La veille sur les weblogs

Henry SAMIER(*), Victor SANDOVAL(**)

samier@istia.univ-angers.fr sandoval@mai.enpc.fr
(*) Enseignant chercheur, (MCF), responsable du DESS Information Stratégique
ISTIA Innovation, Université d'Angers, 62 avenue notre dame du lac, 49000 ANGERS
(**) Professeur à l'Ecole Nationale des Ponts et Chaussés, ENPC, Marne la vallée,

Mots clés :
veille stratégique – veille sur internet – weblogue – blogue

Keywords :
competitive intelligence – internet watch – weblog – blog

Palabra clave :
inteligencia strategic – vigilancia internet – weblog – blog

Résumé

Depuis ces dernières années les weblogs ont eu la croissance la plus forte de toutes les sources d'information humaines sur internet. Avec plus de 10 millions de weblogs, le phénomène des weblogs et de la syndication de contenu révèle une évolution sociale profonde dans ces sources d'information et une évolution technologique naturelle des technologies RSS[1]. La surveillance des weblogs est stratégique pour les entreprises afin de recueillir des informations complémentaires aux forums, aux listes de diffusions et aux « newsletters ». Mais les entreprises doivent créer aussi leurs propres weblogs afin d'animer de nouvelles communautés de partenaires et de clients.

Nous proposons de voir, d'une part, comment la veille sur les weblogs se met en place sur l'internet et d'autre part, comment cette veille s'intègre dans la veille stratégique en entreprise. Après le développement du panorama des weblogs, la définition, l'historique, les technologies et les pratiques, nous les positionnerons par rapport aux autres sources internet. Nous proposons de voir dans quelle mesure les méthodes de veille sont utilisables sur les weblogs.

[1] RSS : Rich site summary & really simple syndication

1. Introduction

Les nouveaux espaces d'expression libre tels que les weblogs, rendent compte aujourd'hui de la richesse des échanges d'idées, permis par les NTIC. Chaque visiteur d'un weblog peut faire un commentaire, lisible par tous, sur ce qu'il lit. Chaque salarié de l'entreprise a la capacité de créer son propre weblog, comme par exemple un weblog-projet, complémentaire aux outils de travail collaboratif de l'entreprise. Avec une croissance de plusieurs milliers de nouveaux weblogs par jour, est-il utile de surveiller ces nouveaux flux d'information pour les entreprises ? Les méthodes classiques de veille sur le web sont-elles adaptées aux weblogs ? Dans cet article, nous nous proposons de répondre à ces questions.

2. Les Weblogs et les sources internet

2.1. Les weblogs et la « syndication »

La relative jeunesse des weblogs induit d'une part une grande variabilité des définitions et d'autre part une richesse et une variabilité terminologique. Les weblogs correspondent à des sites web personnels thématiques d'actualité, qui autorisent les contributions des internautes et qui sont publiés par ordre ante-chronologique. La terminologie abondante et variée que nous pouvons trouver est la suivante : weblog, blog, blogue, blogueur, journal de bord sur le web, web journal, web syndication, diariste, online diarie, personal knowledge publishing, joueb, cybercarnet, carnetier, carnetweb, carnet de bord, moblog (mobile blog), ou encore des weblogs collaboratifs tels que les wikiweblog, wikiblog. Aussi la notion de syndication désigne la diffusion et la publication automatique d'une même information en flux RSS sous différentes formes, supports et autres weblogs.

Le weblog est un phénomène sociologique des communautés virtuelles qui échangent et partagent des idées en utilisant des technologies qui le permettent. Le phénomène des weblogs et de la « syndication de contenu est profond », il est durable et ne relève pas d'un effet de mode. Les weblogs ont un accès libre pour les internautes, les mises à jour sont réalisées par ordre chronologique inversé, les informations qui sont éditées engagent la responsabilité des auteurs et toutes les informations sont archivées sur le site.

