Solutions de veille choix technologiques et stratégiques

par Frédéric Martinet

Si les solutions de veille sur Internet existent désormais depuis plusieurs années, le nombre d'acteurs francophones est toujours relativement limité. Ainsi, quelques uns se partagent la part d'un gâteau dont la taille est restreinte si l'on exclut de son champ les études sur mesure et tout ce qui relève du « Search » et de la Business Intelligence.
© Frédéric Martinet

La révolution du Web 2 n'a pas été sans effet sur les nouvelles directions empruntées par ces acteurs. Le web qui se résumait il y a dix ans à quelques bases de données payantes, sites corporates et forums en guise d'espaces participatifs s'est transformé en un univers hétéroclite et dont la structure documentaire a fortement évolué. L'information n'est plus seulement contenue dans une page mais également dans des espaces annexes tels que commentaires, avis, trackbacks.

Les enjeux ont évolué et la demande également : maîtrisant leurs marques dans les mass medias, les entreprises sont donc confrontées à l'impérieuse nécessité de se préoccuper de ce Web un peu effrayant car souvent incompréhensible dans ses mécanismes de propagation.

Comment les acteurs tels que Digimind, Ami Software, KB Crawl, Spotter, Ixxo, iScope et autres ont-ils fait face à ces évolutions qualifiées d'e-reputation ? Comment ont-ils répondu à l'évolution structurelle du Web depuis 10 ans ? Quels sont les choix technologiques qui ont rythmé leurs « releases » ?

L'objet de cet article ne sera pas de détailler nominativement la stratégie des éditeurs mais bien de lister des choix essentiels auxquels ils peuvent être confrontés et donc, de fait, des conséquences fonctionnelles qui en découlent pour leurs utilisateurs.

SAAS or « in the house » ?

Un des premiers choix qui se pose à un éditeur de logiciel de veille est de livrer une plateforme en accès distant, hébergée sur leurs serveurs, dites SaaS (Software as a Service ou anciennement ASP) ou bien de permettre l'installation sur leurs serveurs.

Ce choix demeure critique à bien des égards.

Par exemple en fonction du secteur adressé : les entreprises désirant effectuer une veille hautement stratégique seront peu enclines à la voir hébergée sur des serveurs extérieurs, ce qui est à mon sens totalement justifiable, quelques soient les certifications ou attestations de sécurité fournies par l'éditeur logiciel. Le secteur de l'Industrie de la Défense fait par exemple partie de ces clients peu réceptifs aux solutions SaaS.

Le mode SaaS permet à l'éditeur une maintenance simple du produit, ne nécessitant pas d'intervention sur site. Pour le client, il comporte de nombreux avantages :

- Le déploiement du projet se révèle en général plus simple car ne nécessitant pas de longues et délicates négociations avec la DSI.
- Il ne nécessitera par ailleurs pas l'achat d'infrastructure informatique (serveur) et donc pas de maintenance 24 / 24 pour les services informatiques de l'entreprise.
- Se situant à l'extérieur du pare-feu de l'entreprise et de ses différents proxys, la consultation du Web est totalement débridée.
- La solution SaaS offre un certain anonymat à l'entreprise dont la signature du proxy sera moins visible.

Pour l'éditeur, le mode SaaS permet de ne pas s'encombrer de processus d'intégration lourds, de s'assurer de la totale maîtrise de son code source de son logiciel et des infrastructures ce qui est bien plus simple en termes de maintenance et d'évolutivité.

Le mode « in the house » comporte également des avantages pour l'éditeur et pour le client :

- Les contenus crawlés et indexés et rediffusés sont totalement extérieurs à l'éditeur logiciel : ce dernier se prémunit ainsi de toute poursuite juridique relative aux droits d'auteur. Bien que la jurisprudence ne soit pas totalement convergente sur ce point, certains jugements ont pu confirmer que celui qui héberge des contenus copiés sans respect des droits d'auteur peut être tenu en partie responsable du préjudice. A contrario, le risque repose alors sur le client.
- Pour le client, ce dernier est assuré qu'en cas de dépôt de bilan de l'éditeur, il restera propriétaire des contenus et de la solution installée sur ses serveurs.
- On peut envisager des projets plus complexes s'interfaçant avec des bases de données ou des solutions logicielles internes de façon plus aisée.

