APPLICATION POUR RAFFINER UN CORPUS ISSU D'UN AGENT DE SURVEILLANCE DE PAGES WEB – VERS UN SUPPORT LOGICIEL MODULAIRE DU PROCESSUS DE VEILLE

Allan ZIMMERMANN, Xavier DELECROIX, Cyrille DUBOIS, Serge QUAZZOTTI

allan.zimmermann@tudor.lu, xavier.delecroix@tudor.lu, cyrille.dubois@tudor.lu, serge.quazzotti@tudor.lu

CRP Henri Tudor, 29 avenue John F. Kennedy L-1855 Luxembourg-Kirchberg (Luxembourg)

Mots-clefs :
Veille sur Internet, agent de surveillance, terminologie, pertinence, système de veille modulaire, coloration lexicale, lecture rapide

Keywords :
Internet observation, Internet monitoring, monitoring agent, terminology, relevance, modular software system, lexical coloring, speed reading

Palabras clave :
Viligancia en Internet, agente de vigilancia, terminología, pertinencia, sistema modular de software, coloración léxica, lectura rápida

Résumé

Internet est devenu une source d'information privilégiée qui met à disposition des entreprises de nombreuses sources d'informations qui concernent à la fois ses compétences, son management ou sa réputation. Cependant, développer un usage professionnel de l'information issue du web peut vite s'avérer décevant ou contre-productif en raison des gros volumes d'information captés par les divers moyens de recherche sur Internet. Même les agents de surveillance sur Internet, outils communément employés par les professionnels impliqués dans une activité de veille, ne proposent pas encore de moyens de filtrage suffisamment adaptés pour cibler avec précision les informations pertinentes issues de la surveillance de pages web. C'est la raison pour laquelle nous avons mis en place un outil de raffinage des corpus issus d'un agent de surveillance, basé sur une terminologie précise du sujet de veille qui autorise une double sélection, automatique et humaine, des informations détectées. L'outil développé s'inscrit aussi dans une démarche qui consiste à construire des systèmes logiciels modulaires, basés sur des solutions techniques existantes, potentiellement interchangeables, pour réaliser des produits ou des prestations de veille à un coût acceptable.

1. Introduction

Le Centre de Veille Technologique et Normative (CVTN) du Centre de Recherche Public Henri Tudor a pour objectif de soutenir l'intégration de la veille, de la propriété industrielle, de l'information normative et réglementaire dans le processus d'innovation des entreprises par la mise à disposition de produits, de services et de formations spécifiques.

Parmi ses produits, le CVTN offre le suivi de sources d'information Internet (veille Internet).

En effet, Internet est devenu une source d'information incontournable de part l'ensemble des informations qu'il contient. Qu'il s'agisse de pages web d'entreprises, de weblogs, de fils d'actualités…., les informations diffusées sur Internet peuvent présenter un grand intérêt pour une entreprise qui cherche aussi bien à déterminer des tendances qu'à identifier les rumeurs autour d'un sujet sensible pour ses activités.

Les problématiques liées à l'utilisation, et surtout à l'analyse de la pertinence de l'information issue d'Internet se complexifient sans cesse du fait de l'accroissement des sources d'information.

Surveiller des sources issues du web se traduit concrètement par la surveillance automatique de pages web avec des agents de surveillance. Ces derniers permettent de détecter les mises à jour réalisées sur un ensemble de pages web sélectionnées en fonction de la problématique à traiter.

Cependant, les moyens de filtrage par mots-clés des agents de surveillance disponibles au CVTN ne sont pas ou peu adaptés pour incorporer tous les mots-clés d'une terminologie multilingue. De plus, malgré ces moyens de filtrage par mots-clés, les mises à jour détectées contiennent encore une forte proportion de bruit, rendant indispensable un filtrage complémentaire, effectué par un agent humain. En l'absence d'un outillage dédié, cette étape de filtrage complémentaire, que nous appellerons raffinage dans la suite de ce document, peut s'avérer fastidieuse, relativement longue et donc coûteuse. C'est pourquoi un travail de réflexion, de recherche et d'expérimentation est mené au CVTN pour disposer d'un outillage permettant, d'une façon simple et rapide, de sélectionner les informations pertinentes issues du processus de surveillance afin de réaliser une veille sur Internet à un coût acceptable.

Cet article a pour objectif d'exposer les travaux menés au niveau technique pour réaliser une première version d'un outil de raffinage basé sur un traitement de coloration lexicale et de montrer comment ce travail s'inscrit dans une approche modulaire du support logiciel au processus de veille. Le CVTN souhaite en effet privilégier une telle approche dans le but de s'affranchir de toute contrainte logicielle [1] et de concevoir des systèmes de veille basés sur des éléments réutilisables ayant fait leur preuve en phase de production.

Ainsi, nous exposerons dans un premier temps les limites des moyens de filtrage des agents de surveillance, qui aboutissent à la nécessité de construire un module de raffinage. Nous détaillerons ensuite la structure de la terminologie utilisée pour décrire un sujet de veille et le principe de fonctionnement du module de raffinage basé sur cette terminologie. Nous décrirons brièvement le système logiciel au sein duquel nous avons articulé un agent de surveillance avec le module de raffinage. Nous ferons ensuite le bilan de son usage dans le cadre d'une veille technico-légale et discuterons enfin des perspectives d'amélioration du système existant, notamment pour effectuer des traitements d'analyse et de visualisation sur les corpus pertinents ainsi construits.

