

XPlor : un outil d'investigation en ligne sur des données relationnelles.

Cristelle ROUX (*), Didier SOSSON (*), Bernard DOUSSET ()**
roux.cristelle@tiscali.fr , didier.sosson@gfi.be , dousset@irit.fr

(*) GFI Bénélux 183, route de Luxembourg, L-8077 Bertrange, Luxembourg

(**) IRIT/Université Paul Sabatier, 118 route de Narbonne 31062 Toulouse cedex 4

Mots clés:

Analyse relationnelle, veille stratégique, base de données relationnelle, Internet, navigation hypertexte, MySql

Key words:

Relational analysis, strategic watch, relational data bases, Internet, hypertext navigation, MySql

Palabras claves:

Analisis relacional, vigilancia estratégica, base de dato relacional, Internet, navegacion hipertexto, MySql

Résumé:

Depuis plus de 10 ans, le logiciel Tétralogie nous permet d'effectuer des analyses stratégiques sur des corpus d'information textuelle issus des sources les plus diverses comme: les bases en ligne, les Cd, le Web visible et invisible, les news, les brevets, la presse, les traces de connexions aux sites, les bases internes,... L'information élaborée qui en est issue représente une synthèse de l'ensemble des documents: identification des acteurs et de leurs relations, sous sujets cohérents, signaux forts et faibles, tendances, composantes stratégiques et, sur demande, études ciblées faites à l'unité et réalisées par des analystes chevronnés. Mais, l'utilisateur final aimerait pouvoir lui-même zoomer facilement sur son propre environnement, pour connaître, par exemple, le positionnement de ses principaux concurrents, les procédés alternatifs connexes à son activité, les marchés potentiels ou il n'est pas encore présent. Nous proposons donc de compléter nos analyses macroscopiques par un système de navigation en ligne au cœur de l'information relationnelle obtenue par des recoupements statistiques, des classifications ou des analyses multidimensionnelles. Le but étant de privilégier l'extraction d'information en fonction du contexte général et non exclusivement par décryptage du contenu intrinsèque de quelques documents pris séparément. Il devient ainsi possible de retrouver, à partir d'un élément connu (acteur, mot clé, date), toute ou partie de l'information qui lui est connexe (équipes, collaborations, concepts, émergences, mots associés,...) et ce par l'utilisation de nombreux opérateurs d'association ou de filtrage et de fonctions de reporting pertinentes.

Afin de faciliter ce type de recherche et de conserver toute l'information disponible, nous avons dû abandonner la notion de matrices (cooccurrences, contingence, présence absence) directement chargées en mémoire vive, pour une structure de base de données relationnelle beaucoup plus souple. Les limitations de taille qui nous obligeaient à tronquer, souvent de façon abusive, nos dictionnaires sont ainsi levées. Le gaspillage de mémoire résultant du stockage de matrices creuses est de la même façon largement évité. De plus, cette base de données est interfacée sur Intranet ou Internet, afin que l'utilisateur puisse lui-même mener ses propres investigations. Chaque champ sémantique peut alors être filtré au moyen de fonctions relationnelles prédéfinies en se servant des liens complexes qu'il possède avec les autres champs de la base. Des statistiques interactives sont alors disponibles pour chaque extrait (fréquences, équivalences, liens pondérés, top n, ...) ainsi que des cartes ou des réseaux (relationnels ou sémantiques). L'extraction des documents pertinents s'effectue alors localement (s'ils n'évoluent pas) ou via le Web pour une mise à jour éventuelle. Enfin, il est toujours possible de reconstituer rapidement n'importe quelle matrice ou sous matrice afin de la traiter par les techniques habituelles du logiciel Tétralogie en étant parfois obligé d'imposer, comme avant, certaines limitations en fonction des capacités du matériel utilisé.

1 Les différents modes d'implantation

1.1 Problème de pertinence pour l'utilisateur

Comme nous l'avons dit précédemment, Tétralogie est un outil particulièrement bien adapté aux analyses macroscopiques, il permet en effet de dégager les signaux forts, les signaux faibles et les tendances à partir d'un ensemble de documents collectés sur un sujet précis. Mais à l'issue des très nombreuses analyses stratégiques que nous avons déjà réalisé avec ce logiciel, il est apparu que les utilisateurs finaux des analyses produites veulent, en complément de l'aspect stratégique, des zooms plus précis sur certains détails et ce afin de satisfaire leur curiosité en matière d'information élaborée autour d'éléments qu'ils ont déjà identifiés (concurrence, marchés, nouveaux produits ou procédés, partenaires potentiels, ...). A posteriori, de nombreux experts ou décideurs ont donc besoin de plus de finesse dans l'approche des éléments constituant traditionnellement leur environnement immédiat. Notamment, pour tout ce qui concerne leur vocabulaire spécifique, les acteurs qu'ils côtoient, les marchés qu'ils convoitent, les alliances qu'ils projettent. Une analyse peut être revisitée par différents spécialistes du domaine et apporter à chacun des réponses précises aux questions qu'il se pose. Le but est ici d'aider l'utilisateur dans sa navigation et dans sa quête de nouveautés ou de compléments d'information ainsi que dans la recherche d'éléments de comparaison avec des connaissances antérieures. La possibilité qui leur est donnée de pouvoir eux-mêmes naviguer sans contrainte dans l'information élaborée est un plus indéniable, car aucun analyste ne peut aller au devant de l'ensemble des préoccupations de chacun, ou alors il faut qu'il soit à leur entière disponibilité, c'est à dire appartenir intégralement à leur structure et très bien connaître leurs problématiques.

1.2 Compilation des matrices dans une base de données

Une première méthode, pour générer la base de données qui sera utilisée pour la navigation interactive, est de partir directement des dictionnaires et des matrices utilisés par Tétralogie pour l'analyse macroscopique. Cette approche présente de nombreux avantages :

- Compatibilité totale avec l'analyse par Tétralogie,
- Ne nécessite pas de système d'extraction complémentaire,
- Seule méthode pour l'instant permettant de prendre en compte les multi-termes,
- Permet de compléter des analyses déjà prêtes,
- Renforce par la navigation la pertinence du rapport d'analyse.

