Création de Gateway par un agent intelligent

Dou Carine [1,2], Leitzelman Mylène [3], Mannina Bruno [1], Giraud Eric [1]

[1] CRRM
Centre de Recherche Rétrospective de Marseille / Université Aix-Marseille III
13397 Marseille Cedex 20
Tel : 04-91-28-87-40, Fax : 04-91-28-87-12

[2] Conseil Régional Provence-Alpes-Côte d'Azur
BP 67, 13441 Marseille Cantini Cedex 06

[3] Marseille Innovation
Technopôle de Château-Gombert
13451 Marseille cedex 20

Résumé :
Le World Wide Web s'accroît de manière exponentielle depuis le début des années 90. Internet qui était une ressource incontestable d'information est en train de devenir un tel foisonnement d'informations qu'il devient difficile de trouver l'information recherchée.
Il faut donc parallèlement à la croissance d'Internet, associer des outils de collecte et de captage de l'information de plus en plus efficace.
Pour des recherches d'informations sur des thématiques précises, l'utilisation de gateway parait idéale, car il permet de rassembler tous les serveurs traitant d'un même thème. Mais le problème réside dans la construction et dans la mise à jour de ces gateways.
L'avantage premier de la construction des gateways par un agent intelligent est bien évidemment le gain de temps.
L'utilisation d'agents intelligents pour la création de gateways est indispensable pour les professionnels de l'information. Ces gateways seront de meilleur qualité, nécessiteront très peu de temps à leur confection, et apporteront une valeur-ajoutée sous formes d'analyses complémentaires.

Mots-clés :

Gateway, Internet, Agent Intelligent, collecte, information

Création de Gateway par un agent intelligent

1. Positionnement du problème

Le World Wide Web s'accroît de manière exponentielle depuis le début des années 90. Internet qui était une ressource incontestable d'information est en train de devenir un tel foisonnement d'informations qu'il devient difficile de trouver l'information recherchée.
Il faut donc parallèlement à la croissance d'Internet, associer des outils de collecte et de captage de l'information de plus en plus efficace.

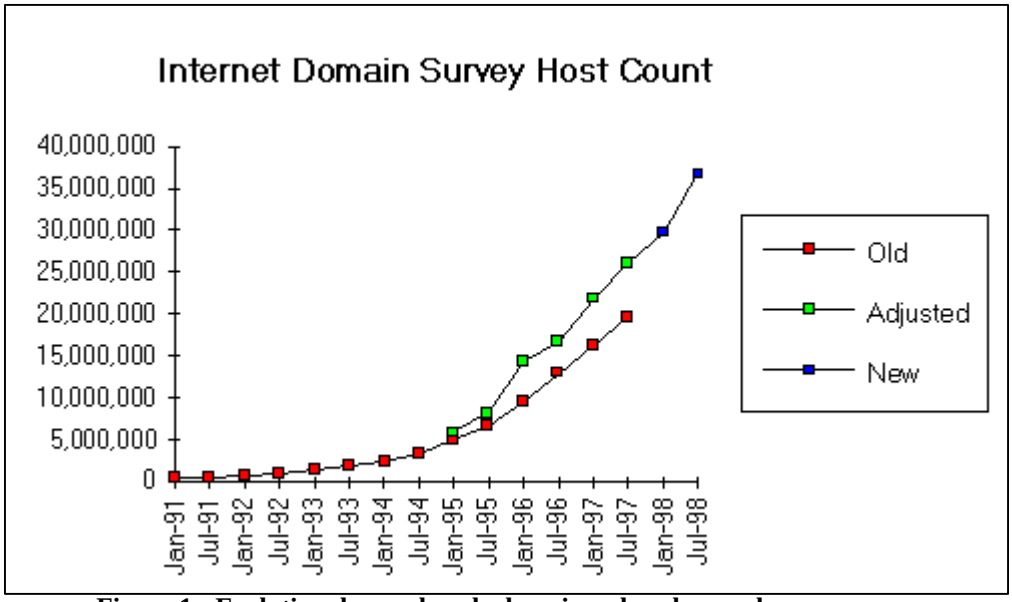

Figure 1 : Evolution du nombre de domaines dans le monde

Pour des recherches d'informations sur des thématiques précises, l'utilisation de gateway parait idéale, car il permet de rassembler tous les serveurs traitant d'un même thème. Mais le problème réside dans la construction et dans la mise à jour de ces gateways.

