INFLUENCE DU PRETRAITEMENT TEXTUEL SUR LA REPRESENTATION GRAPHIQUE DANS UN CONTEXTE D'ANALYSE DE DONNÉES RELATIONNELLES

Eloïse LOUBIER(*), Sabine CARBONNEL(*), Bernard DOUSSET (*)
loubier@irit.fr, carbonne@irit.fr, dousset@irit.fr

(*)Institut de Recherche en Informatique de Toulouse, IRIT-SIG
Université Paul Sabatier,
118 route de Narbonne,
31062 Toulouse cedex 9
(France),

Mots clefs :
analyse des données textuelle , homogénéisation, représentation graphique, interactivité, synonymie, requêtage.
Keywords:
textual data analysis, homogenisation, chart, interactivity, synonymy, query.
Palabras clave :
análisis de datos textual, homogeneización, carta, interactividad, sinonimia, pregunta.

Résumé

Le rendement global de l'analyse des données textuelles est limité par l'homogénéisation et la hiérarchisation de l'information des différentes formes rencontrées dans les données brutes et, d'autre part, la restitution notamment graphique des résultats ainsi que les possibilités d'interactivité qu'elle offre aux utilisateurs.
La clarté des résultats et donc de leur représentation graphique dépend étroitement de la qualité du travail initial réalisé sur les données brutes. Dans ce contexte d'analyse de données, nous proposons d'étudier les dépendances entre le prétraitement et la représentation graphique des données, au sein du logiciel de macro-analyse de données Tétralogie.
Nous proposons un outil générique de prétraitements permettant à l'utilisateur d'introduire des connaissances spécifiques aux données. Nous proposons également d'étendre les fonctionnalités de VisuGraph (module de restitution graphique des résultats de Tétralogie) afin d'homogénéiser et hiérarchiser l'information, en fonction des connaissances introduites, ainsi qu'en augmentant les possibilités d'interactivité aux utilisateurs.
Le prétraitement apparaît ainsi comme une fonctionnalité indispensable pour une représentation graphique signifiante et ergonomique, en ce qui concerne la visualisation de gros volumes de données, mais aussi au niveau de la classification, en choisissant comme critère la granularité plutôt qu'un niveau de similarité.
Du point de vue de l'interaction avec l'utilisateur, il est important de cibler l'analyse de données en fonction de son besoin. Pour cela, nous proposons une méthode de requêtage qui permet de rechercher les informations issues de nos corpus de données, de type union / intersection pour affiner la demande. Les résultats d'une requête sont proposés sous deux formes complémentaires : le retour aux notices et la représentation sous forme de graphe.
Dans cet article nous présentons ces principes, révélant les dépendances entre les prétraitements et la restitution des résultats, et nous les illustrons par des exemples concrets.

1 Introduction

L'efficacité des plate-formes dédiées à l'analyse des données textuelles n'est plus à démontrer, aussi bien pour des informations semi structurées, que pour le texte libre. De très nombreuses méthodes sont déployées allant de l'analyse sémantique à l'analyse statistique, ce qui engendre un champ d'investigation très large, ainsi que de réelles possibilités de déductions sur le contenu informatif direct ou indirect de l'ensemble des sources disponibles. Comme ce point est assez valorisant au niveau recherche, il est particulièrement bien abordé dans la littérature. Par contre, à l'heure actuelle, il reste tout de même deux domaines qui limitent encore le rendement global de l'analyse des données textuelles. D'une part, l'homogénéisation et la hiérarchisation de l'information et notamment le traitement des équivalences et des inclusions entre les différentes formes rencontrées dans les données brutes (noms d'acteurs, termes d'indexation, adresses, périodes, lieux, sigles, multi-termes, multi-linguisme) et, d'autre part, la restitution notamment graphique des résultats et surtout les possibilités d'interactivité. En effet, la clarté des résultats et donc de leur représentation graphique dépend étroitement de la qualité du travail initial réalisé sur les données brutes, car une bonne homogénéisation permet de réduire notablement la complexité des résultats et le traitement hiérarchique permet de jouer sur la pertinence des niveaux de granularité disponibles qu'elle offre aux utilisateurs. Or, ces deux domaines sont fortement liés bien que se trouvant aux deux extrémités de la chaîne de traitement. Dans cet article, nous abordons cette problématique en faisant un état de l'art sur les plate-formes existantes et nous traitons de l'interdépendance de ces deux maillons via deux outils que nous avons développés, l'un pour les prétraitements et l'autre, VisuGraph, pour la visualisation graphique et interactive des données relationnelles. Nous illustrons cette dépendance par de nombreux exemples, afin de montrer que la qualité du résultat final dépend étroitement du soin apporté à ces deux phases : simplification des relations, comparaison entre hiérarchie « sémantique » et hiérarchie « fonctionnelle », requêtage via les visualisations avec retour aux documents quelque soit la forme donnée à un item.

