CREATION D'HYPERTEXTES AUTOMATIQUES APPLIQUES A LA VEILLE

Léveillé Valérie, Doctorante en dernière année
Rostaing Hervé, Maître de conférence au CRRM
Quoniam Luc, Maître de conférence au CRRM

CRRM
Centre scientifique de Saint-Jérôme
Av. Escadrille Normandie-Niemen
13397 Marseille Cedex 20
Tel : 04 91 28 87 40 Fax : 04 91 28 87 12
e-mail : {leveille, rostaing, quoniam}@crrm.univ-mrs.fr

Résumé : Un monde de plus en plus complexe, et ouvert aux innovations, peut offrir de multiples opportunités à ceux qui savent anticiper mais peut également noyer des entreprises qui n'auront pas su saisir à temps ces opportunités. La mission de la veille est précisément de discerner dans une multitude d'informations les signaux faibles, mais pertinents qui préfigurent l'avenir. Cette information est généralement facile d'accès mais surabondante, fragmentaire et donc peu sûre. Trier, recouper et synthétiser l'information devient alors essentiel pour la rendre fiable et efficiente. Pour cela, il est nécessaire de pouvoir favoriser ces recoupements quel que soit son format sous une même interface avec un outil destiné aux utilisateurs du système d'information. Dans la première partie de notre article, nous présentons les différents besoins ressentis par les experts face au dossier de veille. Nous montrons ensuite que la navigation hypertextuelle peut aider l'expert dans sa phase d'analyse et de validation des informations. Nous proposons un outil de navigation relationnelle destiné à faciliter l'analyse. Cet outil, basé sur le principe de la construction automatique hypertexte, propose aux utilisateurs du système de veille les moyens pour analyser, trier et recouper l'information.

Mots - clés : information endogène, hypertexte, veille technologique, aide à la décision, loi de Zipf

Abstract: A world increasingly complex, and opened with the innovations, can offer multiple opportunities to those which can anticipate but can also embed companies which will not have known to seize in time these opportunities. The mission of technology watch is precisely to distinguish in a multitude of information the weak signal, but relevant which precede the future. This information is generally easy to access but superabundant, fragmentary and thus not very sure. To sort, confirm and synthesise information become essential then to make it reliable and efficient. For this, it is necessary to support these stepping under a same interface with a tool intend for end-user of information system whatever the information format. In the first part of our article, we present the various needs felt by the experts to the technology watch file. We show that navigation with hypertext can help them in their analysis phase. We propose a tool for relational navigation intended to facilitate this analysis. This tool, based on the principle of automatic construction hypertexte, proposes to the technology watch's users the means to analyse, sort and confirm information.

Key words: endogenous information, hypertexte, technology watch, decision making, Zipf Law

Colloque VSST'98 ; Toulouse ; 19-23 octobre 1998 ; pp. 339-347

INTRODUCTION

Un monde de plus en plus complexe, et ouvert aux innovations, peut offrir de multiples opportunités à ceux qui savent anticiper mais peut également mettre en péril des entreprises qui n'auront pas su saisir à temps ces opportunités. Dans l'environnement actuel, seule une attitude appropriée de veille est susceptible de déjouer les pièges qui menacent les entreprises de paralysie. La mission de la veille est précisément de discerner dans une multitude d'informations les signaux faibles, mais pertinents qui préfigurent l'avenir.

Cette information est généralement facile d'accès mais surabondante, fragmentaire et donc peu sûre. Elle se présente sous multiples formes : textuelle (références bibliographiques, articles de presse, rapports d'experts), numérique et infographique (résultats d'analyses bibliométriques, tableaux de bord et tableaux prévisionnels), image (plaquettes publicitaires ...). Trier, recouper et synthétiser l'information devient alors essentiel pour la rendre fiable et efficiente. Pour cela, il est nécessaire de pouvoir favoriser ces recoupements quel que soit son format sous une même interface avec un outil destiné aux utilisateurs du système d'information.