Mais aussi de nombreux inconvénients :

- Nécessite la génération préalable de toutes les matrices y compris évolutives,
- Nécessite l'utilisation des mêmes filtres et synonymes tout le long de cette génération,
- Les synonymes ne sont pas validés par la base de données,
- Certaines matrices sont tronquées (perte d'information, signaux faibles),
- Diffère la disponibilité de la base pour une navigation immédiate,
- Ne permet pas des mises à jour faciles de la base (nouveaux documents)

Cette technique est essentiellement destinée à compléter les analyses Tétralogie en permettant à l'utilisateur final de naviguer à sa guise, afin de préciser certains passages de l'analyse et de les ramener dans le contexte et l'environnement requis. Pour certaines analyses généralistes disponibles en ligne, cette approche permet à chacun de compléter son interprétation des conclusions toujours un peu stéréotypées de ce type de démarche globale. Pour des analyses plus pointues sur des sujets très précis, la taille plus réduite des dictionnaires utilisés permet de conserver toute l'information utile, notamment au niveau des champs sémantiques. Dans ce cas, cette approche nous semble la mieux indiquée.

1.3 Génération directe de la base de données

Dans un second temps, nous avons envisagé de court-circuiter les phases d'extraction et de croisement de l'information par Tétralogie, afin de ne plus utiliser de matrices (souvent très creuses) pour stocker le relationnel. Le gain de place est impressionnant, par contre le temps d'extraction reste équivalent.

En effet, Tétralogie utilise des techniques de *Hash Coding* pour optimiser cette phase très lourde, ici, nous bénéficions de l'indexation ce qui revient à peu près au même. Par contre, toute l'information est maintenant retenue, sans être obligé de recourir à des tailles mémoire dignes de gros main frame ou à des techniques sophistiquées de choix de la terminologie pertinente en termes de distribution dans les documents formant le corpus et de typage des relations inter champs. Le principal inconvénient de cette approche est la réécriture du module d'extraction de l'information dans un langage adapté. Par contre, la navigation est immédiate, l'information préservée, le stockage réduit et l'extraction contextuelle opérationnelle. Notre premier prototype ne conservait que les données statiques (ensemble des occurrences ou des cooccurrences dans le corpus), ceci nous est apparu trop restrictif pour les analyses stratégiques ultérieures. Nous avons donc opté pour un stockage différencié sur au plus quatre périodes choisies de façon à présenter des volumes assez homogènes. Nous stockons donc les occurrences et les cooccurrences en fonction de leur appartenance à une des quatre périodes retenues. Dans la navigation, la composante temps est donc toujours disponible et de nombreux opérateurs de calcul ou de visualisation de tendances peuvent ainsi être immédiatement déclenchés.

1.4 Extraction de matrices depuis la base de données

Si nous voulons revenir à Tétralogie, pour réaliser une analyse totale ou partielle du contenu de la base, il faut générer des matrices au format requis. Ceci est effectué par un module d'extraction qui permet éventuellement de réduire la portée des opérateurs de croisement à un sous ensemble du corpus (période, équipe, pays, liste d'items, cluster, relation, ...). Un même corpus peut donc être visité de différentes façons sans avoir recours à des stratagèmes de filtrage peu compatibles avec une utilisation grand public. La générations de matrices d'évolution (3D) est aussi grandement facilité par le stockage dans la base des informations associées à 4 périodes suffisamment homogènes. Cette composante temps étant absolument essentielle pour une analyse stratégique digne de ce nom. Cette technique, qui passe d'abord par la constitution d'une base de données, nous semble parfaitement adaptée à l'utilisation autonome du couple navigation interactive + Tétralogie, les analyses poussées n'étant déclenchées que sur des sous-ensembles ou dans des contextes précis. Une seconde utilisation peut être la gestion des connaissances, sur un sujet plus vaste, où peuvent se côtoyer des concepts généraux, des zooms ponctuels, des signaux forts et faibles, des tendances et des composantes stratégiques découvertes par l'analyse.

2 Nouvelles procédures d'extraction et de stockage

2.1 Implantation des analyses

Chaque analyse est implantée séparément, elle peut être accédée par mot de passe et sa description est consignée dans une table des analyses. Pour chaque analyse nous devons ensuite définir plusieurs entités: les champs, les filtres et les matrices constituant la structure actuelle ou future de l'analyse et définissant les points qui ont été traités et qui sont disponibles pour la navigation. D'un autre côté se trouvent les utilisateurs des analyses. Il sont identifiés dans une table des utilisateurs, leurs accès est sécurisé par mot de passe, les analyses sur les quelles ils ont des droits ainsi que les entités visibles sont aussi consignées dans des tables. Des extensions de droits sont données à l'administrateur, des restrictions peuvent aussi intervenir (données publiques, données privées) aussi bien en lecture, qu'en écriture. Le modèle de données est présenté dans la figure suivante, il tient compte de son implantation future dans un serveur d'analyses accessible sur InterNet ou IntraNet. Comme le plus souvent, est associé à une base de données un rapport d'analyse sous forme électronique (.doc, .html), il est possible de créer des liens entre les différents chapitres du rapport et les fonctions interactives de zoom et de reporting offertes par la base. Cette méthode permet de dynamiser la lecture du rapport et de s'en approprier le contenu de façon très personnelle. Un même sujet peut intéresser plusieurs personnes, d'où l'idée du partage de certaines analyses via le Web. Un corpus global pouvant être revisité de plusieurs manières tout en gardant, comme fil conducteur, la structure de l'analyse macroscopique déjà réalisée. C'est dans cette optique que nous avons conçu l'implantation des analyses dans un portail traitant de la veille et dans lequel se trouvent des espaces publics et des espaces privés suivant les possibilités de partage et les contraintes de confidentialité rencontrées.

