La veille sur Internet

Une avancée dans la recherche de l'information stratégique

Ismaïl Timimi, Jacques Rouault

Jacques Rouault est professeur de sciences de la communication.

Mathematicien de formation, Ismail Timimi est docteur en informatique et communication. Il s'intéresse au traitement de données textuelles par des approches numériques (le logiciel 3AD développé dans sa thèse calcule des distances entre phrases dans un but de recherche d'information). Actuellement, il est Ater en Ingénierie de la Langue à l'université de Lille 3 et membre du comite de pilotage du projet Aupelf (ARC3) sur l'évaluation des extracteurs terminologiques.

Plan

Avant propos

Progressivement, la veille est devenue un vocable « à la mode » pour certains, mais de grande nécessité pour d'autres ; on le trouve fréquemment employé dans le milieu industriel (à des fins décisionnelles ou concurrentielles) et le milieu académique au sein des laboratoires et départements de recherche et de développement. Nous voudrions dans ce premier numéro de la revue, et à travers cet article, non pas soulever une nouvelle problématique de ce sujet ou en résoudre une ancienne, mais plutôt présenter à grands traits à une communauté scientifique peu familiarisée avec la veille un état des lieux des pratiques actuelles. Dans l'équipe Cristal – membre du Gresec – les travaux sur l'indexation et la recherche d'information fondées sur le Traitement Automatique des Langues (analyses linguistiques et statistiques des textes) s'inscrivent pleinement dans une première étape d'un processus de veille, et pourront éventuellement faire l'objet d'une suite.

Qu'est-ce qu'on sur-veille ?

C'est parce que nous sommes en train de vivre quotidiennement un passage inéluctable et massif d'une documentation sur son support traditionnel vers un gigantesque gisement documentaire en ligne, c'est parce que ce gisement avec ses liens hypertexte renouvelables régulièrement représente des documents volumineux et ouverts à l'évolution, et c'est parce que nos besoins en information sont assujettis à cette révolution, qu'il est devenu vital et urgent de prendre conscience de cette révolution documentaire et de l'accompagner des méthodes et des outils techniques appropriés.

La plupart des experts du management de l'information reconnaissent au moins le triplet (collecte, traitement, diffusion ou mise à disposition) comme les maillons de la « chaîne documentaire », aussi bien pour l'information traditionnelle sur support papier que pour l'information numérique véhiculée par des réseaux informatiques. Mais les informations en ligne représentent actuellement une telle quantité de données qu'il est de plus en plus difficile de la maîtriser. Ainsi, au-delà du triplet assez simple précité, certains ont recouru à un traitement plus complexe qui inclut des procédures linguistiques et des approches statistiques pour analyser le contenu de ces données (après un tri et une sélection), les indexer, et les classer par thèmes. D'autres ont suggéré que le management de l'information recouvre d'autres tâches comme l'analyse de l'expression des besoins, la mise en forme (reformatage), la mémorisation et l'actualisation, le résumé de l'information, la détection des signaux d'alerte, l'interprétation de signaux faibles, et l'aide à la décision...

Plus encore, ce genre d'information en ligne évolue si vite qu'il faut mobiliser des ressources humaines et logistiques importantes. Suivant les besoins, trois ordres de surveillance s'imposent : une surveillance pour un renseignement donné (étude de marché, fiche technique d'un produit concurrentiel), une surveillance subordonnée à un besoin immédiat (dossier sur un sujet précis), et enfin une surveillance régulière de fonds qui consiste non seulement à aller chercher intelligemment une information solution, mais dans le sens inverse aussi, c'est l'information qui vient avertir le veilleur de toute évolution, grâce au déploiement d'une panoplie d'outils de veille. Les deux premières ont un caractère d'une recherche rétrospective, la dernière d'une alerte.

L'association de différents traitements (linguistiques, statistiques, économiques, informatiques, etc.) ont fait de la veille une passerelle entre la recherche académique d'une part et l'univers industriel d'autre part. Et plusieurs veilles ont vu le jour : la veille *diplomatique* à caractère politique international, la veille *scientifique* à des fins de coopération, la veille *technologique et stratégique* qui se veut comme un management de connaissances orienté vers l'anticipation, le *marketing* et *l'intelligence économique* pour des raisons évidentes, etc. Toutes ces veilles sont subordonnées à un seul impératif : la surveillance de la *concurrence* qui joint l'anticipation à la prospection.