L'objet de cette communication est de présenter l'élaboration d'un outil de navigation relationnelle. Cet outil, basé sur le principe de la construction automatique hypertexte, propose aux utilisateurs du système de veille les moyens pour analyser, trier et recouper l'information.

BESOINS DE GESTION DE L'INFORMATION EN VEILLE TECHNOLOGIQUE

Le processus de veille technologique peut se résumer à ces trois phases : collecte de l'information, analyse de l'ensemble des informations collectées, diffusion de cette analyse pour action. L'une des principales difficultés de la seconde étape concerne le volume important d'informations à analyser, recouper, synthétiser en un temps très court. Veilleurs et experts du domaine interviennent ensemble sur cette phase essentielle en veille technologique. Le premier est chargé de préparer le dossier de veille à l'analyse de l'expert. En organisant et en structurant le dossier, le veilleur s'assurera ainsi de la pleine collaboration de l'expert.

L'organisation du dossier de veille doit permettre à l'expert [Rousseau 97]:

- ⇨ d'avoir une vision globale du dossier,
- ⇨ de positionner son entreprise, son activité ou ses concurrents dans le dossier,
- ⇨ d'explorer la masse de documents mise à sa disposition. L'exploration est une phase indispensable en veille. L'expert doit pouvoir "entrer" dans le dossier par une information, un mot-clé ou un auteur qu'il connaît et pouvoir connecter d'autres informations qui lui sont inconnues.

Les outils bibliométriques classiques réalisent d'ores et déjà ces opérations de cartographie et de structuration du dossier [White 89], [Dousset 95]. L'exploration et la navigation au sein de la masse de documents relève davantage du domaine des outils de gestion de l'information que des outils bibliométriques. On peut distinguer plusieurs approches dans le cas d'une assistance informatique:

- ⇨ L'approche booléenne. L'utilisateur interroge le fonds documentaire par des requêtes booléennes ou en langage naturel. Cette approche requiert de la part de l'utilisateur une certaine connaissance de la base.

⇨ L'approche hypertextuelle. Elle permet à partir d'un document de départ de naviguer au sein de l'ensemble des documents sans avoir à priori une parfaite connaissance de sa structure.

⇨ L'approche linéaire : la simple transposition du dossier de veille "papier" en version électronique sous Word ou en format PDF qui n'autorise que très peu de liens hypertextes.

L'approche hypertextuelle est très séduisante en veille technologique par sa capacité à produire de l'information endogène : information non présente dans le dossier mais issue du croisement de deux informations. Deux freins majeurs ont cependant retardé la généralisation de l'utilisation d'hypertextes en veille [Lelu 95]:

- Les limites liées à l'édition manuelle des liens. D'une part, lier manuellement deux documents impose une bonne connaissance du contenu de la base. D'autre part, le temps utilisé pour la création manuelle serait inacceptable dans un processus de veille.
- Le manque de vision globale. Le manque de vue d'ensemble du corpus est à l'encontre de ce qui est requis d'un outil de veille. L'hypertexte classique présente uniquement des liens entre documents, sans outils de cartographie ou de classement du corpus.

En utilisant des nœuds hypertexte ayant un rôle de synthèse, la structuration des informations présentée par A. Lelu, permet d'obtenir une vision globale des documents. Cette structuration permet de regrouper des documents en thèmes principaux. Cette méthode de construction, développée tout d'abord par Damashek [Damashek 95], est basée sur une analyse classificatoire, les nœuds de synthèse représentant un ensemble de documents. D'autres méthodes ont également été développées, notamment la méthode Tachir [Agosti 94]. Cette méthode n'est pas basée sur l'analyse classificatoire mais établit les liens entre documents via des termes indexés (mot-clé, auteurs…).

