XPLOR : un portail pour la navigation en ligne dans les analyses stratégiques

Didier SOSSON , Mathias VASSARD
GFI Bénélux, Luxembourg,
didier.sosson@gfi.be, mathias.vassard@gfi.be

1 Introduction

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 experts en analyses. Cependant, 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 où 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 de quelques documents pris séparément. Il devient ainsi possible de retrouver, à partir d'un élément connu (acteur, mot clé), 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, 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 lui même et les autres champs de la base. Des statistiques interactives sont alors disponibles pour chaque extrait (fréquences, équivalences, liens pondérés, …) ainsi que des cartes ou des réseaux (relationnels, 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.

2 Les différents modes d'implantation

2.1 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 utilisateur 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 projètent. Une analyse peut être revisitée par différents spécialistes du domaine et apporter à chacun des réponses précises aux questions stratégiques et parfois confidentielles 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 antérieurs. 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.

2.2 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 remises 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.

2.3 Extraction de matrices depuis la base de données

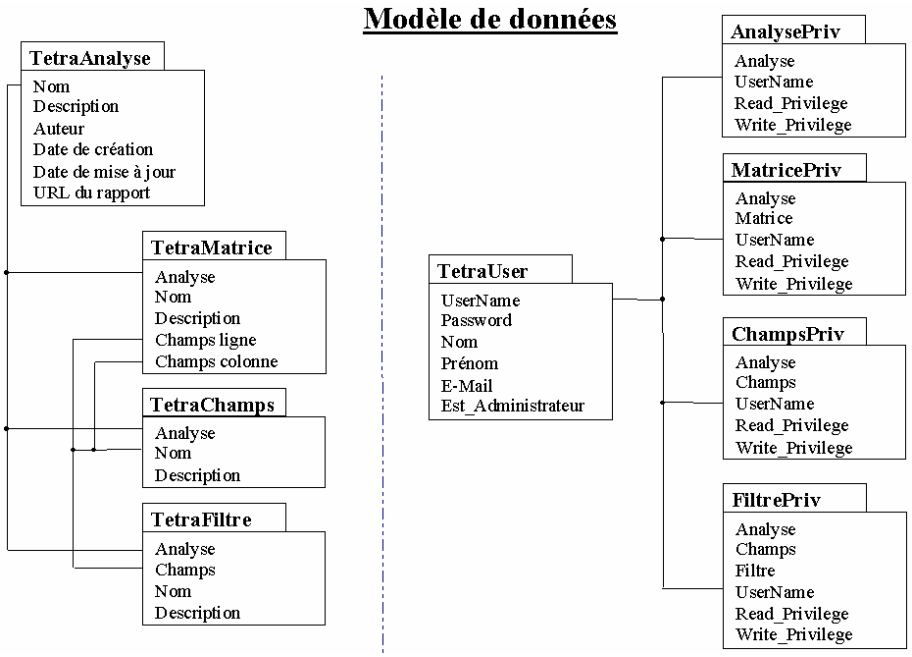
Si nous voulons revenir à Tétralogie, pour réaliser une analyse totale ou partielle du contenu de la base, il faut générer des matrices au format requis. Ceci est effectué par un module d'extraction qui permet éventuellement de réduire la portée des opérateurs de croisement à un sous ensemble du corpus (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 génération 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.

3 Nouvelles procédures d'extraction et de stockage

3.1 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. Ils sont identifiés dans une table des utilisateurs, leurs accès sont sécurisés par mot de passe. Les analyses sur lesquelles 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, un rapport d'analyse sous forme électronique (.doc, .html) est associé à une base de données, 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.



3.2 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,…).

3.3 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.