Par exemple, une veille économique sur un réseau (en particulier d'Internet) sert plusieurs intérêts : découvrir de nouveaux acteurs potentiels sur le marché ; savoir ce qui se dit sur les produits de son entreprise ; mieux connaître son environnement technologique.

De même, voici quelques exemples de l'intérêt des outils de l'*infométrie* et de la *bibliométrie* comme cas particulier de veille. Ces outils se définissent comme l'exploitation statistique ou dynamique des publications, ils permettent de rendre compte de l'activité des producteurs (chercheur, laboratoire, institut...) ou des diffuseurs (périodique, éditeur...) de l'information scientifique, tant du point de vue quantitatif que qualitatif, et ce pour plusieurs applications :

- évaluer le travail d'un chercheur, ou le définir par analyse sémantique ;
- évaluer le fonds de périodiques d'une bibliothèque ;
- suivre l'évolution d'un thème de recherche au cours du temps à travers des publications ;
- apprécier l'impact d'un article ou la qualité d'une revue.

QUELLE INFORMATION DANS QUELLE SOURCE ?

La veille est alors plus qu'une recherche d'information documentaire systématique, c'est une recherche de l'information pertinente, dans des sources pertinentes, à des moments opportuns, avec des outils perfectionnés, puis une analyse pour une diffusion sélective et persuasive, de façon à favoriser une prise de décision (un contrôle) propice. La pertinence de l'information – particulièrement celle sur support électronique – peut se caractériser de plusieurs manières.

Le degré d'accessibilité

Information blanche : c'est une information publique et accessible à tout le monde (pages Web classiques), elle ne fait l'objet d'aucune sécurisation particulière. Pour l'avoir, on utilise les outils « classiques » de recherche d'information.

Information grise : information sensible d'accès légal, elle ne fait pas l'objet de publicité, mais on peut la trouver de manière indirecte ou détournée. Pour l'avoir, on utilise les techniques avancées de recherche et de traitement de l'information. Cette information peut nous parvenir par inscription (gratuite ou payante) dans des groupes de discussion, des listes de diffusion, ou des serveurs d'information spécialisés comme les agences de presse scientifique, *Medline*, etc.

Information noire : elle fait l'objet d'une haute sécurisation, et y accéder indûment relève de l'espionnage industriel ou scientifique, activités qui sont illégales. Ce genre d'information (pour ceux qui le souhaitent) est très difficile à consulter.

La crédibilité

En fait, pour le néophyte, sur Internet, il est quasiment impossible de bien distinguer l'information fiable du reste. Outre l'incrédibilité due à des sites dont l'information est valide mais non actualisée (par exemple des liens hypertexte morts, des informations non mises à jour...), il y a aussi l'incrédibilité due à la désinformation volontaire pratiquée par et sur certains sites à des fins concurrentielles.

Le contrôle exercé par un fournisseur sur l'information qu'il diffuse

Sources classiques (sans contrôle) : il s'agit en général d'une panoplie de sites qui traitent de thèmes très variés, allant d'une page personnelle d'un amateur débutant en langage HTML au site interactif d'une grande entreprise d'informatique prestataire de services. Ces sources sont mises à disposition des internautes sans aucun contrôle ni normes, elles sont en accès libre aussi bien pour consulter que pour éditer : Wais, http://, FTP...

Sources interactives (semi-contrôlées) : il s'agit des listes thématiques de discussion et de diffusion, des Foires aux questions (FAQ), des News... Ces outils sont un moyen simple, souvent gratuit, de se tenir au courant des développements dans un domaine mais aussi d'apprécier des échanges sur un sujet donné. Les informations dans ces sources peuvent être contrôlées par un ou des modérateurs, ou donner lieu à des débats entre les abonnés. Une FAQ standard se présente sous forme structurée (avec un sommaire) de fichiers texte où chaque fichier est concis, daté, avec désignation de son sujet, de son auteur et de son adresse électronique. Une FAQ se caractérise par une mise à jour régulière, mensuelle si possible, avec une diffusion large, visible par le plus grand nombre.

