Xavier Lepage
Fondateur et managing partner, Swield et chargé de cours à l'École de guerre économique.

Frédéric Lucard
Directeur d'Aestigia.

Thibault du Manoir de Juaye
Avocat à la cour et auteur du *Droit de l'intelligence économique*.

Frédéric Martinet
Consultant expert en outil de veille, Help Management, et animateur du blog *www.actulligence.com*.

Lydie Nguyen Van Tu
Consultante en projets de veille, Help Management.

Guillaume Nuttin
Journaliste, *Archimag*.

Gilles Richard
Documentaliste, *Archimag*.

Jean-Paul Taravella
Consultant TI, Atos Origin.

Bruno Texier
Journaliste, *Archimag*.

annonceurs

Ami Software : 45
Antidot : 31
Centredoc : 55
Esmerk : 43
GB Concept : 81
IDP : 3e de couverture
INPI : 2e de couverture
Iscope : 57
Ixxo : 11, 59
KB Intelligence : 35
Linguay : 28
Neotia : 63
OIS : 73
Polyspot : 75
Serda : 69
Sinequa : 4e de couverture

de recherche•••

www.archimag.com

47
veille
jauger l'efficacité de sa cellule

solutions

49
outils et services
offre et demande, recherche d'équilibre

51
solutions de veille
cap sur le collaboratif et le décisionnel

49-62
vue d'ensemble
• des acteurs de toutes origines
• grille de sélection
• fiches produit

64-74
moteurs de recherche
• ne pas perdre de vue ses besoins
• tableau de sélection

76
moteurs de recherche
c'est demain la veille

78-84
audit et conseil
• un bon prestataire vous comprend
• fiches service

85-88
se former pour devenir un...
• professionnel aguerri
• fiches service

retours d'expérience

89
Préparation olympique et paralympique
l'information gagne des médailles d'or

90
Chambre de commerce internationale
des dossiers bien plus légers et rapides

91
Lhermet La Bigne & Remy
le sens des brevets traqué

92
technocentre Renault
la veille conduit l'automobile

93
pôle image Iconoval
mon meilleur concurrent

94
cabinet Blueboat
aider à surveiller pour décider

95
SDV Logistique Internationale
très conseillé pour le business

sources

96
bibliographie associations webographie

veille et web 2.0

aref jdey et frédéric martinet*

nouveau paradigme

Au-delà des discours marketing promouvant le web 2.0, il est intéressant de s'attarder sur les fondamentaux du concept et ses impacts professionnels sur les métiers de l'information et plus particulièrement sur la veille. En quoi cela consiste et quel est le public visé ? Quelles sont les technologies utilisées et les avantages qui en découlent ?

Plus qu'une révolution technologique, le web 2.0, concept inventé par Dale Dougherty (1) en 2004, est une nouvelle manière d'utiliser le web et de le pratiquer. On peut le qualifier de « *web vivant* » (2). Selon le Centre d'expertise de consultation canadien, le web 2.0 est un terme évolutif qui décrit un ensemble d'améliorations au réseau *world wide web* permettant d'accroître la collaboration et le partage de l'information (3). En résulte une métamorphose des usages, facilitée par l'apparition de nouvelles technologies et applications informatiques, tels les blogs, les wikis, l'Ajax, les RSS et les réseaux sociaux.

les blogs, premiers vecteurs

En matière de veille, les outils web 2.0 trouvent un écho très favorable auprès des professionnels. Des spécialistes et des experts publient des contenus intéressants par le biais des blogs dans tous les secteurs des informations de première main, des tests de nouveaux produits, des analyses prospectives et critiques du marché, offrant un angle de vue alternatif. Les blogs permettent aussi de sentir le pouls d'une ou plusieurs communautés rassemblées autour d'un produit, d'un service ou d'une cause. Ils sont les premiers vecteurs de communication de ces groupes et donc les premiers canaux de transmission de signaux faibles. Dans des situations extrêmes, certains blogs, par leur simplicité de création et l'utilisation de RSS, peuvent détruire la réputation d'une société (4) ou d'une marque (5).