2. Définition du Gateway

Grâce à la technique de "l'hypertexte" qu'offre Internet, il est possible de rendre un objet dynamique, que ce soit du texte, du son, de l'image ou de la vidéo. Un simple objet avec un hyperlien peut donc conduire à l'apparition d'un nouvel objet, quelque soit sa nature (textuelle, sonore, vidéo ou graphique). De plus, ce lien peut se faire vers un objet local ou un objet distant.

Cette brève description de l'hyperlien laisse paraître à quel point cette technique a révolutionné la sphère de l'information qui notamment a démultiplié ses applications avec l'utilisation d'Internet.

Le Gateway est un recensement dans une page HTML, d'un nombre assez importants de liens hypertextes internes ou externes concernant une thématique bien précise.
Exemple : Gateway sur les Agents Intelligents : Serveur de Bruno Mannina : *http://ms161u06.u-3mrs.fr/bookagent.html*
En fait, un Gateway est équivalent à un "Bookmark" classique sur un seul thème fabriqué par les utilisateurs d'Internet grâce à leur navigateur.

Certains de ces Gateways peuvent contenir des informations ou commentaires sur ces liens hypertextes. Ceux sont les Virtual Libraries.
Exemple : Virtual Library on Information Sciences sur le serveur du CRRM : *http://crrm.univ-mrs.fr/gateway/gateway.html*

Les premières applications sont apparues dans le domaine de la défense, la recherche et l'enseignement supérieur (les premiers domaines à utiliser Internet). Aujourd'hui, aucun site Internet n'échappe à cette ouverture, quelque soit le vocabulaire utilisé pour nommer ces pages : Gateway, Bookmark, Virtual Library, Favorite Links, Signets...

Ainsi définie, il s'agit maintenant de comprendre le rôle d'un Gateway dans un site.
Un serveur qui néglige de proposer un Gateway dans ses pages peut être critiquable à plusieurs titres. D'une part, cela peut être interprété comme le signe d'un manque d'ouverture sur le monde extérieur ou encore comme une mauvaise connaissance de son environnement.
D'autre part, un tel site ignore la richesse de "co-citation" qu'offre les liens hypertextes. En effet, la co-citation sur Internet fonctionne comme celle faite sur les publications, c'est un facteur d'amplification de la notoriété. De plus, les sites contenant des Gateways verront leur fréquentation augmenter par le biais de la navigation des internautes.
Ce type de fichier est très répandu sur Internet, puisque sur Altavista environ 85,000 de bookmarks sont recensés. (Requête altavista : url:bookmark OR url:signet OR (favorite near link) OR url:gateway))

3. Utilisation d'un Agent Intelligent pour créer un Gateway

La création de ce type de fichiers est fastidieuse, car il faut recenser manuellement tous les liens Internet sur un sujet donné.

Cette opération peut durer plusieurs mois. Il faut rechercher les sites via des moteurs de recherche, puis fabriquer les pages HTML avec les liens vers ces serveurs.

De plus, Internet étant en constant renouveau, les liens hypertextes doivent vérifiés et validés, et sans oublier la possibilité d'émergence de nouveaux sites. Il est donc nécessaire de mettre à jour les Gateways, trimestriellement environ, en les reconstituant entièrement.
Cette manipulation étant trop longue à faire manuellement, vue la durée de vie d'un Gateway, l'utilisation d'un Agent Intelligent devient nécessaire.
En effet, celui-ci permet de faire de manière automatisée les pages de liens hypertextes, et donc de gagner énormément de temps.

Les étapes à la création d'un gateway se décompose en plusieurs phases.
Dans un premier temps, il est nécessaire d'établir une liste de mots-clés correspondant à la thématique étudiée. Pour cela, des tests de requêtes sur des moteurs de recherche classiques, de préférence en mode avancé, doivent être effectués (Altavista, HotBot, Lycos, Yahoo ou Ecila).

