Analyse et visualisation de données relationnelles par morphing de graphe prenant en compte la dimension temporelle

Eloïse Loubier

THÈSE

En vue de l'obtention du

DOCTORAT DE L'UNIVERSITÉ DE TOULOUSE

Délivré par *l'Université Toulouse III - Paul Sabatier*
Discipline ou spécialité : *Informatique*

Présentée et soutenue par Eloïse LOUBIER
Le *09 Octobre 2009*

Titre : *Analyse et visualisation de données relationnelles par morphing de graphe prenant en compte la dimension temporelle*

JURY

Wahiba BAHSOUN	Maître de conférences, Université Toulouse III	*Co-encadrante de Thèse*
Claude CHRISMENT	Professeur, Université Toulouse III	*Président du jury*
Edwin DIDAY	Professeur, Université Paris Dauphine	*Examinateur*
Bernard DOUSSET	Professeur, Université Toulouse III	*Directeur de Thèse*
Brigitte GAY	Professeur, ESC de Toulouse	*Examinatrice*
Michel LAMURE	Professeur, Université Lyon I	*Rapporteur*
José MARTINEZ	Professeur, École Polytechnique de Nantes	*Rapporteur*
Klaus SOLBERG SÖILEN	Professeur, Blekinge Institute of Technology (Suède)	*Examinateur*

École doctorale : *Ecole Doctorale Mathématique Informatique Télécommunications de Toulouse*
Unité de recherche : Institut de Recherche en Informatique de Toulouse
Équipe d'accueil : *Systèmes d'Informations Généralisés - Extraction et Visualisation d'Informations*
Directeur de Thèse : *Bernard DOUSSET*
Co-encadrante : *Wahiba BAHSOUN*

Eloïse LOUBIER

Analyse et visualisation de données relationnelles par morphing de graphe prenant en compte la dimension temporelle

Directeur de thèse :
Bernard DOUSSET, professeur à l'université Toulouse 3 – Paul Sabatier

Résumé

Avec la mondialisation, l'entreprise doit faire face aux menaces de plus en plus fortes de la concurrence et à l'accélération des flux d'information. Pour cela, elle est amenée à rester continuellement informée des innovations, des stratégies de la concurrence et de l'état du marché tout en gardant la maîtrise de son environnement. Le développement d'Internet et la globalisation ont à la fois renforcé cette exigence, et fourni les moyens de collecter l'information qui, une fois synthétisée, prend souvent une forme relationnelle. Pour analyser le relationnel, le recours à la visualisation par des graphes apporte un réel confort aux utilisateurs, qui, de façon intuitive, peuvent s'approprier une forme de connaissance difficile à appréhender autrement.

Nos travaux conduisent à l'élaboration des techniques graphiques permettant la compréhension des activités humaines, de leurs interactions mais aussi de leur évolution, dans une perspective décisionnelle. Nous concevons un outil alliant simplicité d'utilisation et précision d'analyse se basant sur deux types de visualisations complémentaires : statique et dynamique.

L'aspect statique de notre modèle de visualisation repose sur un espace de représentation, dans lequel les préceptes de la théorie des graphes sont appliqués. Le recours à des sémiologies spécifiques telles que le choix de formes de représentation, de granularité, de couleurs significatives permet une visualisation plus juste et plus précise de l'ensemble des données. L'utilisateur étant au cœur de nos préoccupations, notre contribution repose sur l'apport de fonctionnalités spécifiques, qui favorisent l'identification et l'analyse détaillée de structures de graphes. Nous proposons des algorithmes qui permettent de cibler le rôle des données au sein de la structure, d'analyser leur voisinage, tels que le filtrage, le *k*-core, la transitivité, de retourner aux documents sources, de partitionner le graphe ou de se focaliser sur ses spécificités structurelles.

Une caractéristique majeure des données stratégiques est leur forte évolutivité. Or l'analyse statistique ne permet pas toujours d'étudier cette composante, d'anticiper les risques encourus, d'identifier l'origine d'une tendance, d'observer les acteurs ou termes ayant un rôle décisif au cœur de structures évolutives.

Le point majeur de notre contribution pour les graphes dynamiques représentant des données à la fois relationnelles et temporelles, est le morphing de graphe. L'objectif est de faire ressortir les tendances significatives en se basant sur la représentation, dans un premier temps, d'un graphe global toutes périodes confondues puis en réalisant une animation entre les visualisations successives des graphes attachés à chaque période. Ce procédé permet d'identifier des structures ou des événements, de les situer temporellement et d'en faire une lecture prédictive.

Ainsi notre contribution permet la représentation des informations, et plus particulièrement l'identification, l'analyse et la restitution des structures stratégiques sous jacentes qui relient entre eux et à des moments donnés les acteurs d'un domaine, les mots-clés et concepts qu'ils utilisent.