RSS démocrate

Le RSS a démocratisé la veille web 2.0. Le format de syndication de contenu possède trois principaux avantages :

- **L'alerte** dès qu'un contenu utile ou une source intéressante sont mis à jour.
- **La possibilité d'un contenu spécifique** et non condensé.
- **La gratuité** et l'absence de publicité, même si cela reste temporaire.

Avec les flux RSS, toute information peut être mise en surveillance ou presque (6), dans le domaine du web 2.0 : blogs, journaux, wikis, recherches sur internet, etc. De jeunes acteurs de la veille web 2.0 apparaissent et se positionnent notamment sur le marché de la recherche et de l'agrégation du contenu, tels Wikio (7), Synthesio (8), Technorati (9). Ils lancent des outils gratuits ou payants au profit des professionnels de la veille, pour observer l'univers virtuel.

favoris

La veille fait appel à d'autres applications du web 2.0 pour mieux cerner son périmètre, comme le partage des favoris, appelé social bookmarking (10). Il s'agit de la mise en commun des sources et des pages internet annotées par les internautes en tant que favoris. Elle offre deux avantages :

- **Le veilleur peut rapidement identifier de nouvelles sources** adaptées pour mener sa veille, grâce à l'expertise d'un ou plusieurs individus. Il gagne du temps et de la pertinence.
- **Le partage des favoris conduit le veilleur à l'enrichissement** de son marqueur de pages de sources à surveiller, mais aussi à l'identification des spécialistes et des experts d'un domaine, tout en enrichissant son réseau social.

nouvelles contraintes

Le web 2.0 ouvre ainsi de nouveaux horizons aux métiers de la veille et aux professionnels de l'information. La mutation du web implique nécessairement de nouvelles contraintes jusque-là inexistantes ou sous d'autres formes :

- **Problématique de la fiabilité et de la validité de l'information.** Quel degré de confiance accorder aux blogs et aux informations disponibles sur les réseaux sociaux ? Quelle méthodologie de vérification utiliser ?
- **Rapidité et complexité des mécanismes de diffusion de l'information.** On perd plus facilement la traçabilité d'une information.
- **Augmentation exponentielle du volume de l'information disponible**, on parle d'infobésité : Difficulté d'identifier les bonnes sources d'information, temps dédié à la consultation plus important que celui consacré à l'analyse.
- **De nouvelles aptitudes et compétences** individuelles et collectives doivent être acquises. Travail multitâche, veille professionnelle, etc.
- **Convergence toujours plus visible du réel et du virtuel.** Procédures de recrutement sur Second Life, blogosphère politique durant l'élection présidentielle de 2007, etc.

vers des pratiques inédites

Les changements sont notamment dus aux caractéristiques mêmes du web 2.0. Un décryptage des nouvelles pratiques et des besoins en découlant s'impose ici, en considérant le web 2.0 comme un cycle d'évolution systémique du web et de ses usages :

- **L'offre de nouvelles technologies.** Depuis quelques années, nous assistons à une forte croissance du nombre d'applications informatiques et logicielles des-

Le nouveau cycle de la veille dans un environnement web 2.0

DR Help Management

tinées à faciliter la vie professionnelle : e-learning, visioconférence, partage des données, travail collaboratif, etc. L'offre est ciblée en fonction du type de clients, à l'image des moteurs de recherche spécialisés en contenu scientifique, multimédia et immobilier. La maturité de l'offre fait émerger de nouveaux usages.

• **Les nouveaux usages.** L'usage est le fait d'appliquer et d'utiliser un objet ou une application pour obtenir un résultat, un effet. Dans notre contexte, il s'agit de gagner du temps, d'accéder depuis n'importe où à des informations professionnelles ou de rester en contact avec son réseau.

