LES SPECIFICITES DU WEB : UN OBSTACLE A SON EXPLOITATION ?

Peggy Cadel, Eric Boutin

LES SPECIFICITES DU WEB : UN OBSTACLE A SON EXPLOITATION ?

Peggy Cadel,
Enseignant chercheur IUT de Saint Raphaël
cadel@univ-tln.fr +33 4 94 19 66 00

Eric Boutin,
Maître de conférences en Sciences de l'information – communication IUT de Toulon
boutin@univ-tln.fr + 33 4 94 14 23 56

Adresse professionnelle
Laboratoire I3M
Université de Toulon-Var ★ BP 132 ★ F-83957 La Garde Cedex

Résumé

Avec le développement d'Internet et des connexions à haut débit, les modes d'accès à l'information se sont automatisés. La documentation numérique via le réseau a si bien supplanté les autres supports d'information qu'aujourd'hui lorsque l'on a besoin d'un renseignement, le réflexe premier est la connexion au Web. Ce réflexe déjà acquis par les professionnels de la recherche d'information et de la veille intervient désormais dans le cadre de stratégies de surveillance automatisées.

Or, parallèlement à cette systématisation, l'accès automatique au Web est devenu de plus en plus complexe à cause du caractère hétérogène des éléments qui le composent.

La mixité des formats, les contraintes liées aux modes d'accès ainsi que la multitude de pratiques rédactionnelles sont autant d'obstacles à son traitement.

Ses principales exploitations que sont l'acquisition de documents, la surveillance stratégique et technologique et l'analyse des tendances souffrent de manière inégale.

Il convient aujourd'hui de connaître les limites des outils que l'on achète ainsi que les caractéristiques des sources que l'on exploite afin de mettre en place une surveillance Web efficace.

Mots clés : recherche d'information, surveillance Web, contraintes d'accès, Internet.

Introduction

La surveillance Web professionnelle orchestrée par les cellules de veille des entreprises a pour objectif de s'informer des nouvelles tendances ou des nouvelles avancées survenues dans l'environnement de l'entreprise en étant à l'écoute des signaux forts comme des signaux faibles.

Cette surveillance fondée sur une recherche d'information empirique sur Internet se heurte à la difficulté de trouver une information précise, fiable et validée, au temps passé à la recherche et au sentiment naturel face à une source aussi immense de ne pas être exhaustif.

Dans le but de résorber le caractère aléatoire et non exhaustif d'une telle recherche, les cellules de veille tentent de structurer et de systématiser leur surveillance en s'équipant d'outils.

Dans un premier temps, nous allons décrire comment les spécificités du Web entravent l'automatisation de l'accès à son contenu textuel. Dans un second temps, nous montrerons comment ces spécificités interviennent sur les différentes exploitations de la source Web.

Les spécificités de la recherche d'information sur Internet

Les outils spécialisés capables de répondre à des problématiques de surveillance automatique rencontrent un certain nombre de contraintes directement liées à la spécificité du Web tant au niveau de l'hétérogénéité du contenu qu'à celui de la terminologie et des pratiques rédactionnelles.

L'hétérogénéité du contenu Web soulève trois difficultés majeures :

- L'accessibilité au contenu
- La représentation du contenu Web
- Le filtrage du contenu Web

L'accessibilité au contenu

De nombreux documents sont accessibles via des formulaires d'interrogation et stockés dans des bases de données. Pour y accéder automatiquement, il est ainsi nécessaire de développer des modules spécifiques à chaque outil d'interrogation afin de simuler le lancement d'une requête manuelle. Il est ensuite indispensable d'étudier le mode de présentation des résultats pour en extraire les documents à récupérer. Les documents ainsi acquis ne sont alors exploitables que si les problématiques de fin de session sont gérées. La spécificité et l'évolution permanente de ces modules demandent aux outils de surveillance des capacités d'apprentissage qu'il est difficile d'automatiser. A ces contraintes s'ajoutent les problématiques techniques liées aux nouveaux accès sécurisés avec mot de passe à reproduire et des contraintes juridiques de propriété intellectuelle.

Le Web n'est pas uniquement composé de pages html statiques. Le php ainsi que le java script sont des langages de programmation Web très répandus qui, de part leur aspect dynamique, peuvent entraver la lecture et la compréhension de leur contenu par un robot.