En 1995 [Rostaing 1995], le CRRM a évalué l'utilisation de la navigation hypertextuelle dans le cadre de l'analyse d'un dossier de veille. L'emploi d'une méthode d'analyse classificatoire, l'Analyse Relationnelle des Données développée par ECAM-IBM [Huot 93], a permis de créer des nœuds de synthèse en regroupant par thèmes les différents documents du dossier.

CONSTRUCTION D'HYPERTEXTES AUTOMATIQUES APPLIQUES A LA VEILLE

Nous développons un outil informatique offrant une navigation issue conjointement de l'hypertexte et de la bibliométrie. Cet outil, de part cette filiation commune, permet de proposer à l'utilisateur une vision à la fois synthétique et globale du dossier, tout en conservant l'accès à l'information primaire.

Organisation des données collectées

Dans un premier temps, nous avons limité notre outil à l'exploitation des informations collectées sur les serveurs de bases de données. Pour bâtir notre hypertexte, nous avons choisi d'exploiter la qualité relationnelle de ces données. En effet, les champs auteurs extraits de documents issus des bases de données, permettent de relier deux documents issus de la même base mais aussi de deux bases différentes. Associer deux documents provenant de deux bases de structures différentes représente ainsi un atout majeur de notre logiciel, un dossier de veille technologique devant être constitué à partir d'informations de sources différentes. Il est important de pouvoir lier les articles scientifiques écrits par un auteur avec les brevets déposés par celui-ci. Nous avons donc préféré organiser la structure de la base relationnelle de l'hypertexte autour de cette notion de forme : auteurs, descripteurs… La totalité du document

étant conservée pour information, nous avons choisi de n'extraire que les quelques champs permettant de naviguer entre documents d'une même source et documents de sources différentes. Ces champs (auteurs, descripteurs, code de classification documentaire …) sont sélectionnés par l'utilisateur lors de la procédure d'importation.

Principe de navigation

Tout d'abord, l'utilisateur retient un ensemble de documents de départ via le module d'interrogation basée sur l'utilisation de requêtes booléenne. Un fois ce lot retenu, la navigation se fait par le module *contexte*.

L'information manipulée est représentée dans ce module par des rectangles de différentes couleurs suivant la nature de l'entité considérée : document, forme ou suivant le champ associé à cette forme (auteur, descripteur, …). La navigation s'effectue par un simple clique de la souris sur l'entité considérée et par la sélection d'une proposition dans le menu contextuel qui s'affiche alors. Suivant le type de l'entité considérée (document ou forme), le menu contextuel présentera les options suivantes :

- Sélection d'un document :
  - Les différents champs associés au document (auteur, descripteur, …)
    Cette option propose une nouvelle de navigation dans le corpus sur la base du contenu du document.
  - La gestion ergonomique de l'affichage des chemins parcourus
    La possibilité de cacher la branche précédemment développée à partir de cette entité
- Sélection d'une forme :
  - Retour à l'information primaire par l'affichage des documents contenant cette forme
  - Les formes présentes avec la forme sélectionnée dans le même champ : lien interne
    Cette option permet la navigation dans le réseau des relations des formes d'un même champ :
    - Navigation dans le réseau "sémantique" des mots-clés. En présentant les descripteurs présents conjointement avec la forme sélectionné, l'utilisateur accède très rapidement à l'environnement sémantique de ce terme.
    - Navigation dans le réseau de collaborations pour les champs de type auteurs, inventeurs, affiliation.
  - Les formes présentes avec la forme sélectionnée dans un autre champ : liens externes
    Cette option permet la navigation croisée entre les réseaux de relations internes des champs :
    - L'utilisateur pourra ainsi basculer du réseau des auteurs vers celui des mots-clés ou inversement.
  - La gestion de l'affichage des développements des différents réseaux
    La possibilité de cacher la branche précédemment développée à partir de cette entité.