Modèle de données

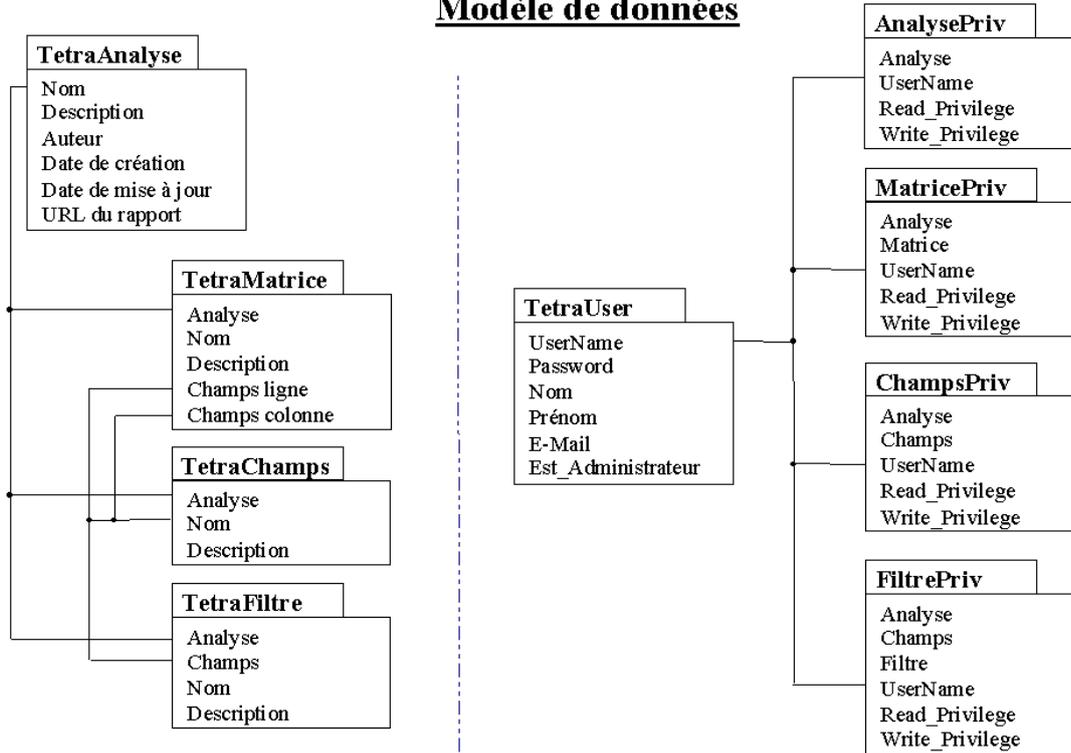


Figure 1 : Le modèle de données..

2.2 Etablissement des dictionnaires

Nous adoptons ici le même principe que celui utilisé dans Tétralogie, à savoir une procédure en trois passes comprenant :

- Une phase de détection des formes brutes des items
- Une évaluation semi-automatique des variations orthographiques
- Un comptage des occurrences rencontrées en tenant compte des synonymies.

Cette procédure peut être complétée, en amont, par l'application de filtres négatifs (mots vides, faux amis, hors sujet, ...) et en aval par une sélection ciblée des termes à retenir (sous sujet, fréquence de coupure, équipes, extraction de codes, sous champs, regroupements,...).

2.3 Remplissage de la base

Ici encore, nous reprenons le même principe que celui de Tétralogie, mais l'ensemble des cooccurrences détectées dans chaque croisement est stocké dans la base. Nous n'avons donc plus de perte d'information, notamment pour les plus grandes applications lors du traitement des champs sémantique et d'une façon générale de ceux dont les occurrences dépassent quelques milliers. Les performances de cette fonction d'extraction sont en partie conservées, car l'optimisation initialement due aux techniques de *hash coding* est maintenant obtenue par indexation de la base de données.

3 Interactivité avec la nouvelle structure de données

3.1 Comment orienter sa recherche d'information ?

Pour faciliter l'accès aux informations, nous avons choisi une double décomposition :

- En fonction de la structure de données (vecteur de fréquence ou matrice de cooccurrence),
- En fonction de la nature des données (acteurs ou éléments sémantiques).

Cet aiguillage permet de déterminer :

- Qui intervient et quels sont les domaines abordés (grâce aux fréquences)
- Quels sont les réseaux d'acteurs ou les concepts (grâce aux matrices de croisement).

Cette fonctionnalité prend, dans le portail, la forme suivante :

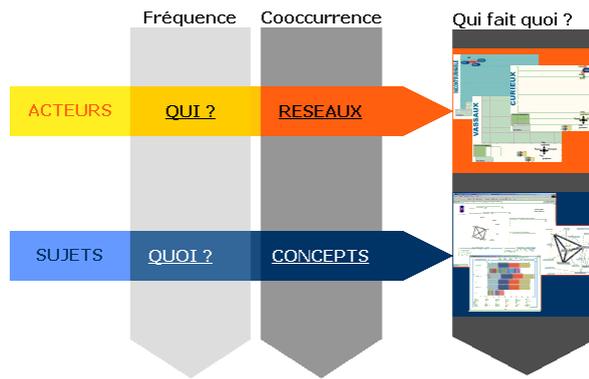


Figure 2 : « L'Aiguillage » qui permet d'orienter initialement la recherche.

3.2 Le filtrage de l'information

Comment arriver à sélectionner de façon interactive, via le web, l'information pertinente pour l'utilisateur. Nous proposons tout un ensemble d'outils de filtrage basés sur l'utilisation des dictionnaires (thématiques, synonymes, hiérarchiques), des matrices (contingences, cooccurrences, présence absence), des tableaux 3D croisant le plus souvent deux variables et le temps. Nous pouvons activer un ou plusieurs filtres par champ afin de ne garder que l'information ponctuelle utile pour l'utilisateur tout en lui permettant de la croiser avec d'autres sur des volumes maîtrisables et compatibles avec les moyens classiques ou innovants des graphiques statistiques et géographiques proposés. Les filtres utilisés sont de deux types: unaires ils ne font qu'intervenir la distribution du champ concerné, binaires ils s'appuient sur les relations avec les autres informations du corpus et font donc intervenir dans leur calcul des opérateurs complexes comme la connexité, les liens transitifs, la consistance, l'équivalence, les coïncidences positives et négatives, les distances et autres métriques.

