Web invisible :
Une nouvelle technologie pour découvrir et exploiter le web profond pour la veille stratégique

Christophe ASSELIN (*)
Christophe.asselin@digimind.com

(*) DIGIMIND, Progiciels de veille stratégique
50, rue de Paradis - 75010 Paris - France
Tel: +33(0)1 5334 0808
FRANCE

Mots clés :
web invisible, web profond, web caché, veille technologique, veille stratégique, métamoteur, collecte d'informations

Keywords:
invisible web, deep web, hidden web, scientific watch, strategic watch, metasearch engine, information gathering

Palabras clave :
red invisible, red ocultado, red profundo, viligancia technologica, vispera estratégica, metamotor de pesquisa, reunir de información

Résumé

Pour rechercher et effectuer une veille sur le web, l'utilisation des seuls moteurs et annuaires généralistes vous privera de l'identification de centaines de milliers de sources.
Parce que des moteurs comme Google, MSN ou Yahoo! Search ne vous donnent accès qu'à une petite partie (inférieure à 10%) du web, le Web Visible. La technologie de ces moteurs conventionnels ne permet pas d'accéder à une zone immense du web. Lors d'une navigation en Antarctique pour prélever des échantillons de glace sur des icebergs, si vous vous limitez à leur partie émergée, vous vous privez de la surface immergée, en moyenne 50 fois plus importante.
Sur le web, c'est la même chose….Mais sur ce réseau, la zone invisible est environ 500 fois plus volumineuse comportant des centaines de milliers de ressources de grande valeur.
Les ressources du Web Invisible sont en effet en moyenne de plus grande qualité, plus pertinentes que celles du web de surface. Pourquoi ? Parce qu'elles sont élaborées ou validées par des experts, faisant autorité dans leurs domaines.

Cette étude vous rappelle le concept de Web Invisible ainsi que ses caractéristiques.
Par ailleurs, il vous explique comment trouver des sites appartenant à ce web "profond".
Enfin, il vous montre pourquoi l'approche unique choisie par Digimind, à travers ses nouvelles technologies, vous permet de mettre en œuvre un véritable process de veille industrielle sur les ressources du Web Invisible.

1 LE WEB INVISIBLE : CONCEPTS ET DEFINITION

1.1 Le web visible

Un moteur de recherche "aspire" puis indexe des pages web. Pour ce faire, ses robots (ou spider) vont parcourir les pages web en naviguant de liens en liens, et passer ainsi d'une page à une autre. Ce robot va archiver sur les serveurs du moteur une version de chaque page web rencontrée.
Ainsi, Google dispose de plus de 11 000 serveurs répartis sur plusieurs états (Californie, Irlande, Suisse…). Ces serveurs connectés entre eux hébergent notamment les milliards de pages web aspirées.
Cette version archivée des pages au moment du passage du robot est d'ailleurs disponible via la fonction "Cache" disponible sur certains moteurs comme Google, Yahoo!, Clusty ou Gigablast

1.2 Le web invisible : Définition et structure

Le Web Invisible est constitué des documents web mal ou non indexés par les moteurs de recherche généralistes conventionnels. En effet, le fonctionnement des moteurs pour "aspirer" le web implique que, d'une part, les pages soient bien liées entre elles via les liens hypertexte (http://) qu'elles contiennent et que, d'autre part, elles soient identifiables par les robots du moteur. Or dans certains cas, ce parcours de liens en liens et cette identification de pages est difficile, voire impossible. Une partie du web est en effet peu ou non accessible aux moteurs de recherche pour plusieurs raisons :

● **Les documents ou bases de données sont trop volumineux pour être entièrement indexés.**
Les moteurs conventionnels n'indexent pas la totalité des contenus de plusieurs milliers de bases de données professionnelles (l'indexation varie entre 5 et 60 % selon les moteurs). Certaines n'étant pas indexées du tout. D'autre part, les moteurs n'indexent pas la totalité du contenu d'une page lorsque celle-ci est très volumineuse : Google et Yahoo! archivent les pages dans une limite de 500k et 505k.