Cet mode de navigation offre une aide d'analyse synthétique puisque en un simple clic, on peut obtenir l'ensemble des thèmes d'un auteur, l'ensemble des journaux où il publie, voire également l'ensemble des numéro des brevets qu'il aurait pu déposé….La figure 1 présente les différentes possibilités de navigation.

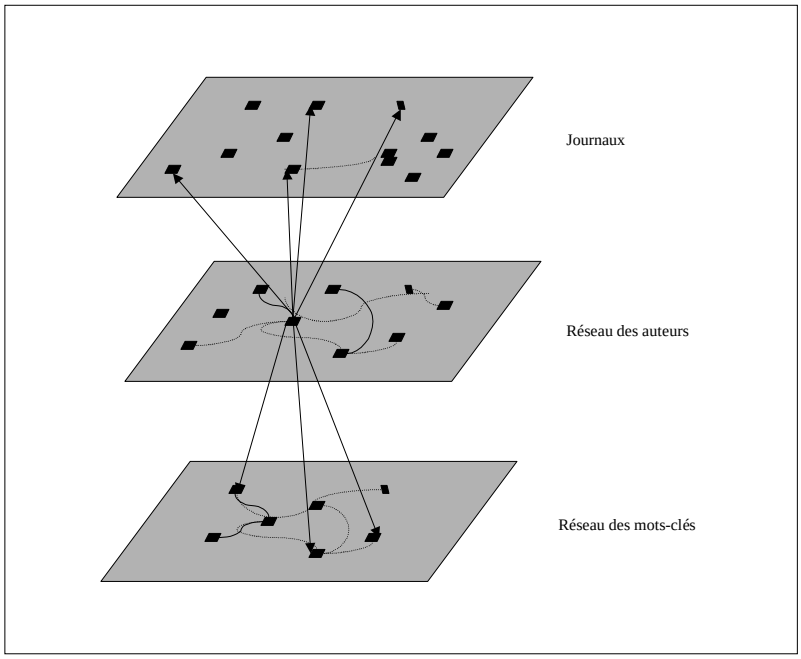

Figure 1 : Les différentes possibilités de navigation

Ainsi, le menu contextuel présentera les options suivantes suite à la sélection de l'auteur Stromberg, AG (figure 2)

- Documents dont l'auteur est Stromberg
- Lien interne : auteurs ayant collaboré avec Stromberg
- Liens externes avec un sous-menu présentant les autres champs : descripteurs associés à l'auteur Stromberg, ou encore les auteurs cités par Stromberg
- Supprimer la branche