Sourcing or not sourcing ?

Un autre choix stratégique pour l'éditeur est de proposer des packages de sources : packages génériques ou sur mesure.

Deux stratégies s'opposaient très clairement.

La première, DIY alias Do It Yourself, imposait au client le paramétrage intégral des sources : de l'indentification jusqu'à la saisie et à l'organisation de ces sources dans la solution de veille. Cette phase, généralement lourde et contraignante pour l'utilisateur nécessitaient à la fois une bonne connaissance des sources nécessaires à son projet (ce qui posait peu de problèmes en général) mais également une connaissance précise des fonctionnements des logiciels de veille.

La seconde solution adoptée consistait à livrer au client des packages de sources, génériques (news, blogs,…) ou spécifiques (pharma, bancassurance, …)

Aujourd'hui le jeu des acteurs apparaît comme beaucoup moins clairs mais l'on peut distinguer plusieurs composantes qui reposent sur certains fondamentaux.

Tout d'abord il faut comprendre que la maintenance d'une base de sources est extrêmement coûteuse : au-delà d'un certain nombre de sources, la qualité chute de façon assez importante, entraînant un taux de mortalité, ou une instabilité pouvant impacter directement la qualité des packages de sources vendues et entraînant donc un coût de maintenance important pour les éditeurs de solutions.

●●●

Plusieurs choix sont alors possibles pour les éditeurs :

- La qualité de leur produit repose sur un corpus de sources internationales, complet, avec des focus métiers. La maîtrise de ces sources est donc un véritable enjeu stratégique. Les projets sont livrés clés en main, sourcing compris, et ils complètent pour chacun des projets avec des sources nécessaires et souvent fournies par leur client. La maintenance de cette base de sources est un poste de coût important pour l'éditeur qui aura tendance à externaliser le paramétrage de ces sources dans des pays où la main d'œuvre est bon marché, le contrôle qualité final se réalisant directement chez l'éditeur. Le bénéfice direct pour le client est la rapidité de mise en place de son projet de veille. Les éditeurs et prestataires de ce type se placent ainsi en très bonne position pour réaliser des études à façon ou pour répondre à ces situations de communication de crise.
- D'autres éditeurs passent au sourcing plus ou moins contraint et forcé, sous la pression de leurs clients. Bien que leurs logiciels ne reposaient pas à la base sur cette logique de sources intégrées, la plupart d'entre eux proposent à minima des packages de source au moins généralistes. On constate sur ce segment des stratégies intéressantes. Par exemple, les éditeurs décident d'externaliser le sourcing via des filiales à l'étranger, sur des marchés solvables. La présence d'une filiale dans l'un des pays du Maghreb par exemple, permet le recrutement en local de personnes et facilite les transactions commerciales à venir avec les acteurs institutionnels et les entreprises ou ex-entreprises d'Etat.
- La dernière option est de permettre l'interrogation de moteurs de recherche qui jouent ainsi le rôle d' « infomédiaire ». Cela permet de livrer un nombre limité de sources paramétrées tout en offrant l'accès à un vaste corpus documentaire.

Le problème est que les moteurs de recherche voient cela d'un mauvais œil et donc se protègent de ces crawls considérés comme du pillage. Liens de redirections, bannières publicitaires, anti-robots, autant d'outils qui compliquent la tâche des éditeurs de logiciels de veille désireux de s'appuyer sur l'infrastructure de ces mastodontes du web que sont devenus Google et Microsoft. Par ailleurs, il est alors délicat pour un client de savoir exactement les sources qui sont interrogées ou pas et comment elles le sont.

Scraper or not scraper ?

Utilisé dans d'autres solutions logicielles telles que iWebScraping (1) ou WebSunDew(2), le Web Scraping (également appelé web harvesting) extrait de l'information d'un site web, de façon ciblée, pour l'utiliser dans un autre contexte.