Sources cumulatives (double contrôle) : les bases de données, particulièrement celles des services bibliographiques et des brevets, sont considérées comme les moyens les plus riches en information stratégique (accès payant pour la plupart). Les éléments de ces sources subissent par contre un double contrôle : une validation à travers une

publication ou un dépôt légal, puis un recensement. Ainsi, la base *Ibiscus* qui contient plus de 115 000 références (dont 80 % avec résumé) d'études, ouvrages, périodiques sur l'économie, l'agriculture, l'industrie, l'urbanisme, l'environnement, les problèmes de société ; la base *Medline* est spécialisée dans la médecine au plan international ; on peut citer aussi les fonds documentaires de l'Inist ou de l'Inria, la base *Patents US* qui recense vingt ans de brevets américains et la base INPI avec deux ans de brevets français.

Outre ces références de nature électronique, on considère que des sources comme les « collèges invisibles », les comptes-rendus de missions, les interviews, les colloques spécialisés comme sources *semi-anticipatives* peuvent être des moyens aussi pertinents pour traquer l'information stratégique d'une manière informelle.

OUTILS CLASSIQUES DE RECHERCHE D'INFORMATION

Dans les outils « classiques » de recherche d'information sur Internet, on recense deux grandes catégories : les outils produits par des humains, appelés répertoires ou *annuaires thématiques*, et ceux qui sont produits automatiquement par des robots et qu'on appelle les *moteurs de recherche*.

Les annuaires thématiques

Pour les *annuaires thématiques*, l'administrateur du site à référencer est tenu de le déclarer préalablement par la commande « *Add URL* » de l'annuaire, en indiquant la catégorie thématique de son site. La recherche se fait par mots-clés, non pas en texte plein, mais simplement par catégorisation sous forme de thèmes hiérarchiques. Les annuaires thématiques présentent l'avantage de guider le demandeur d'information dans ses recherches en accédant successivement à des catégories de plus en plus précises. Certes, les annuaires se caractérisent par une rapidité remarquable dans la recherche, puisque l'exploration ne concerne que des sites référencés, mais l'inconvénient majeur est que ces sites référencés ne présentent qu'une partie mineure du réseau, et le risque de manquer des sites pertinents non référencés dans l'annuaire est assez probable. Comme exemple d'annuaires, on trouve Yahoo (*Yet Another Hierarchically Organized Oracle*) qui essaye d'élargir sa base de recherche en ajoutant dans le menu, des rubriques d'actualités, des forums et des adresses électroniques (mél). On trouve également l'annuaire Nomade qui se veut le concurrent direct du Yahoo français, puis Eurêka qui a été lancé en mai 1996 et revendique 8000 sites classés, QuiQuoiOù le catalogue de Wanadoo-France Telecom, et enfin d'autres annuaires à caractère géographique.

Les moteurs de recherche

Les *moteurs de recherche* proposent eux aussi à *l'offreur* de l'information la possibilité de se déclarer (de se référencer) directement dans le moteur en remplissant un formulaire, ou indirectement à travers des sites à vocation de référencement (*Submit*). Par contre, ici, la recherche se fait par le suivi du trajet des liens hypertextes. Le moteur lance plusieurs *robots logiciels* fédérateurs (*spiders*) pour scruter des pages Web de base, ces *spiders* identifient les liens hypertextes de ces premières pages et vont parcourir les autres pages associées à ces liens. Et sur ces nouvelles pages, ils procéderont de la même façon. Les *spiders* parcourent ainsi rapidement la totalité d'un site, puis d'autres sites en liaison et télédéchargent automatiquement les pages. Ce sont eux qui apportent leur puissance aux divers moteurs de recherche. Il n'est pas rare que deux moteurs utilisent les mêmes robots pour scruter le Web. C'est alors le paramétrage qui fait la différence.

En fonction du taux de rotation et des points d'entrée des *spiders* sur le Web, un moteur de recherche dispose donc d'une gigantesque base de données d'URL alimentée

régulièrement, qui donne une bonne approximation de la totalité du Web à un instant donné. La réponse à une question est une liste classée de documents. Ainsi, le moteur AltaVista paru en décembre 1995 annonce que son robot *Scooter* scrute l'Internet tous les 12 jours, Excite revendique la première place pour la quantité de pages indexées et pour ses méthodes de recherche, Lycos, créé à l'Université Carnegie Mellon, est considéré comme un des pionniers, et l'on trouve aussi des moteurs de recherche français tels que Voila, Ecila, Lokace…

Une étude américaine dans la revue *Science* indique que le Web contiendrait début 98 environ 320 millions de pages, représentant « une encyclopédie de 15 milliards de mots ». Ainsi, HotBot indexerait 34 % de cet univers, AltaVista 28 %, NorthernLight 20 %, Excite 14 %, Infoseek 10 % et Lycos 3 %. Selon les auteurs de cette étude, le fait d'effectuer ses recherches sur ces six moteurs simultanément augmenterait la taille de l'index de base d'un facteur 3,5.

