

AGENTS INTELLIGENTS INTEGRES A L'APPRENTISSAGE AUTOMATIQUE POUR LA VEILLE STRATEGIQUE: VIGIALE

ORTIZ MONTENEGRO, IVETTE, Dra. Ing.
Directora General IALE Tecnología, www.iale.cl
iortiz@iale.cl
Chili

ROJAS MATURANA, VÍCTOR Ing.
IALE Tecnología, www.iale.cl
vrojas@iale.cl
Chili

ESCORSA O'CALLAGHAN, ENRIC.
IALE Tecnología, www.iale.es
enric@iale.es
Espagne

CHAUR BERNAL JAIRO, Dr.
IALE Tecnología, www.iale.es
chaur@iale.es
Colombie

Mots Clés:

Outils open source, crawler, apprentissage automatique, veille technologique, Vigiale.

Résumé:

La conception d'outils facilitant la récupération, la synthèse ainsi que la valorisation de l'information, de manière rapide et adapté au cycle de VT, représente de nos jours un des principaux défis pour les professionnels de la Veille.

La Veille Technologique et l'Intelligence Economique représentent le processus dynamique de surveillance et d'analyse stratégique des progrès scientifico-technologiques, de l'environnement concurrentiel, commercial, environnemental, réglementaire ; visant à améliorer les prises de décisions, réduire l'incertitude dans les marchés et identifier de nouvelles opportunités dans la R + D + i (Recherche, développement et innovations).

IALE Tecnología présente dans ce document les résultats obtenus à partir d'un programme de développement technologique ayant l'objectif de créer un outil répondant aux critères énoncés ci dessous. Le programme de développement inclut notamment la création d'une interface à l'architecture adaptée et personnalisée facilitant l'interaction de l'utilisateur avec l'information. L'interface permet à l'utilisateur d'accéder à l'information classifiée ceci a travers la génération de *tags* adaptés. Elle permet également l'accès au référentiel (serveur de stockage interne), en utilisant les stratégies de recherche des bases de données modernes. Vigiale, qui est le nom de la plateforme, est aussi un réseau social où les utilisateurs peuvent échanger des avis et des documents dans un secteur spécifique. L'interaction entre les utilisateurs est possible grâce à des forums de discussion (par le biais de commentaires) et un système de notation met en évidence les commentaires ou les documents les plus pertinents.

1. Introduction

Les outils de gestion de l'information et de connaissances subissent des changements remarquables. Dans le monde d'Internet, par exemple, des sociétés comme Google, Microsoft, Yahoo! développent continuellement de nouvelles applications qui visent à faciliter l'accès à l'information. Il apparait également de nouveaux logiciels pour le traitement de l'information capturée, composée le plus souvent de centaines ou de milliers de d'enregistrements de bases de données, de publications scientifiques ou bien de brevets, d'une aire déterminée. Il apparait également de plus en plus de logiciel de Text Mining [11], capables de « lire », capturer, classifier et grouper les informations non structurées, c'est-à-dire pas nécessairement disponibles dans le format des bases de données (avec balisage et prêtes à être traitées).

Toutes ces tendances actuelles viennent répondre à l'inquiétude croissante liée la manière de transformer de simples données en information efficace et comment convertir ces données en connaissances dans le but d'améliorer les processus stratégiques internes des organismes (entreprises, instituts ...) notamment en matière de prise de décision.

Être attentif à tous les signaux de l'environnement qui affectent ou ont la possibilité d'influencer le développement futur d'une organisation, que ce soit d'un point de vue concurrentiel, technologique, commercial, environnemental ou politique, nous permettra d'améliorer les processus de prise de décision, et de les réaliser dans un temps plus opportun avec toute l'objectivité et la rationalité possible. Ceci aura pour effet de réduire l'incertitude et les risques associés et de mieux utiliser les possibilités offertes par le contexte.

Les modèles traditionnels de surveillance de l'environnement tels que la participation à des salons professionnels, la lecture de revues spécialisés, l'étude des produits concurrents ainsi que les apport en information des relations clients-fournisseurs, bien qu'ayant une réelle valeur, sont insuffisants pour répondre aux changements d'un environnement technologique. En effet, la situation n'est jamais figée mais au contraire en perpétuelle évolution.

La conception d'un système de Veille et d'intelligence stratégique en réponse aux besoins constant en informations efficaces de la part des entreprises, dans un délais court et le moins coûteux possible, est un pari stratégique ambitieux mais qui peut faire face à l'évolution des marchés et des concurrents. De ce fait, la veille technologique est définie comme le processus de suivi et d'analyse stratégique des progrès scientifiques et technologiques, et comme assurant la détection de nouvelles possibilités émergentes en science et technologie. Les résultats sont généralement utilisés en recherche et développement (R & D) et permettent de prendre les bonnes décisions stratégiques.

Aujourd'hui, le concept d'intelligence économique est en train de remplacer celui de la veille en générale, car il consiste à présenter les informations de manière plus élaborée et associées entre elles ce qui permet une meilleure vision et donc une meilleure prise de décision. De plus, il inclut les résultats de veille d'autres milieux (technologique, financier, concurrentiel, etc.) ce qui donne une vision plus large et plus ouverte du secteur étudié. On peut alors dire que l'intelligence économique est la deuxième génération de Veille technologique. La différence fondamentale entre les deux concepts est la dynamique. En effet la recherche active d'informations pertinentes sur les activités choisies, de manière à fournir une connaissance permanente des évolutions et des tendances émergentes, est essentiellement ce qui distingue l'intelligence économique et stratégique de la simple veille.