La présence d'une filiale dans l'un des pays du Maghreb par exemple, permet le recrutement en local de personnes et facilite les transactions commerciales à venir avec les acteurs institutionnels et les entreprises ou ex-entreprises d'Etat.

Il existe plusieurs niveaux de scraping pour lesquels les éditeurs ont pu opter.

Le niveau 0

consiste à prendre une page Web sans se soucier de cibler une partie du contenu. La page est prise en tant que telle et l'on va surveiller cette page et / ou l'apparition de nouveaux liens sur cette page extraits automatiquement. C'est la solution de facilité. Pas de paramétrage complexe pour les éditeurs ni pour les utilisateurs de ces solutions, mais par contre la qualité des documents collectés s'en ressent.

Le niveau 1

consiste à sélectionne certaines parties de contenus par des filtres automatiques.

Par exemple Website Watcher [3] permet de suivre (ou de ne pas suivre) des liens comportant certaines chaînes de caractères seulement. Il dispose par ailleurs d'un outil d'apprentissage des zones à considérer comme inintéressantes. Les technologies utilisées sur ce niveau 1 sont les expressions régulières, les « algorithmes maison [4] », couplé éventuellement avec une base de connaissances de liens à ne pas suivre (adsense, advertisement, etc …)

Le niveau 2

consiste à cibler précisément la partie des pages à suivre. Pour cela les éditeurs ont opté pour des technologies différentes.

Il est possible de paramétrer à la main les parties de pages à suivre. A partir de chaînes de caractères du code source, l'on extrait la partie de la page se trouvant entre deux marqueurs de ce type.

Ce choix pose des problèmes importants en cassant par exemple des liens hypertextes, ou faisant disparaître des images ou la mise en forme [5]. Peu de clients sont enclins à avoir ce niveau d'intervention sur leur projet et l'on retrouve ces technologies plutôt chez les éditeurs qui font outsourcer la gestion de leurs sources. L'on peut par ailleurs, en cas de besoin impérieux, faire la même chose y compris avec des solutions ne proposant pas ce type de paramétrage en utilisant des outils externes (Yahoo Pipes par exemple).

Sur ce même niveau de découpage, il est possible de reposer sur le modèle DOM[6] de la page afin d'en effectuer une découpe simple. L'explorateur DOM est utilisé depuis plusieurs années par les développeurs et on le retrouver dans de nombreux plugins ou toolbar type, DebugBar, ou DOM Inspector… C'est également la technologie majoritairement retenue pour faire du Web Harvesting[7] . L'utilisation de cette technologie permet aux éditeurs de logiciels de veille d'offrir un ciblage précis des zones à surveiller le tout avec des ergonomies extrêmement simples utilisables par tous.

Le niveau 3

du Web Harvesting permet d'extraire plusieurs champs et de les attribuer en base de données. Le Web Harvesting est massivement utilisé dans la « veille tarifaire » ou la veille technologique. Cette combinaison de technologies se fonde à la fois sur l'utilisation du DOM et sur des paramétrages complémentaires manuels via la description des marqueurs. Le recours à ces technologies est particulièrement efficace lorsque l'on crawle peu de sites mais de tailles très importantes et que l'extraction de la donnée revêt une importance particulière et directement opérationnelle.

Malgré toutes les promesses du « sémantique », malgré tous les algorithmes magiques, elle reste la seule technologie qui assure un passage efficace du web déstructuré à une information structurée.

•••

••• notes •••

(1) iWebScraping est uyne société de services proposant l'extraction de données Web – www.iwebscraping.com

(2) WebSunDew est une solution payante – www.websundew.com

(3) Website Watcher est une solution logicielle monoposte d'entrée de gammes éditée par Maerin Aignesberger – www.aignes.com

(4) Dans un white paper de février 2004, intitulé « Extraction automatisée d'actualités on line pour la veille stratégique », Digimind présentait son algorithme iScrap, permettant de détecter sur une page les parties relevant de ce qui est actualité et donc de ne suivre que ces liens particulier.

