Portail pour la navigation en ligne dans les analyses stratégiques

Bernard DOUSSET (*), Didier SOSSON (**), Mathias VASSARD (**)
dousset@irit.fr , sosson@primesphere.com , vassard@primesphere.com

(*) IRIT/Université Paul Sabatier, 118 route de Narbonne 31062 Toulouse cedex 4
(**) Primesphere, 4 rue Jos Felten BP 1078 L-1010 Luxembourg,

Mots clés:
Analyse de données, analyse relationnelle, veille stratégique, base de données relationnelle, Internet, navigation hypertexte, équations de recherche relationnelles, MySql

Key words:
Data analysis, relational analysis, strategic watch, relational data bases, Internet, hypertext navigation, relational search equation, MySql

Palabras claves:
Analisis dato, analisis relacional, vigilencia estratégica, base de dato relacional, Internet, navegacion hipertexto, ecuacion busqueda relacional, 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 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 toutes 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.

1 Présentation de "Tétralogie"

1.1 Aperçu général

"Tétralogie" est un système de découverte de connaissances dans des bases textuelles ou factuelles. développé à l'Institut de Recherche en Informatique de Toulouse (IRIT). Il propose plusieurs types de méthodes pour détecter et diffuser l'information stratégique cachée dans de grandes masses de données souvent hétérogènes.
Il est essentiellement basé sur des méthodes de statistiques, de théorie des graphes et d'analyse de données appliquées sur des tables de contingence ou de présence-absence croisant deux ou trois entités (acteurs, mots clés, dates) préalablement traitées par une analyse sémantique.
Tétralogie est un système ouvert de sorte que de nouvelles fonctionnalités peuvent être facilement rajoutées. La restitution des résultats est réalisée sous la forme de visualisations graphiques (histogrammes, arbres de classification, cartes factorielles 2D, 3D et même 4D, cartes géographiques, zooms de matrices triées, réseaux de liens,…).
En simplifiant, Tétralogie est essentiellement composé de deux modules :

- Un module d'extraction d'informations (terminologie, auteur, date…) et de création des tables de contingence (fréquence, cooccurrence, présence-absence) à partir des corpus des bases documentaires. Il tient compte des spécificités de chaque base et de chaque format.
- Un module qui applique différentes méthodes statistiques et d'analyse des données sur les contenues de ces tables. Ce système aide aussi à l'élaboration des comptes rendus d'analyses générés par les diverses méthodes utilisées.

1.2 Descripteur de structure

Pour pouvoir traiter les bases documentaires sous leur format initial et sans les reformater, un descripteur paramétrable de la structure des bases a été créé.
Ce descripteur défini le format de lecture d'une base. Il contient la définition de la structure de la base en retenant les champs (auteur, titre, résumé,…) qui vont être utilisés ultérieurement dans l'analyse de données. Ainsi, il existe un descripteur par format de base. Ce descripteur est constitué de cinq colonnes :

- nom : désigne la définition du champ en texte libre.
- abrev : désigne l'abréviation du champ. Elle permettra ensuite de faire référence de façon standard à ce type de champ pour une exploitation multi-bases.
- champ : bannière d'un champ dépendant du format de la base.
- visible : détermine si le champ spécifié doit être pris en compte ou pas durant les études ultérieures (par les valeurs True ou False).
- séparateur : définit les séparateurs utilisés pour l'extraction des éléments pertinents des différents champs de données (les éléments de ponctuation, l'espace ou autre) en fonction du niveau de découpage désiré. Pour un même champ, ces séparateurs peuvent différer d'une base à l'autre.

Exemple de descripteur de structure (Base ELSEVIER/Internet)

Notice
descripteurs de Elsevier internet

# nom	abrev	champ	visible	Separateurs #
MTM	MT	MTM:	True	b"
Titrel	TI	Title:	True	"
Titrec	Ti	Title:	True	.",";":"b"(")"["]"sb"s."s,"s:"s;""|"
Journal	JN	Journal:	True	Vol."
Date	DP	Date:	True	,"-"ORD0"
Dates2	DT	Date2:	False	"
Keyword	DE	Keywords:	True	;"

Classif	CL	Classification:	True	;","
Auteur_lg	AL	Auteur:	True	;"\n"
Adresse	AD	Organisme:	True	"
Organisme	OR	Organisme:	True	,"and";"
Pays	PA	Organisme:	True	,";"
Ref	RE	Reference:	False	"
Resume	AB	Abstract:	True	.","":"b"(")"["]"sb"s."s,"s:"s;"""|"
FIN	FIN	FIN	FIN	"

1.3 Analyse du contenu, correction et filtrage

Pour extraire le contenu de chaque champ, Tétralogie utilise une fonction de dénombrement dont l'objectif est de dresser une liste des items trouvés en précisant leur fréquence d'apparition dans le corpus. Afin de ne garder que l'information utile, ce processus est complété par une sélection pilotée par deux types de filtres :

- négatif, il permet d'éliminer des éléments d'information non souhaités,
- positif, il permet de ne garder que certains éléments pertinents.