● **Les documents ou bases de données sont trop volumineux pour être entièrement indexés.**
Les pages sont protégées par l'auteur (balise meta qui stoppe le robot des moteurs)
Certains sites sont protégés par leur créateur ou gestionnaire, grâce à un fichier robot.txt inséré dans le code des pages, interdit leur accès aux robots des moteurs. L'utilisation de ce fichier robot.txt est effectuée pour protéger. De cette manière, il évite que les moteurs archivent des pages payantes et les mettent à disposition gratuitement via leur fonction "Cache".

● **Les pages sont générées seulement dynamiquement, lors d'une requête par exemple**
De nombreux sites web génèrent des pages dynamiquement, c'est-à-dire uniquement en réponse à une requête sur leur moteur interne. Il n'existe pas alors d'URL (adresse) statique des pages que les moteurs pourraient parcourir puisque les robots des moteurs n'ont pas la faculté de taper des requêtes

● **Les pages sont protégées avec une authentification par identifiant et mot de passe.**
Les robots de moteurs n'ayant pas la faculté de taper des mots dans des formulaires complexes, ces pages ne leur sont pas accessibles.

● **Les pages sont mal liées entre elles ou sont orphelines (aucun lien présent sur d'autres pages ne pointent vers elles).**

1.3 Un Web Invisible, Caché ou Profond ?

Plutôt que de parler d'un Web Invisible, certains spécialistes ont tenté de préciser le concept et de l'illustrer différemment. Aussi, plutôt que le web visible et invisible, l'étude de BrightPlanet [1] préfère évoquer un "Surface Web" et un "Deep web" (web profond). En effet, le problème n'est pas tant la visibilité que l'accessibilité par les moteurs. Il existe donc un web de surface que les moteurs parviennent à indexer et un web profond, que leurs technologies ne parviennent pas à encore à explorer, mais qui est visible à partir d'autres technologies perfectionnées telles que Digimind Finder.

2 CARACTERISTIQUES DU WEB INVISIBLE : TAILLE ET QUALITE

2.1 La taille du web invisible : des estimations

Plusieurs chercheurs ont tenté d'estimer la taille du web invisible.

2.1 1 Le Web Visible

En juillet 2000, une étude de la société Cyveillance, "Sizing the Internet" [2] estime le web visible à 2,1 milliards de pages, augmentant ainsi à un rythme déjà soutenu de 7,3 millions de pages par jour. Selon ce taux de croissance, Cyveillance estime la taille du web de surface à 3 millions en octobre 2000 et à 4 milliards en février 2001 soit un doublement de volume par rapport à juillet 2000 (ce qui correspond à une période de 8 mois).
Si l'on estime que le web visible double de volume tous les ans, il pourrait représenter aujourd'hui, en 2005, au moins 64 milliards de pages. C'est une estimation évidemment très minorée puisque basée sur un taux de croissance stable ce qui est très peu probable. Ainsi, la possibilité d'accès à la publication sur le web à des utilisateurs toujours de plus en plus nombreux (via les weblogs par exemple qui dépasse les 10 millions) ne fait qu'augmenter son taux de croissance mois après mois.

• Capacité des moteurs généralistes actuels
Aujourd'hui, les moteurs généralistes proposant les plus larges index annoncent recenser jusqu'à 20 milliards de pages. En août 2005, Yahoo! annonce indexer 19,2 milliards de pages web.
Or, les chercheurs du centre IBM d'Almaden (Californie) estiment que 30% du web est constitué de pages dupliquées et que 50 millions de pages sont modifiées ou ajoutées chaque jour.

Dans ces conditions, un des meilleurs moteurs actuels indexe moins de 30% du web visible. Mais on peut tempérer ce faible taux de représentation du web visible en précisant que l'index de Google n'est pas équivalent à celui de Yahoo! Search ou de Ask Jeeves. Pour estimer la taille de l'indexation du web visible, il faut donc considérer les index de plusieurs moteurs et non d'un seul.

Ainsi, une étude du printemps 2005 [3] réalisée par le métamoteur Dogpile.com en collaboration avec des universités de Pennsylvanie estime que les résultats des premières pages fournis par les 3 moteurs à large index que sont Google, Yahoo! Search et Ask Jeeves sont de plus en plus différents.

Sur les premières pages de résultats de chacun de ces moteurs, et sur 336 232 résultats analysés :
- 3% des résultats seulement sont partagés par les 3 moteurs
- 12 % sont partagées par 2 des 3 moteurs
- 85 % ne sont proposés que par un seul des 3 moteurs !
L'étude du métamoteur Jux2, sur un échantillon toutefois moins élevé, va également dans ce sens.