2. Limites des moyens de filtrage des agents de surveillance

Un agent de surveillance a pour principal objectif de détecter les mises à jour au sein d'un grand nombre de pages web (ou éventuellement de documents dans d'autres formats, PDF p.ex.). Il permet en outre de déclencher des alertes (son, courrier électronique) lorsque des mises à jour sont détectées sur un ensemble de pages surveillées.

Les mises à jour détectées par un agent de surveillance sont généralement composées de texte et/ou d'hyperliens. Nous appellerons « documents liés » les documents qu'on peut atteindre à partir d'une mise à jour en suivant les hyperliens qu'elle contient (voir figure 1).

Le processus de veille Internet dont il est question dans cet article a pour objectif de diffuser des bulletins d'alertes hebdomadaires et des rapports webographiques mensuels ciblés sur un sujet de veille précis. Un bulletin d'alerte contient l'ensemble des mises à jours détectées sur les pages web surveillées, réduites aux seules parties qui contiennent des informations pertinentes par rapport au sujet de veille. Un rapport webographiques contient la liste des références webographiques des documents pertinents liés aux mises à jour contenues dans les bulletins hebdomadaires.

Les agents de surveillance disponibles au CVTN permettent de détecter des mots-clés dans les pages web surveillées en associant des listes de mots-clés et/ou d'expressions régulières à une ou plusieurs pages surveillées. Ce mécanisme permet de filtrer les mises à jours sur mots-clés, c'est-à-dire d'écarter automatiquement les mises à jour qui ne contiennent pas de mot-clé. Les agents de surveillance disponibles au CVTN permettent donc d'enregistrer les mots-clés qui décrivent un sujet de veille pour ne retenir que les mises à jour susceptibles de contenir des informations pertinentes.

De plus, les agents de surveillance intègrent un navigateur qui permet notamment de consulter une version locale des pages web mises à jour où les éléments nouveaux et les mots-clés sont mis en surbrillance. Ce traitement de coloration lexicale des mots-clés permet de repérer plus rapidement dans les pages web mises à jour les informations susceptibles d'être pertinentes car contenant ces mots-clés.

Cependant, la saisie des mots-clés sous forme de listes n'est pas adaptée pour gérer une terminologie structurée et relativement importante de mots-clés et/ou d'expressions régulières. Par exemple, les mots-clés ne peuvent pas être regroupés par catégories ou encore reliés de façon explicite à leurs différentes traductions. Or nous pensons qu'une terminologie relativement précise est indispensable pour, d'une part, garantir une certaine exhaustivité des informations détectées au sein du processus de veille, et pour, d'autre part, tirer le meilleur parti d'un traitement de coloration lexicale destiné à visualiser le plus rapidement possible les informations susceptibles d'être pertinentes.

De plus, les filtres par mots-clés utilisés dans les agents de surveillance n'empêchent pas les mises à jour détectées de contenir une forte proportion de bruit. Par exemple, il est fréquent de disposer de mises à jour qui contiennent une liste d'extraits de documents dont seulement une partie contient des mots-clés (voir figure 1). Parfois les extraits sans mot-clé peuvent être très majoritaires par rapport aux extraits avec mots-clés.

Pour atteindre les objectifs de veille mentionnés précédemment pour un coût acceptable, il nous est donc apparu indispensable de développer un moyen de cibler de façon précise les informations pertinentes contenues dans les mises à jour produites par un agent de surveillance. Ce moyen, basé sur une terminologie décrivant le sujet de veille, doit aussi nous permettre d'éliminer le bruit contenu dans les mises à jour et d'extraire les références webographiques

des documents pertinents liés à ces mises à jour.

Dans le paragraphe suivant nous allons détailler les principes de construction d'une terminologie descriptive d'un sujet de veille.