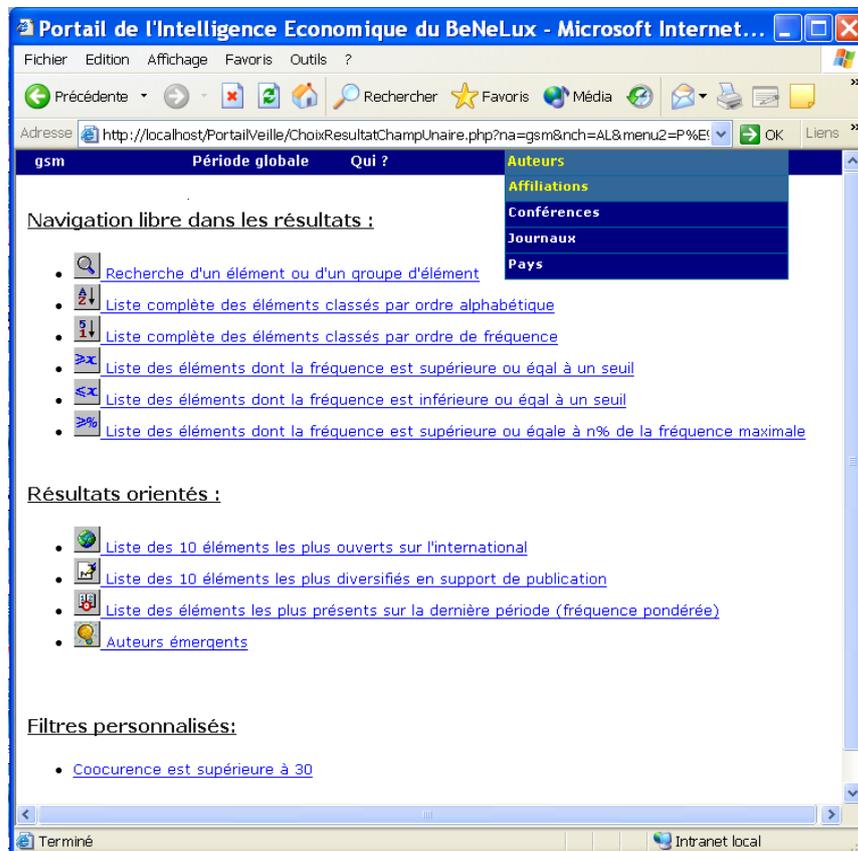


Figure 3 : Opérateurs de recherche d'un élément ou d'un groupe

3.3 Les sorties alpha-numériques

Le portail permet de rechercher par critère une information ou une collection d'informations sur un ou plusieurs champs de la base. Il est possible d'ordonner ces informations :

- Par ordre alphabétique (ascendant ou descendant)
- Par fréquences (croissantes ou décroissantes)

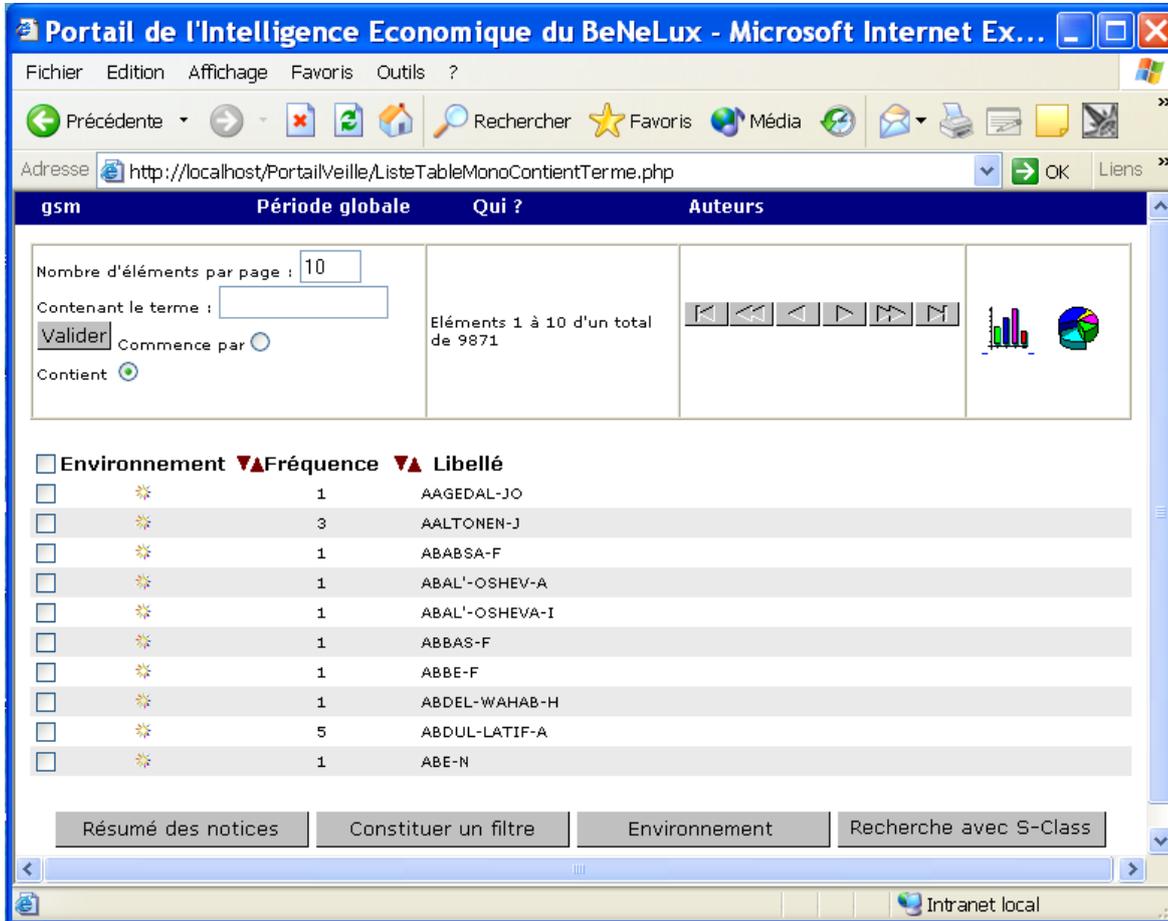


Figure 4 : Liste triée des éléments d'un champ avec accès à l'environnement.

3.4 Les sortie graphiques

Outre les grands classiques (histogrammes 2 et 3D, camemberts, boîtes à pattes, droite de régression, zoom de matrices,...), nous comptons intégrer des techniques de visualisation propres aux fonctions avancées du logiciel Tétralogie comme (cartes factorielles, arbres de classification, zoom 3D de matrices, fish eye, cartes géographiques interactives, ...). Cet ensemble de possibilités doit permettre à chacun de trouver les bons réglages pour découvrir puis communiquer l'information stratégique ciblée à intégrer dans son rapport d'analyse personnalisé. Les fonctions de "reporting" sont essentielles pour réussir la présentation d'un travail de veille et pour convaincre les décideurs par un document lisible, pertinent et concis.

3.4.1 Les histogrammes et camemberts (mono variable)

Ces histogrammes à barres horizontales (pour une meilleure lisibilité des libellés) possèdent une fonction de survol servant à visualiser les valeurs une par une. Une fonction de zoom permet d'avoir une vue d'ensemble ou de se focaliser sur les signaux forts.