Une fois, la requête ciblée sur la thématique, il faut configurer l'Agent Intelligent qui va élaborer automatiquement le gateway. Très peu d'outils sont capables de répondre à cette demande. Auresys, agent intelligent élaboré au CRRM, fait partie de cette catégorie. C'est celui-ci qui sera utilisé par la suite.

Le choix d'Altavista

Le principe de fonctionnement d'Auresys se base sur un interfaçage avec Altavista, moteur de recherche le plus complet. En effet, Altavista propose tous les opérateurs booléens standards (and, or, not, parenthèses et troncatures) et de plus, il est le seul à permettre de faire des requêtes avec des opérateurs de proximité (near).
Afin d'être le plus exhaustif possible, l'utilisation d'Altavista est souhaitable, puisqu'il indexe 28% du Web. Ce taux est un des plus élevé avec HotBot (34 %) mais celui-ci a un taux d'erreurs plus important (les liens ne sont pas vérifiés), et de plus, il ne permet pas la troncature.

Principe général de fonctionnement

Auresys procède en deux temps. Tout d'abord, grâce aux informations fournies par l'utilisateur, Auresys simule une requête à Altavista qui, en retour, lui fournit une liste d'adresses pertinentes. Ensuite, Auresys commence la recherche depuis ces adresses, avec différents caractéristiques (profondeur de recherche, sélection de pages...).
Le résultat obtenu formera le Gateway.

Pour optimiser son utilisation, il faut auparavant configurer Auresys selon le type de résultats que l'utilisateur désire obtenir, et cerner les différentes classes de visualisations possibles.

Configuration

La configuration de l'agent intelligent se décompose en deux phases :

① Première phase : Il faut configurer les paramètres à envoyer à Altavista.

- Sélectionner "Le Web" pour la recherche (nécessaire pour fabriquer un Gateway).
- Choisir le langage.
- Formuler la requête.

[Optionnel]

- Classement des résultats (Ranking).
- Période de recherche (Dates inférieure et supérieure).

② Deuxième phase : Configurer les paramètres nécessaires au fonctionnement d'Auresys :

- Renseignements sur la personne désirant utiliser Auresys (protection par Login).
- Nom du Gateway à fabriquer.
- Deuxième série de mots clés permettant à Auresys d'effectuer un filtrage des réponses.
- Profondeur de la recherche (profondeur hypertexte).
- Critère de sélection d'une page sur le nombre de liens dans celle-ci : Eviter de traiter les pages Gateways (Redondance des liens).
- Caractéristique de parcours de la recherche : Auresys utilisera ou non, les liens hypertextes des pages ne répondant pas à la requête pour effectuer sa recherche.