Le suivi à long terme est la meilleure façon de connaître le mieux son secteur à n'importe quel moment. Il est donc important d'être clair sur le fait que la veille technologique moderne (ou veille concurrentielle) implique le développement de cycles de veille.

2. Le cycle de la veille

Bon nombre des outils de veille qui se développent et qui sont actuellement commercialisés, manquent d'une vision globale, et sont limitées à la collecte et la catégorisation des informations.

Un outil destiné au processus de veille technologique devrait permettre de capturer, classifier et valoriser l'information ainsi que de la communiquer aux les utilisateurs.

Cela devrait permettre d'associer la connaissance globale avec la prise de décision (technologique, commerciale, stratégique...).

Selon la norme UNE-166 006 [7], les processus initiaux impliqués dans le suivi de la technologie sont, comme indiqué dans la figure 1: identification, recherche, traitement et validation, communication des résultats pour prise de décision (valeur de départ.)

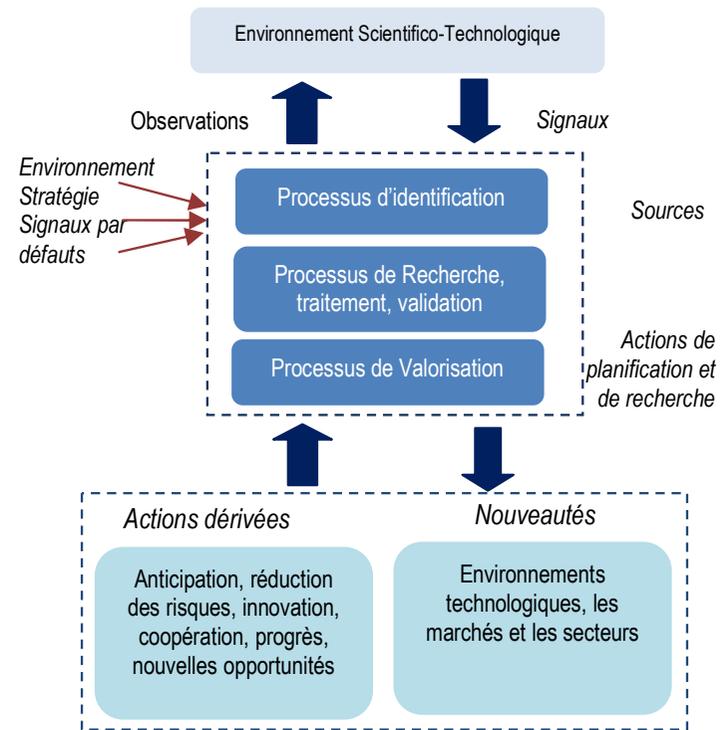


Figure 1. Processus impliqués dans la VT selon la norme UNE 166006:2006

2.1 Identification

L'entreprise a besoin d'identifier bon nombre de choses relatives à son environnement et pour cela il faut qu'elle trouve les bonnes informations. Cette étape consiste à:

- Identifier quels sont les besoins en information.
- Identifier les sources d'informations, internes et externes
- Identifier les moyens d'accès à ces sources

2.2 Recherche, traitement et validation

Catégorisation de l'information qui permet de la filtrer et de l'homogénéiser en suivant une approche stratégique liée à l'activité de l'entreprise. De cette façon, l'information peut être évaluée en termes de pertinence pour la société et donc gardée.

2.3 Processus de valorisation de l'information

Une fois l'information retenue, on l'analyse afin de lui attribuer sa valeur stratégique selon les critères de l'entreprise (selon ce que l'entreprise recherche). On analyse les données obtenues afin de remarquer et d'identifier les signaux de changement de l'environnement, à venir ou déjà opérés, qui pourraient avoir un impact sur l'activité de l'entreprise. C'est évidemment cette identification qui orientera la prise de décision finale.

IALE Tecnología est une entreprise pionnière dans le domaine de la Veille technologique et de l'Intelligence économique, avec des contributions importantes d'un point de vue théorique et pratique, qui s'est attelée à développer un outil informatique pour faciliter la réalisation de ces cycles de suivi, en créant la plateforme VIGIALE.

3. Evolution des technologies associées à la Veille

La conception d'une étude de VT commence avec l'utilisation de méthodes scientométrique, appliquées aux bases de données structurées de publications scientifiques et de brevets. Ces bases sont consultées grâce à des mots clés identifiés préalablement par les experts du domaine concerné, et les résultats sont ensuite analysés à l'aide d'outils de data mining. De cette manière on peut obtenir les cartes géostratégiques qui mettent en évidence les relations qu'entretiennent les différents acteurs ou les évolutions du domaine étudié :

- Evolution du nombre de publications scientifiques sur une période donnée.
- Lignes d'investigation technologique émergentes.
- Pays présent dans les secteurs d'investigations détectés.
- Centres ou Instituts acteurs du marché.
- Cluster ou groupes dans l'aire d'intérêt.