Aujourd'hui, bien que les moteurs génériques de recherche semblent gérer de manière automatique ces phénomènes, ils mentionnent tout de même, lors de leurs conseils, pour un « meilleur référencement » de préférer les pages statiques aux pages dynamiques.

Le caractère dynamique des constructions ainsi que la multiplicité des procédés d'affichage se réalise dans l'exemple ci-dessous. La page 1 est trouvée car elle répond à une surveillance du site www.01net.com sur la problématique des « systèmes de paiement en ligne » or, la lecture de cette page ne répond à première vue pas à la problématique. Si l'on regarde le code source de cette même page, on voit la présence des termes « systèmes de paiement en ligne », si l'on approfondit cet examen, on se rend compte que le titre de l'article pertinent se trouve bien sur la page 1 mais pointe sur la page 2.

Page1 :
http://www.01net.com/outils/PseudoRubV4.php?base=UMac&rub=4368&pseudo=cont

Extrait du code source de la page 1

```
<table width=193 cellspacing=0 cellpadding=0 border=0>
        <tr>
                <td><img src="/img/dot.gif" width=10 height=1></td>
                <td valign=top>
                <table width=155 cellspacing=0 cellpadding=0 border=0>
                        <tr>
                                <td valign=top colspan=3><a href="http://www.01net.com/conso/"><font
face="Arial" size=1 color=#CC0000 style="font-size: 11px; text-
decoration:none;"><b><u>Conso</u></b></a><img src="/img/dot.gif" width=5 height=1><br></font>
                                <img src="/img/v4/flechefine_fondtransp.gif"><a
href="/article/256787.html"><font face="Arial" size=1 color=#000000 style="font-size: 11px; text-
decoration:none;">Les nouveaux systèmes de paiement en ligne</font></a></td>
                        </tr>
        </tr>
</table>
```

Page 2 : http://www.01net.com/conso/

Cet exemple montre la complexité de la recherche d'information textuelle dans des pages Web, complexité difficilement gérable lorsque l'information textuelle que l'utilisateur perçoit comme telle, se trouve techniquement constituée par une image ou

par un élément informatif non standardisé ce qui est le cas pour la datation d'un document. Il est ainsi très complexe de connaître la date d'un document Web.

La vérification de la mise à jour des pages se fonde soit sur la comparaison du texte, soit sur la date de dépôt de la page. La première information est très difficile à extraire automatiquement car elle n'est pas clairement identifiable au sein d'une balise spécifique, la seconde n'est pas toujours exprimée.
L'accès à la date d'une page est un enjeu important car cet élément informatif peut à lui seul juger de la pertinence d'un document.

La représentation du contenu

De part sa nature, le document Web est hétérogène et multimédia car composé de texte, d'image, de son et de vidéo.

Malgré cette diversité, son mode d'accès privilégié reste le texte ce qui suppose que ces documents ont été indexés à partir des données textuelles présentes. L'opération d'indexation est le processus qui permet de représenter sous forme d'index le contenu d'un document. Cette opération implique soit, l'intervention d'un spécialiste capable de représenter à partir de termes les idées présentes de manière explicite ou implicite dans ces documents soit, l'intervention d'un automate capable d'extraire automatiquement ces termes avec une précision fortement liée à la méthode employée et à la finesse des connaissances linguistiques utilisées. Les deux stratégies trouvent leurs limites dans le caractère subjectif et le manque d'homogénéité de l'une, et l'absence de conceptualisation et la faible levée d'ambiguïté de l'autre.

A ces défauts rencontrés lors de l'analyse des indexations de texte, il faut également tenir compte des difficultés supplémentaires que l'on rencontre lorsqu'il s'agit de transposer un contenu non textuel en un index textuel. Ce passage implique une description de l'objet non textuel par un spécialiste qui va disposer d'une liste de critères non exhaustifs et par conséquent décrire partiellement un contenu.