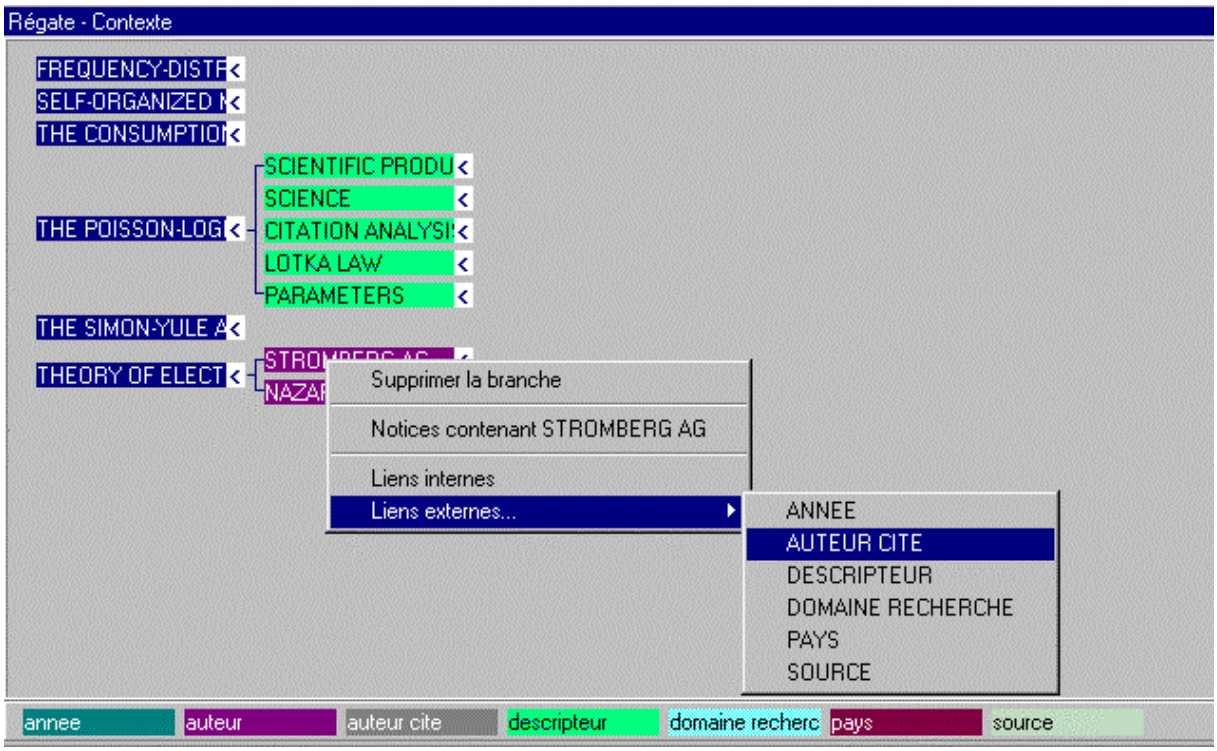

Figure 2 : Navigations possibles à partir d'un auteur

La sélection d’une de ces options provoque l’exécution d’une requête puis l’affichage des résultats de cette requête. L’utilisateur peut ainsi naviguer au sein de la base sans jamais avoir été orienté par des liens pré-définis. Les liens sont créés automatiquement par les relations que les informations entretiennent entre elles de façon intrinsèque. L'utilisateur en fonction de son domaine, de ses pôles d'intérêt, de ses aspirations choisira de développer une voie plutôt qu'une autre. Sans posséder une connaissance du fonds collecté, il peut ainsi par son cheminent explorer et découvrir de nouvelles informations.

Utilisation de la bibliométrie pour aider l’utilisateur dans sa démarche

Nous nous sommes très rapidement heurtés au problème de la multiplication des liens entre documents. Ce phénomène perturbe la vision globale offerte à l'utilisateur. En effet, les termes à haute fréquence, lient pratiquement tous les documents. Les termes à faible fréquence, parasitent également le schéma de navigation en ne liant que très peu de documents.

Pour résoudre ce problème, nous avons choisi de proposer à l'utilisateur un filtre de fréquence, suggérant d'éliminer de la navigation, d'une part les termes triviaux et d'autre part les termes marginaux. Ces filtres sont déterminés à partir de la méthode de l’entropie [Lhen 95]. Cette méthode permet de découper automatiquement la distribution des termes en deux ou trois zones en calculant les fréquences de coupure entre ces zones. Ce système de découpage en trois zones est proposé pour chaque champs de la base, c'est-à-dire pour chaque type de données(figure 3). Il paraît évident qu'étant basé sur des calculs de fréquences, cette méthode ne pourrait pas s'appliquer à l'ensemble des formes de tous les champs confondus. L'amplitude des fréquences entre les formes mots-clés ou auteurs n'ont rien de comparable.