2 Etat de l'art

Les agents de traitement automatisé de l'information sont des logiciels capables d'explorer les banques de données informatiques afin d'en tirer les informations pertinentes et de les adresser de façon claire et précises aux personnes concernées. Ces logiciels traitent aussi bien des banques de données informatiques que de fichiers acquis à partir d'un support papier (articles de presse, rapports...) via un scanner et un logiciel de reconnaissance optique de caractères, mais aussi des informations recueillies sur Internet. Ils sont souvent des analystes linguistiques, statistiques et des agents d'archivage et de tri : une fois les documents acquis, le logiciel les archive selon un procédé d'analyse sémantique. Les outils de veille stratégique peuvent être catégorisés en outils de :

- collecte de l'information
- analyse / synthèse de l'information
- diffusion de l'information

Nous nous intéressons à la 2e catégorie, c'est à dire aux outils de veille permettant l'analyse et la synthèse d'informations à partir de données textuelles. Voici quelques outils, donc certains intègrent aussi la partie collecte de l'information :

- Reseau-Lu [1] est un outil permettant de réaliser des cartographies de réseaux et de liens à partir de données diverses, comportant également un module de text mining;
- Keywatch [2] est un outil intégré (collecte et analyse/synthèse) qui utilise la dimension temporelle pour l'analyse de données textuelles structurées ou non;
- Online miner [3] comporte un module de text mining pour détecter des éléments importants dans des textes à partir de bases de données ou flux de données. Il permet également de faire du clustering de documents et comporte des outils d'analyse et de visualisation des données;

- Lexiquest mine [4] combine des méthodes d'analyse statistique et linguistique et est surtout adapté à l'analyse de documents non structurés;
- Intellixir [5] est un outil d'analyse statistique de documents structurés comme par exemple des notices bibliographiques, il permet entre autres d'établir des cartes de relations entre concepts.

La plupart de ces outils sont également complétés par des moteurs de recherche de documents plus ou moins élaborés.

L'INIST [6] a commencé une étude comparative de ce type d'outils d'où il ressort que les aspects prétraitement des documents sont souvent fastidieux et que la lecture des résultats ainsi que la recherche de documents est parfois difficile. Les principaux problèmes concernent :

- le reformatage des données concernant des acteurs ;
- la radicalisation ne prenant pas en compte les mots composés ;
- la lecture des résultats pour distinguer les éléments des différents champs ;
- la recherche de notice pas toujours simple ;

3 Présentation de Tétralogie

Tétralogie [7] est un logiciel de macro-analyse de données textuelles semi-structurées intégrant la dimension temporelle. Il permet donc de faire des études évolutives.

Les données analysées par la plateforme Tétralogie sont issues :

- Des bases de données : cela constitue une source majeure d'information qui doit cependant être complétée ;
- Des revues, journaux, périodiques : utiles pour une recherche beaucoup plus récente que sur les bases de données ;
- Des revues de veille technologique (ex. technologie et stratégie, french technology survey, comline news service) : permettent d'avoir des informations qui ne sont pas incluses dans les bases de données ;
- Des thèses et Des brevets ;
- Des CDROMS.

Les informations extraites de ces sources sont synthétisées sous forme de matrices de cooccurrence, exploitables dans les différents modules proposés par Tétralogie. Les unités de base de toute analyse sont le terme, le champ (auteur, mots-clefs, adresse, date, ...) et le document. Un champ est une balise prédéfinie de la base de donnée semi-structurée, par exemple auteur, date, adresse, organisme. Un champ, peut être mono-valué (journal) ou multi-valué (auteur, mot-clef,…). Un terme (ou entité) est une unité textuelle correspondant au contenu d'un champ mono-valué ou une partie d'un champ multi-valué délimité par des séparateurs. Une étude consiste à analyser les données sous forme de cooccurrences de termes de plusieurs champs, dont l'un de ces derniers peut être par exemple la date, si on prend en compte l'aspect temporel.

Pour cela Tétralogie est composé de modules :

- prétraitements : synonymie et comptage des termes pour chaque champ ;
- calcul des cooccurrences et analyse des données (méthodes factorielles, clustering) ;
- visualisation des données et interaction : VisuGraph (graphes), cartographies, arbres de classification, ...