Par exemple, un filtre contenant une liste de noms de pays permettra d'extraire les différents termes contenus dans le champ adresse du corpus et ainsi de créer un champ PAYS virtuel qui sera très utile pour évaluer ensuite la géostratégie du domaine étudié. Une homogénéisation des différentes orthographes rencontrées sera alors nécessaire afin d'assurer l'unicité de la terminologie retenue. Sur ce point et malgré nos efforts répétés, il n'a pas été possible d'automatiser totalement cette fonction. Il ne faut pas ici perdre de vue que nos sources sont des bases documentaires hétérogènes, qu'elles n'ont pas été prévue au départ pour un traitement informatique et que toute étude doit passer inévitablement par une phase de correction manuelle.

1.4 Cas particulier du texte libre

Dans l'analyse du texte libre, un terme peut être un radical, un mot simple ou un groupe de mots. Dans ce dernier cas, nous le définirons comme une succession (liste ordonnée) de mots simples et de connecteurs (exemple : le tiret, le blanc,…). Plusieurs travaux de recherches ont montré que la prise en compte de groupes de mots (mots composés, syntagmes, associations de mots simples ou unitermes) permet d'avoir des unités syntaxiques ayant une meilleure valeur sémantique que des mots simples isolés et notamment dans l'analyse scientifique, technique, juridique ou médicale. La notion de groupes de mots (ou multi-termes) est alors un moyen d'expression des associations d'idées et de concepts. Nous pouvons donner comme exemple: informatique de gestion, base de données…
Différentes méthodes de repérage des groupes de mots ont été définies. [LEWIS, 92] et [CHURCH, 88] utilisent la prise en compte de relations syntaxiques. D'autres, utilisent des méthodes statistiques de mesure de cooccurrence [STEIER et al., 93] ou des méthodes de repérage d'associations d'unitermes par positionnement [RAZOUK, 90]. Des méthodes basées sur la fréquence d'apparition des groupes sont également utilisées [DKAKI, 93].

Pour détecter les multi-termes, Tétralogie utilise une approche statistique et éventuellement un dictionnaire prédéfini de groupes de mots. Ce dictionnaire dépend bien sûr du domaine traité et chaque base documentaire ou chaque domaine peut posséder son propre dictionnaire.
Statistiquement, si une chaîne de caractères (mots vides éliminés et synonymes pris en compte) a été répétée plus d'un certain nombre de fois dans les champs textuels (exemple : résumé et titre) de la base traitée, elle est considérée comme un multi-terme. En fait, ce nombre représente un seuil qui doit être fixé par l'utilisateur en fonction de la taille de la collection de documents étudiée. Plus la taille de cette collection est grande, plus le seuil doit être grand. Par contre, les groupes de mots du dictionnaire prédéfini seront codés comme multi-termes même si leur fréquence ne dépasse pas le seuil.
Tétralogie réécrit alors les différents champs textuels pour chaque document de la base. Ainsi, pour un document donné, chaque champ textuel aura l'équivalent de son contenu réécrit en utilisant uniquement les multi-termes détectés. Dans le cas particulier des multi-termes, nous devons encore passer par l'analyse sémantique de Tétralogie pour générer le champ d'indexation MTM:, la base de données ne peut donc, dans ce cas, être entièrement générée par notre nouvel extracteur.

1.5 Tables de contingence et autres matrices

Tétralogie analyse essentiellement des tables de contingence et, éventuellement de présence/absence ou de fréquence absolue. Ces tables permettent de mettre en correspondance différents champs d'une base documentaire pour confronter leur contenu et étudier les corrélations entre leurs items. Pour la majorité des champs rencontrés, l'utilisation de ces tables ne présente aucun inconvénient car elles restent de taille acceptable lorsqu'elles croisent par exemple : pays, villes, journaux, organismes, dates, classifications, codes, …Mais dès qu'il s'agit de prendre en compte les auteurs, les mots clés, les multi-termes ou les documents eux même, la taille des matrices nécessaires explose surtout si ces champs pléthoriques en items sont croisés entre eux. Des choix limitatifs s'imposent : coupure à un niveau de fréquence donné (génération automatique de ces filtres), focalisation sur un sous ensemble (équipe, collaboration, sujet précis, sous corpus, terminologie émergente ou items très typés en terme de distribution dans les documents) et, dans certains cas, découpage en plusieurs matrices. Pour pouvoir conserver toute l'information utile, il devient nécessaire de changer de structure même si on dispose d'un Go de mémoire vive. Une autre limitation dans l'utilisation des matrices de croisement vient du fait qu'on ne peut pas y intégrer la totalité de l'information contenue dans le corpus qui comporte souvent plus d'une dizaine de champs significatifs. Tétralogie est très bien adapté pour le croisement des champs deux à deux en faisant éventuellement intervenir le temps comme troisième variable, d'où une analyse statique ou dynamique de leurs corrélations. En fait, chaque matrice permet de répondre à une question précise, comme dans les exemples suivants :