Cependant, l'information d'intérêt ne se retrouve pas nécessairement toujours dans les bases de données structurées (bases scientifiques ou technologiques), mais peut se situer dans ce que l'on appelle « le web invisible » (*deep web*) [9]. C'est de cela qu'est né le besoin de compter sur de nouvelles technologies permettant d'accéder à cette information tant importante pour la compétitivité de l'entreprise.

En accord avec ce besoin, nous avons développé un système informatique capable de récupérer l'information non structurée du Web. Ceci consiste en le téléchargement périodique des informations disponibles sous forme de documents (html, pdf, word,...) depuis des adresses spécifiques. Les données contenues dans ces documents sont classées et indexées automatiquement dans la plateforme ce qui permet de réaliser des recherches sur l'ensemble des données stockées à partir d'un unique serveur. De plus, les documents stockés peuvent être comparés avec les versions initiales en ligne ce qui permet, au travers d'algorithmes de différenciation, de détecter et reporter les différences rencontrées.

Un moteur de Veille Stratégique se doit de nos jours, d'aller plus loin qu'identifier les informations du domaine scientifique et technologique. Par aller plus loin s'entend ouvrir encore plus son champ de vision en s'intéressant notamment aux informations de type légales. Il s'agit alors d'actualiser et alerter l'utilisateur de nouveaux documents juridiques concernant son secteur industriel. Ainsi on alerte l'utilisateur des divers changements juridiques opérés dans n'importe quel pays et repérés par la plateforme.

La nécessité d'être toujours mieux informé englobe les notifications. C'est en effet un excellent complément aux sources déjà analysées. C'est pour cela qu'il est fréquent d'incorporer la syndication des contenus Web pour la lecture permanente de notifications avec des filtres thématiques. Ceci se produit lorsque le contenu d'une page Web est en lien avec d'autres sites ou souscripteurs individuels. Les programmes informatiques compatibles avec chacun de ces standards, consultent périodiquement une page web possédant les titres de plusieurs documents qui les relient avec les articles complets présents dans le site original.

Les brevets peuvent également être surveillés au même titre que les autres documents. En utilisant un client http, comme il est « recommandé » et en complétant les champs de recherche dans les bases de données brevets automatiquement, il est possible d'extraire et alerter de n'importe quelle nouvelle apparition grâce aux canaux thématiques préalablement définis par mots clés ou par taxonomie.

Dans ce cas, la syndication RSS peut également être utilisée pour obtenir une actualisation des sites de brevets à partir de certains systèmes d'agrégations de contenus, sans avoir besoin de consulter le site de manière répétée.

Ce type de technologie représente un formidable outil pour les activités de veilles. Il permet de surveiller en permanence les informations de type normative, brevets et d'une manière générale la plupart des sources Web, en détectant les changements même minimes. Les sources Web non-structurées peuvent être publiques ou requérir d'une authentification. Dans les deux cas cela nécessite des procédés de récupération adéquats.

4. Un outil pour assurer la Veille du web 2.0

VIGIALE a été créée pour la gestion de la veille technologique. Elle est conçue selon les critères de la deuxième génération Web (Web 2.0) qui permet le suivi des sources sélectionnées et la notification des changements détectés.

Elle permet la gestion intégrée des diverses sources d'information, de trier, de classer et mettre à jour les données, en utilisant des technologies modernes de capture, de catégorisation, d'indexation et de filtrage de toutes sortes. Ces fonctions sont personnalisables selon les besoins et les exigences de chaque entreprise ou organisme.

En ce sens, VIGIALE constitue un point d'accès à des documents (moteur de recherche vertical), des actualités, des événements, des projets et des informations d'intérêts pour l'entreprise et fournit également des mises à jour de surveillances actives des sources en ligne, des mises en garde sur les changements dans la politique, des marchés et des brevets par syndication RSS.

Elle permet de gérer les commandes personnalisées de surveillance d'études, facilitant la communication entre l'équipe assurant la veille et les utilisateurs. On pourrait dire que c'est dans ce sens, un outil de gestion des connaissances.

4.1 Structure fonctionnelle de VIGIALE

La figure 2 montre la structure fonctionnelle de VIGIALE. La plate-forme dispose d'un *référentiel*¹ de suivi (serveur interne de stockage), qui stocke l'information réglementaire, des brevets, l'actualité, etc. En outre, il y a aussi des dépôts de rapports de veille technologique effectués par des experts en Intelligence économique et veille concurrentielle sur des thèmes variés.

Le système dispose d'une interface Web et bien qu'elle ait une structure initiale définie, elle peut être personnalisée selon les préférences de l'entreprise utilisatrice (un classement personnalisé des documents est possible).

¹ Référentiel est le terme utilisé pour désigner le serveur central qui stocke et gère les informations au format numérique.