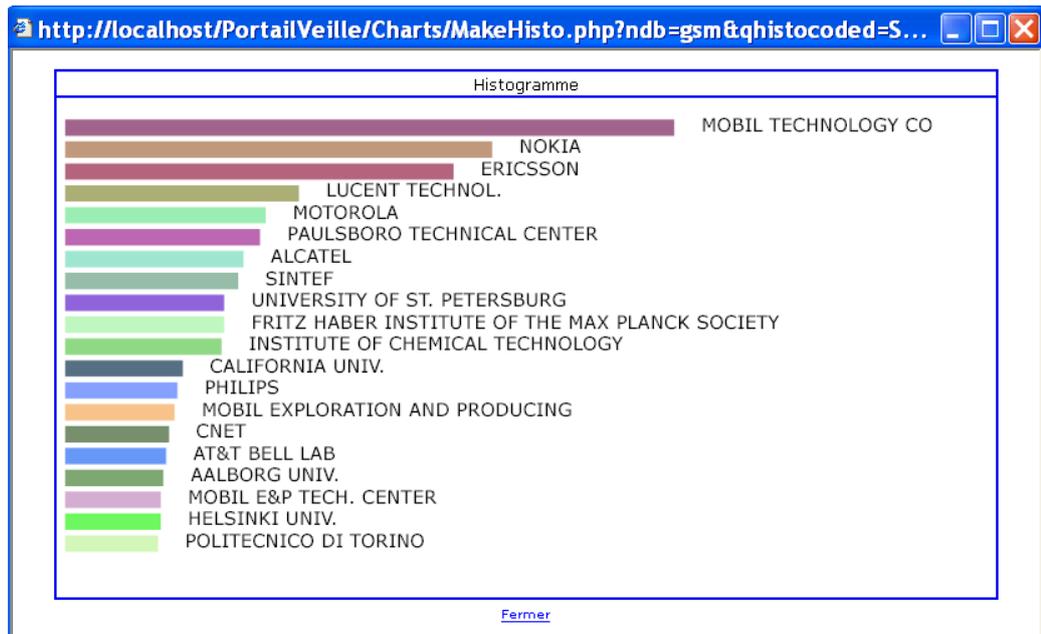


Figure 5 : Représentation sous forme d'histogramme des données extraites

Même fonctionnalité pour les secteurs ou camemberts.

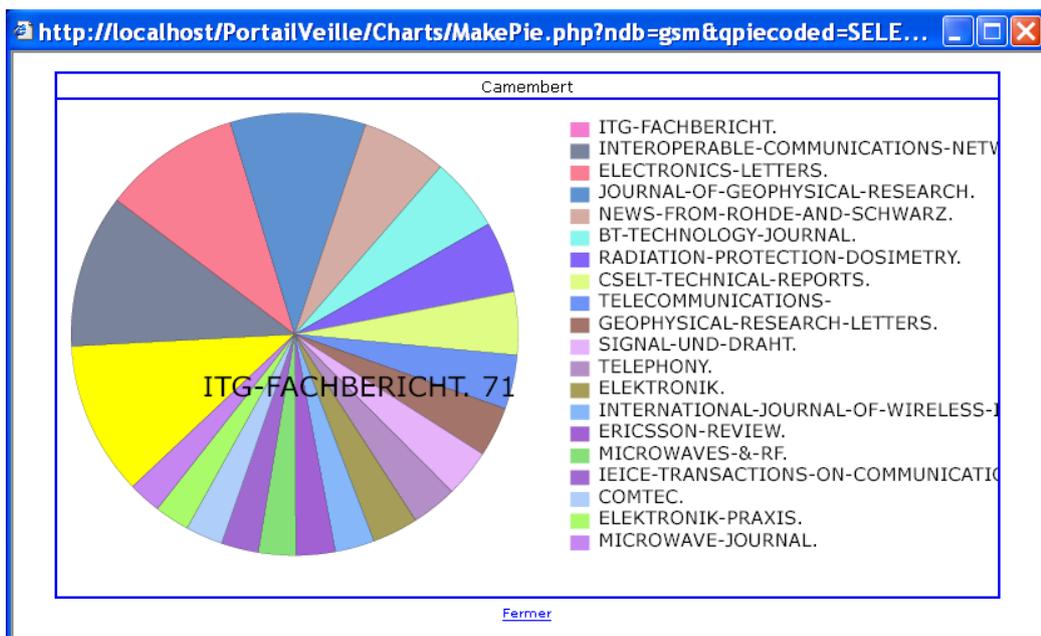


Figure 6 : Représentation des données filtrées sous forme de camembert

3.4.2 Le dessin de graphes et ses variantes (bi-variables)

Dans le cas de matrices (de réseaux ou de concepts), il est possible de visualiser les liens sous forme de graphes avec un niveau de transitivité réglable. Les noms des sommets et le graphe lui-même peuvent être zoomés, les arêtes soumises à des seuils haut et bas et la forme du graphe peut, au choix, être proposée sous forme circulaire, d'arbre ou de graphe 2D. La répartition de graphe est elle-même modifiable à l'aide de la souris, afin de pouvoir explorer des zones à forte densité.