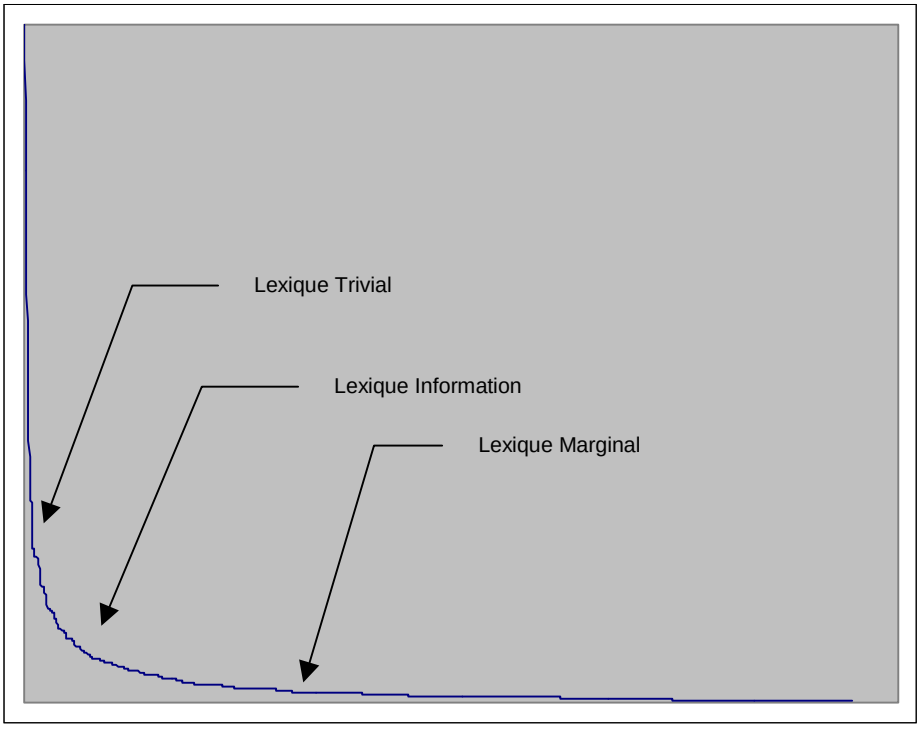

Figure 3 : Découpage du lexique d'un champs en trois parties

Chaque forme (auteur, mot-clé, code de classification documentaire..), est qualifié comme appartenant à la catégorie HF (haute fréquence, termes triviaux), MF (moyenne fréquence) ou BF (basse fréquence, termes marginaux). Les termes qualifiés MF peuvent être les seuls sélectionnés par l'utilisateur pour associer deux documents. Cette méthode nous permet ainsi « d'alléger » la représentation hypertextuelle et d'offrir ainsi une meilleure vision du dossier à l'expert. Bien évidemment, le logiciel propose à l'utilisateur d'ajuster à la suite de son expertise les seuils de fréquences de ces trois zones.

Exemple de navigation

L'exemple de navigation proposé figure 3, représente la navigation au sein d'un corpus extrait de la base de l'ISI. La base comporte environ 200 documents et concerne les articles publiés en bibliométrie. Nous avons considéré les champs suivants : année, auteur, auteur cité, descripteur, domaine de recherche, pays et source. Nous avons constitué l'ensemble de départ en sélectionnant uniquement les articles publiés en 1994.

Nous avons choisi pour continuer la navigation le document dont le titre est *"Theory of electroanalytical chemistry - developpement in the last 5 years, Current state, and prospects - Scientometric Aspect"*. Nous désirions obtenir très rapidement une première vision des communautés virtuelles qui s'étaient créées part le biais des citations.

Pour cela, nous avons choisi de développer la branche des auteurs de cet article (voir figure 3) puis nous avons développé la branche des auteurs cités par l'un d'eux (Stromberg). Nous avons fait apparaître la branche des auteurs citant Braun (auteur cité par Stromberg), par un simple clique sur Braun. Enfin pour connaître les descripteurs utilisés par l'un de ces auteurs, nous avons présenté les descripteurs utilisés par Arunachalam.