Dans cet article, nous étudions les liens entre les modules 1 et 3. Nous ne présentons donc pas le module 2, utilisé notamment pour la génération des matrices de cooccurrences. Nous détaillons dans la suite, section 4, la notion de détection des synonymes que nous étendons pour permettre l'intégration de connaissances extérieures. Ensuite, la section 5 présente la visualisation graphique des données sous forme de graphes. Enfin dans la section 6 nous étudions l'impact de la détection des synonymes sur la représentation des données.

4 Le prétraitement des données : la synonymie

4.1 Problématique et objectifs

Nous nous intéressons à l'intégration de connaissances liées à la notion de synonymie de termes : c'est-à-dire un ensemble de connaissances permettant de déterminer des classes d'équivalence entre termes et donc de réaliser des analyses plus pertinentes.
Sans ces prétraitements, l'analyse statistique des données est biaisée, et les résultats sont difficiles à interpréter. De nombreuses études ont concerné la détection de synonymes, la normalisation de données [8] avec des approches plus ou moins automatisées. Cependant, les premiers résultats de l'étude de l'INIST [6] portant sur des outils d'analyse montre que de nombreux problèmes persistent. De plus, l'OST (Observatoire des Sciences et Techniques) travaille à l'édition de recommandations pour la saisie d'adresses et de noms en prévision de la construction de futures bases de données et afin de limiter les ambiguïtés (par exemple, toujours nommer un organisme de la même façon). Nous proposons un outil générique de prétraitements permettant à la fois à l'utilisateur de choisir parmi des traitements prédéfinis et d'introduire des connaissances spécifiques aux données qu'il étudie. L'avantage d'un tel outil est qu'il est très paramétrable et utilisable quelles que soient les données, de plus, le format des connaissances à introduire est relativement simple (sous forme de règles ou de listes). Ces règles peuvent être communes aux prétraitements d'une base complète ou spécifiques à chaque champ.

Les difficultés de la détection des synonymes sont liées au fait qu'un dictionnaire des termes valides n'est pas toujours disponible car les domaines d'étude sont très divers et que le vocabulaire peut être ouvert. De plus, les données doivent souvent être nettoyées ou normalisées. Certaines équivalences sont donc simplement liées à des écritures différentes d'un même terme : il est important de pouvoir s'abstraire des erreurs de typographie, des erreurs de saisie, des abréviations, des variation morphosyntaxique, etc…, et de proposer un représentant par classe d'équivalence (le plus fréquent ou généré suivant certains critères). Pour l'instant nous n'utilisons pas d'algorithme de radicalisation (Porter [9], Carry [10]) car les analyses que nous effectuons sont le plus souvent sur des acteurs, notions spatio-temporelles ou mots-clefs. Nous nous basons sur une notion de proximité des termes (d'après une distance d'édition) et il serait intéressant de combiner ces deux approches pour un regroupement en classes d'équivalences notamment pour mieux exploiter les champs en texte libre comme le titre ou le résumé d'un article par exemple. D'autres classes d'équivalences peuvent regrouper des termes synonymes sémantiquement, et c'est dans ce cas que l'introduction de connaissances extérieure est la plus importante.
L'objectif est donc d'automatiser au maximum les prétraitements textuels, tout en permettant à l'utilisateur d'introduire des connaissances spécifiques à son étude.

4.2 Les connaissances liées aux variations de forme

Ces connaissances sont des règles de :

- radicalisations très simples (féminin, pluriel) ;
- nettoyage des données ;

- normalisation des termes.

À ces connaissances s'ajoute une distance d'édition permettant de s'abstraire d'inversion, d'oubli ou de rajout de caractère pour regrouper des termes en fonction de leur proximité.

4.3 Les connaissances liées à la synonymie sémantique

Les connaissances que peut définir l'utilisateur pour regrouper des termes sémantiquement permettent notamment de bien définir les notions spatio-temporelles d'une base de donnée comme la spécialisation/généralisation de termes géographiques et l'inclusion de dates dans des intervalles temporels : notions primordiales dans le cadre d'études évolutives.
Un autre aspect intéressant est la possibilité de rendre des termes synonymes à des valeurs (par exemple positif/négatif, score, etc...) pour une analyse plus globale.
En fonction des spécificités des données et de l'étude, il est possible d'introduire les connaissances suivantes :

- dictionnaire des termes possibles ;
- règles de normalisation ou d'homogénéisation de termes (adresse, organisme, date) ;
- pour les dates : degré de précision, date de départ, durée d'un intervalle et recouvrement ;
- dictionnaire de synonymes comme par exemple : nom de lieux géographiques (France~Francia, États-Unis~USA) ; mots-clefs (automobile~voiture) ; entreprise, institution (Université Paul Sabatier~Université Toulouse 3) ;
- relation d'ordre partiel décrivant une hiérarchie de termes.