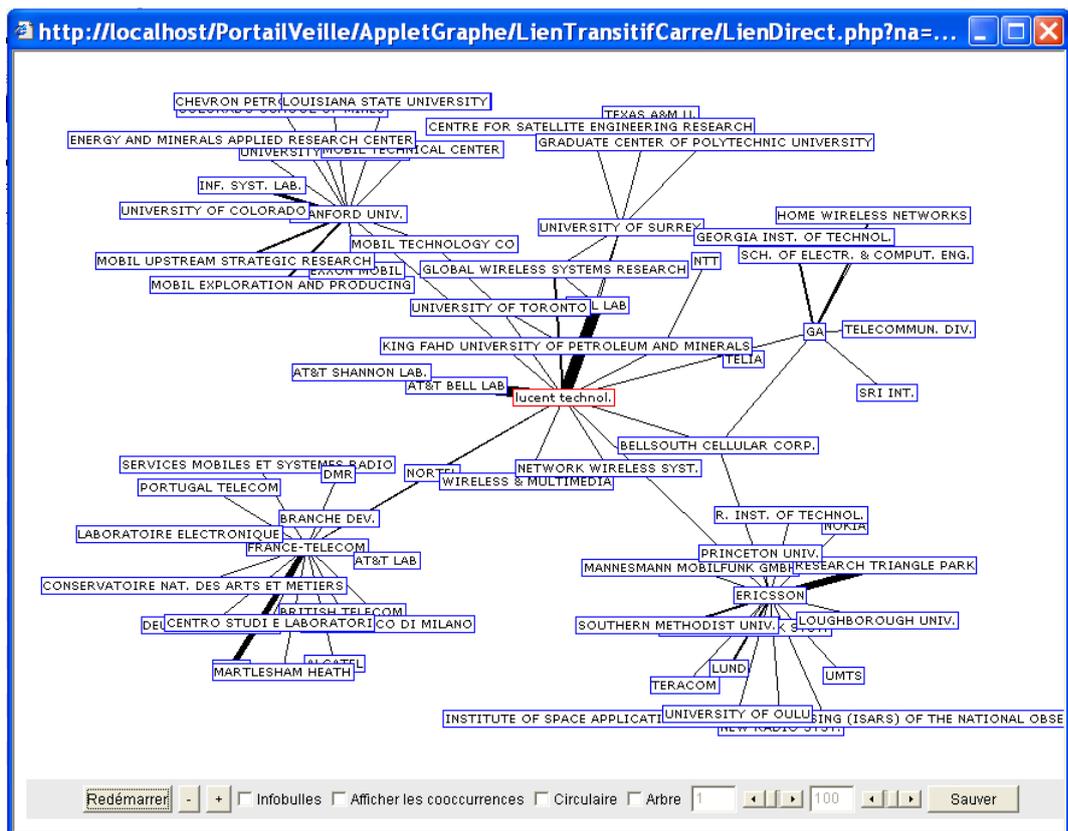


Figure 7 : Représentation d'un extrait de matrice sous forme de réseau de liens

3.4.3 Les cartes géostratégiques

Lorsqu'une des variables d'un tableau représente les pays, il est possible de demander une visualisation de la répartition de l'autre variable sur une carte du monde. Cette carte possède aussi des fonctionnalités de zoom et de survol afin de pouvoir identifier signaux forts et faibles à partir des données numériques et sur des pays de petite surface. La coloration de la carte est fonction de la valeur relative dans la sélection (en % du maximum), son échelle de dégradés peut être choisie linéaire ou logarithmique.

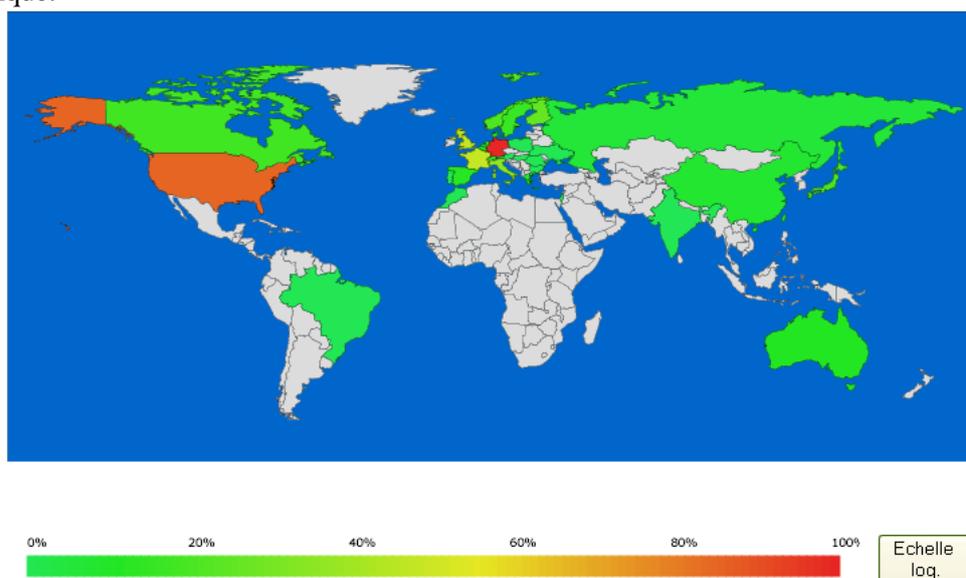


Figure 8 : Représentation d'un paramètre sous forme de carte géostratégique

3.5 L'analyse multidimensionnelle

3.5.1 L'accès aux différentes dimensions

Afin de pouvoir interagir avec toutes les dimensions d'un élément (ou d'un groupe d'élément) préalablement sélectionné, nous avons conçu une fonction appelée le Soleil, fonction qui nous permet de sélectionner tour à tour chaque dimension en vue d'identifier ses composantes les plus marquantes. Dans la figure ci-dessous, nous pouvons jouer avec les 10 dimensions de l'auteur sélectionné.