Eloïse LOUBIER

Analysis and visualization of relational data by graph morphing taking temporal dimension into account

Supervisor:
Bernard DOUSSET, Professor at Toulouse 3 University – Paul Sabatier

Abstract

With word wide exchanges, companies must face increasingly strong competition and masses of information flows. They have to remain continuously informed about innovations, competition strategies and markets and at the same time they have to keep the control of their environment. The Internet development and globalization reinforced this requirement and on the other hand provided means to collect information. Once summarized and synthesized, information generally is under a relational form. To analyze such a data, graph visualization brings a relevant mean to users to interpret a form of knowledge which would have been difficult to understand otherwise.

The research we have carried out results in designing graphical techniques that allow understanding human activities, their interactions but also their evolution, from the decisional point of view. We also designed a tool that combines ease of use and analysis precision. It is based on two types of complementary visualizations: statics and dynamics.

The static aspect of our visualization model rests on a representation space in which the precepts of the graph theory are applied. Specific semiologies such as the choice of representation forms, granularity, and significant colors allow better and precise visualizations of the data set. The user being a core component of our model, our work rests on the specification of new types of functionalities, which support the detection and the analysis of graph structures. We propose algorithms which make it possible to target the role of the data within the structure, to analyze their environment, such as the filtering tool, the *k-core,* and the transitivity, to go back to the documents, and to give focus on the structural specificities.

One of the main characteristics of strategic data is their strong evolution. However the statistical analysis does not make it possible to study this component, to anticipate the incurred risks, to identify the origin of a trend, and to observe the actors or terms having a decisive role in the evolution structures. With regard to dynamic graphs, our major contribution is to represent relational and temporal data at the same time; which is called graph morphing. The objective is to emphasize the significant tendencies considering the representation of a graph that includes all the periods and then by carrying out an animation between successive visualizations of the graphs attached to each period. This process makes it possible to identify structures or events, to locate them temporally, and to make a predictive reading of it.

Thus our contribution allows the representation of advanced information and more precisely the identification, the analysis, and the restitution of the underlying strategic structures which connect the actors of a domain, the key words, and the concepts they use; this considering the evolution feature.

Eloïse LOUBIER

Analyse et visualisation de données relationnelles par morphing de graphe prenant en compte la dimension temporelle

Mots-clés

- Modélisation, approximation, simulation, identification, optimisation, systèmes experts, extraction et gestion des connaissances.
- Bases de données relationnelles, gestion bibliographique, indexation automatique.
- Analyse de données, analyse exploratoire, analyse textuelle, analyse relationnelle, visualisation évolutive de données, graphe statique, graphe dynamique, fouille textuelle, fouille numérique, découverte de connaissances, morphing de graphe, bibliométrie, scientométrie, infométrie, veille scientifique et technologique, intelligence économique, aide à la décision.

Remerciements

« *Il y a mille moyen d'encourager les fausses vocations,*
aucun moyen de décourager les vraies. »
Edmond et Jules de Goncourt

Tout travail de recherche en thèse de doctorat nécessite toujours une collaboration multiforme. Dans le cas de ma recherche, nombreux sont ceux qui m'ont apporté une contribution scientifique, logistique ou morale. Que chacun trouve dans l'accomplissement de cette thèse l'expression de ma reconnaissance pour sa contribution quelle qu'elle soit.

Cependant je voudrais exprimer ma gratitude à messieurs Gilles Zurflhu et Claude Chrisment, directeurs de l'équipe des Systèmes d'Information Généralisés pour m'avoir accueillie. Je remercie tout particulièrement ce dernier pour m'avoir encouragé et pour avoir toujours eu les bons mots.

J'exprime aussi ma reconnaissance à mon directeur de thèse, le professeur Bernard Dousset sans qui ces travaux n'auraient pas la portée qu'ils ont. Son encadrement, son expérience, son savoir faire et ses conseils m'ont permis d'atteindre mes objectifs et je tiens à l'assurer de ma gratitude.

Madame Wahiba Bahsoun, maître de conférences et responsable de la filière Statistique et Informatique Décisionnelle m'a accompagnée durant une grande partie de mon cursus scolaire. Après avoir été ma directrice d'IUP, elle est devenue co-encadrante de cette thèse et je l'en remercie. Je tiens à lui exprimer ma reconnaissance pour son soutien, son investissement dans mes travaux et sa compétence.

Je tiens à remercier les professeurs Michel Lamure et José martinez pour l'intérêt qu'ils ont apporté à mes travaux de recherche et pour avoir accepté d'être rapporteurs de ce mémoire.

Au cours de ces années, de nombreuses rencontres ont été riches et je suis consciente de leurs conséquences bénéfiques. Le plus mémorable reste celle avec Brigitte Gay, professeur à l'ESC que j'admire pour ses compétences ainsi que son professionnalisme. Un immense merci pour ses encouragements, son écoute et surtout pour avoir cru en mes travaux.

Merci à Ilhème Ghalamallah et Anass Elhaddadi pour avoir égayé l'ambiance du bureau durant nos collaborations studieuses ou encore autours d'un bon café.

Mes pensées vont aussi pour tous les membres de l'équipe avec qui j'ai pu travailler ou encore simplement partager de bons moments.