Les weblogs sont déjà des sources d'informations « humaines » complémentaires au web qui ont leurs annuaires[2] et leurs moteurs de recherche[3] spécifiques. La « technologie » utilisée se présente sous le format (flux, fil ou canal) RSS (rich site summary : sommaire de site enrichi) ou le format ATOM[4]. Un format RSS, qui utilise le format XML, est simplement une description de fichier ayant des balises spécifiques RSS, basée sur la norme RDF. Le fonctionnement des weblogs passe, par l'utilisation d'agrégateur[5] en ligne (gratuit ou payant), ou par l'utilisation d'une logiciel éditeur[6] ou agrégateur[7] à installer sur son serveur. De l'étudiant à l'enseignant, au chercheur, des réseaux sociaux ou économiques, tout le monde peut aujourd'hui créer ou contribuer à un weblog et partager et valoriser ses propres travaux[8].

---

[2] par exemple www.scienceport.org, www.annublog.com www.rssreporter.net, www.lamoooche.com, www.retronimo.com, www.newsisfree.com/sources/bycat/

[3] par exemple www.feedster.com, www.blogdigger.com, www.daypop.com, www.completerss.com, www.rssfeeds.com, www.popdex.com, blawgs.detod.com, www.boogieplay.com, www.search4blogs.com, www.technorati.com, www.blogdump.com

[4] se reporter à www.ietf.org

[5] par exemple : www.rss4you.com, www.my.syndicnation.com,

[6] se reporter à www.cafelog.com, www.nucleus.org

[7] se reporter à www.newsisfree.com, www.newsgator.com, www.feedreader.com

[8] par exemple le site et weblog remarquable de Rachel Gibert : http://gibert.rachel2.free.fr/stage/

2.2. Positionnement et tendances

La toute première forme de « weblog » est publiée en 1992 par de Tim Berners Lee avec sa rubrique « what's new ». Mais c'est en 1996 que nous trouvons les premières formes abouties de weblogs avec les sites des pionniers tels que John Barger, Dave Winer, William Quick, Cameron Barrett et Rob Malda. Dans ces années, la première vague des weblogs se fonde sur des partages d'informations personnels, communautaires et thématiques. La seconde vague voit apparaître dans les années 2000 le développement des weblogs regroupés par professions[9] tels que les journalistes, informaticiens, enseignants, documentalistes, politiques, etc. Depuis ces deux dernières années, nous assistons à la troisième vague des weblogs développés par les entreprises (« corporate blog »), organismes et institutions. Aujourd'hui, le développement est « tout azimut » et ces trois vagues de weblogs se confondent dans la « blogosphère » qui a subi la plus forte croissance des sources web. La « blogosphère » a d'ailleurs ses propres valeurs et ses propres outils de recherche. Fort de plusieurs dizaines de millions de sites, les weblogs continueront leur croissance et dans les années à venir ils atteindront un nombre proche, voire équivalent au nombre des sites web puis chaque site web aura son propre weblog.

Les weblogs sont actuellement complémentaires aux sources d'informations humaines telles que les forums de discussion, les listes de diffusion et les « newsletters ». Toutes ses sources co-existeront, car elles répondent à des besoins et à des fonctions différentes. Par ailleurs, la technologie des weblogs permet des mises à jour par courriers électroniques, « mail » ou SMS, ce qui laisse augurer un bel avenir pour ce type de communication.

3. La veille sur les weblog

3.1. Démarche (détection et repérage)

La démarche de veille sur les weblogs débute classiquement par la première détection des sites. Premièrement, l'utilisation des « annuaires web » permet de trouver les rubriques thématiques et donc des listes d'annuaires et de moteurs de recherche spécifiques à la blogosphère. Par exemple, sur yahoo.fr dans la l'onglet « rubrique web » nous recherchons le terme « blog ».

Le résultats de la rubrique « blogs > portails et annuaires » contient une dizaine d'adresses dont, blogonautes.com qui recense plusieurs milliers de weblogs francophones. Remarquons que sur yahoo.com la rubrique weblog compte plus de 2000 sites classés par sous rubriques. Cette démarche itérative doit se poursuivre sur plusieurs « annuaires web» afin de trouver d'autres « annuaires de weblogs[10] ».

Nous trouvons ainsi des rubriques de blog sur l'actualité, la politique, les photos, les outils de publication, etc. A cette étape, nous consultons les sites afin de valider leur pertinence sur des thèmes de veille. Ensuite, nous utilisons les « annuaires web » pour trouver les moteurs de recherche spécifiques aux weblogs tel que www.feedster.com par exemple.