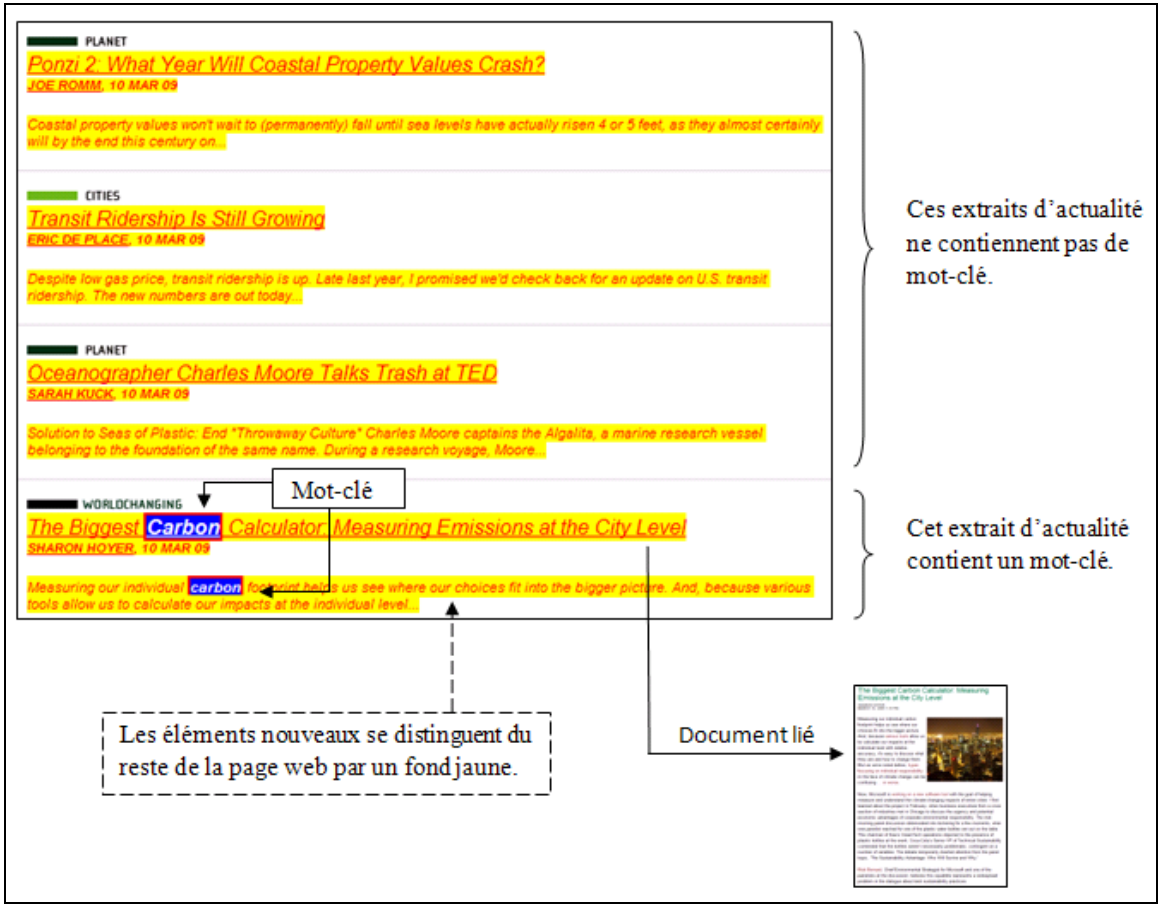

Figure 1 : extrait d'une page web mise à jour (source : www.worldchanging.com)

3. Principes de construction d'une terminologie à 3 niveaux

A partir d'outils simples et éprouvés, comme un système de gestion de base de données et un mécanisme d'expressions régulières, nous avons cherché à construire une terminologie multilingue qui soit précise tout en restant simple à réaliser.

Pour atteindre ce double objectif de précision et de simplicité, nous avons conçu un modèle de terminologie composé de 3 niveaux : les thématiques, les mots-clés dans la langue de référence et les expressions régulières (voir tableau 1).

3.1 Niveau 1 : les thématiques

L'utilisation de thématiques permet de structurer le sujet de veille en regroupant les termes jugés équivalents, c'est-à-dire les termes qui ont un sens proche les uns des autres (exemple : des synonymes) ou les termes qui se réfèrent, directement ou indirectement, à une même réalité (exemple : le nom d'un groupe de travail sur une norme et la désignation de la norme elle-même).

Deux types de thématiques coexistent :

- Les thématiques de recherche.
- Les thématiques secondaires.

Une thématique de recherche regroupe les mots-clés (mots-clés de recherche) qui décrivent les principaux éléments du sujet de veille. Les thématiques de recherche permettent au veilleur de disposer d'une vision synthétique d'un sujet de veille. Une telle vision est particulièrement utile lorsque le veilleur évolue entre plusieurs sujets de veille différents, dont il n'est pas spécialiste, et facilite aussi la transmission de l'exécution de la veille à un collaborateur tiers. En outre, ces thématiques font l'objet d'un traitement de coloration lexicale permettant d'identifier tout ou partie du sujet de veille dans un texte.

Une thématique secondaire regroupe des mots-clés (mots-clés secondaires) qui décrivent les éléments connexes au sujet de veille ou des type d'information particulier (exemple : il est possible de définir une thématique secondaire « Evénement » pour désigner les événements, foires, salons...). Les thématiques secondaires ont pour objectif de faciliter la lecture des mises à jour via un traitement de coloration lexicale. Elles peuvent notamment conduire à détecter de nouveaux mots-clés. Par exemple, dans le cadre de la veille technico-légale réalisée, la thématique « Evènement » peut servir à détecter, dans les mises à jour, les noms des évènements susceptibles de fournir des documents ciblés sur le sujet de veille. Ces noms peuvent alors être intégrés à la terminologie en tant que mots-clés de recherche.

3.2 Niveau 2 : les mots-clés dans la langue de référence

Un mot-clé dans la langue de référence est une forme grammaticale particulière (exemple : un substantif) écrite dans la langue de référence pour la veille (la langue maternelle du veilleur ou la langue de travail du client).

L'ensemble des mots-clés représente donc tous les mots que le veilleur déduit de la demande de son client, de sa culture générale et de terminologies spécialisées existantes pour décrire le plus précisément possible le sujet de la veille (et éventuellement des thématiques secondaires), sans se soucier à ce niveau de leurs traductions ou de leurs variantes grammaticales. En revanche, c'est à ce niveau que le veilleur doit trouver le plus de termes jugés équivalents pour décrire le sujet de veille afin de garantir une certaine exhaustivité.

3.3 Niveau 3 : les expressions régulières

Au moins une expression régulière est utilisée pour reconnaître un mot-clé ou la racine d'un mot-clé dans la langue de référence et éventuellement dans une ou plusieurs des autres langues cibles de la veille. Il est possible d'utiliser une seule expression régulière pour définir à la fois le verbe, l'adjectif, les participes, le pluriel et le féminin en français ou en anglais. Les éventuelles fautes d'orthographe peuvent aussi être modélisées à l'aide des expressions régulières. De plus, étant donné les nombreuses racines latines communes au français et à l'anglais et les nombreuses racines germaniques communes à l'allemand et à l'anglais, il est assez fréquent de pouvoir utiliser une expression régulière pour désigner une ou plusieurs variantes grammaticales d'un mot-clé dans au moins 2 de ces langues, parfois dans les 3 à la fois.