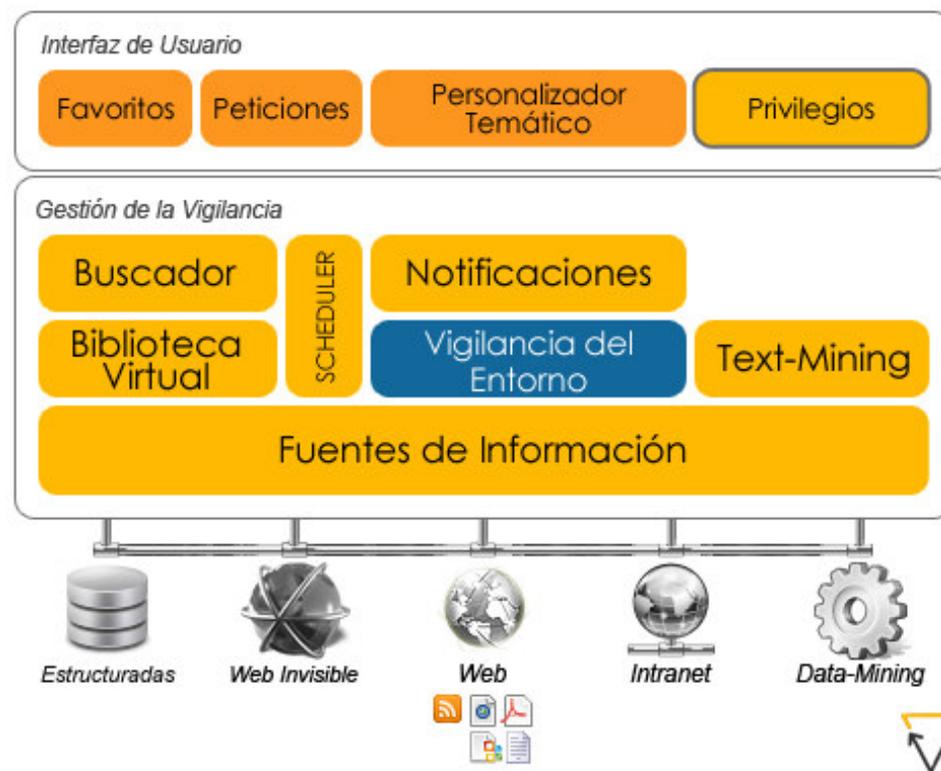


Figure 2. Structure fonctionnelle de la plateforme Vigiale

La structure fonctionnelle ainsi définie, comporte trois types d'utilisateurs: *les administrateurs*, chargés de l'agrégation d'information non-automatisable, de la programmation de la veille automatique, entre autres ; *les experts* sectoriels ou thématiques qui peuvent donner des conseils ou élaborer des rapports en matière de Veille concurrentielle (en option) ; et *les utilisateurs abonnés* capables de consulter, échanger et naviguer sur la plate-forme.

Chaque utilisateur abonné (autorisé par la société) a des exigences et des attentes différentes de par la nature du secteur qu'il veut étudier. C'est pourquoi VIGIALE permet une personnalisation de manière à obtenir l'information adaptée aux attentes particulières, à travers la mise en œuvre de filtres configurables pour chacune d'elles.

VIGIALE prend également en compte des concepts avancés comme le développement de réseaux sociaux professionnels, en créant des espaces de collaboration et de discussion de thèmes ou de sujets d'intérêts communs entre les utilisateurs de la plateforme, et la possibilité pour eux de participer au classement de l'information à travers l'utilisation de tags (étiquettes, mots clés) générées dynamiquement par la communauté utilisatrice.

4.2 L'interface

La figure 3 montre l'interface² telle que l'utilisateur la voit lorsqu'il utilise la plateforme après s'être identifié.

Dans l'exemple pris ici, la plateforme correspondrait à une entreprise (hypothétique) positionnée dans le secteur alimentaire, dont l'une de ses spécialités (zone de commercialisation) est la viande de mouton. Par conséquent, ceci apparaît sur le canal gauche identifié avec le mot «mouton» (Notons qu'un utilisateur de la plateforme peut avoir accès à autant de canaux qu'il le désire, c'est l'abonnement qui change). La structure du canal se partage en plusieurs aires thématiques et d'étiquettes (type tags), qui sont

regroupées dans le centre de l'interface sous le titre de « Nube de Etiquetas » (Production, traçabilité, Omega-3, Consommateur, Halal, etc.).

Sur la partie gauche on trouve un espace appelé « Lo último... », qui répertorie les dernières actualités recueillies par la plateforme, c'est un flux d'informations générales (actualités, articles, brevets, normes...) à partir de sources qui sont contrôlées et identifiées quotidiennement. C'est en cliquant sur chaque titre qu'on obtient le texte intégral.

Sur la partie centrale de l'interface apparaît un moteur de recherche qui permet d'avoir accès aux informations stockées par la plateforme. Cette barre de recherche vous permet d'effectuer des recherches générales ou au contraire spécifiques (titre, description, mots clés, autres) avec toutes les fonctionnalités associées à un moteur de recherche classique c'est-à-dire la possibilité d'utiliser des troncations (truncation searching), opérateurs booléens, guillemets, etc.



Figura 3. Interface de la plateforme Vigiale

² Dans cet exemple, on montre l'interface "générique" de la plate-forme. Comme indiqué plus haut, la plateforme est personnalisable en fonction des préférences des clients.

Sur le coté droit ce trouve l'espace «Actualité scientifique et technique», qui comprend plusieurs catégories (brevets, revues, etc.) qui fonctionnent grâce à des filtres. En cliquant sur l'une des catégories, comme par exemple « revues » on obtient les publications issues des journaux spécialisés dans l'aire thématique qu'on a choisie au départ. Tous les documents répertoriés dans VIGIALE, passent par un traducteur (lecteur automatique) qui sélectionne/défini 4 mots clés en espagnol, même si le document original est en anglais, en français, en italien, en portugais ou en allemand.