Deuxièmement, nous utilisons les moteurs de recherche à l'aide des méthodes Xfind en croisant les termes blog ou weblog avec les expressions de la recherche. Troisièmement, nous construisons les carnets d'adresses (signet) des weblogs, pour leur mise en surveillance ultérieure.

Nous proposons de voir dans quelle mesure les méthodes Xfind s'appliquent à la veille manuelle des weblogs. Les quatre méthodes pratiques RapidFind, DetectFind, LinkFind et ExtractFind [SAM 99] [SAM 03] sont donc appliquées sur les sources d'information de type weblog afin de mesurer le gain de temps et la pertinence.

---

[9] voir par exemple http://www.cyberjournalist.net/cyberjournalists.php,

[10] par exemple sur les sites : www.rssreporter.net, www.lamoooche.com, www.retronimo.com, www.newsisfree.com/sources/bycat/

Ces méthodes utilisent systématiquement les options avancées des moteurs de recherche exploitant la structure des documents de l'internet. Les recherches peuvent donc s'effectuer à la fois dans : le titre, le résumé, l'adresse de la page (URL), le nom du site, le nom de domaine et l'adresse d'un lien.

3.1.1. Méthode RapidFind

La méthode Rapidfind vise à trouver l'existence d'une expression précise dans le titre d'un document afin d'obtenir initialement des réponses pertinentes mais en faible nombre. La logique de recherche est une logique de détection et non une logique de recensement. A cet étape, l'analyse des résultats dépend du temps consacré par l'équipe de veille.

Par exemple, sur le thème de la maladie de la vache folle, nous consultons les moteurs de recherche suivants :

alltheweb.com

1. url :weblog and title : « mad cow disease » : 75 résultats [méthode rapidfind]
2. url :blog and title : « mad cow disease » : 413 résultats [méthode rapidfind]
3. url:weblog and "mad cow disease" : 406 résultats
4. url:blog and "mad cow disease" : 1400 résultats
5. blog and "mad cow disease" : 27700 résultats

google.com

1. allintitle: weblog "mad cow disease" : 1 résultat [méthode rapidfind]
2. allintitle: blog "mad cow disease" : 24 résultats [méthode rapidfind]
3. weblog "mad cow disease" : 6430 résultats
4. blog "mad cow disease" : 22300 résultats

hotbot.com

1. weblog "mad cow disease" <in the title> : 25 résultats [méthode rapidfind]
2. blog "mad cow disease" <in the title> : 48 résultats [méthode rapidfind]
3. weblog "mad cow disease" : 3557 résultats
4. blog "mad cow disease" :7270 résultats

En complément, nous réalisons les recherches sur les moteurs de recherche dédiés aux weblogs.

feedster.com

1. "mad cow disease" : 11398 résultats
2. "prion disease" : 890 résultats
3. "maladie de la vache folle" : 41 résultats
4. "maladie du prion" : 6 résultats

blogdigger.com

1. "mad cow disease" : 565 résultats
2. "prion disease" : 68 résultats
3. "maladie de la vache folle" : 2 résultats
4. "maladie du prion" : 2 résultats

3.1.2. Méthode « DetectFind »

La méthode « Detectfind » vise à trouver l'existence de carnets d'adresses web sur les weblogs afin de trouver des weblogs complémentaires à ceux trouvés précédemment.

Le principe est le suivant :

allteweb.com

1. url:favorite AND url:weblog : 142 résultats
2. url:bookmark AND url:weblog : 22 résultats
3. url:favorite AND url:blog : 329 résultats
4. url:bookmark AND url:blog : 80 résultats

google.com

1. favorite weblog <dans l'adresse> : 54 résultats
2. bookmark weblog <dans l'adresse>: 23 résultats
3. favorite blog <dans l'adresse>: 6480 résultats
4. bookmark blog <dans l'adresse>: 141 résultats

3.1.3. Méthode « LinkFind »

La méthode « Linkfind » vise à trouver l'existence de liens HTML entre les weblogs (link) en excluant les liens circulaires (url).