Matrice	Utilité
Auteurs - Auteurs	Elle fait apparaître l'ensemble des collaborations, leurs structures, les personnes qui les animent ainsi que les différentes équipes isolées du domaine.
Auteurs - Dates	Elle permet de connaître l'évolution de la productivité scientifique de chaque auteur dans le domaine étudié et de détecter les auteurs émergents ou ceux qui s'éloignent du sujet.
Mots clés - Dates	Elle montre l'évolution des problématiques de recherche, c'est à dire si un sujet a suscité beaucoup d'intérêts, s'il est innovant.
Auteurs - Pays	Elle fait apparaître les collaborations internationales des auteurs et indirectement entre Pays
Multi-termes - Journaux	Elle permet de retrouver les thématiques abordées dans les différents journaux
Auteurs - organismes	Elle permet de connaître les collaborations entre les organismes de recherche et éventuellement leur concurrence.
Mots clés - Auteurs	Elle fait apparaître les domaines de recherche des différents auteurs ainsi que les collaborations ou concurrences entre groupes d'auteurs sur des problématiques de recherche

Mais cette technique devient plus difficile à appliquer lorsqu'on veut travailler simultanément sur plusieurs variables peu homogènes entre elles comme, par exemple, le croisement des multi-termes avec, en simultané, les variables Auteurs, Journaux et Organismes. D'une part, la matrice obtenue est très volumineuse et le manque d'homogénéité nuit à l'interprétation. De plus une analyse globale d'une telle masse d'information n'a pas un réel intérêt. Par contre, trouver en une seule étape l'ensemble des auteurs, des journaux et des organismes liés à un thème défini par un nombre limité de termes spécifiques est beaucoup plus pertinent. De plus, cette question très ciblée ne peut en général pas être prévue par l'analyste aussi subtil soit-il, elle émane en fait de la démarche intellectuelle de l'utilisateur final (expert, ingénieur, technicien, décideur) et c'est donc lui qui doit pouvoir la poser directement au système. C'est pour cela que nous avons prévu, en parallèle aux analyses traditionnelles faites par Tétralogie, une nouveau système d'investigation plus proche de l'utilisateur et qui doit permettre de mieux tirer parti du travail de préparation, d'homogénéisation et d'analyse de l'information stratégique.

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 problématiqu larivet@esa.upmf-grenoble.fr

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 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 dynamique des tables de la base 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 ordinateurs 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 quatre périodes maximum et 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.