Certes, les moteurs de recherche considérés comme des logiciels d'exploration et d'indexation sont plus puissants que les annuaires thématiques et permettent au demandeur de l'information d'avoir toutes les pages comportant un mot ou une partie de sa requête. Leur point fort réside dans le volume de données signalées (exploration en texte intégral), et ils sont d'autant plus efficaces qu'on utilise leurs commandes de recherche avancée :

- restriction de recherche aux pages créées depuis une date donnée, ou seulement dans des sources données (Usenet, adresses e-mail...) ;
- possibilité de lier les mots significatifs d'une requête par des connecteurs booléens (et, ou, sauf). Ainsi « veille ET faq » donne les références des pages où apparaissent ces deux termes ;
- faculté de recherche par mots exacts ou mots tronqués. Ainsi « pays* » représente une requête sur les chaînes de caractères « pays », « payse », « paysan » (et variations morphologiques : « paysans », « paysanne(s) », « paysannat », « paysannerie », « paysage », « paysagiste »...) assorties implicitement de l'opérateur booléen OU. Notons que la troncature dans certains moteurs de recherche ne peut restituer qu'un nombre limité de caractères ;
- possibilité de formuler des requêtes avec des mots adjacents ou voisins ; on peut ainsi chercher « veille technologique » comme une seule entité pour la démarquer des autres veilles, et on peut aussi chercher les mots « pays » et « développement » séparés au maximum par « 4 » mots ;
- intégration des opérateurs de balisage (URL, Title, Link...) dans la recherche. Par exemple, avec AltaVista, la commande « link:u-grenoble3.fr » retrouve toutes les pages qui comportent des liens hypertexte pointant vers le site « http://www.u-grenoble3.fr/ ». Pareillement, avec le moteur Infoseek, on peut cocher des cases d'option pour sortir les pages Web dont l'URL (ou les documents dont le titre) contient le mot « Cristal » mais pas « Grenoble ».

Ces commandes et d'autres (éventuellement couplées), font des moteurs de recherche des outils indispensables et fonctionnels dans la recherche d'information. Mais elles restent cependant insuffisantes pour les recherches routinières de surveillance ou

l'exploration approfondie de sites. Ainsi on relève certaines imperfections des moteurs de recherche (1) :

- aucune mémorisation des sessions passées dans la recherche d'information ;
- certains moteurs ne tolèrent pas, dans les requêtes, des mots débutant par un nombre ;
- les recherches sont toujours longues ;
- l'exploration textuelle implique un taux de bruit élevé ;
- il y a également un silence non négligeable dans les réponses à une requête, dû au fait que les moteurs de recherche même unis, ne peuvent pas explorer tout l'Internet (cf. l'étude américaine citée supra), d'une part, et aux limites inhérentes à la recherche d'information par appariement de chaînes de caractères d'autre part. Il y a aussi la question des langues ;
- un autre type de silence provient du fait que les moteurs de recherche n'indexent que les pages HTML « brutes » sur le Web, et non les sites accessibles par formulaires dynamiques tels que des bases de données ou des conversations dans les listes de diffusion (*mailing-lists*) ;
- un troisième type de silence est du genre « Liste rouge d'indexation ». Il existe en fait une commande dans le code HTML (ou au niveau du serveur hébergeant) qui entrave l'exploration d'une page par les moteurs de recherche. Mais cette situation de barrage reste un peu particulière, due probablement à l'usage privé du site ou au fait que ce dernier est en cours de validation ;
- il arrive que des réponses à une requête ne soient pas à jour car les recherches sont effectuées seulement dans les stocks indexés depuis la dernière exploration intégrale. Évoluant très rapidement, le contenu des fichiers sur Internet doit être vérifié régulièrement. Or ce n'est pas toujours le cas bien qu'AltaVista déclare explorer quotidiennement environ trois millions de pages.