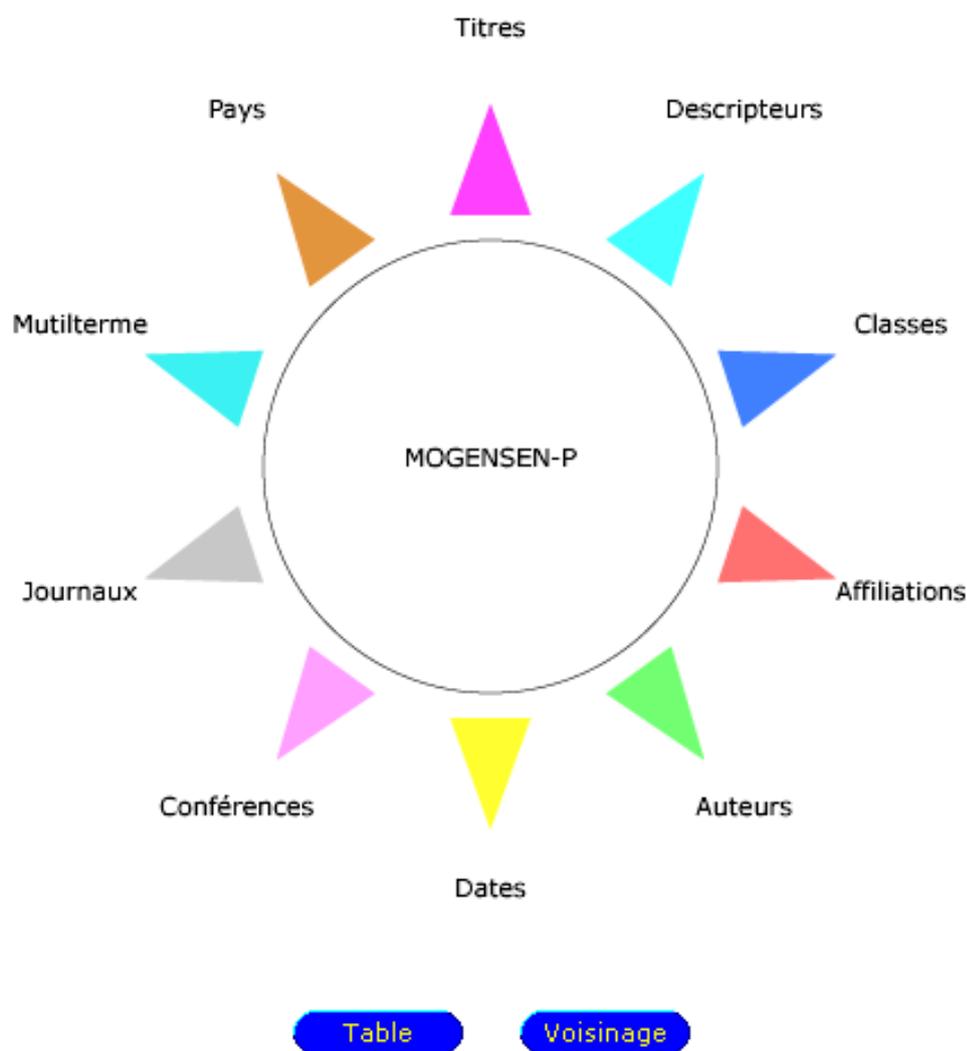


Figure 9 : « Le Soleil » : accès aux composantes multidimensionnelles d'une sélection

3.5.2 La synthèse comparative

Comme l'évaluation est souvent un des motifs des analyses pratiquées, nous proposons un comparatif, dimension par dimension, de la sélection étudiée avec les autres données contenues dans la base. Ainsi, un comparatif avec la moyenne, le minimum, le maximum, les n plus importants est proposé afin de situer la sélection par rapport à l'ensemble dont elle est extraite. En fonction de la demande, ce comparatif peut être aménagé en fonction des critères ou des indicateurs spécifiques de l'utilisateur. Nous présentons, ci-dessous, un exemple possible de ce comparatif déclenché par le bouton *Table*.

http://localhost/PortailVeille/TablePedigreeShort.php?na=gsm&nch=AF&ids=...

Affiliations	Fréquence de présence	Nombre de pays	Nombre d'auteurs	Nombre d'affiliations	Nombre de supports	Nombre de mots émergeants
MOBIL TECHNOLOGY CO	220	22	83	108	?	327
NOKIA	154	10	62	17	6	495
ERICSSON	140	12	65	17	3	422
LUCENT TECHNOL.	84	9	47	15	3	315
MOTOROLA	72	9	22	10	3	250
PAULSBORO TECHNICAL CENTER	70	18	18	48	?	97
MOBIL OIL	70	8	27	33	?	170
ALCATEL	64	9	24	5	6	226
SINTEF	62	18	13	41	-	97
UNIVERSITY OF ST. PETERSBURG	57	20	22	47	?	85
FRITZ HABER INSTITUTE OF THE MAX PLANCK SOCIETY	57	18	13	41	?	83
INSTITUTE OF CHEMICAL TECHNOLOGY	56	18	13	41	?	83
CALIFORNIA UNIV.	42	1	29	7	3	148
PHILIPS	40	6	7	3	1	119
MOBIL EXPLORATION AND PRODUCING	39	7	30	29	?	115
CNET	37	1	18	5	3	130
AT&T BELL LAB	36	2	36	3	3	182
AALBORG UNIV.	35	2	16	6	3	143
MOBIL E&P TECH. CENTER	34	2	17	12	?	90
HELSINKI UNIV.	34	2	23	10	3	134

[Fermer](#)

Figure 10 : La fonction table comparative

Une seconde méthode d'investigation est l'analyse multidimensionnelle qui permet, pour chaque élément sélectionner, de connaître son voisinage c'est à dire l'ensemble des autres éléments qui lui sont connexes dans chaque dimension. On active cette fonction par le bouton *Voisinage*.