(5) Pour les techniciens, l'on trouve parfois dans la partie head d'une page HTML une balise « BASE HREF » qui permet de résoudre les URL relatives. Supprimer ce HEAD supprime cette base de référence brisant les liens hypertextes. De même dans ce HEAD se trouve de plus en plus souvent le lien vers une feuille de style (CSS) : sa suppression dégrade la mise en forme des contenus collectés et parfois même toutes les images.

(6) Document Object Model

(7) On pourra par exemple regarde rla solution open source Web Harvest qui repose typiquement sur cette technologie : http://web-harvest.sourceforge.net/screenshots.php

(8) L'extraction d'entités nommées permet de dégager automatiquement à l'intérieur d'un document le nom de personnes, de sociétés, de lieu… Cette fonctionnalité repose à la fois sur de l'analyse morpho-syntaxique (en général on écrit nom et prénoms avec une majuscule à la suite), une reconnaissance de certains termes spécifiques (verbe de prise de parole, d'action, de situation,…) et parfois est couplée avec une base de connaissance. On pourra par exemple voir de ce type la solution Open Calais qui procure des API et également plugin Firefox, Gnosis (https://addons.mozilla.org/fr/firefox/addon/3999).

Sémantique ou pas ?

C'est un choix cornélien qui a par ailleurs déjà fait de nombreuses victimes. Le sémantique est-il viable ? Fonctionne-t-il ? C'est un choix sur lequel est revenu Alain Beauvieux, PDG d'Ami Software, dans le numéro 116 de Veille Magazine dans l'article intitulé : « Il y a sémantique et... sémantique »

Nombre d'éditeurs de logiciels de veille l'ont aujourd'hui compris : le sémantique n'est pas facilement applicable à la veille. Pas sans ontologies. Pas sans base de connaissances. Certains résistent encore (difficilement) et peinent à prouver leur valeur si ce n'est sur des projets dont les ontologies métiers sont déjà bien maîtrisées.

Le choix du sémantique limitera par ailleurs fortement le nombre de langages aux quels la technologie pourra s'appliquer.

Le choix du morpho-syntaxique, du statistique et de l'extraction d'entités nommées (8) semble toutefois se dégager de plus en plus fortement chez les éditeurs.

L'extraction d'entités nommées permet d'enrichir les contenus collectés par la solution de veille, de rapprocher des documents par leur contenu, et d'envisager des navigations interactives.

La communication autour de ces technologies sémantiques est toutefois aujourd'hui plus que floue. Il est délicat pour ces éditeurs de communiquer sur l'e-reputation, le « sentiment analysis » ou « opinion mining » champs d'application dont la sémantique est absolument indissociable tout en sachant la complexité, voire l'impossibilité à implanter ces technologies sur des corpus de documents webs multilingues dans le cadre d'un processus de veille.

Tous les choix technologiques présentés précédemment ont un impact : un impact direct sur les possibilités de la solution, sur son mode d'implémentation en entreprise, sur les compétences nécessaires aux utilisateurs, sur le montant de la facture annuelle mais également et surtout sur la qualité du logiciel de veille et sur ce qui reste essentiel dans un dispositif de veille automatisé : la qualité de l'information collectée.

Frédéric Martinet, Actulligence

Frédéric Martinet – Actulligence Consulting – Consultant indépendant dispositifs de veille
Frédéric exerce désormais depuis une dizaine d'années dans le secteur de la veille stratégique et de l'intelligence économique. Consultant indépendant spécialisé dans la mise en place de systèmes de veille, il a travaillé pour de grandes entreprises françaises sur les phases de conception et de déploiement de leurs dispositifs de veille. Frédéric tient l'un des blogs francophones les plus lus sur la veille et l'intelligence économique : www.actuligence.com . Il intervient par ailleurs dans des troisièmes cycles spécialisés en intelligence économique et est Maître de conférences associé de l'IUT de Montluçon (Université Blaise Pascal).