L'intérêt (mais aussi l'originalité) de cette dernière connaissance est de permettre le choix de la granularité de l'étude ainsi qu'une homogénéisation des termes. Par exemple, une relation d'ordre intéressante concerne des informations géographiques avec villes, départements, régions, pays, continent...

Toulouse	<	Haute-Garonne
Haute-Garonne	<	Midi-Pyrénées
Midi-Pyrénées	<	France
France	<	Europe
Montpellier	<	Hérault
Hérault	<	Languedoc-Roussillon
Languedoc-Roussillon	<	France
Ariège	<	Midi-pyrénées

Dans cette relation d'ordre x<y signifie que x est plus spécifique que y, et que la notion y recouvre la notion x.

L'utilisateur doit fournir une liste décrivant la précision qu'il choisit, par exemple pour une étude par région :

- Languedoc-roussillon ;
- Midi-pyrénées.

Cela génère les synonymes suivants (tous les termes < à Languedoc-Roussillon deviennent synonymes a Languedoc-Roussillon et les termes > ne sont pas pris en compte) :

terme	**synonyme**
Toulouse	Midi-Pyrénées
Haute-Garonne	Midi-Pyrénées
Ariège	Midi-Pyrénées
Montpellier	Languedoc-Roussillon
Hérault	Languedoc-Roussillon

Ainsi les termes Toulouse, Haute-Garonne, Midi-Pyrénées et Ariège seront considérés comme équivalent pour cette étude, et le représentant choisi pour cette classe sera le terme midi-pyrénées.
Cette étape de prétraitement permet d'obtenir des informations organisées et homogènes, répondant davantage aux besoins de l'utilisateur. Nous nous intéressons alors à leur représentation.

5 La représentation graphique de données relationnelles

Le prototype VisuGraph est un module de visualisation, offrant, à la plateforme, Tétralogie la possibilité de représenter les données matricielles sous forme de graphe.

5.1 Les données analysées

Les données analysées sont générées sous forme matricielle, dont le nombre de dimensions prises en compte est variable, selon les spécificités de l'analyse.
Les informations étudiées peuvent être issues d'un croisement entre deux champs, sous champs ou groupes de champs afin d'obtenir des matrices présence-absence ou de cooccurrence sur lesquelles porteront ensuite les analyses. Pour chacune de ces matrices, le croisement entre deux entités révèle la valeur de la métrique du lien entre ces deux dernières. Quel que soit le type d'entité (auteur, journal, …), il est possible de faire intervenir jusqu'à trois champs simultanément.
Pour prendre en compte l'aspect évolutif, la troisième dimension représente alors le temps. Les croisements entre les deux premières entités s'effectuent sur plusieurs segments temporels (ou périodes) homogènes, afin d'analyser les différences induites dans le temps comme : différences absolues, relatives, vitesses, accélérations, implosions, explosions de clusters, ….

5.2 La représentation graphique

Les recherches portant sur la visualisation de l'information montrent clairement que le recours à la représentation graphique de données facilite grandement leur analyse. En effet, les représentations textuelles de gros corpus de données offrent des informations difficilement exploitables.
Basé sur ce principe, VisuGraph [11] est développé en java[1] et nous proposons d'en étendre certaines fonctionnalités, telles que la restitution graphique des résultats, l'homogénéisation et la hiérarchisation de l'information et surtout des possibilités d'interactivité offertes aux utilisateurs.

[1] Disponible sur station SUN pour les systèmes d'exploitation SUN/OS et Solaris, et accessible par le réseau aussi bien à partir de terminaux X que de PC ou de Macintosh.

5.2.1 Le codage de l'information

VisuGraph offre la possibilité de représenter les entités en les assimilant à des sommets représentés sous forme de cercles, dont le diamètre est relatif à la valeur de la métrique utilisée, habituellement le nombre d'unités textuelles contenant l'entité : document, paragraphe, phrase, …
Les liens existants entre les différents sommets sont représentés sous forme d'arêtes, traduisant la relation existant entre deux entités et dont l'intensité est relative à la valeur de la métrique du lien c'est à dire le nombre d'unités textuelles contenant simultanément les deux entités croisées : cooccurrence.
Ainsi, dans le cas d'une matrice symétrique, les entités croisées (issues du même champ) représentent les sommets du graphe et les éléments de la matrices les arêtes (simultanément présence et valeur de chaque arête). Si la matrice est asymétrique, les entités croisées en lignes et en colonnes appartiennent à deux champs différents (graphe biparti) dont les sommets sont différenciés par des couleurs différentes.