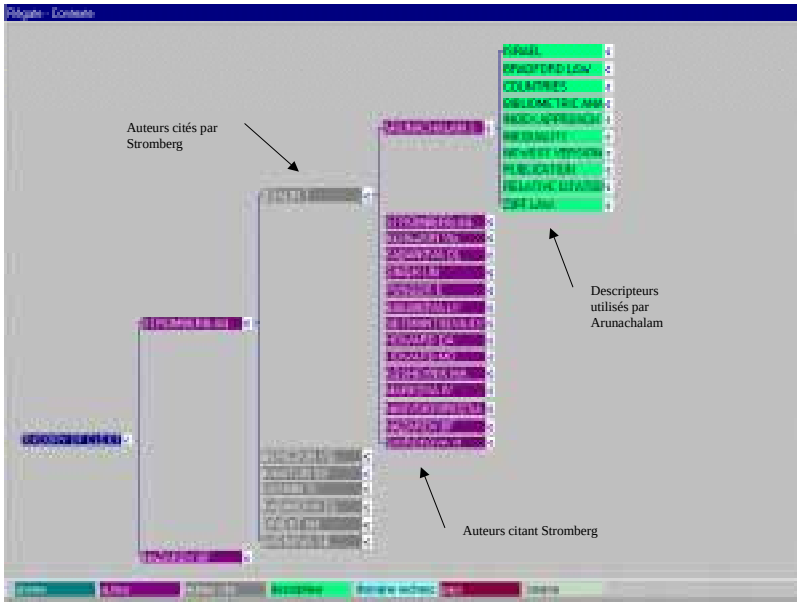

Figure 4 : Exemple de navigation

Pour limiter l'apparition d'un nombre trop important de formes, nous avons choisi d'appliquer des filtres à la navigation (figure 4). Nous avons sélectionné uniquement les auteurs cités dont la fréquence dans la base est comprise entre 2 et 10. Ainsi des auteurs très souvent cités comme Garfield (39 fois), Rosenthal (14 fois), ou encore Egghe (11 fois) n'apparaissent pas dans la représentation, tout comme ceux qui ne sont cités qu'une fois. Nous avons procédé de même pour les descripteurs.