2.1 2 Le Web invisible

C'est l'étude "The Deep Web: Surfacing Hidden Value" qui propose les estimations de la taille du web les plus avancées. Cette étude de 2001 propose des ordres de grandeur permettant de mieux mettre en perspective le web profond à l'égard du web de surface.- l'information publique sur le web profond est considérée comme de 400 à 550 fois plus volumineuse que le web de surface (web visible) :
- le web profond est constitué de plus de 200 000 sites web.
- 60 % des sites les plus vastes du web profond représentent à eux seuls un volume qui excède de 40 fois le web de surface.
- le web profond croit plus vite que le web visible.
- plus de la moitié du Web Profond est constitué de Bases de données spécialisées.
- 95% du contenu du web profond est accessible à tous (gratuit ou à accès non restreint à une profession)

Si l'étude date de 2001, les proportions restent valables (en estimation basse) compte tenu des taux de croissance très élevés du volume du web profond, c'est à dire une multiplication par 9 chaque année (estimations IDC).

2.2 Contenu et qualité du web invisible

2.2. 1 Contenu et couverture des sites du web invisible

• Typologie.
A travers l'analyse de 17000 sites appartenant au web profond, l'étude de Bright Planet a élaboré une typologie du contenu du web invisible.
Ces sites ont ainsi été répartis en 12 catégories :

1. Les Base de données spécialisées : il s'agit d'agrégateurs d'informations, interrogeables, spécialisés par sujet. Ce sont par exemple les bases de données médicales, de chimie, de brevets. Exemples : Base de donnée de la National Library of Medicine interrogeable via PubMed, Esp@ceNet, la base de données européenne de brevets ou GlobalSpec , base dédiée à l'information technique et d'ingénierie .
2. Les Bases de données internes à des sites web volumineux. Ces pages sont générées dynamiquement. Exemple : la base de connaissance des sites Microsoft

3. Les publications, c'est-à-dire les Bases de données requêtables (via un moteur interne) donnant accès à des articles, des extraits d'ouvrages, des thèses, des livres blancs. Exemple : FindArticles

4. Les sites de ventes en ligne, de ventes aux enchères. Exemple : Ebay, Fnac.com, Amazon.fr.

5. Les sites de petites annonces. Exemples : de particulier à particulier, VerticalNet (places de marché industrielles)

6. Les portails sectoriels. Portes d'entrées sur d'autres sites, ces sites agrègent plusieurs types d'information : articles, publications, liens, forums, annonces interrogeables via un moteur et une base de données. Exemple : PlasticWay, le portail dédié à la plasturgie.

7. Les bibliothèques en ligne. Bases de données issues essentiellement de bibliothèques universitaires ou nationales. Exemples : Bibliothèques du Congrès US, BNF-Gallica, les bases de bibliothèques recensées par l'OCLC …

8. Les pages jaunes et blanches, les répertoires de personnes morales et physiques. Exemples : Yellow Pages, ZoomInfo (répertoire des dirigeants et salariés d'entreprises)

9. Les Calculateurs, Simulateurs, Traducteurs. Ce ne sont pas des bases de données à proprement parler mais ces bases incluent de nombreuses tables de données pour calculer et afficher leurs résultats. Exemples : Traducteur Systran Babelfish, Grand Dictionnaire Terminologique.

10. Bases de données d'emploi et de CV. Exemples : Monster, APEC, Cadresonline…

11. Site de messages ou de Chat

12. Bases de données de recherche généraliste. A la différence des bases de données spécialisées, elles recensent des thèmes éclectiques. Exemple : Weborama.

Par ailleurs, l'étude révèle que 97,4 % de ces sites du web profond sont d'accès public, sans aucune restriction. 1,6% ont un contenu mixte : résultats disponibles gratuitement côtoyant des résultats nécessitant un enregistrement gratuit ou un abonnement payant.

Seulement 1,1 % des sites du web Invisible proposent la totalité de leur contenu par souscription ou abonnement payant. Ainsi, les bases de données à très forte notoriété telles que Lexis Nexis, Dialog Datastar, Factiva, STN International, Questel….ne constituent qu'à peine plus de 1% du Web Profond!