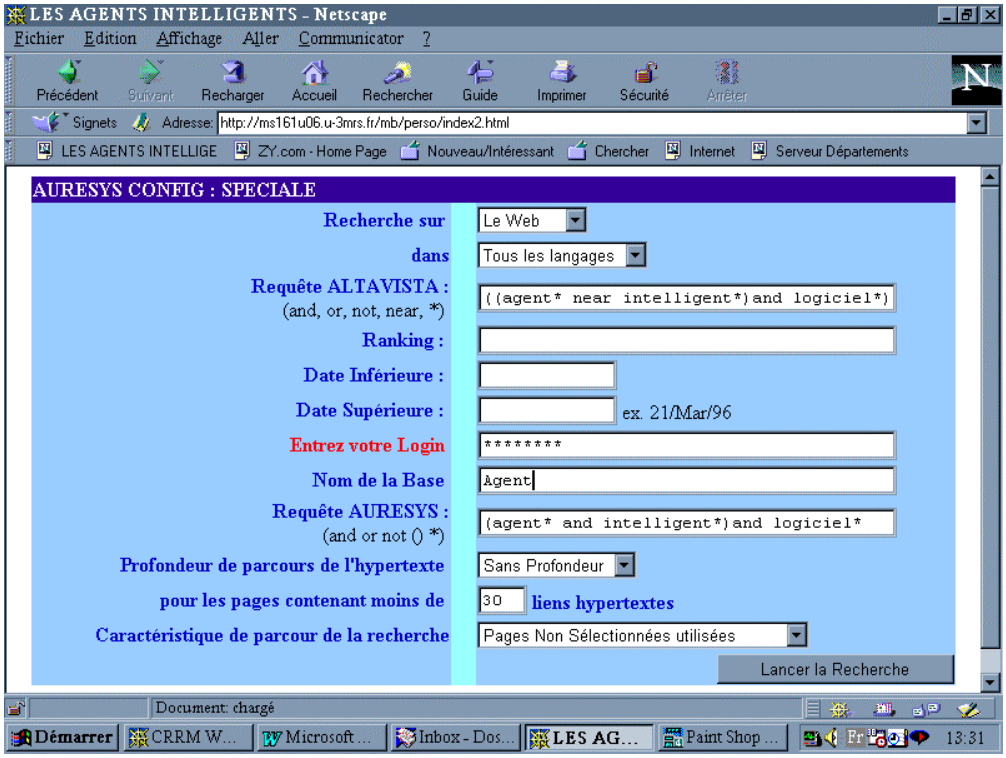

Figure 2 : Première étape : Configuration de l'Agent Intelligent AURESYS

Présentation des résultats : Le Gateway

Les résultats sont présentés dans une fenêtre divisée en cinq parties. Ces pages sont liées entre elles par des liens hypertextes qui permettent de développer les différents types de renseignements, et ainsi de naviguer dans le gateway.
Cette structure peut paraître complexe, mais elle permet de n'avoir aucune perte d'information.

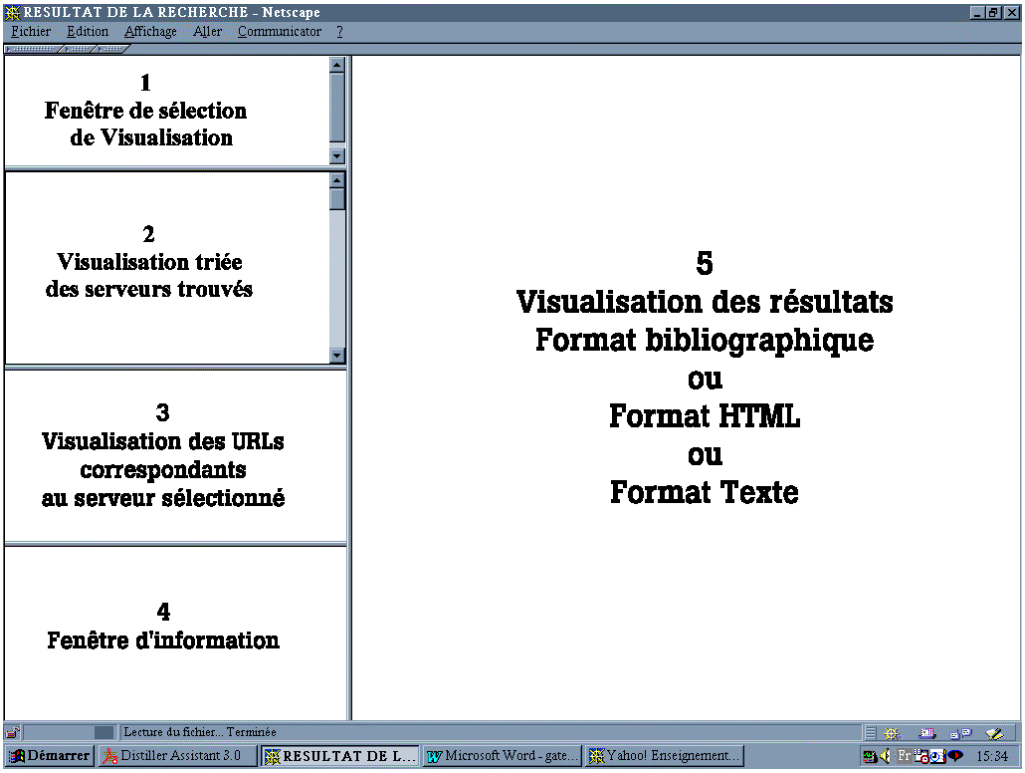

Figure 3 : Descriptif de la structure du Gateway

Fenêtre 1 : Sélection du choix de la visualisation :
Quatre choix de visualisation sont offerts :