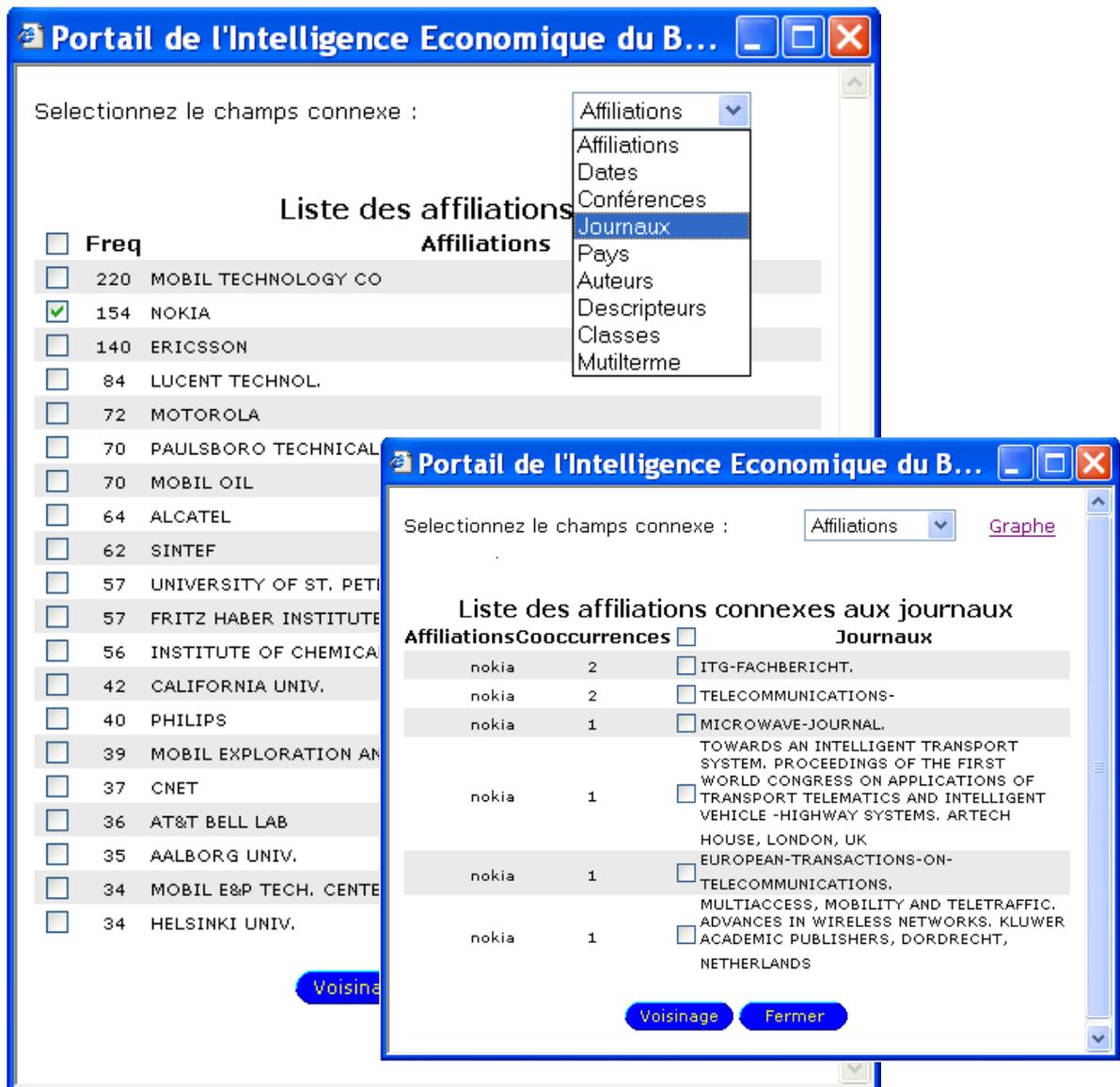


Figure 10 : La fonction voisinage

4 Conclusion

Avec ce nouvel outil, nous couvrons l'ensemble des sources disponibles à l'heure actuelle : à savoir: les bases documentaires en ligne ou sur CD/rom (comme Medline, Inspec, Current contents, Biosis, Pascal, Sci, Chemical abstract, ...), les pages web, les news groups, les traces de sites, la presse, les brevets (Ibm, Uspto, Derwent, Inpi, ...), les dépêches d'agences, le non structuré, ... Les utilisateurs vont ainsi pouvoir naviguer dans leurs analyses par des techniques qu'ils maîtrisent maintenant très bien (InterNet, les statistiques descriptives, le filtrage, les fonctions de reporting). De plus, les analyses papier issues de Tétralogie et qui sont élaborées par des équipes d'experts (documentation, informatique, analyse de données et statistiques, domaine étudié) vont pouvoir être connectées par des liens hypertexte à ces bases de données en ligne et devenir de véritables documents hypertextes avec toutes les possibilités de raffinement voulues par l'utilisateur. La macro analyse ne servant plus que de fil conducteur aux investigations du lecteur qui, tout en ayant pris connaissance du contexte général de l'étude, ira chercher lui même les informations personnalisées et pertinentes qu'il est le seul à pouvoir débusquer dans la masse de données qui est maintenant correctement indexée et commentée.

De plus, cet outil doit permettre de s'affranchir des problèmes de volume rencontrés sur des corpus géants comme par exemples ceux que l'on constitue pour analyser exhaustivement un domaine (classes de brevets, littérature scientifique pour un laboratoire) ou pour évaluer le positionnement d'un

organisme de recherche (Inra, Cnrs, Inserm, Cea, Université) ou d'un grand groupe industriel ou commercial. En effet, dans ces cas extrêmes, les volumes des matrices sont tels que la mémoire vive des machines modernes est dépassée et donc que les temps de réponse deviennent absolument prohibitifs. Un autre avantage est de pouvoir plus facilement valider l'information brute obtenue et notamment réaliser une validation des propositions de synonymes (comme ceux des adresses ou des noms d'auteurs) en s'appuyant sur des coïncidences faisant simultanément intervenir l'ensemble des champs significatifs du corpus étudié.

5 Bibliographie

J. Mothe, C. Chrisment, T. Dkaki, B. Dousset, D. Egret

"Information mining: use of the document dimensions to analyse interactively a document set." " 23rd BCS European Colloquium on IR Research: ECIR, Darmstadt. BCS IRSG, pp 66-77, 4-6 avril 2001.

J.-L. Multon, G. Lacombe, B. Dousset

"Analyse bibliométrique des collaborations internationales de l'INRA". Veille stratégique, scientifique et technologique : VSST'01, Vol 1, pp 261-270, (Barcelone, Espagne), octobre 2001.

D. Sosson, M. Vassard, B. Dousset

"Portail pour la navigation en ligne dans les analyses stratégiques." " Veille stratégique, scientifique et technologique : VSST'01, Vol 1, pp 347-358, (Barcelone, Espagne), octobre 2001.

B. Dousset, S. Karouach

"Collaboration interactive entre classifications et cartes thématiques ou géographiques." " 9^{èmes} rencontres de la société francophone de classification. (Toulouse France), 16-18 septembre 2002.

J.-L. Multon, G. Lacombe, B. Dousset

"Analyse bibliométrique des collaborations internationales de l'INRA". 9^{èmes} journées d'études sur les systèmes d'information élaborée: Bibliométrie - Informatique stratégique - Veille technologique, (Ile Rousse Corse France), CD-ROM, 14-18 octobre 2002.

J. Mothe, C. Chrisment, B. Dousset, S. Karouach,

"Représentation des documents textuels : étude d'un domaine à travers des publications". 5^{ème} Congrès de la société française de recherche opérationnelle et d'aide à la décision ROADEF. (Avignon France), pp 130-131, 26-28 février 2003.

S. Karouach, B. Dousset

"Les graphes comme représentation synthétique et naturelle de l'information relationnelle de grande taille". Workshop sur la recherche d'information : un nouveau passage à l'échelle, associé à INFORSID'2003, (Nancy France), 3-6 juin 2003.

J. Mothe, C. Chrisment, B. Dousset, J. Alaux

"DocCube : multi-dimensional visualization and exploration of large document sets". Journal of the American Society for Information Science and Technology JASIST, Special topic section: "Web Retrieval and Mining". Guest Editor: Hsinchun Chen, 54(7), pp 650-659, March 2003.

S. Karouach, B. Dousset

"Visualisation de relations par des graphes interactifs de grande taille". ISDM Journal of Information Sciences of Decision Making, vol. 6, Mars 2003.

S. Karouach, B. Dousset

"Analyse d'information relationnelle par des graphes interactifs de grandes tailles". 4^{èmes} journées d'EGC (Extraction et Gestion de Connaissances), Clermont Ferrand, 20-23 janvier 2004.

C. Chrisment, B. Dousset, S. Karouach, J. Mothe

"Information mining : extracting, exploring and visualising geo-referenced information". Workshop on Geographic Information Retrieval, SIGIR 2004.