2.4 Extraction de matrices depuis la base de données

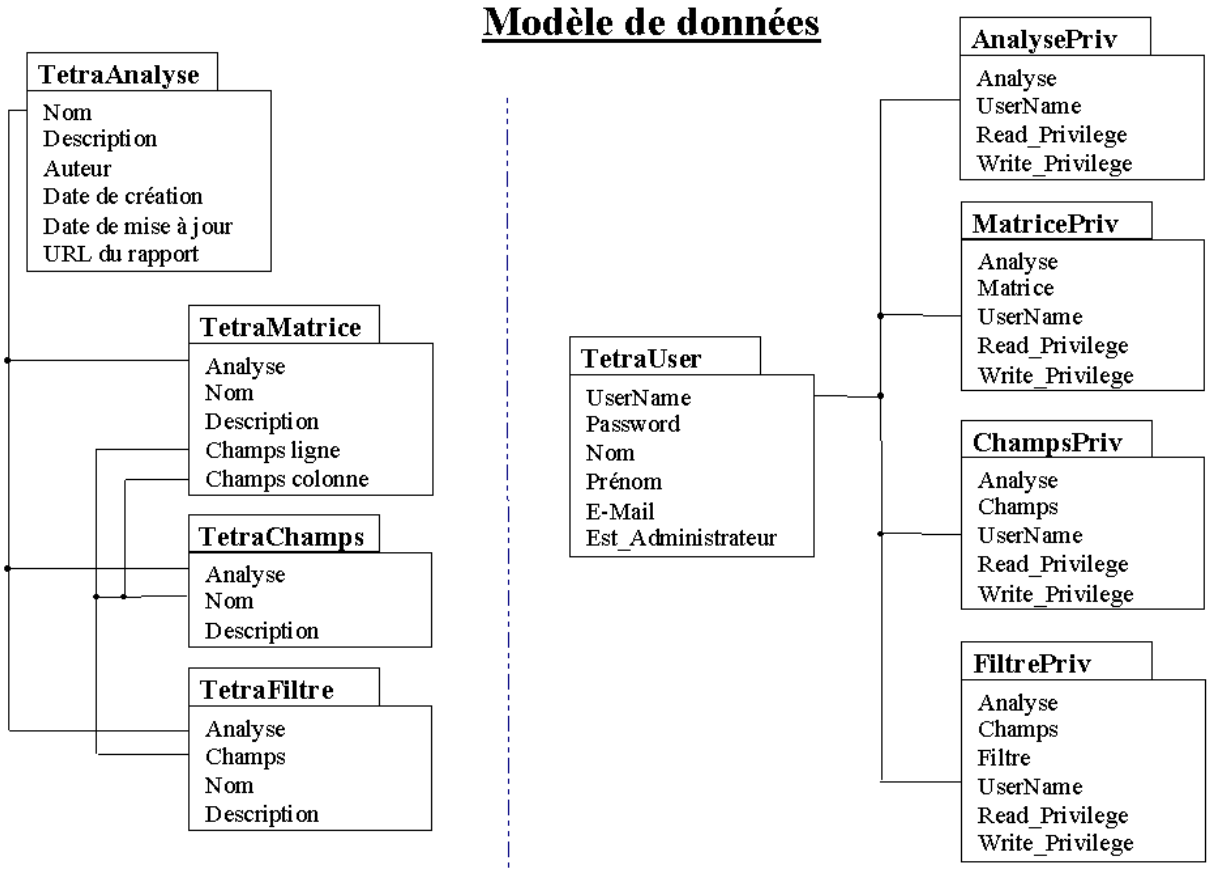
Si nous voulons revenir à Tétralogie, pour réaliser une analyse totale ou partielle du contenu de la base, il faut générer des matrices au format requis. Ceci est effectué par un module d'extraction qui permet éventuellement de réduire la portée des opérateurs de croisement à un sous ensemble du corpus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

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. Il 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 fonction 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

