., B. Dom, D. Gibson, J. Kleinberg, P. Raghavan, and S. Rajagopalan., « Automatic Resource Compilation by Analyzing Hyperlink Structure and Associated Text », *Proceedings of the 7th World-Wide Web conference*, Elsevier Sciences, Amsterdam, 1998.
http ://www7.scu.edu.au/programme/fullpapers/1898/com1898.html

[CIT] Base de publications en libre accès CiteSeer, http ://citeseer.nj.nec.com

[CON] Lettre Confidentiel-Défense, juillet 2000.
www.confidentiel-defense.com/anciens/numero%201/nsa.htm

[COU 96] J.P. COURTIAL, « Construction des connaissances scientifiques, construction de soi et communication sociale », revue en ligne *Solaris* n° 2, 1996.
www.info.unicaen.fr/bnum/jelec/Solaris/d02/2courtial.htm

[CYB] Site Cybergeography, www.cybergeography.org

[DAM 95] DAMASHEK M., « Gauging Similarity with N-grams : Language-Independent Categorization of Text », *Science*, vol. 267, p. 843-848, 1995.

[DEE 90] DEERWESTER, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). « Indexing By Latent Semantic Analysis », *Journal of the American Society For Information Science*, vol. 41, p. 391-407.
http ://lsa.colorado.edu/papers/JASIS.lsi.90.pdf

[EPI] Electronic Privacy Information Center (EPIC).
www.epic.org/privacy/carnivore

[GAR 98] GARFIELD E., « Mapping the World of Science », Institute for Scientific Information®, *The Scientist*® Publisher, Philadelphia, 1998.
http ://www.zbp.univie.ac.at/gj/citation/mapsciworld.htm

[GOOa] Moteur GOOGLE/Groupes/Forums de discussion Usenet.
www.google.fr/grphp

[GOOb] Moteur GOOGLE, www.google.fr

[HAG 96] HAGER N., Secret Power : New Zealand's role in the International Spy Network, N.Z. : Craig Potton, Nelson, 1996.
Cf. aussi www.bullatomsci.org/issues/2000/ma00/ma00richelson.htm

[HAL 01] M. Hallab, Hypertextualisation automatique multilingue à partir des fréquences de N-grammes, Thèse de l'université Paris 8, 2001.

[IKO] Démo IKONA (INRIA),
http ://www-rocq.inria.fr/cgi-bin/imedia/ikona/exec

[JAD] Journées Internationales d'Analyse des Données Textuelles 1998 et 2000.
http ://www.cavi.univ-paris3.fr/lexicometrica/jadt/jadt1998/JADT1998.htm
http ://www.cavi.univ-paris3.fr/lexicometrica/jadt/jadt2000/tocJADT2000.htm

[KAR] Méta-moteur KARTOO, www.kartoo.com

[KRU 94] C. KRUMEICH, Intervention au 3e Forum de l'Intelligence Economique et Concurrentielle, Sophia-Antipolis, 8 décembre 1994.
www.scipfrance.org/documents.htm

[LIN] Site : Linux Security, http ://echelon.linuxsecurity.com/keywords.htm

[LEL 00] LELU A., Hallab M., « Consultation floue de grandes listes de formes lexicales simples et composes : un outil preparatoire pour l'analyse de grands corpus textuels », *actes de : JADT'2000*, coord. : M. Rajman, EPFL, Lausanne, mars 2000.
www.cavi.univ-paris3.fr/lexicometrica/jadt/jadt2000/pdf/81/81.pdf

[LEL 01] LELU A., Aubin S., Vers un environnement complet de synthèse statistique de contenus textuels, séminaire ADEST du 13/11/2001.
www.upmf-grenoble.fr/adest/seminaires/lelu02/ADEST2001_SA_AL.htm

[LEL 02] LELU A., « Comparaison de trois mesures de similarité utilisées en documentation automatique et analyse textuelle » – *JADT'2002*, coord. IRISA, St. Malo, 13-15 mars 2002.

[LEL 94] LELU A., « Clusters and factors : neural algorithms for a novel representation of huge and highly multidimensional data sets », *New Approaches in Classification and Data Analysis,* E. Diday, Y. Lechevallier & al. eds., p. 241-248, Springer-Verlag, Berlin, 1994.