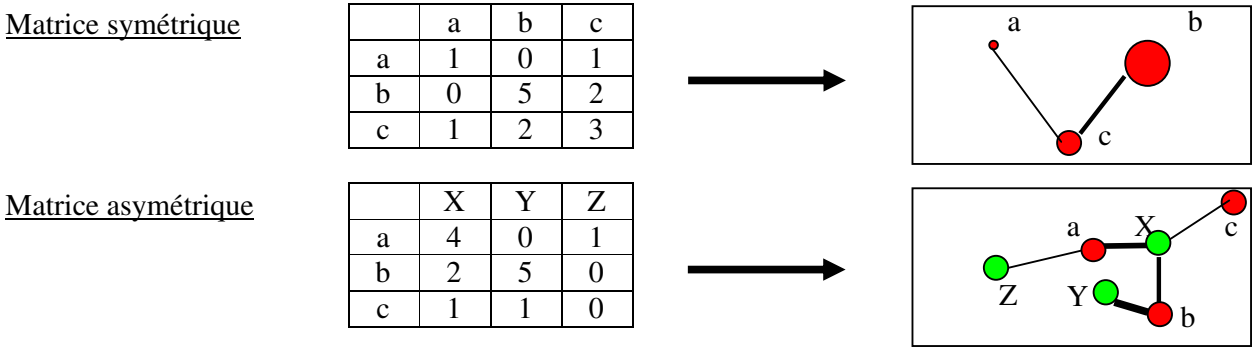

Figure 1 : Représentations graphiques d'une matrice symétrique et d'une matrice asymétrique

5.2.2 Le dessin de graphe

Afin d'obtenir un graphe lisible, dont le nombre d'entrecoupements d'arêtes est minimisé et dont les sommets sont organisés en fonction de l'importance de leur lien, nous utilisons un algorithme de forces d'attraction et de répulsion dont le principe a été introduit par Eades [12].
Notre travail se basant sur des graphes valués, nous avons proposé [14] une extension des formules proposées par Fruchterman [13], pour prendre en compte la valuation des arcs et la pondération des sommets.
Les sommets sont assimilés à des objets alors que les arcs le sont à des ressorts. La force d'attraction permet de rapprocher, graphiquement, les sommets fortement liés, alors que la force de répulsion éloigne les sommets faiblement ou non liés. Les sommets sont alors animés et se déplacent de façon à s'organiser de façon claire en rendant le graphe plus lisible.

A partir d'un état initial de forte énergie, nous laissons se relaxer le système afin que les sommets se positionnent harmonieusement les uns par rapport aux autre sans se superposer. Ces forces sont paramétrables, par le biais d'un slider qui permet d'augmenter ou de diminuer ces forces, laissant ainsi à l'utilisateur la possibilité d'influencer l'organisation du graphe selon sa convenance.

6 Les dépendances entre les prétraitements et VisuGraph

Remarque : nous développerons et illustrerons plus amplement cette section dans la version finale de l'article.

6.1 L'homogénéisation des données

Comme nous l'avons vu précédemment, la fonction de synonymie permet de regrouper des termes équivalents en fonction des connaissances introduites par l'utilisateur. Cette fonctionnalité est indispensable pour une représentation graphique signifiante et ergonomique. En effet, plus le nombre de sommets représentés est important, plus le nombre de liens l'est aussi et plus le graphe est dense, perdant en lisibilité ainsi qu'en possibilité d'interprétation et de synthèse de l'information traduite.

Le regroupement de termes en classes d'équivalences permet de représenter sous un seul sommets plusieurs termes et de limiter ainsi la taille du graphe à étudier. Si nous prenons l'exemple du terme «automobile», dans de nombreux documents, il est présent sous la forme «voiture», «véhicule», «berline», ... La classe d'équivalence : (automobile, voiture, véhicule, berline, …) est représentée par le terme «automobile». Lors d'une étude des publications portant sur l'automobile, comme dans la figure suivante, un graphe biparti permet de représenter les cooccurrences croisant les entités «document» et «mot-clés». Pour l'utilisateur, il est fortement intéressant que le graphe soit le plus simple possible, avec un sommet assimilé à chaque terme, incluant tous les synonymes de ce dernier (cf figure 2). Ainsi un lien entre un document et un mot-clé révèle le nombre d'occurrences de ce dernier, mais aussi de ses synonymes, dans la publication. Le lien entre «automobile» et le document «x» révèlera le nombre de fois où ce mot clé mais aussi « voiture », «berline», «véhicule» apparaissent dans «x». L'utilisateur peut visualiser tous les termes d'une classe d'équivalence en sélectionnant le représentant.