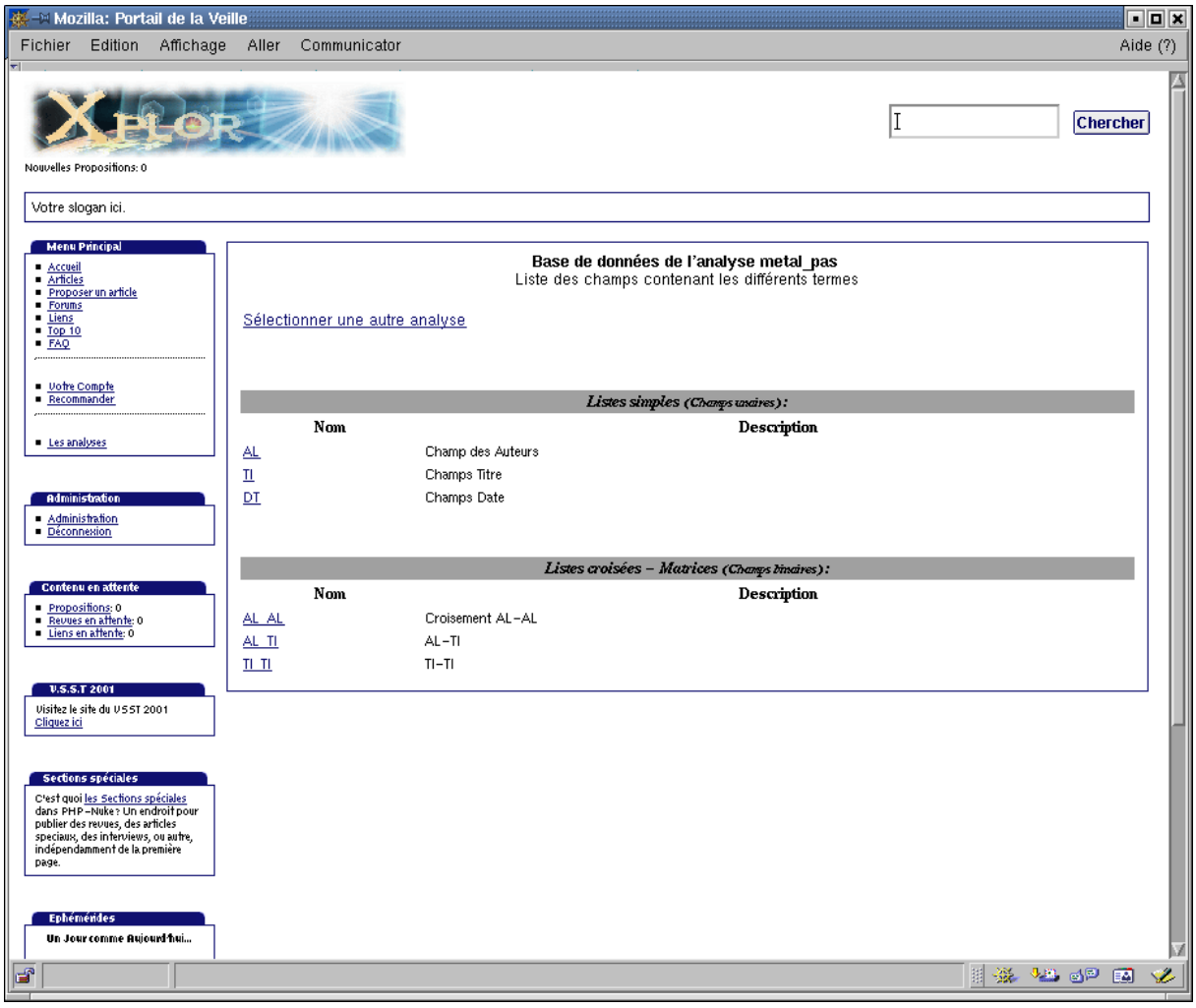
Afin de conserver la possibilité d'analyser le corpus, comme avant, par l'ensemble des méthodes implantées dans Tétralogie, nous générons les matrices nécessaires depuis la base de données. Dans la grande majorité des cas, l'extraction de matrice ne pose pas de problème, ni de taille, ni de contenu. 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.
La figure ci-dessous nous montre comment choisir l'objet (liste ou matrice) sur lequel portera le filtrage et éventuellement les sorties graphiques demandées en ligne :