La partie supérieure de l'interface propose divers onglets : *législation, études VT, liens et documents* sur la gauche et *mes préférences, mes demandes et paramètres* sur la droite. Chacune d'entre elle donne accès à la page correspondante de la plateforme.

Mes préférences est une option très importante qui permet à l'utilisateur d'éviter d'être constamment connecté à la plateforme pour consulter les informations que génère la plateforme. Cela peut être intéressant pour certain utilisateur de n'être averti que des évènements ou des actualités les plus important(e)s, défini(e)s préalablement selon ses propres critères.

Une dernière option très utile est la génération automatique sous format PDF de bulletins périodiques regroupant toutes les informations d'intérêt relatives à l'aire thématique choisie (pour chaque utilisateur enregistré). La fréquence de génération de ces bulletins est également configurable (tous les mois par exemple).

Cette option est très importante car elle permet à l'utilisateur, s'il le désire, d'éviter d'être constamment devant son ordinateur à vérifier les informations répertoriées par la plateforme. En fait, cette option a conduit à une version simplifiée de la plate-forme, appelée *Vigiale Newsletter*, qui possède la capacité de produire ces bulletins d'alerte basés sur deux types de ressources: les flux RSS (Actualité scientifique et technologique) et les robots d'indexation (ressources contrôlables, capable de détecter des changements dans le contenu).

En concentrant l'attention de la plate-forme vers le contenu de ces ressources, l'accès sera modifier par rapport à une plateforme Vigiale normale, en fournissant à l'utilisateur un seul point d'entrée (page d'accueil ou Home) pour visualiser le contenu du bulletin.

5. Intégration des agents web et apprentissage automatique

Afin de répondre à la demande croissante en informations actualisées et attenantes aux aires thématique sélectionnées, VIGIALE a intégré un outil pour l'exploration et la collecte de données à partir des ressources web. Ces outils sont populairement connus sous le nom de Crawlers. Ce sont des robots qui parcourent le web pour en connaitre sa structure et éventuellement son contenu.

5.1 Crawler

Un crawler (également connu sous le nom Spider Web, Web Harvester ou robot d'indexation) est un logiciel qui inspecte l'espace Web de manière automatisée et méthodique. Les crawlers ont été créés pour le stockage et la récupération d'une page web à partir d'une source de stockage locale. Cette source locale ou référentiel peut être utilisée à des fins diverses en fonction des objectifs des applications qui font usage de celui-ci, comme par exemple les moteurs de recherche qui sont indispensables et très couramment utilisés de nos jours. Ce processus est connu sous les noms de crawling ou spidering. Actuellement, la plupart des principaux moteurs de recherche tels que Google, Yahoo, utilisent le crawling comme moyen de fournir les informations les plus actualisées/actuelle du Web.

Les crawlers peuvent également être utilisés pour automatiser les mises à jour d'un site web (analyse périodiques des liens hypertextes) ou afin de recueillir certain type d'information à partir des pages web visitées (emails, etc.).

Un crawler commence généralement par visiter une liste d'URLs, appelées « semences ». Pour chaque URL visitée et traitée, le crawler identifie tous les liens hypertextes contenue dans la page et fabrique une nouvelle liste d'URL qui à ce niveau s'appelle « la frontière ». Les URLs qui font parties de la “frontière” son visitées de manière régulière par le crawler selon un système de programmation (fréquence paramétrable).

Les collecteurs Web stockent une copie de toutes les pages visitées et celles ci sont ensuite traitées par un moteur de recherche. Le moteur de recherche à son tour va créer un indice de chaque page téléchargée pour fournir un système de recherche rapide.

Le but de VIGIALE est de proposer/fournir de l'information efficiente issue de divers sources (pages web, documents, sites web avec lecteur RSS, rapport de VT/IE=veille techno/intelligence économique) validés par des experts et en accord avec les axes thématiques choisi par les clients. C'est pour cela qu'il est nécessaire d'ajouter un processus de révision des pages collectées pour identifier celle qui apporteront une réelle contribution pour les utilisateurs de la plateforme et éliminer les autres.

En outre, le processus de crawling ne fournit pas de mécanisme de classification des pages collectées pour que s'organise par ordre de pertinence les informations recueillies. Il en existe néanmoins un sur la plateforme Vigiale. Pour cette raison et compte tenu des nouvelles fonctionnalités, nous avons développé un classificateur automatique des sources d'information appelé Classificateur (catégoriseur).

5.2 Classificateur (catégoriseur)

En parallèle du défi de l'alimentation automatique en informations à travers un crawler, s'ajoute la nécessité de pouvoir classifier automatiquement toutes ces données collectées.

Le volume de données collectées peut être un nombre important de pages web, par conséquent, le développement et l'intégration d'un processus de classification devient une pièce maitresse pour l'efficacité du fonctionnement de la plateforme.

Les informations Meta-définies issue des ressources gérées par Vigiale sont divisées en trois groupes. Le premier et le plus important de tous correspond au *Modèle Conceptuel* que représente la classification de haut niveau d'une ressource. C'est le top niveau de la classification.

Le second groupe correspond aux métadonnées de caractère général, présents dans toutes les ressources et qui servent à définir les données spécifiques pour chacun d'eux. Enfin, le dernier groupe réunit les métadonnées spécifiques selon la nature de la ressource. Exemple: les URLs pour les ressources de type Web (Actualité, législation, liens), ou année de publication pour les ressources de type documents.