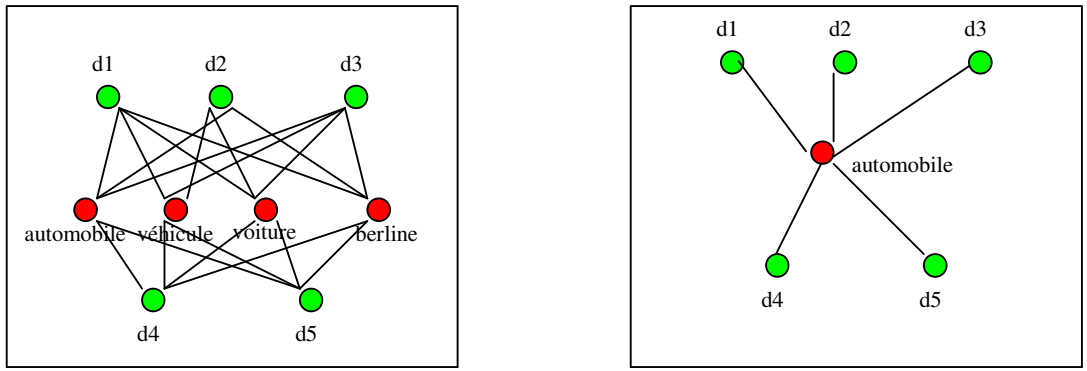

Figure 2 : Graphe sans utilisation de la synonymie (à gauche) et avec utilisation de cette fonctionnalité (à droite), un sommet est alors étiqueté par le représentant de sa classe d'équivalence.

6.2 La notion de hiérarchie

La synonymie permet, comme nous l'avons vu dans le section 4, de définir des relations d'ordre exprimant des hiérarchies entre les termes. Par exemple, dans un contexte géographique, la ville de Toulouse est inférieure à la région « Midi Pyrénées », elle-même inférieure à la région « Sud-ouest », elle-même inférieure à « France ». Cette distinction permet ainsi de manipuler les données, en prenant en compte un certain niveau de granularité. Dans un contexte de classification, il est possible de choisir comme critère la granularité plutôt qu'un niveau de similarité. En effet, plutôt que d'affecter chaque sommet à la classe dont le centre est le plus proche, (cas des KMeans), les sommets peuvent être regroupés d'après la granularité choisie.
Par exemple, étudions le nombre de contrats signés entre villes. Si les données sont volumineuses, le nombre de sommets obtenus est important. Chaque sommet représente une ville et chaque arête le nombre de contrats signés entre les villes. L'application des forces d'attraction et de répulsion, vues précédemment, permet de rapprocher les sommets des villes, ayant un nombre important de contrats communs. Inversement, les villes n'ayant pas de contrat commun ne sont pas liés et donc ne sont pas proches, d'un point de vue graphique, les unes des autres. Si l'exemple est pris à une échelle internationale, il est intéressant de pouvoir changer de granularité. Pour approfondir cette étude il est donc possible de choisir une représentation graphique par région. Les connaissances introduites lors des prétraitements permettent d'obtenir pour chaque ville la région à laquelle elle appartient. Un codage par la couleur est utilisé afin d'attribuer à chaque région une couleur spécifique sans modifier le positionnement des villes. Cela permet de visualiser à la fois l'appartenance d'une ville à une région ainsi que l'étendue des régions et de déterminer lesquelles collaborent le plus, la présence / absence de contrats entre elles…

6.3 Le retour aux notices

Nous ajoutons à VisuGraph, la fonctionnalité permettant le retour aux notices. Cette dernière permet, à partir d'un sommet sélectionné, d'afficher dans un éditeur de texte, les documents contenant l'entité choisie. Ainsi, en reprenant l'exemple portant sur l'automobile, vu dans la section 6.1, si l'utilisateur clique sur le sommet «automobile», tous les documents comportant spécifiquement ce terme (dans le champ mot-clef) s'afficheront dans un éditeur de texte [15].
Les connaissances de synonymie sémantique permettent alors d'élargir la recherche effectuée. En effet, dans un contexte de recherche d'information, l'intérêt est d'étendre la recherche à l'ensemble des termes proches de l'étiquette du sommet choisi. Ainsi, dans l'exemple précédent, la recherche effectuée sur le terme «automobile» retourne aussi les documents comprenant les synonymes de ce mot.
Un autre aspect important de la synonymie, dans un contexte de « retour aux notices », est la proximité des termes. En effet, un même terme peut être écrit sous des formes différentes (erreur de saisie, abréviation …). Par exemple, le nom d'un auteur peut être écrit sous la forme « M. Dupont » ou « Dupont » ou « Dupont Marc ». Or, dans le cas d'un croisement entre auteurs et documents, lors de la recherche des documents où cet auteur apparaît, le système doit retourner ceux comportant au moins une de toutes les formes possibles d'écriture de ce nom. Pour cela, les classes d'équivalence déterminées automatiquement lors de la recherche des synonymes permettent de retrouver toutes les formes présentes dans les documents de la collection.