OUTILS PRIMAIRES DE FILTRAGE

Au-delà de ces imperfections liées aux moteurs de recherche, il en existe d'autres liées au système de filtrage qui consiste à trier l'information entrante et à la catégoriser suivant des critères précis. Si ce mode s'avère très efficace pour le filtrage du courrier électronique, il présente deux inconvénients majeurs dans le cas d'articles de *news* ou des listes de diffusion :

- l'usager (en tant qu'offreur, demandeur, ou simple consultant d'information) doit s'inscrire de lui-même aux groupes qui l'intéressent, avec le nombre de groupes qui naissent ou disparaissent chaque jour, un effort de veille serait nécessaire ;
- si l'usager s'inscrit à plusieurs groupes, le nombre d'articles peut être considérable même après filtrage. On pourrait bien sûr fixer un seuil pour limiter ce nombre, mais ces outils n'étant fondés que sur une indexation automatique des documents

. . . . . .

1. Imperfection liée aux opérateurs booléens : si l'on cherche des documents en linguistique sur les syntagmes nominaux, avec « syntagme » ET « nominal » adjacents, on perd « syntagmes nominaux », « SN », « nominal », « nominaux », « nominalisation », et surtout, « nom(s) », très couramment utilisé par les linguistes. En revanche, avec la commande « nom* », on récupère « nommer » et toute sa conjugaison, « nomade », « nombre » (+ dérivés), « nombril », « nomenclature », « nominalisme », « nomination »... et éventuellement leurs dérivés.

> *Page 6*

sans aucun traitement sémantique, on prendrait le risque de passer à côté d'articles pertinents.

Pour remédier à cela, de nouveaux services en réseaux (*online filtering)* sont apparus sur le principe de la DSI (diffusion sélective d'information). Un tel service scrute les différents groupes de discussion en indexant leur contenu et recense tout nouveau groupe. L'usager s'inscrit au service en fournissant son profil, et est identifié par son adresse électronique et un mot de passe attribué à la première connexion, ainsi que d'autres paramètres qui peuvent varier d'un service à un autre (durée d'inscription, partie de l'article à envoyer, le nombre maximum d'articles à envoyer, etc.). L'usager reçoit ainsi périodiquement, par voie de courrier électronique, les en-têtes, parfois les premières lignes des messages qui correspondent à son profil. Si un article l'intéresse, il pourra en récupérer l'intégralité en retournant (*Forward*) un message au serveur. Un tel système évite ainsi à l'usager l'effort de recherche et d'inscription aux différents groupes de discussion.

Le succès d'Internet étant lié pour une bonne part à son ouverture et à son accès libre, tant en consultation qu'en diffusion, il reste souvent difficile de mesurer la pertinence des informations par des moyens classiques tels que les moteurs de recherche même en s'appuyant sur des services de filtrage. D'ailleurs cette pertinence elle-même reste subordonnée suivant le type de veille, à certains traits tels que le caractère anticipatif de l'information, son utilité, sa fiabilité, son originalité, sa structure, sa portée...

OUTILS INTÉGRÉS DE DÉPOUILLEMENT

Depuis peu de temps, sont apparus de nombreux logiciels de type *méta-moteurs* pour combler les défauts précités des moteurs de recherche. Installés sur un poste de travail, ces méta-moteurs interagissent avec les moteurs de recherche pour effectuer des recherches plus élaborées. Ce sont des outils très puissants qui, à partir d'une seule interrogation, vont consulter une fédération de moteurs de recherche et présentent les résultats sous forme synthétique : réponses triées, doublons supprimés (« dédoublonnage »), sites mémorisés (stockage), etc. La plupart des méta-moteurs peuvent être programmés pour une veille continue ou travailler à intervalle régulier (tel jour à telle heure). D'autres pratiquent une indexation des pages Web très variée : ils n'indexent que les occurrences des mots non vides, prennent en compte leur place dans les pages, la taille des pages, le nombre de liens hypertextes (internes et externes au site). Certains peuvent même élaborer des thésaurus, etc. Mais l'inconvénient majeur des méta-moteurs est technique : ils sont très grands consommateurs en bande passante et supposent donc l'utilisation de liaisons à très haut débit, bien que des liaisons spécialisées puissent aussi se révéler notoirement insuffisantes. Dans cette série de méta-moteurs, on trouve ProFusion, MetaCrawler...