Nous suivons le principe ci-après :

allteweb.com

1. link:blog.simmins.org AND NOT url:blog.simmins.org AND "mad cow disease": 77 résultats
2. link:www.scienceblog.com AND NOT url:www.scienceblog.com AND "mad cow disease" : 17 résultats
3. link:purplemedicalblog.blogspot.com AND NOT url:purplemedicalblog.blogspot.com AND "mad cow disease": 10 résultats
4. link:www.worldmagblog.com AND NOT url:www.worldmagblog.com AND "mad cow disease": 42 résultats

3.1.4. Méthode « ExtractFind »

La méthode « Extracfind » vise à extraire d'un weblog, les informations pertinentes sur un sujet. Sans parcourir le site, nous connaissons le nombre de documents contenus dans un site sur un sujet donné.

google.com

1. "mad cow disease" site:www.scienceblog.com : 173 résultats
2. "mad cow disease" site:purplemedicalblog.blogspot.com : 2 résultats
3. "mad cow disease" site:www.worldmagblog.com : 4 résultats
4. "mad cow disease" site:blog.simmins.org : 1 résultat

3.2. Analyse et discussion

La première méthode Rapidfind est totalement adaptée aux weblogs puisque d'une part elle conduit à trouver rapidement 100 % d'informations pertinentes et d'autre part, elle est complémentaire aux recherches effectuées sur les moteurs de recherche spécifiques aux weblog (feedster.com, blogdigger.com). Cette méthode permet de trouver rapidement plusieurs centaines de documents, pertinents à 98 %, après analyse, parmi plus de 27700 documents.

La deuxième méthode Detecfind permet de trouver des carnets d'adresses web (favoris et bookmarks) sur les weblogs, mais il n'est pas possible de croiser le thème « mad cow disease » avec les blogs. Cette méthode, pourtant utile, se limite à des recherches plus généralistes.

La troisième méthode Linkfind a permis de trouver les liens existants entre une page pertinente sur un site et d'autres sites. En règles générales, les liens ont un taux de pertinence supérieur à 90%, même si dans les cas des weblogs, nous avons eu 100% de pertinence dans les résultats.

Enfin, la quatrième méthode Extracfind a permis d'extraire toutes les pages pertinentes des sites de Weblogs. Cette méthode fonctionne au même titre que sur le web, et même si la plupart des weblogs ont leur propre moteur de recherche interne pour accéder aux archives, cette méthode autorise la recherche sur plusieurs sites en même temps.

La limite actuelle des méthodes manuelles « rapidfind », « detectfind », « linkfind » et « extractfind » porte sur le fait qu'elles ne peuvent pas encore s'appliquer dans les moteurs spécifiques aux weblogs tels que feedster.com et blogdigger.com, par exemple. En effet, ces moteurs n'ont pas encore des options de recherche avancée qui permettent de réaliser les recherche sur les titres, les adresses et les liens. Mais nous sommes convaincus que cette limite n'est que temporaire.

Ainsi nous avons validé les méthodes rapidfind, linkfind et extractfind sur les weblogs. A l'issue de ces recherches manuelles, nous avons obtenu des listes de sites que nous allons surveiller heures par heures avec les outils de veille automatique.

4. La veille stratégique sur internet

La mise en place de la veille stratégique sur l'internet nécessite d'une part, l'utilisation de logiciels ou de systèmes informatiques de veille et d'autre part des outils graphiques de pilotage stratégique.

Les logiciels de veille actuels les plus à mêmes de surveiller les weblogs sont Kbcrawl, Digimind, Weformancewatch et Orbiscope. Pour notre expérimentation nous avons utilisé Weformancewatch et Kbcrawl. La programmation et le paramétrage de ces outils ne diffèrent pas des pratiques du web. Afin de suivre l'évolution des thèmes de veille, nous proposons des outils graphiques sous forme de tableau de bord.

Les tableaux de bord de veille sur l'internet (TBVI) fournissent une traçabilité de la stratégie de surveillance. Par exemple, nous proposons un tableau de bord sur le thème des weblogs et la maladie de la vache folle.

Dans le tableau de bord, nous utilisons d'une part les critères des dossiers de surveillance et d'autre part, nous définitions des critères utiles à la traçabilité du processus de veille. Au total, nous obtenons 13 critères qui constitueront les tableaux de bord :

1. Adresse du site (url) : exemple : scienceblog.com
2. Objectifs de la surveillance : ce que l'on cherche réellement
3. Fonction de surveillance : aspirer, surveiller, mettre à jour, etc.
4. Outils de surveillance : agents intelligents utilisés.