6.4 Le requêtage

6.4.1 Construction de la requête

Afin de cibler l'analyse de données en fonction du besoin de l'utilisateur, nous proposons une méthode qui permet de rechercher les informations issues de nos corpus de données, répondant à un ou plusieurs critères de la requête, de type union / intersection pour affiner la demande.

Nous proposons deux manières de sélectionner les informations utiles à la conception de la requête :

- sélection graphique de sommets ;

- sélection via des listes dans un menu.

La sélection graphique permet d'obtenir une liste des sommets sélectionnés organisée par champ. Cette liste est éditable par l'utilisateur, en effet il peut interagir avec sa sélection afin de supprimer ou ajouter des éléments. Dans le cas de champs multi-valués, l'utilisateur peut choisir l'intersection ou l'union entre les éléments sélectionnés. Par exemple, si l'utilisateur a sélectionné les auteurs A1, A2, A3, A4 et les journaux J1 et J2 dans un graphe représentant les cooccurrences entre auteurs et journaux, il choisit le lien entre les auteurs ce qui lui permet de générer une des requêtes suivantes:

(A1 et A2 et A3 et A4) et (J1 ou J2)
(A1 ou A2 ou A3 ou A4) et (J1 ou J2)

Dans la première requête la sélection concerne les documents contenant les quatre auteurs simultanément dans le champ correspondant et l'un des deux journaux. Seule une union est possible entre plusieurs journaux puisque ce champ est mono-valué, en effet un article ne peut être publié que dans un seul journal. Dans la seconde, la sélection concerne au moins un des quatre auteurs.

La sélection via des listes s'effectue dans un menu composé des différents champs représentés dans le graphe. L'utilisateur peut générer une requête en cochant des éléments dans les listes et choisir les liens (intersection ou union) entre ces derniers. Le paramétrage de cette fonctionnalité s'effectue par l'affichage d'une fenêtre tierce, contenant les différents types d'entités croisées dans le graphe (date, auteurs, journaux, termes, titres…). Si l'utilisateur privilégie une dimension dans son analyse, il sera libre de déplacer en conséquent les sommets la représentant, afin d'obtenir un graphe organisé. La fonctionnalité de requêtage permet donc de restreindre le domaine de recherche en fonction des besoins de l'utilisateur, afin de focaliser l'analyse sur une cible spécifique.

6.4.2 Restitution des résultats

Le requêtage a pour finalité de restreindre les informations visualisées dans le graphe afin de mieux correspondre aux besoins de l'utilisateur et donc de préciser l'analyse. Les résultats d'une requête sont proposés sous deux formes complémentaires : le retour aux notices et la représentation sous forme de graphe.
Le retour aux notices restitue les documents répondant à la requête par le même procédé que celui détaillé dans la section 6.3. Les documents répondant aux critères de la requête sont retournés à l'utilisateur sous forme de fichier texte. Ils peuvent provenir de diverses bases et peuvent donc être de formats différents. Cet aspect est pris en compte lors de l'affichage des données sous forme textuelle.
La représentation par défaut, proposée par VisuGraph, permet de visualiser les structures des différentes organisations concernant la totalité des données traitées. Afin de visualiser graphiquement ces résultats, le graphe partiel, composé des éléments sélectionnés graphiquement ou dans la liste de choix, est extrait de la représentation graphique globale (initiale). Cette restitution graphique s'effectue par masquage des sommets et des liens non concernés par la requête. La visualisation finale est plus claire, puisque le nombre de sommet ainsi que le nombre de lien sont diminués, rendant l'analyse plus simple et plus précise. La structure, au sein du graphe global est conservée, permettant ensuite à l'utilisateur d'organiser le graphe des résultats comme il le désire puisqu'il a la possibilité de déplacer avec la souris les sommets. Ainsi l'utilisateur maîtrise pleinement son analyse dans chacune des étapes de requêtage, que ce soit dans la sélection d'information ou encore dans la manipulation des résultats, restitués sous forme textuelle et graphique.