3.4 Extraction des matrice pour Tétralogie

Afin de conserver la possibilité d'analyser le corpus, comme avant, par l'ensemble des méthodes implantées dans Tétralogie, nous générons les matrices nécessaires depuis la base de données. Dans la grande majorité des cas, l'extraction de matrice ne pose pas de problème, ni de taille, ni de contenu. On peut en effet générer des matrices 2D de contingence et de présence absence, et dans la version 3D, des matrices comportant jusqu'à quatre plans consignant les données sur quatre périodes initialement prédéfinies au moment de la constitution de la base. Cette dernière particularité est essentielle, car il est impossible de faire de la veille si le paramètre temps n'est pas omniprésent dans la démarche d'analyse et notamment pour l'étude des corrélations et des croisements entre deux entités. Cette procédure présente tout de même un inconvénient: on ne peut, en effet, générer pour l'instant des matrices basées sur d'autres mesures que les cooccurrences (comme: proximités de portées variables, coïncidences totales, ordres d'apparition, …). A terme, nous comptons aussi stocker ce type d'information dans la base, afin de pouvoir proposer de nouveaux opérateurs et de ne pas limiter la portée de nos analyses.

4 Interactivité de la nouvelle structure de données

4.1 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. Les filtres utilisés sont de deux types: unaires ils ne font intervenir que 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.

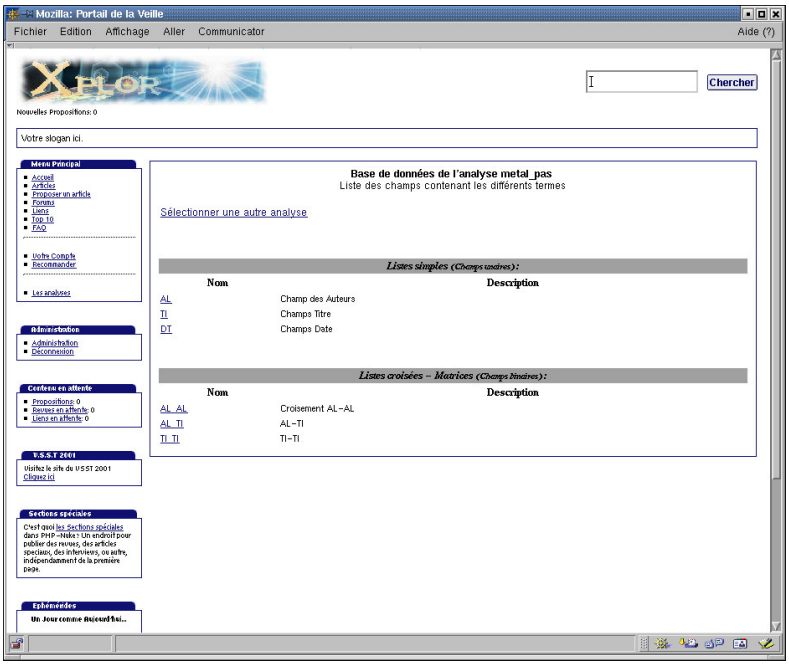


Figure 1. : Sélection d'un champ ou d'une matrice

4.2 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)
- Par numéros d'ordre dans la base