• **Les nouveaux besoins.** Les différents usages nés de cette évolution augmenteront le niveau d'exigence des utilisateurs en terme de réactivité, de transparence et d'originalité. Comme tout internaute est potentiellement capable de produire du contenu, la communauté cherchera ceux proposant un contenu original, se différenciant du reste des publications par leur esprit analytique, ton et argumentation. La mondialisation et la large couverture du haut débit entraînent quant à eux un besoin irrépressible d'informations récentes. On privilégiera alors la réactivité des producteurs d'information.

• **Les nouvelles pratiques.** La panoplie de besoins est également créatrice d'idées et donc de segments de marché pour pouvoir y répondre. Des nouveaux métiers et pratiques professionnelles ont vu le jour avec le web 2.0 : e-documentation, e-publishing, e-edition, e-marketing, etc. Le contenu numérique et sa maîtrise restant au centre de ces compétences.

• **Les nouvelles cultures.** Ces changements vont nous amener à identifier de nouveaux éléments et composantes d'une culture socioprofessionnelle inédite, caractérisée par la mobilité, la transversalité et la pluridisciplinarité, que l'on appelle la « *net génération. À en juger par l'évolution récente du marché du travail, de nombreux métiers demanderont de plus en plus de savoir être polyactif* », selon Sylvie Octobre, sociologue (11). •

**Consultants chez Help Management.*

(1) www.oreillynet.com/pub/au/26
(2) cf. Mary Hodder, directrice de Dabble
www.dabble.com
in Courrier international, HS octobre 2007
(3) www.tpsgc.gc.ca/onlineconsultation/text/glossary-f.html
(4) Cas de Dell, contructeur informatique,
www.buzzmachine.com/?tag=dell
(5) Cas de Vichy, produit cosmétique,
amiel.typepad.com/marketing_relationnel/2005/05/vichy_blog.html
(6) 101 idées pour utiliser le RSS
vtech.canalblog.com
(7) www.wikio.fr
(8) www.synthesio.fr
(9) www.technorati.com
(10) del.icio.us
(11) Les enfants de la net génération, Le Monde, 6 octobre 2007
www.lemonde.fr/web/article/0,1-0,36-963839,0.html

repères

glossaire web 2.0

Ajax : Acronyme de Asynchronous Javascript and XML. Assemblage de languages informatiques permettant de produire du contenu dynamique avec une économie de la bande passante.

Blogs : Journal personnel animé par un ou plusieurs internautes et traitant d'un sujet ou d'un thème bien défini. Il se caractérise par sa fréquence de mise à jour, sa structure en deux tiers colonnes et l'ordre antichronologique de publication.

RSS : Acronyme de Really simple syndication. Format de syndication de contenu alertant l'utilisateur à chaque mise à jour d'un contenu internet.

Social bookmarking : Activité de partage de bookmarks ou favoris en ligne entre plusieurs individus. Tout individu peut mettre à disposition de la communauté ses sites favoris, grâce au tagging.

Tagging folksonomie : Mode de classement collaboratif et asynchrone de contenu totalement régulé par les internautes, ne dépendant pas d'un lexique normalisé comme les thésaurus.

Wiki : Site web collaboratif où chacun apporte les modifications et mises à jour souhaitées et publie de nouveaux contenus.

moteur de recherche

frédéric martinet*

évaluer la performance

On n'attend pas d'un moteur de recherche qu'il révèle une vérité. On souhaite soit des réponses à des questions, appréciées selon la qualité des sources et par recoupements, soit des faisceaux de signaux. Donc pas de record absolu, mais une performance. Tout est question d'évaluation : chaque moteur a ses points forts et ses points faibles.

Quels sont les critères vous permettant de savoir si le moteur de recherche que vous utilisez peut répondre à votre besoin ? Les moteurs grands-public de recherche sur internet indexent différents types de ressources, comme le texte, les pages web, les documents bureautiques, d'autres s'intéressent aux ressources multimédias. Au-delà de cette typologie, on différencie les moteurs de recherche web des moteurs de recherche d'entreprise indexant des fichiers stockés sur le système d'information de l'entreprise. Ils travaillent sur des corpus restreints.