• Couverture sectorielle du contenu des sites du web invisible
Les 17000 sites de l'étude révèlent une distribution relativement uniforme entre les différents domaines sectoriels. En fait, le web profond couvre tous les secteurs d'activités principaux.

2.2.2 Qualité des sites du web invisible

La qualité est une notion évidemment subjective. Aussi, cette étude part d'un précepte simple : lorsque l'on obtient exactement les résultats que l'on désire via un site du web profond, on considère que la qualité dudit site est très bonne.
A l'inverse, si vous n'obtenez pas de résultats satisfaisants, la qualité est considérée comme nulle. En terme de pertinence, la qualité du Web Profond est estimée comme 3 fois supérieure à celle du web de surface.
Pourquoi ? Essentiellement parce que la majeure partie des sites du web invisible sont des sites spécialisés, dédiés à une activité, une technologie, un métier et que leur contenu émane ou est validé par des professionnels, spécialistes et experts.

3 DIGIMIND ET LA VEILLE SUR LE WEB INVISIBLE

3.1 Les specificités d'une veille automatisée sur le web Invisible et les approches technologiques de Digimind

Le web profond présente un certain nombre de caractéristiques qui nécessitent des technologies de surveillance adaptées, si l'on souhaite pouvoir effectuer une veille efficace.
En effet, un veilleur qui doit consulter régulièrement une vingtaine de moteurs sur plusieurs dizaines de thématiques risque d'y passer ses journées.
D'où la nécessité d'utiliser des outils de "méta-recherche" et de surveillance de ces moteurs de recherche pour automatiser autant que possible le travail répétitif d'identification et de rapatriement de nouvelles informations pertinentes apparues sur le web invisible.

Afin d'apporter un maximum de productivité dans le travail du veilleur, ces outils doivent surmonter les obstacles suivants :

3.1.1 L'interfacage avec des moteurs internes de tout type

La surveillance du web profond nécessite d'interroger les moteurs internes de sites: moteur de portail, moteur de bases de données. Ces moteurs utilisent des modes de requêtage très variés, depuis le type de requête (GET, POST, mixte), le format des formulaires d'interrogation (HTML, Javascript), la méthode de soumission de la requête (formulaire ou redirection Javascript), ou encore la syntaxe de la requête. L'interfaçage avec ces différents moteurs est un processus qui peut s'avérer particulièrement complexe afin de mémoriser puis automatiser les requêtes posées.
En conséquence, la plupart des métamoteurs du web invisible ne recherchent que dans leur propre sélection de moteurs pré paramétrés. Un métamoteur annonce ainsi un chiffre de 1000 moteurs regroupés en 120 catégories. Un choix qui reste néanmoins limité en regard des 200.000 sites du web invisible répertoriés en 2001 ! Utile pour une veille généraliste, ces outils montreront vite leurs limites pour le veilleur professionnel ou l'expert de l'entreprise.
De leur côté, les logiciels de surveillance de sites web permettant d'accéder au Web Invisible contournent la difficulté de création d'un "connecteur générique" à chaque moteur en mettant en surveillance la page de résultats de chaque requête – ce qui nécessite le paramétrage manuel de chaque requête sur chaque moteur, travail qui peut s'avérer extrêmement fastidieux pour notre veilleur interrogeant une vingtaine de moteurs sur plusieurs dizaines de thématiques (20 moteurs x 90 thématiques = 1800 requêtes à paramétrer…et administrer).

• L'approche de Digimind : configurer des connecteurs en trois clics
Pour permettre de mettre rapidement en surveillance toute nouvelle requête sur sa propre sélection de moteurs, Digimind a développé des technologies exclusives de configuration de connecteurs en trois clics, utilisées dans le nouveau module Finder de Digimind Evolution, héritier du célèbre métamoteur "Strategic Finder" en version serveur.
La première reconnaît automatiquement les formulaires html et javascript, tandis que la deuxième permet l'apprentissage sémantique de la requête http. Résultat : le veilleur souhaitant ajouter un nouveau moteur à son métamoteur a juste besoin de simuler une recherche sur ce moteur pour que le connecteur soit activé…et le moteur immédiatement disponible pour des recherches temps réel ou les surveillances en cours.