Certes, ces outils sont très puissants par rapport aux annuaires thématiques et aux moteurs de recherche, mais une recherche très pertinente dans la veille fait appel à des outils plus élaborés appelés *agents intelligents* de la veille (ou *vigiciels*) : ce sont des composants logiciels ou matériels servant à des degrés différents comme des fouineurs, récupérateurs, synthétiseurs, traducteurs, résumeurs, filtreurs, indexeurs, prospecteurs, etc. Une tâche d'un simple agent consiste par exemple à enrichir la requête par des procédures linguistiques, rechercher avec plusieurs moteurs les réponses à cette requête, rapatrier les pages en local, analyser, filtrer, supprimer les doublons, puis donner seulement les bonnes pages en les gardant dans une mémoire consultable ultérieurement.

Ces vigiciels ont trois propriétés fondamentales : l'autonomie, la communication et l'apprentissage. Muni d'un fonctionnement automatique et autonome, chaque agent peut échanger des informations avec d'autres agents, ou programmes, et même entrer en interaction avec l'usager ; en outre, il peut avoir la faculté de réagir avec un environnement, de s'adapter aux circonstances, de prendre une décision ou d'enrichir lui-même son propre comportement, sur la base d'observations qu'il effectue.

En employant des agents intelligents dans une démarche de veille, un veilleur peut bénéficier de plusieurs avantages, listés ci-après.

- Gestion des problèmes liés à la grande quantité d'information reçue par le veilleur.
- Accroissement de l'exhaustivité et de la pertinence.
- Élimination des doublons, création des résumés à la volée.
- Économie de temps de connexion et de traitement pour le balayage des serveurs par un traitement des pages en local (*off line searching*).
- Gestion des informations résultats par constitution de bases de données thématiques consultables *off line*.
- Mémorisation du profil de l'expert et de ce qui est hors de son profil.
- Mise en évidence des pages nouvelles par rapport à la dernière recherche.
- Surveillance des mises à jours des sites (suppression, ajout, modification, validité des liens, etc.) avec une périodicité au choix.
- Paramètrage des pages Web copiées localement : copier un site Web sur un poste local/respecter l'arborescence du site d'origine/possibilité d'inclure ou d'exclure certains formats/possibilité de spécifier, dans l'arborescence, le niveau des pages que l'on veut inclure...
- Détermination de manière automatique des signaux faibles (genre d'information éparse et fragmentaire).
- Diffusion des résultats sous différentes formes (mél, fenêtres volantes, économiseur d'écran, etc.)

Sur Internet, on peut trouver plusieurs agents intelligents. Entre autres, Copernic 98, un outil gratuit qui scrute plus de 30 sources regroupées en trois domaines (Web, groupes de discussion, e-mails) ; WebFerret, qui est aussi gratuit, mais ne peut pas éliminer les doublons ; DigOut4U qui est spécialisé dans les annuaires thématiques, il reformule la requête posée en langage naturel, fait de l'analyse sémantique des pages récupérées, classifie par hiérarchisation selon des degrés de pertinence et détecte les signaux faibles.

Parmi ces agents intelligents de recherche d'information, il en existe certains destinés à exécuter des missions et des tâches très spécifiques, ce sont des *agents sectoriels* spécialisés dans des domaines précis tels que les finances, les sciences et techniques, ou la littérature... Par exemple, les *agents pour le commerce électronique* permettent de faciliter aux consommateurs la comparaison des prix-services, et la sélection de boutiques, de marques ou de produits, et permettent également aux vendeurs de mieux connaître la demande, les consommateurs et de gérer des profils clients.

Cela dit, les agents intelligents ne sont pas, au moins pour le moment, vraiment intelligents, et le recours aux experts humains est souvent nécessaire. Cependant ils deviennent de plus en plus efficaces, quand on leur fait prendre en considération quelques enjeux tels que l'intégration des outils sophistiqués d'analyse linguistique pour l'enrichissement de la requête et la compréhension du contenu des textes d'une part, et l'exploitation des outils statistiques d'analyse multidimensionnelle pour l'élaboration des

cartographies dynamiques et interactives aidant à la visualisation des résultats d'autre part.