Les expressions régulières constituent le moyen technique pour effectuer une reconnaissance des variantes grammaticales (principe de lemmatisation) des mots-clés dans une des langues cibles du sujet de veille.

Nous parlerons respectivement de mots-clés de recherche et de mots-clés secondaires pour qualifier les mots-clés reconnus par des expressions régulières associées respectivement à des thématiques de recherche et à des thématiques secondaires.

Tableau 1: extrait d'une terminologie

Thématiques	Mots-clés	Expressions régulières	Langues
Ecologie	Ecologie	[éeö][ck]olog	FR, EN, DE
	Développement durable	dévell?opp?ement durable	FR
		sustainable development	EN
		nachhaltige Entwicklung	DE
Carburant	Carburant	carburant	FR
	Essence	essence	FR
		gasol[ei]n	EN
		benzin	DE

4. Le module de raffinage

Le module de raffinage (voir figure 2) a pour objectif d'identifier des documents pertinents à partir des mises à jour produites par un agent de surveillance.

Ce module se décompose en 4 phases :

- Phase 1 : sélection automatique des mises à jour produites par un agent de surveillance.
- Phase 2 : sélection visuelle des informations pertinentes dans les mises à jour sélectionnées lors de la phase précédente.
- Phase 3 : identification des documents pertinents associés aux mises à jour sélectionnées lors de la phase précédente.
- Phase 4 : traitement des mises à jour sans mot-clé de recherche.

4.1 Phase 1 : Sélection automatique des mises à jour

Une mise à jour est sélectionnée si elle contient au moins un mot-clé de recherche reconnu par une expression régulière définie dans la terminologie. Cette sélection conduit à définir un sous-corpus C1 des mises à jour avec mots-clés de recherche et un sous-corpus C2 des mises à jour sans mot-clé de recherche. Cette sélection automatique est rendue indispensable par le choix de s'affranchir des filtres par mots-clés ou expressions régulières disponibles au niveau des agents de surveillance, dans la mesure où ceux-ci ne permettent pas de gérer aisément une terminologie relativement importante et multilingue.

4.2 Phase 2 : Sélection visuelle des informations pertinentes

Les mises à jour contenues dans C1 font ensuite l'objet d'un traitement de coloration lexicale inspiré de celui disponible dans les agents de surveillance utilisés au CVTN. Un code couleur unique est associé à chaque thématique et permet de mettre en surbrillance les mots-clés qui lui sont associés via des expressions régulières. Il est donc possible, pour un agent humain, non seulement de détecter rapidement les informations qui contiennent des mots-clés au sein des mises à jours mais aussi d'évaluer rapidement si ces informations sont pertinentes dans la mesure où il devient possible de distinguer les différentes thématiques de la veille parmi les mots-clés mis en surbrillance (voir figure 3).

Au cours de cette phase, les mises à jour font aussi l'objet d'un double traitement de réduction, automatique et manuel, afin de les intégrer dans un bulletin d'alerte. Ce traitement de réduction consiste à supprimer d'une mise à jour toutes les informations sans mot-clé et effectivement sans rapport avec le sujet de veille.

4.3 Phase 3 : Identification des documents pertinents

Le module de raffinage intègre un navigateur Internet qui permet de localiser, par navigation hypertexte, les documents pertinents qu'il est possible d'atteindre à partir des informations sélectionnées en phase 2. Le traitement de coloration lexicale peut-être rappelé sur les pages web qui contiennent les documents liés pour faciliter l'évaluation de leur pertinence. Les documents de types autres que des pages web (PDF, DOC, RTF, XLS ...) ne sont pas concernés par ce traitement de coloration lexicale.

Les documents jugés pertinents sont alors identifiés et documentés par leurs références webographiques. Ces références correspondent à l'ensemble de métadonnées suivant : {Nom de la source, date de collecte, titre du document, format électronique du document, type du document}. Cet ensemble de métadonnées, simple et concis, a pour but de permettre à un client d'évaluer rapidement l'importance relative d'un document par rapport à un autre.

Les métadonnées des documents identifiés sont alors en partie saisies automatiquement à partir des métadonnées des documents électroniques et/ou de leurs URL. Le titre d'un document contenu dans une page web peut par exemple être déduit de la balise <TITLE> du code HTML de la page web. Cependant, une saisie manuelle complémentaire est souvent nécessaire pour corriger les erreurs ou les manquements de la saisie automatique.

4.4 Phase 4 : Traitement des mises à jour sans mot-clé de recherche

Les mises à jour contenues dans le corpus complémentaire C2 sont rassemblées dans un document unique et font l'objet d'une lecture rapide, après coloration lexicale, pour identifier d'éventuels nouveaux mots-clés de recherche. Dans ce cas, seuls les mots-clés secondaires sont mis en évidence dans les mises à jour de C2. Lorsqu'un nouveau mot-clé est détecté, une expression régulière est ajoutée à la terminologie pour assurer sa reconnaissance. Lorsque la lecture du corpus complémentaire est achevée, le processus de raffinage est alors à nouveau exécuté pour prendre en compte la ou les mises à jour qui contiennent le ou les nouveaux mots-clés détectés.