- Visualisation des pages trouvées dans un format bibliographique. Utile pour effectuer des analyses plus poussées. (résultat sur la fenêtre 5)

- Visualisation des serveurs trouvés classés par le nombre d'URLs trouvés. Seules les pages des serveurs répondant à la requête sont visualisables dans la fenêtre 5.

- Visualisation des serveurs trouvés classés par leur type de domaine (.fr, .com, .edu, .org ...).

- Visualisation sous forme de tableau des différents types de domaines trouvés classé par leur fréquence.

Fenêtre 2 : Les Serveurs Trouvés : Fenêtre qui contient tous les serveurs correspondants à la requête. Tous les serveurs sont liées à la fenêtre " Les URLs du serveur trouvé " (3), ce qui permet de voir instantanément les pages.
II est aussi indiqué le nombre de pages ayant un rapport avec la requête.
Fenêtre 3 : Les URLs du serveur trouvé : La liste des liens d'un même serveur ayant un rapport avec la recherche. La visualisation de cette liste d'adresses ce fait par une simple sélection d'un serveur dans la fenêtre "Les Serveurs trouvés".

Pour chaque page, il est possible de :

- Connaître la pertinence de la page.
- Accéder à un format condensé de la page HTML, comprenant des renseignements sur son contenu.
- Accéder à une visualisation en local de la page HTML.
- Se connecter directement à la page du site.

Fenêtre 4 : Les renseignements sur la recherche : Renseignements concernant le résultat de la requête : Nombre de fichiers visités, Nombre de fichiers sélectionnés, nombre d'erreurs de connexion, nombre d'erreurs de protocole (ne traite que le protocole HTTP).
De plus, cette fenêtre rappelle la stratégie de la requête.
Fenêtre 5 : Visualisation des résultats : Visualisation des différents format des pages trouvées (bibliographiques, HTML, format texte condensé).
Cette visualisation est rapide puisque locale; Il n'y a pas de connexion aux différents sites. C'est un moyen très efficace d'avoir un aperçu du contenu de chaque serveur. L'utilisateur garde quand même la possibilité de se connecter directement à chaque serveurs.

4. Exemple de Gateway

Voici un exemple de gateway qu'il possible de consulter sur Internet : *http://ms161u06.u-3mrs.fr/Aint001R.html*
La thématique porte sur les agents intelligents, et plus précisément les logiciels, dans les adresses de serveurs français.

Dans un premier temps, une requête sur Altavista a été effectué. La requête était :

((agent* near intelligent*) and logiciel*) and domain : fr

C'est de cette page de résultat qu'Auresys démarrera la collecte d'informations. Il aura une autre requête, car il n'accepte pas l'opérateur "near" :

(agent* and intelligent*) and logiciel*

De plus, la profondeur de parcours est de 1, c'est-à-dire qu'Auresys va parcourir les liens des pages renvoyées par Altavista, et pas au-delà.