DÉMARCHE LOGIQUE DANS UNE VEILLE

Si se contenter d'un seul outil pour une recherche d'information semble une tromperie, utiliser trop d'outils pourrait en être une autre. En fait, un seul outil (si puissant soit-il) est souvent insuffisant et peut induire un silence considérable. Le plus judicieux est de faire coopérer un ensemble restreint d'outils pour un traitement progressif et linéaire de l'information. Par exemple, une première étape consiste à utiliser un annuaire thématique comme Yahoo ou Looksmart, car ils donnent des éléments de réponse rapides et précis ; puis, en fonction des résultats obtenus, on s'oriente simultanément vers deux ou trois moteurs de recherche pour élargir la recherche aux sources non référencées dans les annuaires. Dans une troisième étape, on utilise des méta-moteurs pour le traitement des informations collectées et des agents intelligents pour une question de suivi d'évolution de l'information.

Dans une démarche de veille plus globale, ce traitement linéaire se transforme en traitement itératif, notamment après l'étape de la diffusion (« *push* ») ou la mise à disposition. Dans cette étape, la plupart des organismes concernés par la veille sont en train d'évoluer progressivement pour passer des moyens traditionnels de diffusion tels que les notes de recherche, les fiches navettes, les dossiers synthétiques (thématiques), vers des moyens réseau telles que la messagerie électronique, l'Intranet (à usage interne), l'Extranet (semi-ouvert), ou l'Internet (ouvert au public).

CONCLUSION

Si le concept de la veille, et particulièrement celui d'intelligence économique, n'est pas si récent, son développement concret en Europe date de la fin des années 80 (des ingénieurs, commerciaux et directeurs ont toujours pratiqué une forme de veille informelle et non organisée, l'activité de veille ayant souvent été rattachée aux centres de documentation, certainement parce que les deux activités sont liées). Actuellement en France, il y a une forte prise de conscience de l'importance de la veille (édition des rapports officiels, émergence d'associations, de sites Web, de *mailing-lists* et de magazines spécialisés, rédaction d'une norme Afnor XP X50-053 « prestations de veille » en avril 1998), mais sa mise en œuvre n'est pas encore assez forte. On remarque qu'il y a encore absence d'une vraie veille dans les PME-PMI, due à l'état d'esprit pour certains et plus probablement aux contraintes budgétaires pour d'autres. Par contre dans de grandes entreprises, les responsables commencent à abandonner l'idée d'une surveillance statique qui consiste à accumuler des données pour construire un réseau dynamique, composé de personnes motivées. Et des cellules de veilles ont émergé d'une coopération entre les différents services (communication ou services commerciaux, documentation, recherche et développement...), avec des stages de formation presque obligatoires. En parallèle, plusieurs instituts universitaires ont intégré dans leur programme pédagogique des cours sur la veille, et d'autres ont même instauré des filières spécialisées dans le domaine.

Pour conclure, nous empruntons une citation à P. Aron et C. Petit, parue le 29 août 97 dans *Le Monde Informatique* sous le titre « L'info, nerf de la guerre » : « L'humanité a produit au cours des trente dernières années plus d'informations qu'en deux mille ans d'histoire, et ce volume d'informations double tous les quatre ans. La qualité du filtre est donc essentielle ». On peut extrapoler cela, et avancer que la qualité du post-filtre lié à

une veille sur l'information devient obligatoire avant de se trouver noyé (si ce n'est déjà fait) dans les déluges électroniques d'information.

LIENS HYPERTEXTE

Pour référencer les produits cités dans l'article, dans l'ordre d'apparition (ces liens ont été toujours valides jusqu'au 12 novembre 99).

Yahoo http://www.yahoo.com/

Nomade http://www.nomade.fr/

Eureka http://www.eureka-fr.com/

Wanadoo http://www.wanadoo.fr/qqo/

Submit http://www.submit-it.com/

Altavista http://www.altavista.digital.com/

Excite http://www.excite.com/

Lycos http://www.lycosuk.co.uk/index.html

Voila http://www.voila.fr

Ecila http://france.ecila.com/index-french.html

Lokace http://www.lokace.com

Revue *Science* http://www.sciencemag.org

Infoseek http://www.infoseek.com/

Profusion http://www.profusion.com/

Metacrawler http://www.metacrawler.com/

Copernic http://www.copernic.com/

Webferret http://www.ferretsoft.com/netferret/products.htm

Digout4u http://www.arisem.com