Afin de pallier ces problèmes de traduction, des méthodes d'indexation fondées sur les caractéristiques même des objets se développent avec, pour exemple, dans le cadre du traitement de l'image, la mise en œuvre de processus d'analyse fondés sur des comparaisons de vecteurs. Cette approche par similarité permet de rechercher des images proches dans des domaines de spécialité. De la même façon que les tentatives de compréhension du langage ne sont applicables que pour des exploitations spécifiques sur des thématiques spécifiques, les algorithmes de reconnaissance d'images ne peuvent être généralisés et appliqués à tous types d'exploitation pour tous types d'images. La reconnaissance d'images médicales n'utilisera pas les mêmes algorithmes que la reconnaissance d'empreintes par exemple car les éléments distinctifs ne seront pas décelés sur les mêmes critères. En utilisant les traitements adaptés, il est possible d'indexer et de rechercher des images en posant des « requêtes images » mais si l'on doit accéder à des documents composites contenant texte et image, il est nécessaire d'utiliser des codes différents pour chaque type de contenu.

L'accès à ces données est ainsi fragilisé à cause des difficultés de transposition de contenu et des combinaisons de requêtes sur formats différents.

L'analyse de site permet de travailler sur des documents de type bureautique insérés sur des pages Web qui ont une vie propre et des documents au format Web qui constituent un segment de site et qui ont une valeur informationnelle à la fois intrinsèque et

extrinsèque. Aujourd'hui un grand nombre de documents de synthèse sont proposés dans des formats de bureautique avec une préférence pour les formats .doc et .pdf. L'accès à ces documents présuppose l'utilisation de modules de conversion capables d'extraire le contenu textuel mais également de rendre compte de la structure du document qui, contrairement au format html, n'est pas directement repérable grâce à des balises. Ce passage doit conserver la structure originelle du document qui (TOUT)en marquant sa structure véhicule des données informationnelles non négligeables.

Le filtrage du contenu

Le filtrage du contenu fondé sur le repérage d'unités linguistiques est confronté à la diversité des pratiques rédactionnelles. La recherche d'information met en évidence un contraste entre les termes employés par les utilisateurs dans le cadre de leur formulation de requête et les termes susceptibles de « répondre » à cette requête dans les documents. Ce contraste est d'autant plus important sur Internet que les pratiques rédactionnelles sont extrêmement variées du fait de la diversité des auteurs et de l'hétérogénéité des sources.
Comme l'expliquent M-F. Bruandet et J-P. Chevallet, il y a une distinction réelle entre la *pertinence système* et la *pertinence utilisateur*.
L'amélioration de cette pertinence passe avant tout par une adéquation entre les termes utilisés pour questionner un système ou une base de connaissance et les termes réellement présents dans les documents capables de véhiculer les idées présentes dans la requête.
L'utilisation de thesaurus, dans une base de connaissances, permet de réduire les distances lexicales dans un domaine de spécialité en tentant de circonscrire sa terminologie. Sur Internet, la délimitation de la terminologie est plus difficile à envisager car on est en présence de style et de champs d'intervention extrêmement larges.

L'influence des spécificités du Web sur son exploitation

La source Web se voit attribuer 3 modes d'exploitation distincts :
- un mode exploitation lié à l'acquisition de documents
- un mode exploitation lié à la surveillance stratégique et technologique
- un mode exploitation lié à l'analyse des tendances des signaux faibles

L'acquisition de documents

Les outils d'acquisition (grabbeurs ou aspirateurs) sont utilisés pour s'approprier le contenu partiel ou intégral de sites de manière automatique avec ensuite une mise à jour fréquente des documents récupérés. Parallèlement à cette acquisition, des systèmes de catégorisation, classification, indexation et recherche se développent afin de gérer l'accès aux documents ainsi acquis ; l'objectif étant de se constituer un référentiel, une base, un historique ou une mémoire.

L'ensemble de la chaîne est ainsi confronté directement aux spécificités du Web et les conséquences en sont un manque d'exhaustivité lié aux sources non exploitées pour des raisons techniques de la base de connaissances.

La surveillance stratégique et technologique

Une surveillance ciblée nécessite une analyse préalable structurelle des sites stratégiques qui doit permettre de déterminer le niveau d'exhaustivité de la veille afin de s'assurer que l'ensemble des données est accessible automatiquement. Cette analyse doit être par la suite réitérée de manière régulière afin d'examiner les éventuels changements de structure.

Il s'agit du mode de surveillance le plus touché par les spécificités du Web car les sites à analyser sont généralement équipés de moteurs de recherche interne ou créés dans des formats difficilement exploitables automatiquement. Lorsqu'il s'agit de sites commerciaux, leur accès est géré par les processus de référencement et non par la mise à disposition du contenu.