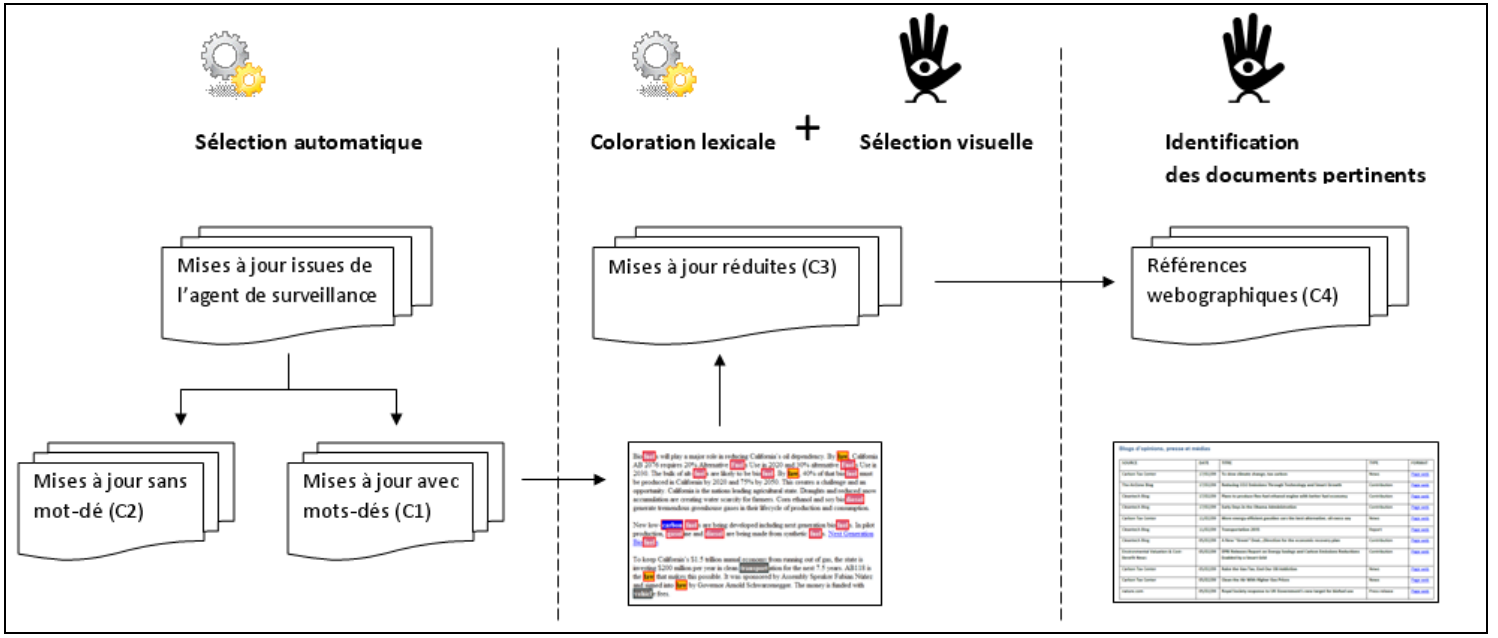

Figure 2: Aperçu du module de raffinage

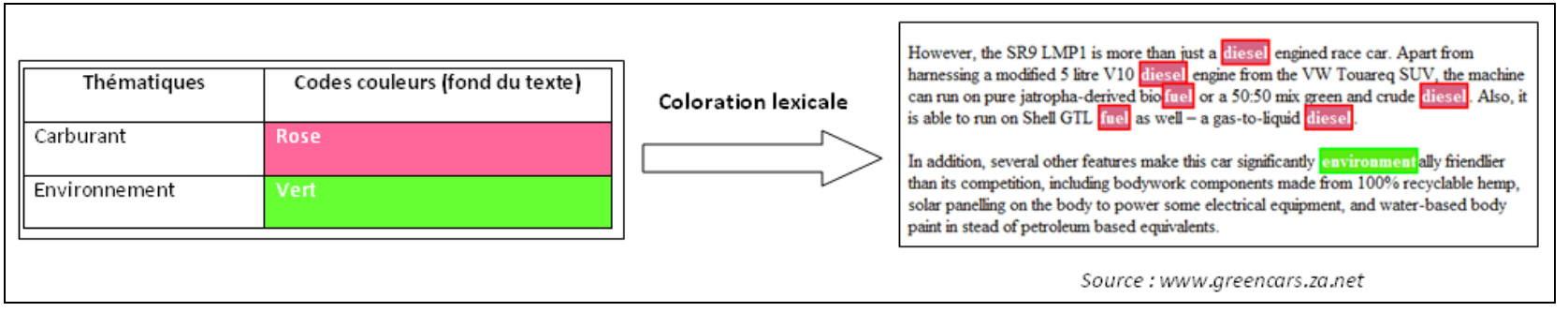

Figure 3 : exemple de coloration lexicale

5. Un système logiciel spécifique

Nous avons mis en place un système logiciel spécifique pour supporter un processus de veille en 2 phases qui produit 3 types de livrables (voir figure 4) :

- Un rapport de mise en place qui comporte notamment une cartographie des sources surveillées et la terminologie employée.
- Un bulletin d'alerte hebdomadaire.
- Un rapport mensuel de références webographiques.