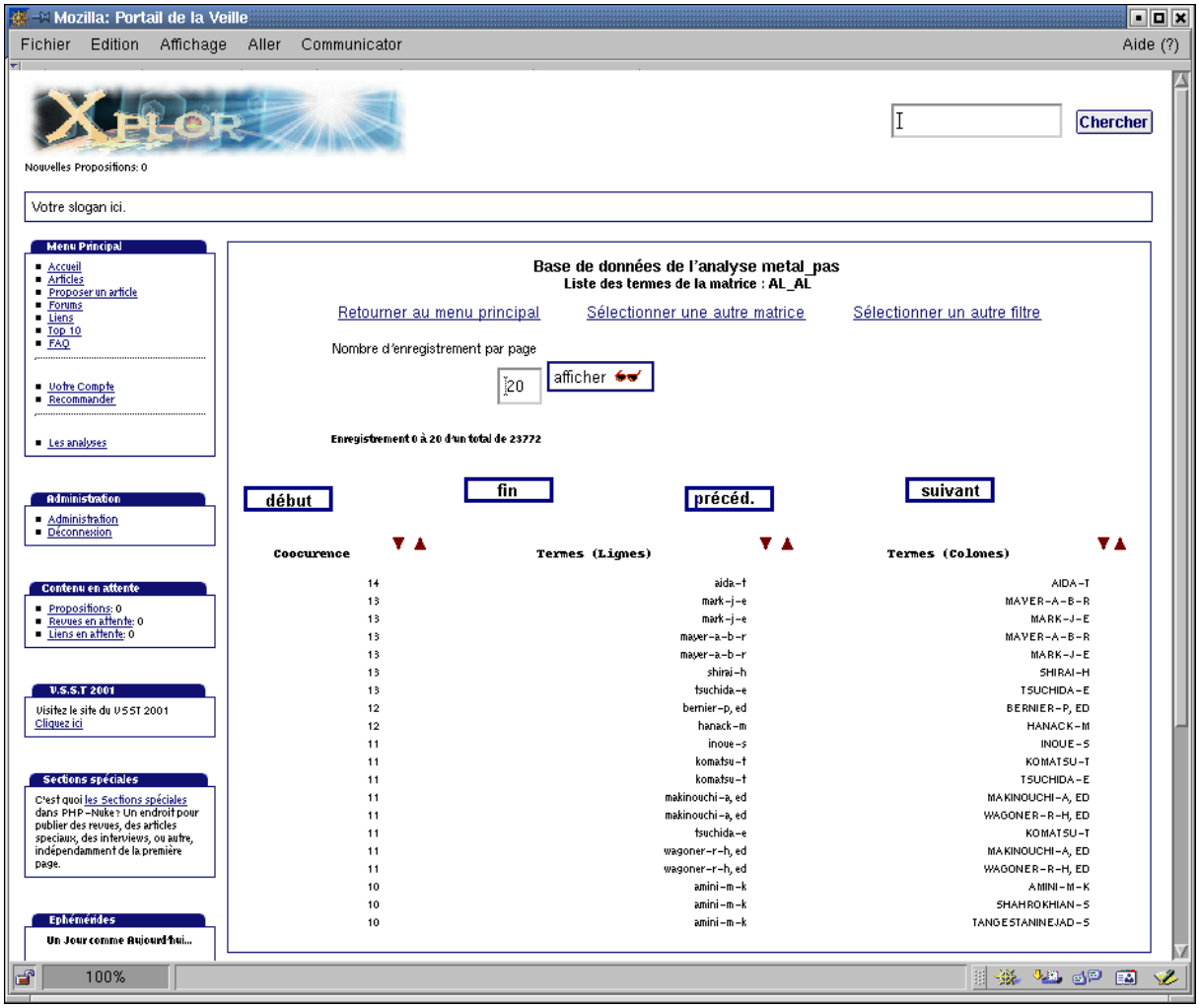


4.2 Les sorties alpha-numériques

Le prototype du 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 dans la base

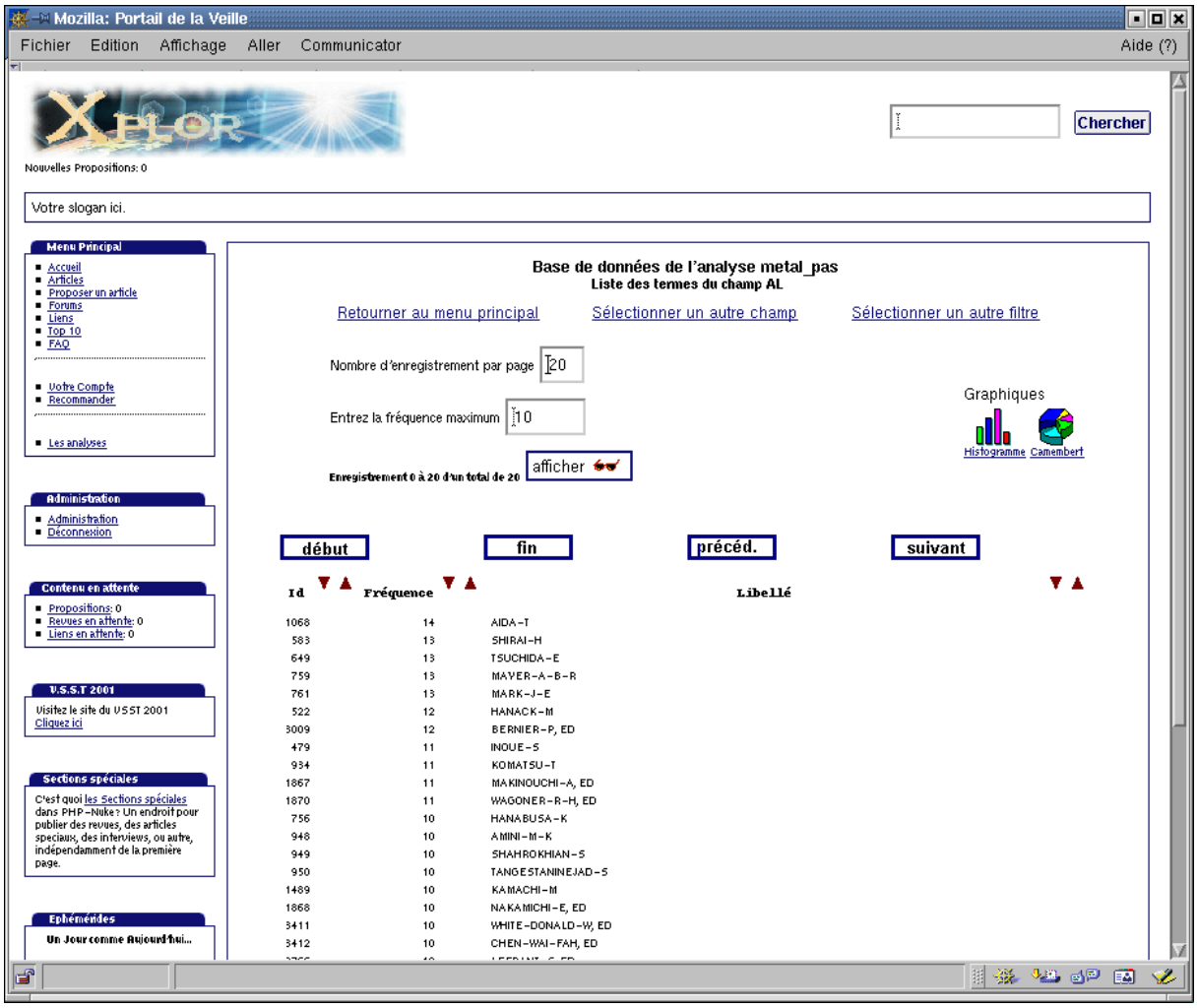
La figure suivante illustre cette fonction sur un champ simple :