Cette classification automatique va principalement aider les travaux de l'utilisateur administrateur quant à l'affectation des éléments des Meta Informations aux nouvelles sources, à savoir, « Sectores » (Canaux d'information), Axe thématique « Ejes Temáticos » (sub-sector/sub- canal) appartenant au model conceptuel ci-dessus.

Les processus intervenant dans la catégorisation automatique prennent leur source dans les systèmes d'Apprentissage Automatique. Sur la plateforme Vigiale, s'exécute une phase d'apprentissage et d'entraînement durant laquelle, sont effectués le calcul des paramètres et des coefficients de classification suivant une sélection de groupes de sources d'informations qui constituent la base d'apprentissage pour le système. Une fois générées, les données de formation en termes de qualité et de quantité, le système est préparé à explorer le Web, collecter des données et les catégoriser en accord avec le model de classification requit (celui qu'a choisit l'utilisateur/administrateur lors du paramétrage de son classement personnalisé).

Comme l'objectif de la plate-forme est d'assurer la qualité de l'information, on ajoute un procédé de validation qui a pour objectif d'attribuer une cote de classement automatique aux informations recueillies. La classification s'établie en fonction de 3 états de qualité de jugement de l'information : recommandée, à revoir, rejetée.

Ce processus est invisible pour les utilisateurs de la plateforme car il se déroule au sein de la séquence informatique du programme de la plateforme. L'utilisateur profite de cette fonctionnalité à travers la recherche, la navigation et la surveillance de l'espace de stockage des données constamment actualisées de plus en plus précisément.

5.3Le Web invisible

Internet invisible est la partie du web accessible en ligne, mais non indexée par des moteurs de recherche classiques généralistes.

Le "Web invisible" donne à l'utilisateur la possibilité d'accéder à une donnée située dans un emplacement web déterminé non accessible directement par les mécanismes de recherches conventionnels. Cela correspond à des sites qui possèdent l'information de manière dynamique mais qui éventuellement peuvent générer différentes barrières de sécurité avant de donner l'accès à cette information. On comprend que ceci gêne le travail de collecte des robots. Afin de passer ces éventuelles barrières, cela consiste à écrire un ou plusieurs termes/mots afin que plus tard, pour ces mots, la plateforme génère une requête adéquate pour les sites « barrés ». S'il y a d'autres connecteurs logiques d'autre nature, comme par exemple les dates, les pays, ceux ci sont également prit en considération pour générer l'équation de recherche.

Compte tenu des caractéristiques de la veille technologique, l'accès n'est pas seulement fait en temps réel par un utilisateur en particulier, mais les requêtes sont générés et permettent d'accéder automatiquement aux sites à tout moment, enregistrer les résultats, et l'affichage d'éventuelles différences entre chacune des entrées, avec l'objectif d'être totalement à l'avant-garde de cette information.

L'accès à certains sites web est effectué grâce à un remplissage automatique des fenêtres de recherche, en sautant les protocoles basiques comme par exemple remplir champs par champs. L'utilisateur ne se soucie que d'entrer des termes liés à la consultation et fait abstraction des actions qui sont étrangères à ses besoins. Afin de récupérer l'information propre, plusieurs étapes sont à passer dont celle des barrières artificielles qui gênent l'accès direct à une donnée, tels que les questions, les requêtes (des champs à remplir par exemple), les paramètres ou autre.

Une fois qu'on contrôle les résultats de ces tâches programmées pour compliquer l'accès à l'information, il devient possible d'émuler la consultation et de la comparer avec les résultats antérieurs et ceci à n'importe quel moment.

Il est possible de gérer des sessions en les maintenant actives. L'objectif est d'optimiser au maximum l'efficacité de consultation en assurant une connectivité permanente afin d'éviter de possibles refus du serveur et d'améliorer le temps de réponse à certaines demandes/requêtes. (C'est ce qui se produit par exemple lorsqu'on veut être tenu au courant des dernières innovations en matière de brevet)

La gestion de session envisage que plusieurs utilisateurs soient connectés au même endroit au même moment. Dans le cas où la visite du site nécessite une authentification (password ou login), l'utilisateur n'est pas au courant que le site la demande et l'accès est transparent. L'accès automatique à un site avec des barrières de sécurité nécessite que le système maintienne les sessions actives durant le temps que prend le processus de collecte d'informations.

Ces données clés de l'accès d'un site dépendent du site en question. Il y a 3 ensembles de données clés telles que :

- Les termes appartenant à la consultation en tant que telle et éventuellement constituant une équation de recherche. Ces termes peuvent être liés par exemple avec le titre, le résumé, la description ou l'auteur, pour n'en nommer que quelques-uns.
- Les données de type "formulaire" préalable à la consultation, par exemple, les antécédents administratifs, les données liées à une entreprise, les mails, enquêtes ou n'importe quels autres types de demandes qui sont en dehors de l'objectif de la recherche.
- Les données liées à l'authentification d'un utilisateur sur un site, tel que le login ou les mots de passe.

Cela permet à l'utilisateur d'effectuer à tout moment une consultation avec les données déjà entrées, et également de permettre au système d'effectuer des recherches automatisées grâce à des tâches planifiées afin de vérifier les différences entre les produits de chacune des consultations avec la base de données de brevet, et ce à n'importe quel moment.