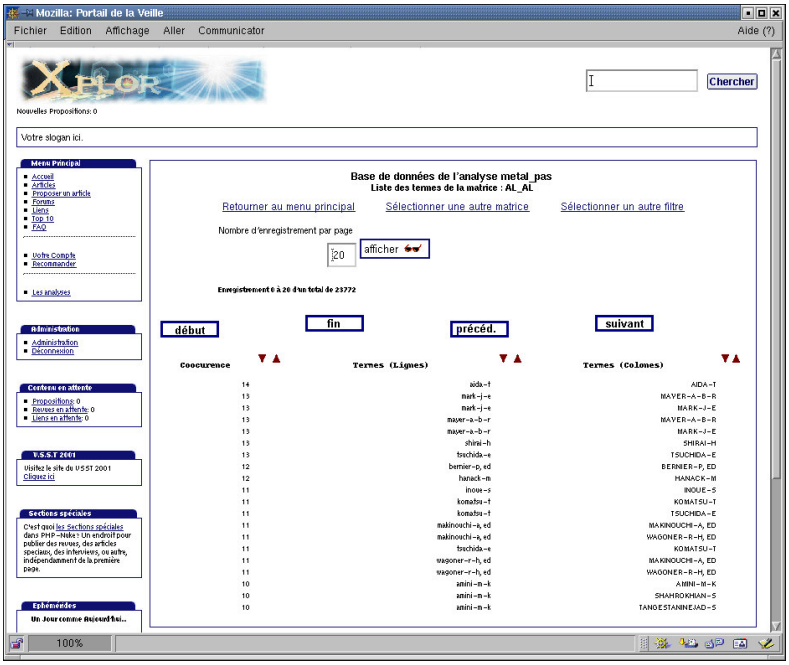


Figure 2. : Modes de présentation des dictionnaires

4.3 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.