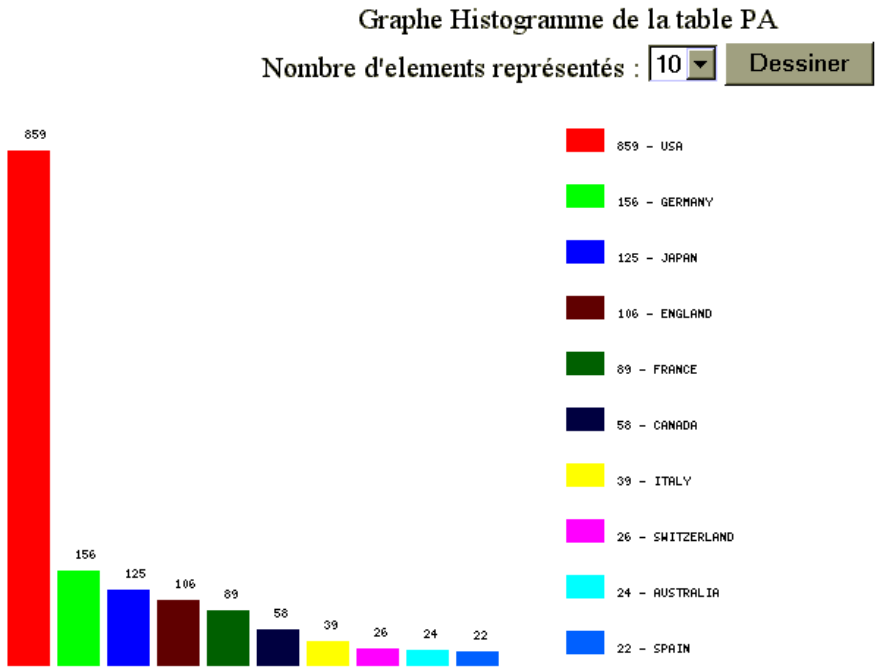
4.3 Les sortie graphiques

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 essentielle pour réussir la présentation d'un travail de veille et pour convaincre les décideurs par un document lisible, pertinent et concis.

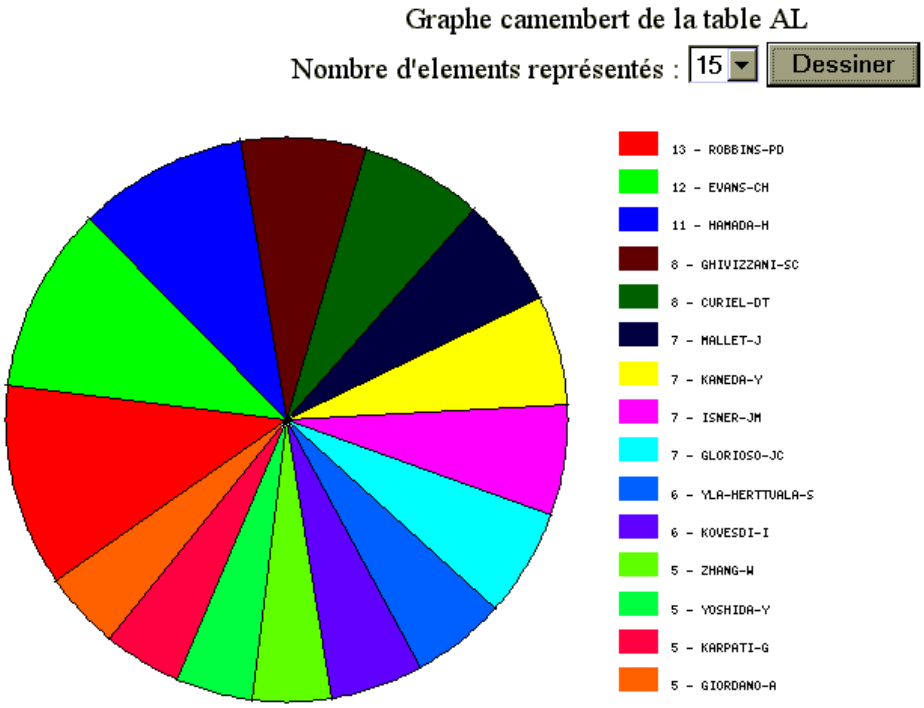
Dans la figure suivante, nous présentons la fonction de filtrage du prototype de notre portail qui permet de se concentrer sur les données utiles et la fonction de choix de la représentation graphique de ces données.