L'analyse de tendances

L'intérêt de la surveillance Web ne se limite plus à l'analyse d'informations scientifiques, commerciales ou concurrentielles. Il s'agit de prendre en compte deux phénomènes difficiles à quantifier, à s'avoir l'image de marque et les rumeurs. Pour ce faire, on étend la surveillance Web standard à « l'écoute » des forums et des blogs.

L'analyse automatique de ces tendances est extrêmement délicate car leur expression diffère de celle d'un Objet ou d'une Action représentés par un ensemble lexical définissable, une nouvelle approche se dessine avec l'examen de ce que l'on va appeler des modalités et qui peuvent être représenter à la fois par des termes, par des locutions et par des marqueurs morphologiques comme le conditionnel. L'ensemble de ces marqueurs est ainsi amené à constituer des « concepts » susceptibles de permettre une reconnaissance automatique de tendance.

Certaines informations sont ainsi difficilement repérables sauf en travaillant sur des signaux très faibles portées par des indices linguistiques capables de véhiculer des notions de [tendance][1], de [développement] ou de [projets] en les couplant à des modalités telles que la [probabilité] ou le [futur].

Si l'on regarde les propositions simples suivantes :

1. La société X a racheté la société Y.
2. La société X rachète la société Y.
3. La société X envisage de racheter la société Y.
4. La société X rachètera la société Y.
5. La société X pourrait racheter la société Y.
6. La société X projète de racheter la société Y.

Si l'on opère une surveillance sur la société Y, que l'on s'interroge sur sa santé financière, l'accès aux 4 propositions suivantes n'apporte pas les mêmes informations Or, elles comportent toutes les 4 le concept de [rachat] et l'objet [société] instancié par le même acteur.

Cet exemple simplifié montre la diversité linguistique avec laquelle un fait concret (un rachat) peut se distinguer d'une probabilité (un hypothétique rachat).

[1] Les notions ont été écrites entre crochets par respect de notation en vigueur lors de représentation de concepts.

Perspectives

Face aux spécificités du Web, les trois principales exploitations de la source se trouvent inégalement affaiblies. L'acquisition manque d'exhaustivité mais permet de cerner les principaux champs d'un domaine. L'analyse des tendances est plus affectée par la difficulté de repérer des signaux faibles que par l'accès aux informations dont les spécificités sont moins contraignantes que celles des sites Web L'exploitation la plus sérieusement affectée est la surveillance stratégique et technologique.
Ce constat établi sur l'accessibilité des données présentes sur Internet et sur leurs exploitations mériterait d'être complété par la perception qu'à l'utilisateur de l'information qu'il reçoit. L'analyse des biais cognitifs qui entrent en jeu lorsque l'internaute doit gérer cette information massive pose alors les questions suivantes : pourquoi privilégie-t-il une information qui confirme un jugement antérieur sur une information allant à l'encontre d'un jugement ? pourquoi a-t-il du mal à regarder une même information sous différentes perspectives ? pourquoi ne perçoit-il pas les changements graduels et incrémentaux ?

Bibliographie

BACHIMONT B., 2003, L'indexation multimédia in GAUSSIER E. et STEFANINI M-H, 2003, Assistance intelligente à la recherche d'informations (Traité STI), p153-184

BOUTIN E., «L'exploitation de l'âge d'une page web : quelles perspectives pour l'analyse cybermétrique », acte du colloque, VSST 2004, tome B,p.439-449, Octobre 2004

BOUTIN E. «Qualifier la présence d'une ville sur le web par des indicateurs cybermétriques dynamiques : une expérimentation sur 10 villes françaises» Actes du colloque TIC et territoire, quels développements ?, Lille, Mai 2004

BRANCIER C., 2004, Décision Micro, 2004, Veille sur le Net : des besoins variés
Http://www.01net.com/article/240376.html

BRUANDET M-F, CHEVALET J-P, 2003, Utilisation et construction de bases de connaissances pour la recherche d'information in GAUSSIER E. et STEFANINI M-H, 2003, Assistance intelligente à la recherche d'informations (Traité STI), p99-132.

MEINGAN D., LEBO I., 2004, Livre blanc : "Maîtriser la veille pour préparer l'intelligence économique"
http://www.knowledgeconsult.com/fr/article.php3?id_article=37

SAMIER H., SANDOVAL V.,1999, La recherche intelligente sur l'Internet et l'intranet, 2ème édition revue et augmentée, Hermes, Paris.