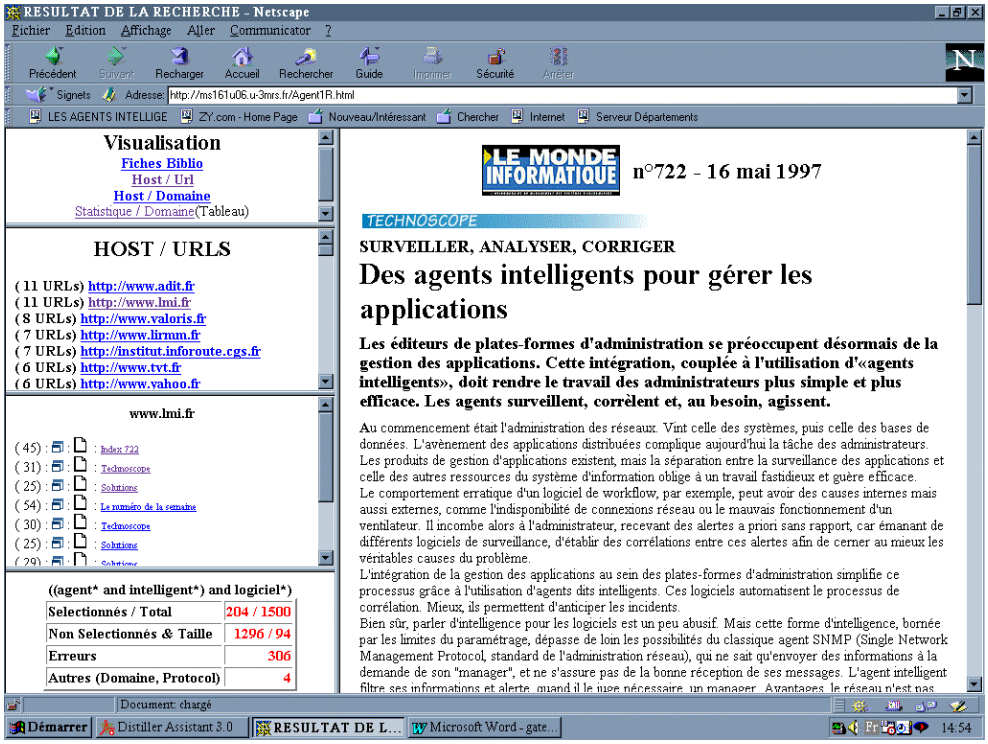

Figure 4 : Exemple de Gateway

La figure ci-après propose donc un gateway de 204 pages HTML correspondant à ce thème. Auresys a donc sélectionné ces pages parmi les 1 500 qu'il a visitées. Ceci est du au fait que soit la page contenait les mots clés dans les Tags HTML, soit elle contenait plus du nombre de liens paramètrés dans la configuration d'Auresys (c'est-à-dire plus de 30 liens).

5. Avantages de ce type d'outils

L'avantage premier de ce type d'outils est bien évidemment le gain de temps. La rapidité d'accès à l'information réside dans le fait que les serveurs du gateway correspondent déjà aux exigences de l'utilisateur. Il n'aura donc aucune recherche à effectuer sur les moteurs de recherche. L'information qu'il a à sa disposition est déjà ciblée. De plus, les pages des serveurs peuvent être prévisualisées en local et donc pas de perte de temps en connexion.

En ce qui concerne la gateway en tant que tel, Auresys permet de faire quelques analyses et statistiques supplémentaires. En effet, il est capable entre autre de faire un comptage des différents types de domaines et de représenter graphiquement les liens entre les serveurs répondant à la requête sous forme de réseau.