• **Le moteur de recherche internet** visite les différents sites web accessibles, en vue de construire un index. Les internautes peuvent l'interroger pour trouver les pages web susceptibles de les intéresser.

• **Le moteur de recherche d'entreprise** est installé sur les serveurs de l'entreprise et indexe les différents serveurs de fichiers et de serveurs de messagerie, afin que les utilisateurs puissent y retrouver plus aisément des informations.

indicateurs de performance

• **La pertinence** demeure le critère de performance. Son calcul se fonde sur la

silence, bruit et pertinence

DR Help Management

qualité du contenu, suite à une demande formulée par un utilisateur. Lorsqu'une recherche renvoie plusieurs centaines de milliers de résultats, la pertinence devient difficilement estimable. L'utilisateur doit se limiter à l'évaluer sur une centaine de résultats devenant ses éléments de comparaison.

L'étude menée par Jean Véronis en décembre 2005 (1) montre que la pertinence des principaux moteurs de recherche perçue par l'utilisateur demeure moyenne. Du point de vue utilisateur, la pertinence est un critère hautement subjectif. L'étude de l'université de l'État de Pennsylvanie démontre l'impact de la marque du moteur sur la perception de la pertinence.

Du point de vue du moteur de recherche, la pertinence se veut un indice comptabilisant les nombres de liens entrants sur une ressource, le nombre d'occurrences d'un mot, sa place dans le contenu, etc.

Les algorithmes de calcul de pertinence des moteurs de recherche tentent de mettre en relation le contenu produit et diffusé sur internet, en valorisant ses points essentiels, et le contenu cherché, en se fondant sur le texte du document et sur sa structuration : titre, sémantique, occurrence de mots et taux de densité d'un terme.

La mise en adéquation est plus ou moins réussie selon les outils. On constate des variations dans la pertinence de moteurs, car ils évoluent sans cesse, en vue de rapprocher les deux visions de la pertinence et de réconcilier chercheur d'informations et outils de recherche.

- **Le taux de précision** atteint, au sein de son index, les résultats relatifs à la demande. On parle de taux de précision et de taux de rappel. Le taux de précision retranscrit la capacité d'un outil de recherche d'information à ramener uniquement les résultats pertinents pour l'utilisateur. Il est égal au ratio des résultats pertinents retournés sur le nombre de résultats retournés.
- **Le taux de rappel** évalue la capacité d'un moteur de recherche à ramener l'intégralité des résultats correspondants à une demande. Il est égal au ratio des résultats pertinents retournés sur le nombre de résultats pertinents au sein de l'index.
- **Les notions de bruit et de silence** sont associées à ces deux indicateurs. Moins un moteur est précis plus les résultats comporteront du bruit. Plus le taux de rappel d'un moteur de recherche est faible, plus le silence est présent. Le taux de précision et le taux de rappel évoluent à l'opposé l'un de l'autre, rendant l'objectif d'un moteur idéal difficilement accessible en l'état des technologies actuelles.

Les indicateurs restent toutefois limités à une application sur des moteurs de recherche en entreprise et dans le cas de corpus documentaires maîtrisés. Il est impossible de les appliquer à des corpus inconnus, tels que ceux de Google et de Yahoo.

indicateurs techniques

En dehors des critères d'évaluation liés au contenu même des documents retrouvés, il existe des critères techniques évaluant les moteurs de recherche. L'un de ces critères est le temps de réponse. En combien de temps le moteur de recherche parvient-il à balayer son index et construire une liste de résultats ? Le type des documents indexés reste un critère fort de différenciation entre les moteurs de recherche. Si la plupart sont à même aujourd'hui de parcourir les pages web, les documents bureautiques et les fichiers PDF, les fichiers multimédias sont plus rarement pris en compte.