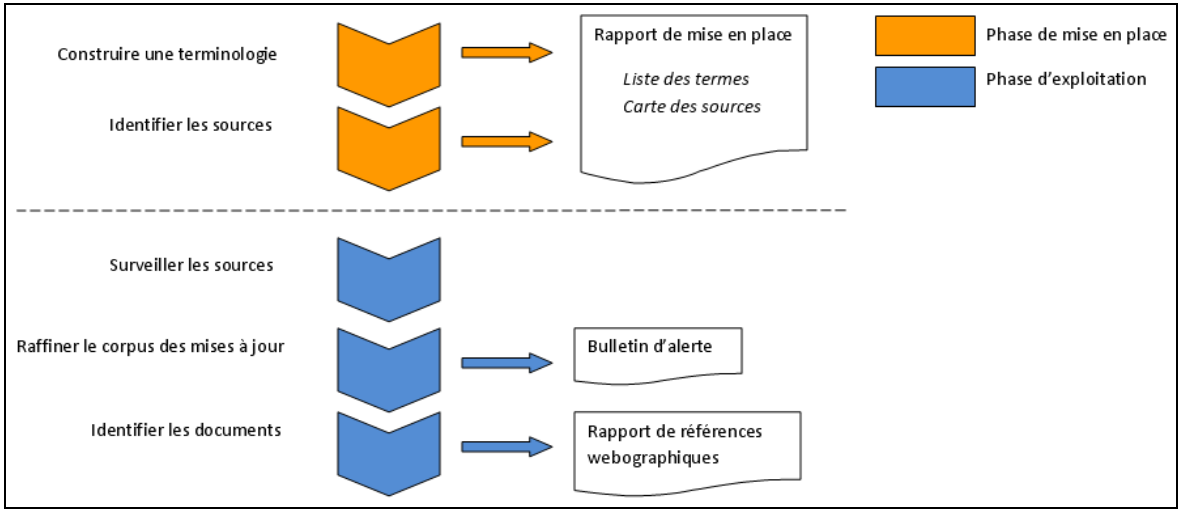

Figure 4 : processus de veille et livrables associés

Ce système fait intervenir 4 logiciels distincts (voir figure 5) :

- Un agent de surveillance.
- Un outil de cartographie conceptuelle.
- Un système de gestion de bases de données (SGBD).
- Un traitement de texte utilisé pour construire un bulletin de références webographiques.

Comme nous l'avons déjà mentionné, l'agent de surveillance permet d'enregistrer les sources surveillées, des pages web, pour détecter les éventuelles mises à jour sur ces pages. L'agent dispose d'une fonction d'exportation qui permet de décharger la plupart de ses données, en particulier les données qui permettent d'identifier une source ainsi que le texte des mises à jour détectées sur ces sources.
L'outil de cartographie conceptuelle est utilisé pour proposer au client une carte des sources de la veille classées selon la hiérarchie utilisée dans l'agent de

surveillance. La carte conceptuelle propose au destinataire de la veille une vision synthétique et relativement lisible de l'étendue des sources surveillées (couverture linguistique, types de sources). La carte est essentiellement réalisée en important dans l'outil de cartographie un fichier au format XML obtenu par reformatage du fichier XML des sources exporté de l'agent de surveillance.

Une procédure intégrée au SGBD permet d'importer les mises à jour contenues dans les fichiers XML de mises à jour exportés de l'agent de surveillance. La base de données intègre une terminologie pour décrire le sujet de veille. Un module de raffinage et un module de création automatique de bulletins d'alertes au format HTML ont été implémentés directement dans le SGBD utilisé. Le module de raffinage est muni d'une interface graphique qui permet de réaliser les opérations manuelles et/ou visuelles liées à la sélection des informations pertinentes dans les mises à jour et à l'identification des documents pertinents liés. Les références webographiques des documents pertinents sont enregistrées dans la base de données.

Le traitement de texte permet de mettre en page les références webographiques contenues dans la base de données via une fonction de publipostage.