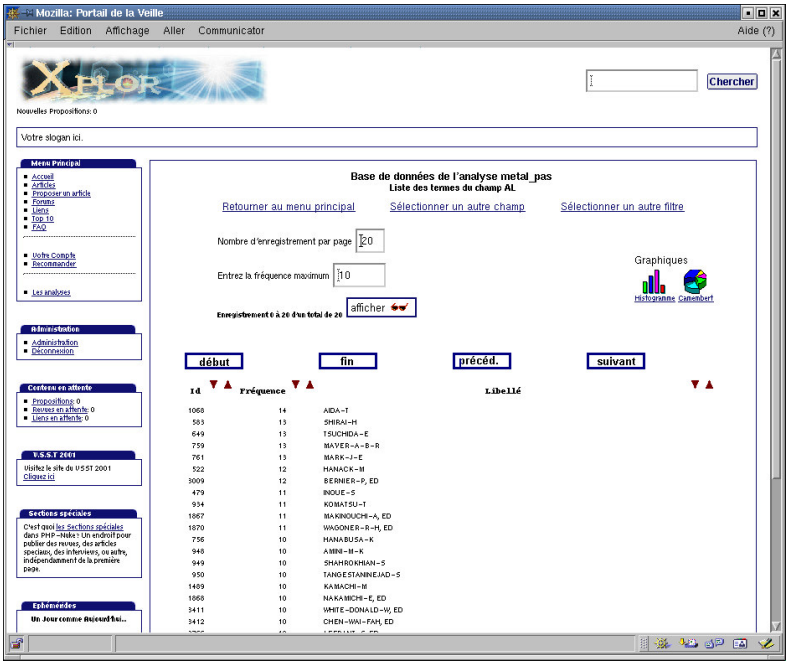


Figure 3. : Fonction de filtrage sur un champ de la base

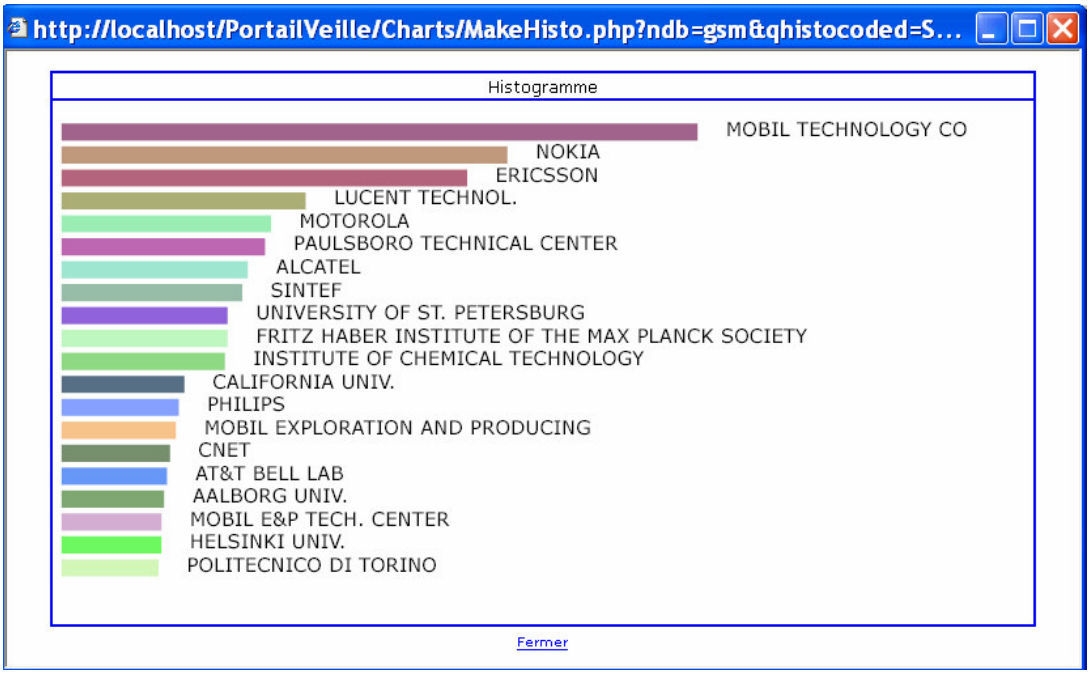


Figure 4. : Représentation sous forme de barres directement issue de la base de données