3.1.2 La nécessité d'extraire uniquement – mais intégralement - les résultats des moteurs

Tout comme leur interface de recherche, les pages de résultats des différents moteurs varient grandement d'un moteur à l'autre en terme d'affichage:
-éléments constituant les résultats (titre, date, résumé pour des actualités, nom de brevet, inventeur, numéro de brevet pour des bases de brevets, etc…) ;
-position des résultats dans la page (voire listing des résultats dans une nouvelle page) ;
-informations non pertinentes apparaissant autour des résultats (rubriques, publicités, aide, contacts,…).

Là encore, les métamoteurs les plus courants auront préparamétré leur propre sélection de moteurs en créant un fichier descriptif pour chaque page de résultats. Ce "masque" indique clairement au "robot" du métamoteur à quelles balises HTML commence et se termine le listing de résultats, et entre quelles balises HTML se situent les éléments constitutifs de chaque résultat. Ce fichier descriptif étant fastidieux à créer, on comprend là encore pourquoi ces métamoteurs ne permettent pas à leurs utilisateurs d'ajouter leurs propres moteurs.
Le problème corollaire de la configuration "manuelle" de chaque page de résultats est qu'en cas de modification de la structure HTML de la page de résultat par l'éditeur du site, le fichier descriptif cesse de fonctionner…et doit être mis à jour.

L'autre difficulté à surmonter est de suivre, le cas échéant, les multiples pages de résultats (« page 1, page 2, page 3, pages suivantes… »). Là encore, les métamoteurs nécessitent la configuration manuelle d'un fichier décrivant le mode de navigation entre les pages de résultats, navigation qui peut s'avérer facilement identifiable lorsque les numéros de pages ou de résultats apparaissent en clair dans l'adresse de la page de résultats (URL), mais beaucoup moins évidente dans les autres cas.
Gérée moteur par moteur à l'aide de fichiers descriptifs par les métamoteurs, l'extraction des résultats n'est absolument pas gérée par les logiciels classiques de surveillance de pages. Ceux-ci sont en effet incapables de distinguer la zone de résultats du reste des informations de la page – générant en conséquence beaucoup de "bruit". Qui plus est, ils crawleront les autres pages de résultats comme n'importe quel autre lien sur la première page de résultat, rapatriant pour chaque nouvelle page de résultats crawlée de multiples pages non pertinentes.

• L'approche de Digimind : la reconnaissance et l'extraction automatique
Là encore, Digimind a développé des technologies exclusives de reconnaissance et d'extraction automatiques des résultats de moteurs de recherche. Basée sur des algorithmes avancés de reconnaissance de forme, cette technologie présente le grand avantage de s'adapter automatiquement à tout moteur de recherche – et même à toute modification d'un moteur de recherche.
Notre veilleur peut ainsi consulter et surveiller à l'aide du métamoteur Finder de la plateforme Digimind Evolution tous les résultats, et uniquement les résultats, des moteurs interrogés – sans aucune intervention de sa part.

3.1.3 Pouvoir explorer l'information "derrière les résultats"

Comme indiqué précédemment, les résultats des moteurs de recherche sont constitués de quelques éléments les plus représentatifs de l'information complète à laquelle ils mènent : titre, date, auteur, résumé dans la plupart des cas…ou encore notice synthétique dans le cas de bases documentaires structurées.
Dans le cadre d'une surveillance, ce résumé compte bien sûr, mais moins que l'intégralité de l'information vers laquelle il renvoi. En effet, les mots clés de votre surveillance peuvent tout à fait se situer au fin fond d'un article scientifique de 90 pages, sans que la notice seule puisse vous le révéler – voire sans que ce document ait été indexé dans son intégralité (voir page 8 les limites d'indexation des moteurs).
Il est donc indispensable dans le cadre de l'automatisation d'une veille du Web Invisible que les outils utilisés soient capables de lire l'intégralité des informations "derrière les résultats".