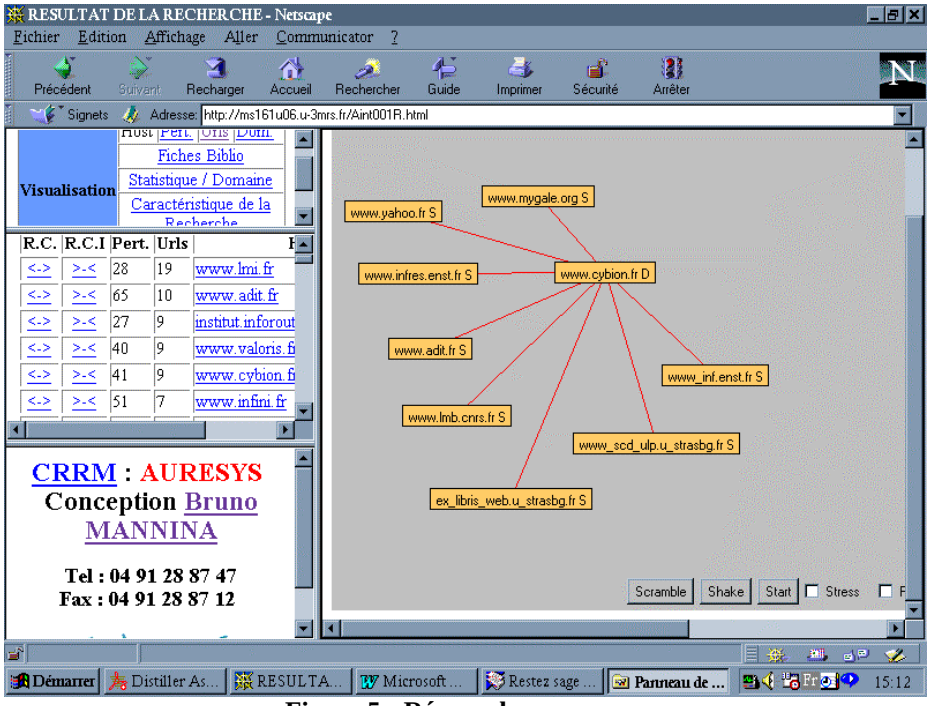

Figure 5 : Réseau de serveurs

De plus, possédant des formats de visualisation bibliographiques, il est possible d'importer ces données dans un logiciel de gestion de données (type ACCESS) ou de faire de la Bibliomètrie. Ce type d'analyse est favorisé par le fait qu'Auresys crée plus d'une vingtaine de champs descriptifs de la page HTML (Mots-Clés, Environnement des Mots-Clés, Présence ou non de MétaTags, Langue, Nombre de liens internes et externes...).

Certes, un gateway est une photographie instantanée de l'état de l'art sur Internet d'une thématique. Il n'est donc ni dynamique, ni évolutif. Le seul palliatif à ce problème est la reconstruction de celui-ci. Il faut donc le refaire souvent (trimestriellement), et c'est là qu'intervient le concept des agents intelligents pour éviter la perte de temps.
Comme explicité précédemment, la construction des gateways par les agents intelligents est très rapide. Pour un thème comme celui des *logiciels sur les agents intelligents*, le temps d'élaboration est d'environ 12 heures.
Cela permet donc de refaire très souvent le gateway, et donc de le mettre à jour de manière complète. La mise à jour et l'intégration de nouveaux liens dans le gateway de façon manuelle, est une tâche trop fastidieuse pour avoir un résultat efficient (recherche sur les moteurs des nouveaux sites, et tests des anciens serveurs).

En conclusion, l'utilisation d'agents intelligents pour la création de gateways est indispensable pour les professionnels de l'information. Ces gateways seront de

meilleur qualité, nécessiteront très peu de temps à leur confection, et apporteront une valeur-ajoutée sous formes d'analyses complémentaires.

6. Bibliographie

1- Journal Science, *Steve Lawrence, Lee Giles,* NEC Research, Institut de Princeton

2- Network Wizards Internet Domain Survey [en ligne]
http://www.nw.com/zone/hosts.gif
(consulté le 29-11-98)

3- "AURESYS 2.0 : Un agent Intelligent au service de l'information stratégique"
Quoniam Luc, Bruno Mannina, Dou Henri, CRRM, SFBA'97, Ile Rousse

4- "Construction automatique de réseaux : un outil pour mieux appréhender l'information provenant de l'Internet"
Eric Boutin (Univ. De Toulon et du Var, IUT TC), Bruno Mannina, Hervé Rostaing et Luc Quoniam (CRRM, Fac. Saint Jérôme, Marseille) JADT, Février 1998

5- "Measuring the social structure of the usenet"
M.A. Smith, University of California, Los Angeles

6- Citations : an exploratory study"
R. Rousseau, KHBO, Faculty Industrial Sciences and Technology

7- "Nouvelles Technologies. Pas si bêtes, ces agents…" [en ligne]
M. Baccar, Génération Internet : http://www.entreprises-virtuelles.ch/art_12.htm
(consulté le 28-09-98)

8- "A smart Itsy Bitsy Spider for the Web"
Hsinchun, Yi-Ming Chung, Marshall Ramsey, C.C. Yang, Journal of the american society for information science, May 15, 1998