5. Date de début de surveillance : jour / mois / année.
6. Mots clés et expressions : profils de recherche et équations logiques.
7. Niveau de surveillance : horaire, quotidien.
8. Nom des Cyberveilleurs : équipe ou individu en charge de la surveillance.
9. Nom des destinataires : équipe ou individus concernés.
10. Fréquences de diffusion : différentes de la fréquence des recherches.
11. Dates et heures de diffusion : adaptée aux rythmes des destinataires.
12. Mode de diffusion : optimisée par destinataires.
13. Localisation des informations : le serveur où se trouvent les informations.

Projet : MCD Watch

Thème : Les recherches sur la maladie de la vache folle

Critère **Nom**	science	Angleterre	Europe
C1 : Adresse du site	scienceblog.com	Liste A1	Liste E5
C2 : Objectifs de la surveillance	Programmes	Cas	Traitement
C3 : Fonction de surveillance	Tracker	Tracker	Aspirer
C4 : Outil de surveillance	KBCrawl	Webformancewatch	Teleportpro
C5 : Date de début de surveillance	jj/mm/aaaa	jj/mm/aaaa	jj/mm/aaaa
C6 : Mots clés et expressions	Profil N6-13	Profil N26-05	Profil N16-08
C7 : Niveau de surveillance	N1 : temps réel	N1 : temps réel	N2 : hebdomadaire
C8 : Nom des Cyberveilleurs	V. Sandoval	H.Samier	M. Corsi
C9 : Nom des destinataires	Comité de pilotage	Comité de pilotage.	Comité de pilotage
C10 : Fréquences de diffusion	Quotidien	Hebdo	Hebdo
C11 : Date et heure de diffusion	8h30	Vendredi 15h	Vendredi 15h
C12 : Mode de diffusion	Mail, Intranet	Intranet	Intranet
C13 : Localisation d'information	Serveur de mail,	Serveur de Veille	Serveur de Veille

Tableau n°1 : Tableau de bord de veille « MCD Watch »

Ces tableaux de bord dédiés aux weblogs sont à intégrer aux autres tableaux de bord du web, les forums, les listes de diffusion et les IRC Chat[11].

5. Conclusion

Nous observons que la veille sur les weblogs apporte des informations complémentaires et non redondantes à la veille classique sur le web. L'analyse des résultats nous permet de dire que sur le thème de « la maladie de la vache folle », il est indispensable de surveiller les weblogs. Il reste à valider et à transposer la démarche sur d'autres thèmes. Mais au regard de la vivacité des weblogs, il sera bientôt indispensable de surveiller ces flux d'information. Nous pouvons imaginer l'apparition dans un futur proche, de nouvelles fonctions en entreprise telles que les responsables de blogs, « corporate blogger », et les responsables de veille sur les blogs, « watchbloggers ». Nous sommes convaincus que cette veille sera d'autant plus efficace si elle fonctionne de manière collaborative et que les sites tels que Vcoop.net et fooxx.com proposeront dans un proche avenir la fonction de veille sur ces flux d'information.

[11] Fonction existante seulement sur Webformancewatch

6. Bibliographie

[1] AIMEUR E., BRASSARD G., PAQUET S., Using personal knowledge publishing to facilitate sharing across communities. In M. Gurstein (Ed.), *Proceedings of the 3rd International Workshop on (Virtual) Community Informatics: Electronic Support for Communities - Local, Virtual and Communities of Practice*, May 2003.

[2] BLOOD R., Weblogs: a history and perspective. http://www.rebeccablood.net/essays/weblog_history.html, 7/9/2000, [consulté le 19/09/2002].

[3] BRADLEY P., Search Engines: Weblog search engines, www.ariadne.ac.uk/issue36/search-engines/ [consulté le 22/05/04].

[4] CyberJournalist.net. Another Journalist Blogger Shut Down. Retrieved April 17, 2003, site www.cyberjournalist.net/news/000240.php

[5] DRUCKER P., *The new society of organizations*, Harvard Business Review, Sept-Oct 1992, p 95-104

[6] FACCA M. F., LANZI L. P., Mining interesting knowledge from weblogs : a survey, data and knowledge engineering review,

[7] Into the blogosphere, site http://blog.lib.umn.edu/blogosphere/

[8] JOHNSON A. R., JohnsonCompetitive Intelligence Weblogs : Building Market Monitoring Capabilities with Rapid Analysis Tools, KMWorld & Intranets 2003, october 14-17, 2003.