Cela suppose également que ces outils soient capables de lire des documents en d'autres formats que le html, car ceux-ci sont très fréquemment publiés directement en pdf, word, excel, powerpoint ou encore postscript.
Les principaux métamoteurs du web invisible ont une limitation majeure par rapport à ces spécificités de la surveillance du web invisible : ils se contentent la plupart du temps de rapatrier les résultats des moteurs de recherche et bases interrogées, sans aller crawler et indexer les informations derrière les liens des résultats. Du coup, ils sont limités par l'indexation effectuée par les moteurs de recherche, elle-même généralement non exhaustive (500 Ko par document pour Google et Yahoo ! par exemple).
Quant aux logiciels de surveillance du web, ils seront confrontés aux mêmes limitations que précédemment : étant incapables de distinguer ni les résultats, ni les pages suivantes des résultats, des autres liens de la page, ils crawleront sans distinction tous les liens, rapatriant pour chaque document pertinent, de multiples informations non pertinentes.
L'approche de Digimind : explorer, extraire et unifier automatiquement l'intégralité des informations derrière les résultats
Là encore, l'approche de Digimind se distingue par des technologies uniques d'exploration et d'homogénéisation de tout type de documents présents derrière les liens des résultats. Une fois la zone des résultats et le mode de navigation entre les pages identifiés par le 1er algorithme, et les résultats décomposés en leurs éléments constitutifs par un 2ème algorithme (voir plus haut) - dont les liens vers les documents cibles des résultats - chacun de ces derniers est alors exploré par un puissant robot. Celui-ci en effet :
-gère les formats bureautiques les plus répandus (pdf, doc, rtf, ppt, xl, postscript, en plus du html, xml, et txt)
-gère tous les types d'encoding de documents, dont non-occidentaux (japonais, chinois, russe, arabe,…)
-explore l'intégralité du document cible, sans limitation de taille
-génère à la volée un résumé des parties les plus pertinentes par rapport à la requête mise en surveillance
Résultat : une surveillance profonde et exhaustive du web invisible, quelles que soient les langues ou les formats des documents cibles.

3.1.4 La nécessité de filtrer les résultats redondants

L'interfaçage de plusieurs moteurs internes de sites du Web Invisible peut vous amener des informations redondantes, en double voire en triple. En effet, certains documents, publications et analyses peuvent être présentes sur plusieurs bases ou portails différents.
Gérée par la plupart des métamoteurs en regroupant les résultats identiques – mais rarement les informations similaires, puisqu'ils ne les re-explorent pas (voir plus haut) – la redondance n'est pas du tout prise en compte par les logiciels de surveillance du web. Dans leur cas, chaque requête effectuée sur différents sites du web invisible rapportera son lot d'alertes redondantes, sans possibilité de regroupage de ses informations autre que manuellement.

• L'approche de Digimind : regrouper automatiquement les informations similaires
Re-explorant les documents cibles des résultats des recherches, le robot de Digimind est capable de regrouper entre eux tous les documents identiques.

3.1.5 La nécessité de traiter de gros volume de données

Nous l'avons vu, le Web Profond est constitué de sites en moyenne plus volumineux que les sites du web visible (60% des sites les plus vastes du Web Profond représentent à eux seuls un volume qui excède de 40 fois le Web de Surface). De plus, le volume du web profond à un taux de croissance très élevé (multiplication par 9 chaque année). Enfin, nous avons compris l'importance de pouvoir explorer l'intégralité des documents issus des résultats de recherches, afin de ne pas rater une information clé située au fin fond d'un fichier PDF de plusieurs mégaoctets.

Dès lors, la puissance nécessaire à la surveillance du web invisible devient hors de portée de logiciels monopostes installés sur son PC de bureau. Copernic limite ainsi à 700 le nombre de résultats par moteur interrogé - et il ne s'agit d'afficher que les résultats, sans explorer les documents indexés. Même les métamoteurs grands publics tournant sur les index de Google, Yahoo ! ou MSN Search limitent le nombre de requêtes simultanées ou le nombre de moteurs interrogés simultanément.

• L'approche de Digimind : puissance, ciblage et profondeur
Les logiciels de Digimind sont hébergés sur des serveurs puissants et leur conception en J2EE autorise une montée en charge infinie en reliant les serveurs en clusters.
D'autre part, permettant de se constituer ses propres bouquets de sources du web invisible, le nouveau module Finder évite la recherche simultanée dans des milliers de moteurs. A une veille horizontale large et généraliste, l'approche préconisée est celle d'une veille ciblée (sur les sources réellement pertinentes pour le veilleur concerné) et profonde (en explorant l'intégralité des informations).