[LEL 96] LELU A., « De l'émergence des concepts : réflexions à partir du traitement 'neuronal' des bases de données documentaires », revue en ligne *Solaris*, n° 2.
www.info.unicaen.fr/bnum/jelec/Solaris/d02/2lelu.htm

[LEL 98] LELU A., M. Hallab, B. Delprat, « Recherche d'information et cartographie dans des corpus textuels à partir des fréquences de N-grammes », *Actes de JADT'98*, coord. S. Mellet, UPRESA « Bases, corpus, langages », Nice, 1998.
www.cavi.univ-paris3.fr/lexicometrica/jadt/jadt1998/lelu.htm

[LSA] Démos Latent Semantic Analysis, http ://lsa.colorado.edu

[LSI] DUMAIS S., Telcordia Technologies, démo. Latent Semantic Indexing.
http ://lsi.research.telcordia.com/lsi-bin/lsiQuery

[MED] Base bibliographique MEDLINE PubMed, fonction « Related articles ».
http ://www.ncbi.nlm.nih.gov/entrez/query.fcgi

[MEM] Journal of Memetics, www.cpm.mmu.ac.uk/jom-emit/

[NEU] Prototype NeuroWeb (A. Lelu *et al.*).
http ://hdyn.hymedia.univ-paris8.fr/neuroweb

[NOM] Moteur Nomino, www.gouv.qc.ca/gouv/gouvqc/index.asp

[NOY 96] NOYER J.M., « Utilisation d'un outil Infométrique,"CANDIDE" dans le contexte d'une réflexion stratégique », revue en ligne *Solaris*, n° 2, 1996.
www.info.unicaen.fr/bnum/jelec/Solaris/d02/2noyer_2.htm

[NOY] Site des travaux scientométriques d'Ed. Noyons, www.cwts.nl/ed

[NOY 98] NOYONS E.C.M. and A.F.J. van Raan, « Mapping Scientometrics, Informetrics, and Bibliometrics », CWTS Working papers, June 1998 (article interactif), http ://www.cwts.nl/ed/sib/home.html

[PAC] Démos de visualisation d'informations du Pacific National Laboratory. www.pnl.gov/infoviz

[PAS] Bases bibliographiques PASCAL et Francis (CNRS/INIST, Nancy). www.inist.fr

[PER] logiciel PERTIMM, de SYSTAL S.A. www.systal.com, démo (12 ans de Journal officiel) http ://pertimm.ensmp.fr

[POL 96] POLANCO X., « Aux sources de la scientométrie », revue en ligne *Solaris*, n° 2, 1996, www.info.unicaen.fr/bnum/jelec/Solaris/d02/2polanco.htm

[REI 00] REINERT M., « La tresse du sens et la méthode "Alceste". Application aux "Rêveries du promeneur solitaire" », *JADT2000*, Lausanne, 2000. www.cavi.univ-paris3.fr/lexicometrica/jadt/jadt2000/pdf/31/31.pdf voir aussi www.image.cict.fr/alceste.html

[ROB 02] numéro spécial « Les nouveaux robots », *La Recherche*, n° 350, Paris, 2002.

[SCH 97] SCHONE P., Townsend J.L., Olano C., « Text Retrieval via Semantic Forests », *The Sixth Text REtrieval Conference (TREC-6),* NIST Special Publication, Gaithersburg, mar., 1997, http ://trec.nist.gov/pubs/trec6/papers/nsa-rev.ps

[SIM] SIMBAD, navigateur pour bases bibliographiques en astronomie, Université de Strasbourg, http ://simbad.u-strasbg.fr/A+A/map.pl

[SOM] Site web Self-Organizing Maps, Université d'Helsinki. http ://websom.hut.fi/websom

[TRE] Text Retrieval Evaluation Conference, http ://trec.nist.gov

[VIP 97] VIPREY J.M., *Dynamique du vocabulaire des* Fleurs du Mal, Champion-Slatkine, Paris, 1997.

[VOI] Moteur VOILA/recherche thématique. http ://themes-search.voila.fr/ ?theme=2047

[VOL 00] VOLLE M., *E-conomie*, Paris, Economica, 2000. www.volle.com/e-conomie/table.htm

[WAL] Visualisation hyperbolique 3D Walrus, www.caida.org/tools/visualization/walrus http ://www.graphics.stanford.edu/~munzner/papers

[ZAM 99] ZAMIR O., ETZIONI O., « Grouper : A Dynamic Clustering Interface to Web Search Results », *8th WWW Conference*, 1999. www.cs.washington.edu/research/projects/WebWare1/etzioni/www/papers/www8.pdf