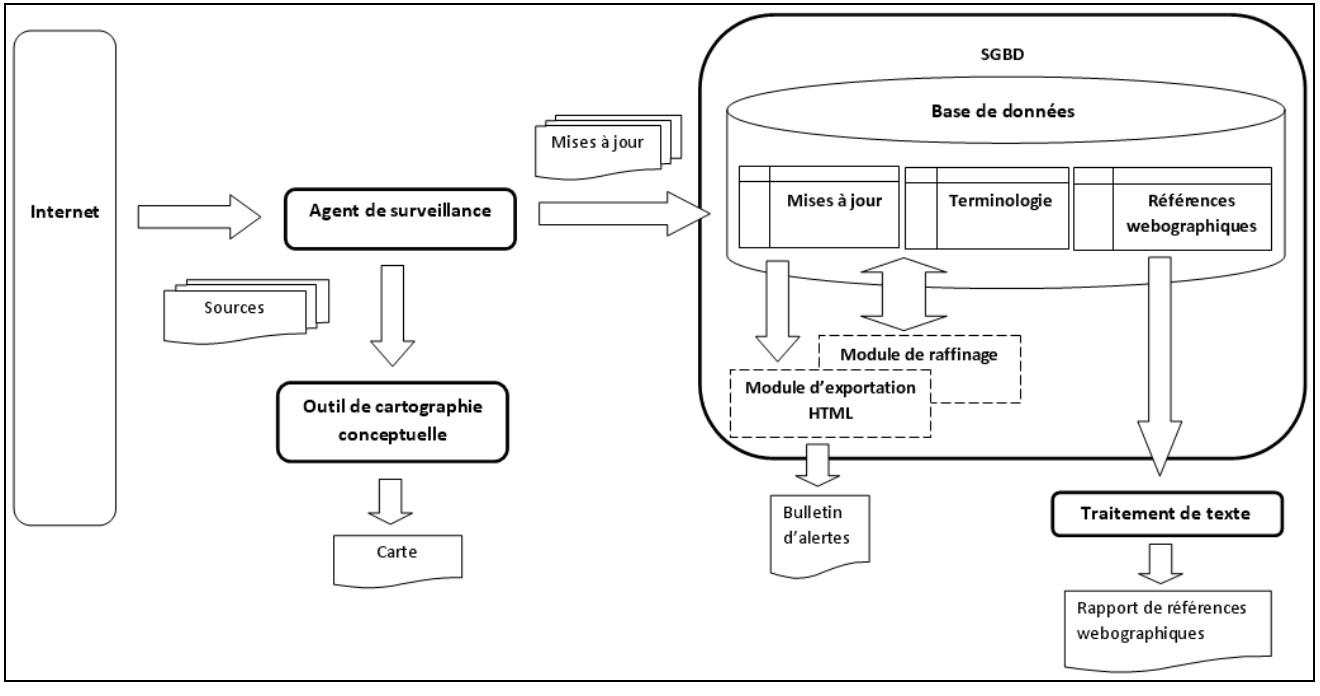

Figure 5 : système composé de 4 logiciels

6. Expérimentation et limites de l'application

L'outil a été initialement développé et utilisé avec succès pour réaliser des bulletins de références webographiques dans le cadre d'une veille technico-légale.

Cette veille peut être caractérisée par les dimensions suivantes:

Tableau 2 : les dimensions de la veille

Terminologie	○ 3 Langues cibles : anglais, français, allemand. ○ 9 thématiques. ○ 63 mots-clés (niveau 2). ○ 112 expressions régulières.
Fréquence des livrables	○ Bulletins d'alertes hebdomadaires. ○ Bulletins mensuels de références webographiques.
Sources	○ 400 pages web surveillées.
Mises à jour hebdomadaires	○ 203 mises à jour collectées chaque semaine en moyenne : ▪ 125 en moyenne sont automatiquement sélectionnées et font l'objet d'une sélection visuelle. ▪ 30 en moyenne sont sélectionnées et réduites.
Documents pertinents	○ 70 documents en moyenne font l'objet d'une sélection visuelle chaque semaine. ○ 39 documents en moyenne sont sélectionnés chaque semaine pour figurer dans le rapport mensuel.

L'outil permet de réaliser la veille pour une durée moyenne de l'ordre de 14 heures/mois dont 9,5 heures sont principalement consacrées à la sélection visuelle des mises à jour et des documents.

Ce temps de réalisation reste cependant trop élevé et impose de futures améliorations de l'application de raffinage (voir § 7.1 Réduction du temps de construction d'un corpus de veille pertinent).

7. Perspectives de développements futurs

Au-delà de bulletins de veille réalisés à partir d'Internet, qui doivent atteindre un objectif de pertinence maximal pour un coût acceptable, le CVTN souhaite rationaliser la réalisation de produits à plus forte valeur ajoutée basée sur des traitements d'analyse complémentaire dont les résultats soient relativement aisés à interpréter par ses utilisateurs finaux.

Ainsi, 2 axes de travail sont envisagés pour atteindre cet objectif.

7.1 Axe 1 : Réduction du temps de construction d'un corpus de veille pertinent

Cet axe de travail a pour but de perfectionner l'application de raffinage décrite dans cet article de 2 façons :

- En filtrant directement des documents plutôt que des mises à jour.
- En mettant en place un système de raffinage par séquence de requêtes exclusives afin de permettre de sélectionner automatiquement une partie des documents filtrés, sans avoir recours à une sélection visuelle complémentaire.

Pour filtrer directement les documents, nous chercherons à exploiter les travaux et/ou les techniques existant en matière d'extraction de données en ligne.

Interroger un document avec une séquence de requêtes exclusives (voir figure 6) consiste à définir un jeu de requêtes sur la terminologie en les classant de la plus précise à la moins précise. Cet ordre définit une séquence d'exécution où chaque requête est appliquée au sous-corpus rejeté par la précédente de telle sorte à garantir qu'aucun document ne fasse l'objet de plus d'une sélection, que cette sélection soit automatique ou humaine. Ce système doit aussi permettre de définir un seuil dans la séquence d'exécution (au moins la première étape) à partir duquel et en-deçà duquel tous les sous-corpus sélectionnés automatiquement peuvent être sélectionnés sans intervention humaine. C'est la précision des requêtes, d'une part, et la faible voire l'absence de polysémie des mots-clés combinés entre eux, d'autre part, qui seront garants de la pertinence des documents sélectionnés automatiquement.

Au-delà du seuil de précision, les documents devront être sélectionnés visuellement en ayant recours au traitement de coloration lexicale présenté précédemment. D'autres moyens de lecture rapide ou de sélection par lots seront éventuellement envisagés en complément de la coloration lexicale comme par exemple des outils de résumé automatique pour réduire le volume de lecture.

Nous envisagerons enfin de rationaliser le traitement du corpus des mises à jour rejetées par toutes les requêtes (corpus complémentaire) pour identifier des documents pertinents qui contiendraient de nouveaux termes, lesquels seraient alors incorporés à la terminologie de recherche.