[9] MERELO-GUEROS J.J., PRIETO B., RATEB F., Mapping weblog, http://arxiv.org/abs/cs.NE/0312047, [consulté le 13/07/2004].

[10] PIKAS K. C., Trends in Blog Searching, www.llrx.com/features/trendsblogs.htm [consulté le 03/07/04].

[11] SAMIER H., SANDOVAL V., "La veille stratégique sur internet", Editions Hermes Sciences, Paris 2002.

[12] WINER D., What makes a weblog a weblog ? blogs.law.harvard.edu/whatMakesAWeblogAWeblog Fri, May 23, 2003 [consulté le 13/09/2003].

[13] ZIMMERMAN K.A., Blogging the competition – weblogs take center stage in competitive intelligence, KMWorld Vol n°12, Issue 10 november/december, http://kmworld.com/publications/magazine/

7. Adresses

B2	www.cafelog.com
Berkeley IP Blog	journalism.berkeley.edu/projects/biplog/
Blizg	www.blizg.com
Blogarama	www.blogarama.com
Blogcensus	www.blogcensus.net
Blogdex	www.blogdex.com
Blogger.com	www.blogger.com
Bloglines	www.bloglines.com
Blogolist	www.blogolist.com
Blogonautes	www.blogonautes.com
Blogroots	www.blogroots.com
Blogs d'infoworld	weblog.infoworld.com
Blogsearchengine	www.blogsearchengine.com
Blogstreet	www.blogstreet.com
Blogwise	www.blogwise.com
Bloogz	www.bloogz.com
Bog City	www.blog-city.com
Competitive intelligence	www.aurorawdc.com/ci/
Crao wiki	http://wiki.crao.net
Cre8pc	www.cre8pc.com/blog/
Daypop	www.daypop.com
EduBlogInsights	http://anvil.gsu.edu/EduBlogInsights/
Educational blogs	http://educational.blogs.com
Fastcompany	blog.fastcompany.com
Feedster.com	www.feedster.com
Google Blogspace	google.blogspace.com
Googleblog	www.google.com/googleblog/
Harvard Law	http://blogs.law.harvard.edu

Hautetfort.com	www.hautetfort.com
Instapunti	www.instapunti.com
Intelligence économique	http://refer-ie.joueb.com
Joueb	http://pages.joueb.com
KM Blog	www.voght.com/cgi-bin/pywiki?KmBlogger
Knowledge workers	http://outilsfroids.joueb.com
L'œil de mouche	http://mouche.blogspot.com
Lesblogs	www.lesblogs.com
Leweblog	www.leweblog.com
Library Weblog	www.libdex.com/weblogs.html
Livejournal	www.livejournal.com
Loiclemeur	www.loiclemeur.com
Mediatic	http://mediatic.blogspot.com
Movable Type	www.movabletype.org
Nucleus	www.nucleuscms.org
Overblog	www.over-blog.com
Pointblog	www.pointblog.com
Politique internationale	www.netlexfrance.com/weblogs
Quebeblogues	www.quebeblogues.com
Radio weblogs	http://radio.weblogs.com
Recherche sur le net	http://influx.joueb.com
Scripting.com	http://www.scripting.com
serendipit-e	http://www.serendipit-e.com
Skyblog	www.skyblog.com
Slashdot	www.slashdot.org
Stervinou	www.stervinou.com
The future of work of blog	www.thefutureofwork.net/blog/
U-Blog	www.u-blog.net
Waypath.com	www.waypath.com
Weblog de l'ESC de Pau	www.esc-pau.fr/weblog/
Weblog de l'ISTIA	http://shiva.istia.univ-angers.fr/~blogistia/blog/index.php
Weblog de l'UTT	http://utt.leweblog.com
Weblog Handbook	www.rebeccablood.net/handbook/
Weblogues	www.weblogues.com
Yahoo Blog	www.ysearchblog.com/
Yulblog	www.billegible.org/yulblog/
Zillman.	http://zillman.blogspot.com/