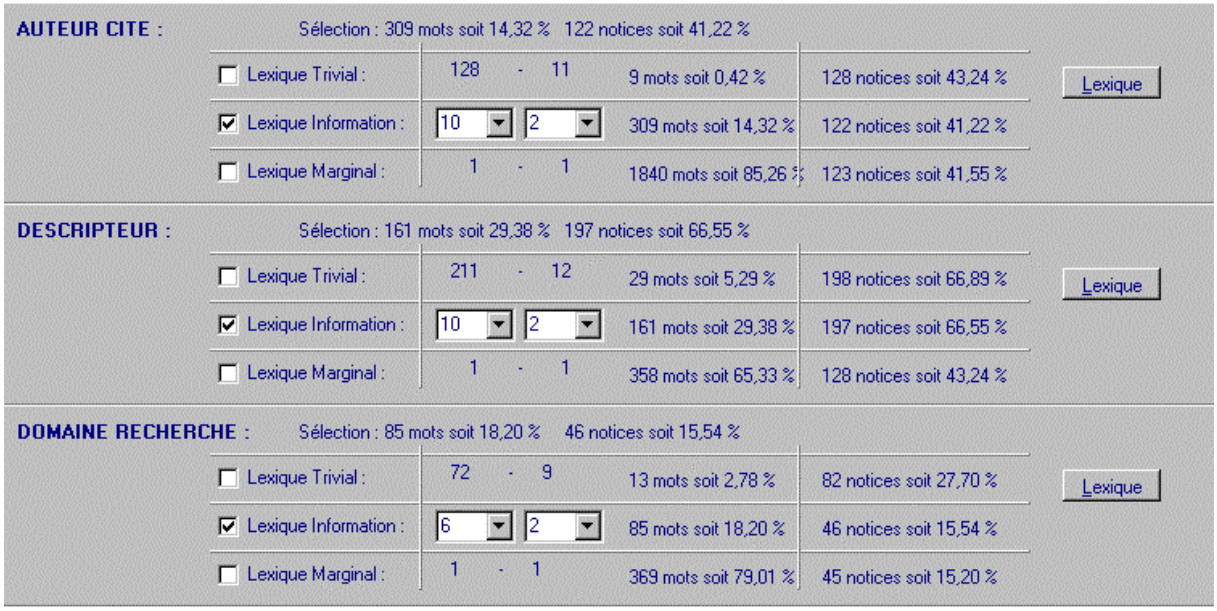

Figure 5 : Paramêtrage des filtres de navigation

CONCLUSION

En répondant aux besoins d'analyse des utilisateurs du système de veille, nous proposons un outil informatique adapté à la veille offrant à la fois une vision synthétique et globale à l'utilisateur ainsi que des possibilités de navigations objectives au sein du dossier de veille.

Le mode de présentation de l'information choisi offre la possibilité de visualiser les liens entre documents, mais aussi entre termes et documents ainsi qu'entre termes. Le logiciel montre graphiquement quels sont les mots-clés associés à tel inventeur ou à telle société. En utilisant un hypertexte classique, il est nécessaire pour cela de parcourir le document pour connaître quels termes sont associés à un auteur particulier. Notre logiciel permet très rapidement d'obtenir ces renseignements tout en conservant l'accès à l'information primaire. L'expert a ainsi une vision globale du dossier, tout en ayant les moyens de recouper, de croiser les informations en naviguant au sein du dossier.

Ces différentes fonctionnalités facilitent ainsi l'expertise du dossier en favorisant le passage de l'information brute à l'information élaborée. En outre l'organisation des données, entièrement indépendante de la structure initiale des informations, en fait un outil réellement ouvert.

BIBLIOGRAPHIE

[Agosti 1994] : Agosti M., Melucci M., Crestani F.; "TACHIR : a tool for automatic construction of hypertexts for information retrieval", *RIAO 94 Intelligent multimedia information retrieval systems and management*, Rockefeller University New-York, N.Y.-USA, October 11-13, 1994, pp. 338-357

[Damashek 1995] : M. Damashek; "Gauging similarity with n-grams : language-independent categorization of text", *Science*, vol. 267, février 1995

[Dousset 1995] : Dousset B., "Le logiciel d'études bibliométriques TETRALOGIE de l'IRIT", *Colloque VSST'95 Veille stratégique, scientifique et technologique*, Toulouse, 25-27 octobre, 1995, pp. 431-471

[Huot 93] : Huot C, Coupet P, Bedecarrax C, "MARS : station d'analyse automatique de documents scientifiques et techniques", *Deuxièmes Journées Internationales de l'Analyse des Données Textuelles*, Montpellier, 21-22 octobre 1993, pp. 199-212

[Lelu 1995] : Lelu A; "Hypertextes : la voie de l'analyse de données", *Troisièmes Journées Internationales de l'Analyse des Données Textuelles*, Rome, 13-15 décembre 1995, pp. 85-96

[Lhen 1995] : Lhen J., Lafouge T., Elskens Y., Quoniam L., Dou H., "La "statistique" des lois de Zipf", *Colloque Les journées d'information élaborée*, Ile Rousse, 1995, pp. 135-146

[Rostaing 1995] : Rostaing H., Djaouzi S., La Tela A., Avignon T., Quoniam L., "Analyse bibliométrique multi-bases pour l'élaboration d'un dossier électronique de veille technologique", *Colloque VSST'95 Veille stratégique, scientifique et technologique*, Toulouse, 25-27 octobre, 1995, pp. 153-168

[Rousseau 97] : Rousseau F., Thil J., "Veille et informatique : des besoins aux solutions", *Technologies Internationales*, n° 39, novembre 1997, pp. 33-36

[White 1989] : White H. D., Mc Cain K. W., "Bibliometrics", *Annual Review of Information Science and Technology (ARIST)*, vol. 24, pp. 119-186, 1989