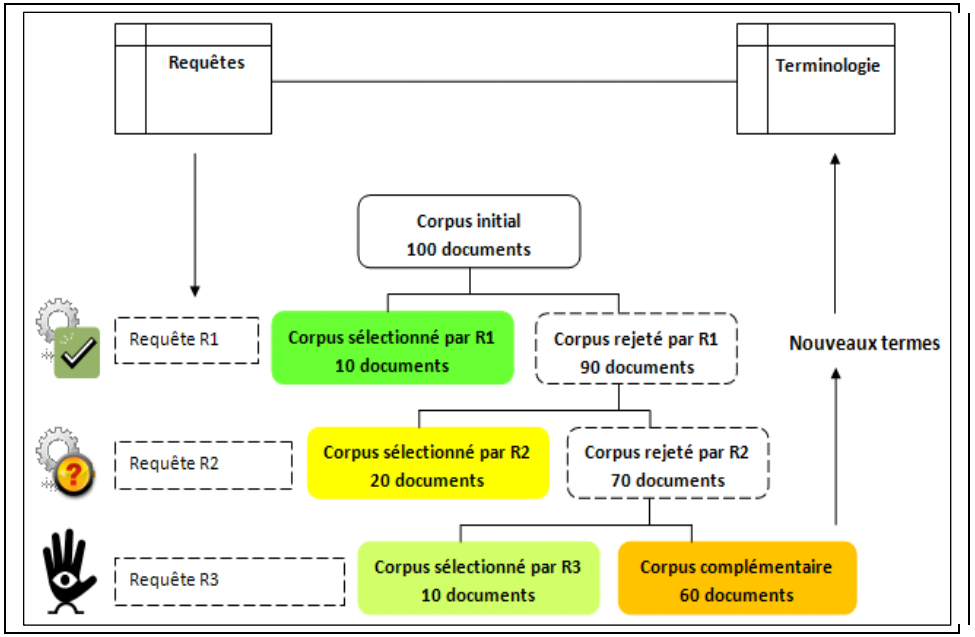

Figure 6 : exemple de raffinage par séquence de requêtes exclusives

Ce travail sera aussi l'occasion de porter une attention particulière aux ontologies [2] pour discuter des éventuelles améliorations que ces outils pourraient apporter au module de raffinage, notamment pour modéliser une terminologie et pour l'enrichir semi-automatiquement à partir d'ontologies existantes.

7.2 Axe 2 : Une approche modulaire pour intégrer de nouveaux outils

Cet axe de travail s'inscrit dans une démarche déjà initiée dans le métier pour construire un système de veille basé sur des combinaisons de modules logiciels distincts (système multi-agents) [3] qui puissent évoluer en fonction des besoins de ses utilisateurs.

Cet axe de travail consiste d'une part à tirer parti de la possibilité d'obtenir des corpus pertinents en provenance d'Internet et d'autre part à tirer parti de l'offre logicielle actuellement disponible, payante ou gratuite, pour articuler ces différentes solutions entre elles afin de proposer des produits ou des prestations de veille à un coût acceptable et prévisible.

Nous envisageons notamment d'intégrer dans le système présenté précédemment des outils d'analyse de type text-mining qui nous permettraient de mettre à jour des relations inédites ou peu accessibles à travers les documents d'un corpus.

Nous porterons aussi une attention particulière aux outils de visualisation pour améliorer la compréhension des résultats d'une analyse et pour transmettre à nos clients des livrables dont le contenu soit visuellement attractif, notamment par recours à des graphiques ou des métaphores visuelles de l'information.

Nous chercherons également à intégrer des éléments provenant de bases de données en ligne dans la perspective de comparer les vues synthétiques qu'on peut obtenir de corpus de documents différents (articles scientifiques, normes, brevets, articles de presse...), ciblés sur un sujet identique.

Cependant, pour être intégré dans un système de veille, un outil de traitement de l'information, quelle que soit sa fonction, doit être en mesure de recevoir et/ou de produire des informations dans un format directement ou indirectement exploitable par un autre outil. Cette exigence pose le problème de la compatibilité des outils de traitement de l'information.

Pour répondre à cette exigence de compatibilité, nous rechercherons ou nous développerons les opérations de reformatage qui s'avéreront nécessaires pour transmettre des informations d'un outil à un autre.

Ce faisant, nous construirons un référentiel d'outils compatibles pour le traitement d'information. Ce référentiel documentera en priorité les capacités d'échange de données de chaque outil. Celui-ci documentera également la ou les fonctions de chaque outil au sein du processus de veille et fournira une évaluation de son coût d'exécution en fonction de paramètres de référence. Ce référentiel permettra alors d'améliorer le système existant par l'ajout de nouveaux outils ou l'usage de fonctions non exploitées d'un logiciel. Adapter le processus de veille sur Internet décrit dans ce document pourrait notamment se révéler pertinent pour construire des documents renouvelables de type état de l'art ou business plan.

Dans le cadre de cet axe de travail, nous chercherons également à tirer parti des travaux réalisés dans le domaine de la « Business Intelligence » [4] dont les solutions techniques, initialement conçues pour traiter des données de gestion internes à l'entreprise, sont désormais aussi tournées vers l'exploitation de données issues d'Internet.

8. Bibliographie

[1] **PERBAL S., DUBOIS C., SCHOSSELER P.,** *Exemple de mise en œuvre modulaire d'un processus de veille*, VSST'2004

[2] **HERNANDEZ N., MOTHE J.,** *TtoO: Mining a thesaurus and texts to build and update a domain ontology, Data Mining with Ontologies: Implementations, Findings, and Frameworks*, Idea Group Inc., 2007

[3] **DENIS X., SIMON G.,** *Utilisation collaborative d'outils de text mining pour la veille sur Internet*, VSST'2004

[4] **BAUMGARTNER R., FRÖLICH O., GOTTLOB G.,** *Web Data Extraction for Business Intelligence : the Lixto Approach*, BTW 2005