6. Les améliorations prévues

Ces dernières années, de nombreux chercheurs ont conçu de nouveaux modèles dans le but de transformer le Web, le faire passer d'un espace d'information à un espace de connaissance par la standardisation des langues et des outils pour appliquer le Web sémantique (s'affranchir des langues, des password et de toutes les barrière s'opposant à l'exploration de l'information désirée). L'idée est que les données peuvent être utilisées et «comprises» par des machines ne nécessitant pas de surveillance/supervision humaine. Le but recherché est que des Agents Intelligents puissent traiter les informations affichées sur des pages web de manière semi-automatique.

Les agents intelligents trouveront non seulement l'information précise, mais pourront également l'interpréter directement en recherchant l'information en relation avec celle-ci situé quelques pages avant. Pour que cela se réalise, il est nécessaire que le web soit représenté de telle manière qu'il soit lisible par les ordinateurs, consensuel et soit réutilisable.

Prochainement est prévu d'incorporer le traitement sémantique au référentiel/serveur de veille de la plateforme, ce qui représente un défi technologique important (figure 4). Les données stockées seront transformés en information efficiente par un processus de traitement technologique de l'information ce qui générera une connaissance susceptible d'être consultée directement. En d'autres termes, créer un système capable les d'interpréter les requêtes effectuées « par les humains » et y répondre le plus précisément possible en offrant des résultats de hautes valeurs ajouté en connaissance et complémentaire aux critères de recherche. Les mécanismes utilisés dans le traitement et la représentation des connaissances sont des ontologies.

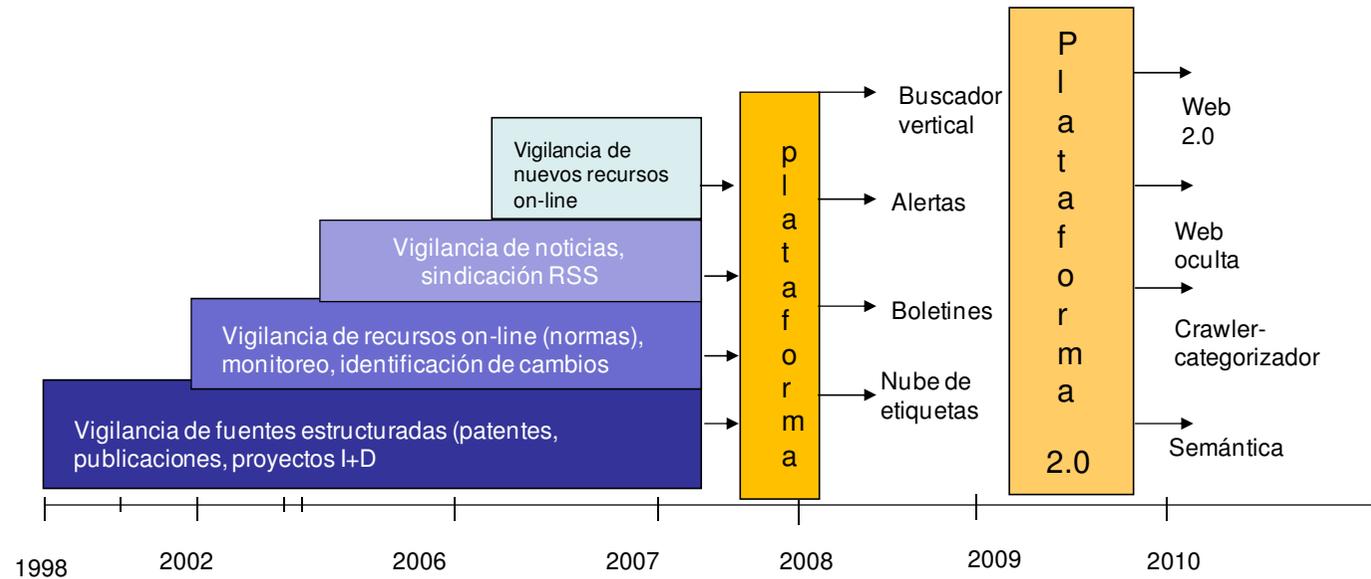


Figure 4. Evolution des innovations appliquées au système informatique Vigiale

Parmi les autres projets d'amélioration de Vigiale, se trouve l'incorporation d'un outil d'exploration de données pour le traitement de l'information structurée à partir de recherches dans des sources telles que les bases de données brevets ou de publications scientifiques.

L'incorporation d'outils de data mining on-line, sur la plateforme constitue une autre évolution prévue pour Vigiale. Le traitement de l'information structurée, retirée des bases de données de publications scientifiques ou de brevets, pourra être fait directement sur la plateforme.

L'objectif principal de cette option est de permettre l'élaboration d'études d'Intelligence Economique par l'utilisateur lui-même car auparavant seules apparaissaient les études issues de la collaboration de professionnels qui possédaient en plus des informations déjà traitées.

L'utilité de cette plateforme de veille web peut se définir comme servant à faciliter la collecte de l'information pertinente à caractère stratégique ; conférant à l'utilisateur une information toujours actualisée grâce à des alertes et des systèmes de diffusions, rendant possible d'identifier à temps les innovations technologiques et les possibles opportunités d'entreprendre.