7 Conclusion

Dans cet article, nous avons montré la dépendance existante entre les prétraitements liés à la synonymie et la représentation des données proposées par la plateforme Tétralogie. L'amélioration du rendement global de l'analyse des données textuelles est effective par l'homogénéisation et la hiérarchisation de l'information et notamment le traitement des équivalences et des relations d'ordre entre les différentes formes rencontrées dans les données.
De plus, la visualisation des résultats est dotée de l'interactivité avec l'utilisateur, lui permettant de sélectionner les éléments à afficher et de paramétrer sa représentation graphique, Les besoins de l'utilisateur sont alors ciblés et l'analyse est davantage précise et pertinente.
Ainsi, un seul terme est représenté, plutôt que plusieurs équivalents, simplifiant le graphe et synthétisant davantage l'information. Le retour aux notices est effectué par la recherche, sous toutes leurs formes, des termes correspondant à l'entité sélectionnée. De plus, la synonymie permet de choisir d'une part les informations à visualiser, afin de restreindre le volume de données analysées et de cibler les besoins de l'utilisateur, et d'autre part d'appliquer un niveau de granularité supérieur par un procédé de regroupement de données.
Dans nos perspectives, il serait intéressant de proposer de réaliser des filtres, dans le contexte de la sélection d'information via des requêtes, portant sur des dimensions qui ne sont pas représentées dans un graphe mais qui ont été prise en compte lors de l'analyse du corpus. Ainsi, l'utilisateur pourrait cibler davantage son besoin et sélectionner les informations les plus pertinentes, par davantage de critères de sélection lors de l'expression de la requête. De plus, nous nous intéressons à la durée de validité des connaissances introduites lors des prétraitements, c'est-à-dire à la façon de représenter et utiliser des dates pour étiqueter ces connaissances. Cela permettrait de modéliser le fait que certains termes peuvent n'être synonymes qu'à une période donnée.

8 Références

[1] BAYEN R., BENARD M., MARRASSE JM., RAGOT F. Réseau'lution N°5, mai 2000 -.http://resolution.chez-alice.fr/pdf/reseau5.pdf

[2] MARCELLI L., REGNIER T., ANDREI P. *Présentation de Keywatch.* ISCOPE : Solutions pour la veille & l'information d'entreprise http://www.ie-veille.com/PDF/iScope.pdf. Janvier 2007.

[3] TEMIS, Online miner : Text Mining solutions. http://www.temis.com/index.php?id=88&selt=1&lg=en

[4] CROCHET DAMAIS A. *LexiQuest apporte de la pertinence à la recherche plein texte*. JDNet, http://www.journaldunet.com/solutions/0106/010606_lexiquest.shtml, 6 juin 2001.

[5] INTELLIXIR l'infométrie décisionnelle. *Traitement et analyse de l'information structurée dans le monde industriel.* http://www.intellixir.com/livreblanc.htm.

[6] L'INIST. *Benchmarking Outils de Veille*. http://outils.veille.inist.fr/index.html.

[7] DOUSSET B., BENJAMAA T., *Trilogie logiciel d'analyse de données*, Conférence sur les systèmes d'informations élaborées : Bibliométrie – Information Stratégique – Veille technologique, 1988.

[8] JOLIBOIS S., NAUER E., CHOUANIERE D., DUCLOY J., GRANDJEAN F., MOUZE-AMADY M. *Adaptation des normes et formats documentaires à la gestion informatisée de corpus bibliographiques.* Bulletin des Bibliothèques de France 45, 2000.

[9] VAN RIJSBERGEN C.J., ROBERTSON S.E. AND PORTER M.F. *New models in probabilistic information retrieval.* London: British Library. (British Library Research and Development Report, no. 5587), 1980

[10] M. PATERNOSTRE, P. FRANCQ, J. LAMORAL, D. WARTEL ET M. SAERENS, *Carry, un algorithme de désuffixation pour le français.* Technical report, Université libre de Bruxelles, http://beams.ulb.ac.be/beams/documents/carryfinal.pdf. 2002.

[11] KAROUACH S., DOUSSET B., Manipulation de graphes de grande taille pour l'étude des réseaux d'acteurs et des réseaux sémantiques". 10ièmes journées d'études sur les systèmes d'information élaborée: Bibliométrie - Informatique stratégique - Veille technologique, (Ile Rousse Corse France), CD-ROM, 13-17 juin 2005.

[12] EADES P. *A heuristic for Graph Drawing*. Congressus Numerantium, vol. 42, p. 149-160, 1984.

[13] FRUCHTERMAN TMJ., REINGOLD EM. *Graph drawing by force_directed placement.* Software – Practice and experience, 21, p. 1129-1164, 1991.

[14] LOUBIER E., CARBONNEL S. *VisuGraph : Un outil d'exploration de données relationnelles évolutives*. INFORSID 2007, à paraître, 2007.

[15] LOUBIER E., BAHSOUN W. *La visualisation de données relationnelles au service de la recherche d'informations*. Conférence francophone en Recherche d'Information et Applications (CORIA 2007), Saint-Etienne, 29/03/2007-30/03/2007.