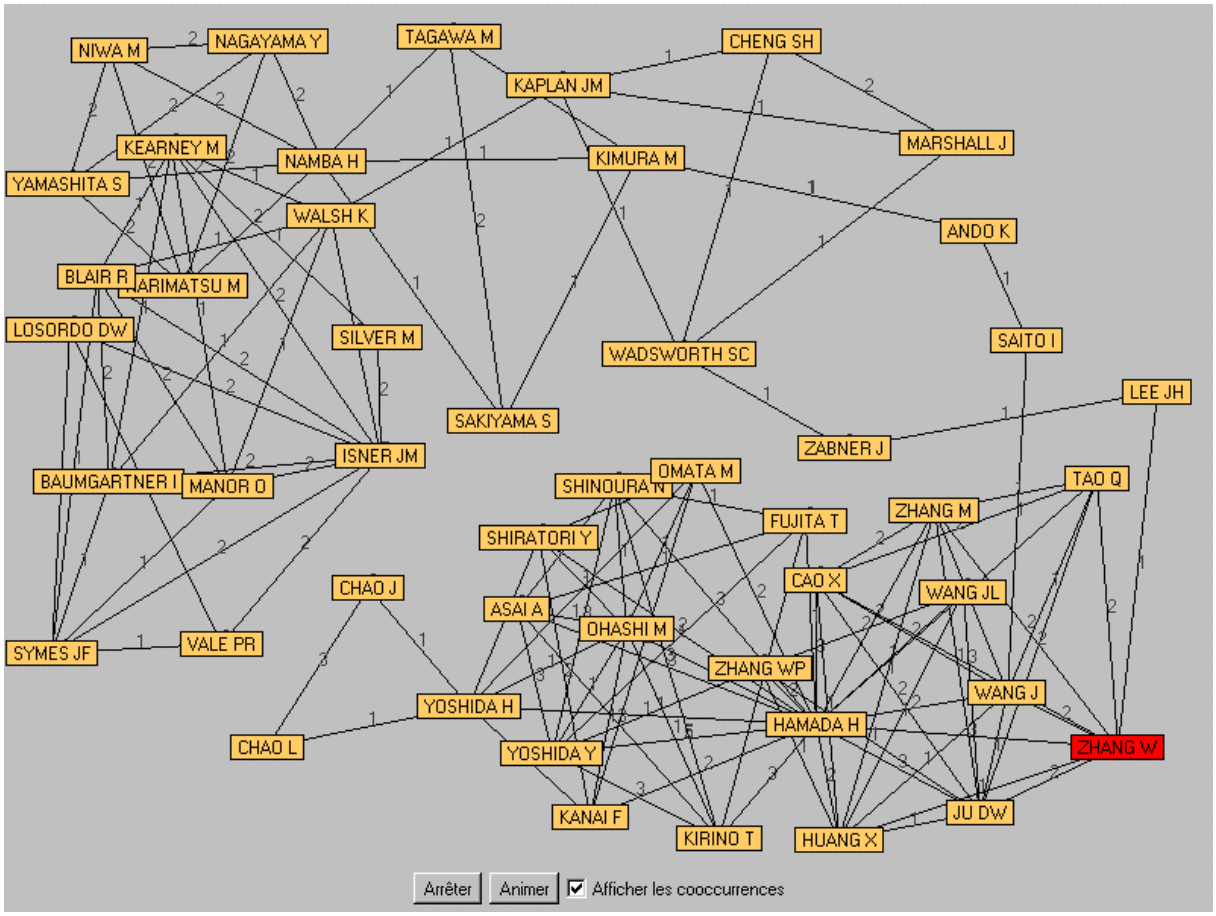
Représentation en ligne, sous forme d'histogramme, directement issue de la base de données :



Représentation en ligne des données filtrées sous forme de camembert :



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



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.

Bibliographie

[1.][DOUS98] Dousset B., Kanoun S.
Optimisation du choix de la terminologie pour la reformulation de requêtes: cas des multi-termes
VSST'98, pp 107-119, octobre 1998 (Toulouse, France).

[2.] [BENA99] Ben Ammar A., Dousset B.
Les métriques et l'analyse relationnelle: Visualisation en quatre dimensions.
7ème Conférence sur les systèmes d'information élaborée: Bibliométrie - Informatique stratégique - Veille technologique. Ile Rousse, 27 septembre au 1° octobre 1999 (Corse France).

[3.] [SALL99] Salles M., Dousset B.
La Veille Scientifique par l'Analyse des Informations Ouvertes.
(Crimée Russie).

[4.] [DOUS00] Dousset B., Dkaki T., Mothe J.
Information mining in order to graphically summarize semi-structured documents
17th International CODATA Conference, 15-19 octobre 2000 (Baveno Italie).

[5.] [DKA00] Dkaki T., Dousset B., Egret D., Mothe J.
Information discovery from semi-structured sources - Application to astronomical literature
Computer Physics Communication, 2000.