3.2 Les avantages de l'approche de Digimind pour la surveillance du web invisible

3.2.1 Une recherche temps réel et sur mesure du Web Invisible

La possibilité d'ajouter ses propres moteurs à ses recherches et surveillances permet au veilleur d'effectuer une veille vraiment ciblée sur ses problématiques clés, évitant ainsi le bruit issu soit de moteurs non pertinents (pour les métamoteurs classiques du marché), soit de modifications non pertinentes de pages surveillées (pour les logiciels classiques de surveillance du web).
Il peut ainsi lancer des recherches sur des moteurs généralistes tels que Yahoo!, des moteurs spécialisés comme Scirus (recherche scientifique) ou des moteurs internes de bases de données, qu'elles soient médicales, juridiques ou scientifiques (USPTO, PubMed, GlobalSpec), et récolter immédiatement des résultats extrêmement pertinents, qu'il pourra ensuite mettre en surveillance en un seul clic.
Pour l'expert, mais aussi pour tout collaborateur de l'entreprise, c'est la possibilité de puiser en temps réel dans des ressources autrement fastidieuses à exploiter une par une – et de pouvoir répondre de manière beaucoup plus réactive et pertinente aux questions « urgentes » de sa hierarchie ou de ses collègues.

3.2.2 Une surveillance profonde du Web Invisible

La possibilité d'explorer en profondeur les documents vers lesquels pointent les résultats assure de plus de ne pas passer à côté d'une information clé qui n'aurait pas été indexée par un moteur classique. Moins important peut-être dans une logique d'état de l'art sur un domaine, où les documents les plus représentatifs d'un sujet sont recherchés en priorité, et contiennent donc forcément les principaux mots-clés recherchés, l'exploration intégrale des documents devient vitale lorsqu'il s'agit d'identifier les « signaux faibles », par nature plus parcellaires, ambigus, et incomplets (Lesca – Caron, 1986).
L'avantage clé pour le veilleur ou l'expert de l'entreprise est une veille beaucoup plus fine et pertinente que celle qu'il pourrait effectuer avec des solutions "généralistes" du marché.

3.2.3 Une industrialisation de la surveillance

Cette veille plus pertinente débouche automatiquement sur un gain de temps important pour le veilleur : plus besoin de consulter des centaines de résultats – ou d'alertes – pour n'en retenir qu'une fraction.

Cette productivité est renforcée par une automatisation de l'intégralité du processus de recherche et surveillance du Web Invisible : application de requêtes à des bouquets de sources, surveillance

profonde quels que soient les formats, langues ou encoding des sources, extraction et dédoublonnage automatiques des résultats, génération de résumés automatiques.

Enfin l'architecture J2EE et les fortes capacités de traitement mis à disposition par le centre serveur de Digimind permet d'assurer une veille sur un nombre de sources, de moteurs et de pages illimité.

Au final, le retour sur investissement de l'automatisation de la surveillance de sources électroniques, et en particulier du Web Invisible, dépasse les 400% pour atteindre dans certains cas 1000% (voir à ce sujet le White Paper de Digimind sur « Evaluer le ROI d'un logiciel de veille », par Edouard Fillias, Consultant Veille Stratégique chez Digimind).

Une démocratisation de la recherche et de la surveillance du web invisible
De part sa simplicité de paramétrage, le module de recherche et de surveillance du web invisible est accessible à des non spécialistes des outils et de l'informatique, et en particulier aux experts métiers de l'entreprise (documentation, marketing, R&D, Produits, Finance,…).
La distribution de capacités de surveillance du web invisible dans l'entreprise à tous les experts métiers, qui vont pouvoir surveiller leurs propres sources, avec leur propre vocabulaire, et analyser les résultats avec leur expertise, enrichit fortement la qualité totale de la veille de l'entreprise.

Une logique de simplicité et de puissance que l'on retrouve dans tous les modules de la plateforme Digimind Evolution, et qui permet de déployer rapidement des projets de veille sur plusieurs dizaines et centaines de veilleurs.

Bibliographie

[1] MICHAEL K. BERGMAN., "*The Deep Web: Surfacing Hidden Value*", juillet 2000-février 2001
http://www.brightplanet.com/technology/deepweb.asp

[2] BRIAN H. MURRAY Cyveillance, "*Sizing the Internet*" , 10 juillet 2000
http://www.cyveillance.com/web/downloads/Sizing_the_Internet.pdf

[3] Les chercheurs de University of Pittsburgh and The
Pennsylvania State University "Missing Pieces", 2005, par le métamoteur Dogpile.com en
collaboration avec. http://dogpile.com