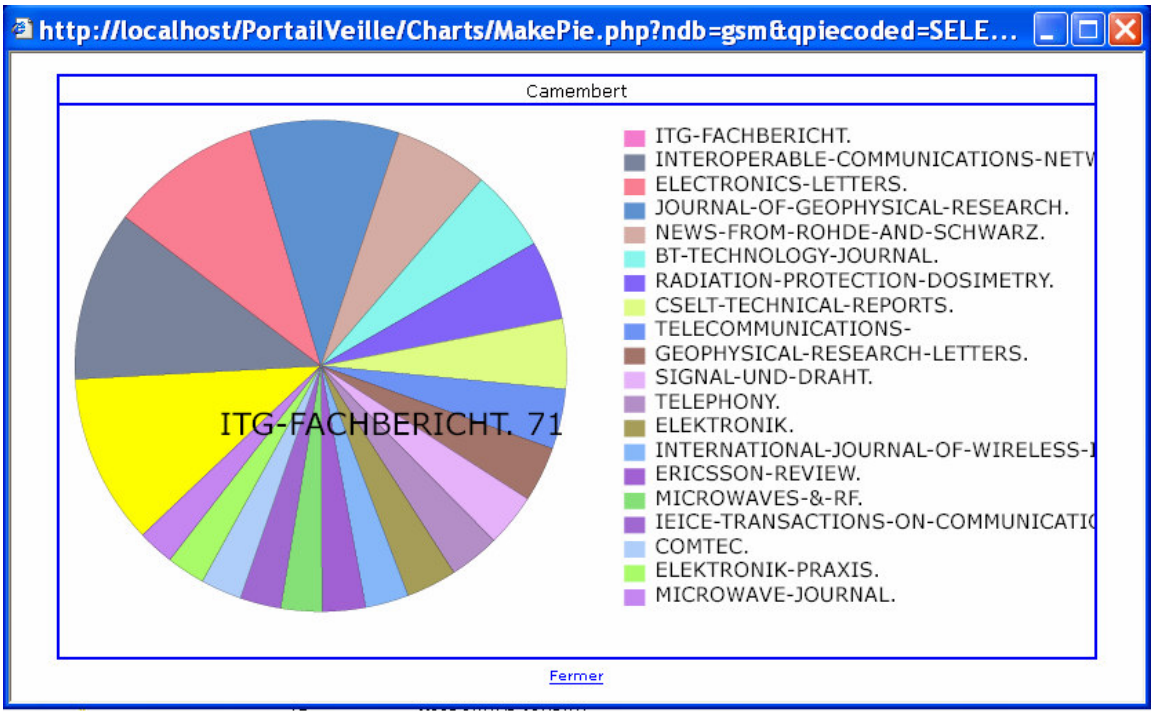


Figure 5. : Représentation en ligne des données filtrées sous forme de camembert

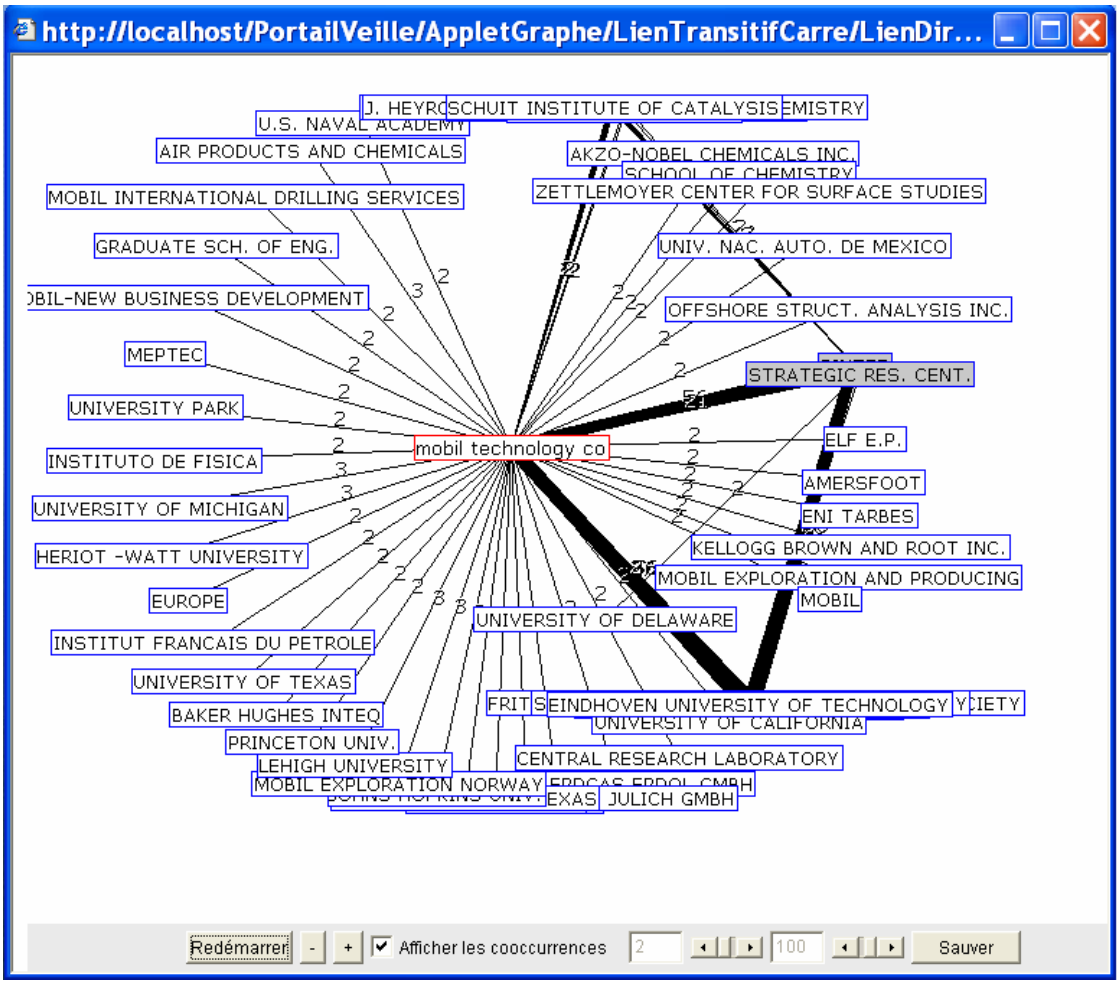


Figure 6. : Représentation d'un extrait de matrice sous forme de réseau de liens

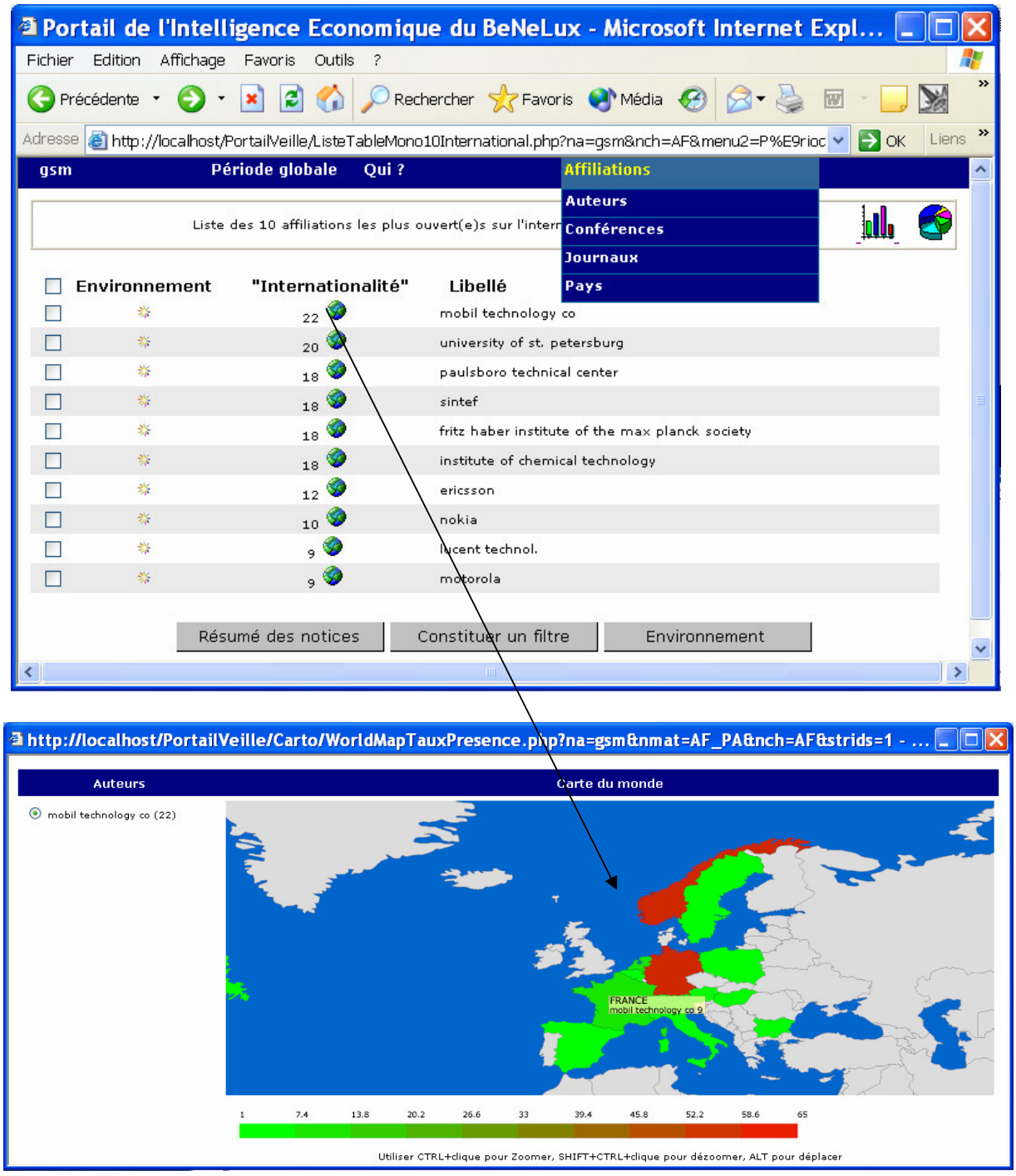


Figure 7. : Génération de cartes stratégiques

5 Conclusion

Le prototype de ce nouvel outil est en cours d'expérimentation sur un très large panel d'analyses que nous avons déjà effectuées à partir du logiciel Tétralogie. 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 hypertextes à ces bases de données en ligne et devenir de véritables documents hypertextes avec toutes les possibilités de raffinage 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 proposition 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.