équipe d'évaluation

L'évaluation d'un moteur reste complexe et exige la participation de plusieurs profils :

- **Un informaticien** sur les aspects techniques : temps de réponse, montée en charge, volume des flux d'information.
- **Un profil documentation** sur la facilité d'interrogation : présence de recherche avancée ou non, construction de l'index documentaire, bon fonctionnement des requêtes booléennes, bruit et silence.
- **Les experts,** sur la pertinence de l'outil au regard d'une thématique.

points forts et faibles

Enfin, il faut considérer que le test d'un moteur de recherche doit se faire sur des problématiques de recherche différentes car chacune des technologies de recherche a ses points forts et faibles, en fonction des usages. Une réflexion approfondie permet d'opter pour un moteur correspondant aux besoins des utilisateurs dans le cadre d'une acquisition pour l'entreprise. ●

* *Consultant de projets de veille pour Help Management.*

(1) Comparative de six moteurs de recherche. J. Véronis. Février 2006.
(2) The effect of brand wareness on the evaluation of search engine results. *Jansen et Zhang. Mai 2007.*
(3) Conception de système d'information multimédia. *Sabbi Sidom. Université Nancy 2. Site Loria.*

repères

méthode d'évaluation

Déterminer la pertinence d'un moteur de recherche exige :
- De préparer un ensemble de problématiques de recherche. Il peut s'agir de recherches factuelles ou portant sur un document en particulier ou un fonds.
- D'interroger les différents moteurs de recherche.
- D'extraire les résultats de leur environnement graphique, soit le même nombre pour tous les moteurs interrogés.
- De soumettre les résultats et la problématique de recherche correspondant à un panel d'utilisateurs jugeant de leur pertinence.
- Un taux de pertinence des résultats extraits par moteur établit un classement.

Pour les aspects techniques, il est nécessaire de :
- Calculer et comparer les indicateurs de temps de réponse et l'utilisation des ressources informatiques.
- Établir une liste détaillée des formats de fichiers indexés et la taille occupée par l'index du moteur pour un corpus documentaire donné.

moteurs de recherche

aref jdey et frédéric martinet*

c'est demain la veille

Les moteurs de recherche d'entreprise associeront demain les divers modes d'interrogation et parcourront l'index multimédia en retournant les réponses appropriées sur une seule liste. Les pages souples se pliant à l'utilisateur et dissociant l'affichage du contenu collecté seront quant à elles la prochaine étape ergonomique.

Aux débuts du web, la lenteur du débit empêchait l'accès aux éléments multimédias et seuls les documents textuels s'échangeaient. L'infrastructure du réseau elle-même ne pouvait supporter le transfert de données massives, contrainte importante pour les outils de veille. Depuis les choses ont bien changé.

images et mots-clés

Les moteurs de recherche d'images restent de simples adaptations de la technologie standard des moteurs, se concentrant sur le texte et les mots-clés situés dans le code HTML de la page contenant l'image. Mais de nouvelles technologies de reconnaissance de forme, de texture et de couleurs sont apparues. Elles déterminent les contours à l'intérieur de l'image et ses tendances colorimétriques. Appelées CBIR (recherche d'images par le contenu), elles rendent possible la recherche de similarité entre un objet cherché et le contenu d'une image. Deux options sont envisagées :

- **La construction d'une base d'images** qualifiées textuellement et préchargées, amenant à l'interrogation au format texte du moteur de recherche.
- **Le chargement d'une image,** afin de parcourir la base de données et d'identifier les similarités.

Les moteurs de recherche Like.com et Riya.com appliquent cette technologie. Le second recherche des objets similaires en privilégiant les formes, les couleurs et les motifs. L'interface proposée par System One, fondée sur la technologie Fast multiresolution image querying, en est un bel exemple. Si l'interface d'interrogation de Flickr.com se limite à l'interrogation textuelle sur les tags associés aux images, sur le nom de l'image et sa description, Retrievr de System One y ajoute la recherche en dessinant ou en envoyant une image.

recherche audio

Les technologies performantes de recherche de fichiers son :

- **Interrogent une base de données** de fichiers multimédias textuellement.
- **Listent les documents** comportant le terme recherché.
- **Identifient le passage** à l'intérieur du document comprenant le mot désiré.