7. Conclusion

L'avenir de la plate-forme est axé sur l'externalisation et la production de connaissances stratégiques dans des environnements de collaboration en utilisant les nouvelles technologies associées au Web 2.0, par exemple.

Ces instruments visent à favoriser la créativité, le partage d'informations et la collaboration entre les utilisateurs, en mettant en avant les principales caractéristiques telles qu'un contenu dynamique, les métadonnées, les standards du Web et l'évolutivité.

En outre, ces nouvelles lignes technologiques permettent non seulement d'afficher des informations en format texte, mais rendent possible l'intégration des informations dans de multiples formats (vidéo, audio, cartes interactives numériques, etc.) ouvrant la possibilité d'incorporer de nouveaux concepts tels que le Mashup (ou application composite), qui consiste à utiliser les données de diverses sources afin de créer un nouveau contenu. N'oublions pas le Web sémantique conçu pour donner plus de sens au web 2.0 pour que les utilisateurs puissent trouver ce qu'ils cherchent plus précisément, rapidement et sensiblement grâce à une information mieux définie, résolvant de cette manière les problèmes que ramènent la surcharge d'information et l'hétérogénéité des sources d'informations présentes sur la toile.

Un des grands défis est d'apporter des solutions technologiques aux petites et moyennes entreprises, en particulier dans les pays qui comme le Chili, se trouvent dans une période de consolidation de cette discipline appelée Veille Technologique.

8. Bibliographie

- [1] Campaña, F. Diseño y desarrollo de un sistema de vigilancia tecnológica. Proyecto de grado presentado a la Universidad Politécnica de Cataluña, para optar el grado de Ingeniero Industrial, 2005.
- [2] Escorsa, P, Cruz, E, y Ortíz, I. Plataformas de tecnologías asociadas a la Vigilancia Estratégica "VIGIALE". En Congreso Internacional sobre vigilancia e inteligencia sistemática para la innovación en las organizaciones. San Sebastián, Octubre de 2007.
- [3] Escorsa, P, Maspons, R. De la vigilancia tecnológica a la inteligencia competitiva. España: Prentice Hall, 2001.
- [4] IALE Tecnología. Guía de vigilancia tecnológica: sistema de información estratégica en las pymes. Bilbao: DZ Centro de Diseño. Disponible en <http://www.zaintek.net/ebizkaia/servlet/EbizkaiaServlet?P=gLDB>
- [5] Muñoz, J, Marín, M, y Vallejo, J. La vigilancia tecnológica en la gestión de proyectos de I+D+i: recursos y herramientas. El profesional de la Información 2006, 15 (6): 411-419.
- [6] Norma 166000:2006. Gestión de la I+D+i: terminología y definiciones de las actividades de I+D+i. Madrid: Aenor, 2006.
- [7] Norma UNE 166006 EX. Gestión de la I+D+i: sistema de vigilancia tecnológica. Madrid: Aenor, 2006.
- [8] O'Reilly, T. What is Web 2.0. 2005. Disponible en <http://www.oreillynet.com/pub/a/oreilly/tim/news/2005/09/30/what-is-web-20.html> Consultado Mayo 12, 2007.
- [9] Rey, B. Internet invisible. Disponible en <http://internetinvisible.galeon.com> Consultado Junio 2, 2008.
- [10] W3Schools. Semantic web tutorial. Disponible en <http://www.w3schools.com/semweb/default.asp> Consultado Mayo 11, 2007.
- [11] Wikipedia. Text mining. Disponible en http://en.wikipedia.org/wiki/Text_mining Consultado Mayo 11, 2008.
- [12] Palop, F, y Vicente, JM. Vigilancia Tecnológica e Inteligencia Competitiva. Su potencial para la empresa española, Fundación COTEC, Madrid, 1999.
- [13] Escorsa, P, y Martínez, V. La detección del avance de la tecnología mediante mapas. Boletín de

- Estudios Económicos 1994, nro. 152. Universidad de Deusto, Deusto, Bilbao.
- [14] Cruz, E, Escorsa, E, y Ortiz, I. Herramientas y métodos para los ejercicios de Inteligencia Competitiva, Inteligencia Tecnológica y Vigilancia tecnológica. En: Escorsa, P, y Lázaro P. La Inteligencia competitiva: factor clave para la toma de decisiones estratégicas en las organizaciones. Madrid: Madri+d; 2007. p. 158-196.
- [15] Vargas, P, Ortiz, I, y Rojas, V. Vigilancia Tecnológica Aplicada a Nanociencia y Nanoecnología en Países de Latinoamérica. Journal of Technology Management & Innovation 2006, 1(4): 83-94. Disponible en: <http://www.jotmi.org/index.php/GT/article/view/art24/44>
- [16] IALE Tecnología. Identificación de Oportunidades de Negocio Globales en Base a Tecnologías para el Cluster Acuícola y Relacionados. Chile: Programa Bicentenario de Ciencia y Tecnología (PBCT), 2005. Disponible en http://www.pbct.cl/574/articles-11665_pdf_2.pdf
- [17] Gruber T. Toward Principles for the Design of Ontologies Used for Knowledge Sharing. Technical Report KSL-93-04, Knowledge Systems Laboratory, Stanford University, CA, 1993.
- [18] Studer S, Benjamins R, y Fensel D. Knowledge Engineering: Principles and Methods. Data and Knowledge Engineering, 25, 161-197, 1998.