La technologie se fonde sur la reconnaissance vocale (1). D'autres techniques travaillent par reconnaissance de sons. Il suffit de sifloter une mélodie pour engendrer l'identification du passage similaire à l'intérieur d'une base de fichiers musicaux. Les moteurs Midomi.com et Everyzing.com utilisent cette méthode à destination du grand-public. Les évolutions préfigurent les moteurs en entreprise de demain. Ils combineront les modes d'interrogation – texte, son, image – et parcourront leur index multimédia en restituant les résultats pertinents à travers une seule liste, à l'image des travaux de Google Universal Search.

traitement multimédia

La majorité des outils de veille sur le marché demeurent déficients quant à l'identification des variations liées au contenu multimédia. Si l'on désire inclure la surveillance des images, la détection de changement se fait par rapport au nom du fichier ou à son poids, le contenu de l'image n'étant pas pris en compte. Un logiciel ne peut nous avertir de

la présence d'une image comportant un avion par exemple, sur une page donnée. Dans le meilleur des cas, l'image se nomme *avion.jpg* ou comprend dans l'attribut *alt* ou *title* le mot *avion*, assurant son identification. Les prestations de veille multimédia reposent sur une surveillance humaine et un corpus limité aux principaux médias. Les technologies déployées par les moteurs pourraient être intégrées à des logiciels de veille, mais le rapport entre la valeur ajoutée du service et le coût d'intégration demeure un frein non négligeable.

nouveautés fonctionnelles

La majorité des outils de veille actuels s'adaptent aux utilisateurs n'ayant ni le temps, ni l'envie de réaliser des paramétrages complexes, alors qu'ils attendent un service immédiat. Les logiciels de surveillance nécessitent de moins en moins l'intervention de l'usager. Ce dernier se contente de coller dans un formulaire l'adresse de la page à mettre en surveillance, charge l'outil de détecter son format, d'identifier les portions de texte intéressantes et d'y appliquer des filtres de façon transparente.

évolutions ergonomiques

Les interfaces des solutions de surveillance deviennent plus conviviales, grâce à l'impact du web 2.0. Les pages riches et flexibles se pliant à l'utilisateur et dissociant l'affichage du contenu collecté par l'application seront la prochaine étape ergonomique.

capacités supplémentaires

Le web 2.0 a également touché les solutions proposant la veille et la génération de flux RSS pour les informations collectées. Le logiciel se positionne tel un gestionnaire de flux, autorisant la diffusion sélective des flux d'information qualifiée, pour les injecter dans les applications de gestion de contenus.

Les outils actuels se dotent de capacités supplémentaires, proposant l'aide à la lecture et à la compréhension des contenus en intégrant une couche de text mining. Apparaissent des fonctionnalités de navigation et de représentation cartographique des corpus de documents, mais aussi des fonctionnalités d'analyse statistique. Les frontières entre les logiciels de text mining et de surveillance disparaissent peu à peu.

vers une fusion des technologies

L'intégration de ces nouvelles fonctionnalités utiles au management crédibilise les applications auprès des décideurs. Les logiciels de surveillance se rapprochent des outils décisionnels acceptant le reporting sur des informations textuelles. La fusion des technologies reste prévisible. Les données issues de la surveillance web seront traitées comme le niveau des stocks, les parts de marché et autres données purement chiffrées. ●

* *Consultants pour Help Management*

(1) Technologie Coveo audio and video search et moteur de recherche Pluggd sur les podcasts.
www.pluggd.com/demo

l'intégration de nouvelles fonctionnalités utiles au management crédibilise les applications auprès des décideurs