Ainsi, je souhaite remercier Mesdames Josiane Mothe, Florence Sèdes, Chantal Soulé-Dupuy, Monsieur Mohand Boughanem, professeurs effectuant leur recherche dans l'équipe SIG. J'ai beaucoup appris en suivant leurs enseignements au cours de mon Master 2 Recherche, mais aussi en les cotoyant au cours de ces dernières années. Ma reconnaissance va aussi à l'égard des maîtres de conférences rattachés à l'IRIT pour leurs encouragements, leur soutien et pour tout ce qu'ils m'ont inculqué. Ainsi, merci à Messieurs Max Chevalier, Taoufiq Dkaki, Gilles Hubert, Franck Ravat, Olivier Teste, Guillaume Cabanac Benoît Encelle, Franck Morvan, Fabrice Gomboa, Sébastien Gadat, ainsi que Mesdames Maryse Salles, Lynda Tamine-Lechani et plus particulièrement Karen Pinel-Sauvagnat, qui reste mon modèle de réussite que ce soit au niveau social ou encore professionnel.

Un grand merci à Cécile Laffaire, Ingénieur d'Etudes, CNRS, pour sa bonne humeur et ses encouragement quand la chance me tournait le dos.

Toute mon amitié va vers les doctorants, actuels ou anciens, de l'équipe et en particulier Dana Kunhkun, Bouchera Soukkarieh , Ronan Tournier, Désiré Kaomparé, Estella Antoni, Nissou, …

Ma pensée va aussi pour tous les chercheurs, apprentis ou confirmés, rencontrés au cours de toutes ces années, ainsi qu'à tous les étudiants auprès desquels j'ai effectué mon monitorat et en particulier ceux qui ont accepté d'évaluer l'outil résultant de mes travaux.

Je remercie l'équipe de choc : Kro, Luc, Mira et Myriam pour toujours avoir été là, même dans les mauvais moments. Un grand merci à tous les autres qui ont suivi toute cette épopée, du début à la fin : Sam, Catherine, Leïla, Ivan, MyriamBis, Lise-Marie, JC, Rémi, Marie, Laurent.…

Toute ma reconnaissance va vers mes parents à qui je dois tout. Merci de m'avoir donné autant pour que je puisse aboutir à ce que je suis aujourd'hui.

Enfin, merci Guillaume pour être et rester Guillaume, pour avoir supporté « *la p'tite élo dans le mauvais temps* ».

Il ne me reste plus qu'à souhaiter de nouvelles rencontres aussi profitables que les précédentes, ainsi que l'approfondissement des techniques sur le terrain, et de nouvelles confrontations dans l'objectif d'avancer davantage dans mes recherches.

❦❦❦

Table des matières

Résumé 1

Abstract 3

Mots-clés 5

Remerciements 7

Table des matières 9

Introduction générale 13
Contexte de travail 13
Problématiques 14
Contributions 15
Organisation du mémoire 17

Chapitre 1. De la découverte de connaissance à son utilisation en veille stratégique 19
1.1. Préambule 20
1.2. Processus de veille 21
1.2.1. Intelligence économique et veille 21
1.2.2. Concept de veille stratégique 23
✓ Définition 23
✓ Objectifs de la veille stratégique 24
1.3. Découverte de connaissance 24
1.3.1. Définition 24
1.3.2. Information et connaissance 25
✓ Qu'est-ce que l'information? 25
✓ Qu'est-ce que la connaissance? 26
1.3.3. Présentation des étapes du processus de découverte de connaissance 26
1.3.4. Sélection des données cibles 27
✓ Sources formelles 27
✓ Sources informelles 28
1.3.5. Prétraitement 29
✓ Metadonnées 29
✓ Niveaux d'information 30
✓ Multi-termes 30
✓ Synonymie 31
✓ Filtrage 34
1.3.6. Transformation 34
1.3.7. Fouille de données 35
✓ Entités textuelles 35
✓ Différentes mesures du relationnel 36
✓ Matrices et cubes 37
1.3.8. Evaluation du modèle 44
1.3.9. Validation 45
1.3.10. Discussion 45
1.4. Synthèse 47

Chapitre 2. Visualisation de données : vers le temporel 49
2.1. Introduction 50
2.1.1. Spécifications 50
2.1.2. Présentation du domaine 50
2.1.3. Enjeux de la visualisation de données 51
2.1.4. Définitions 51

2.2. Domaines d'application 52
2.2.1. Réseaux sociaux 52
✓ Réseaux d'acteurs 53
✓ Réseaux de citation 53
✓ Réseaux de co-auteurs 53
✓ Réseaux de contacts 53
✓ Réseaux d'appels téléphoniques 53
2.2.2. Réseaux sémantiques 53
✓ Graphes conceptuels 53
✓ Ontologies 54
2.2.3. Autres réseaux 54
✓ Réseaux de neurones 54
✓ Réseaux trophiques 54
✓ Réseaux d'interactions entre protéines 54
✓ Réseaux de distribution électrique 55
✓ Réseaux informatiques 55
✓ Réseaux de transports 55
2.3. Principes de la visualisation de graphes 55
2.3.1. Complexité de la visualisation 55
2.3.2. Sémiologie graphique 56
2.3.3. Problèmes d'ergonomie 57
2.3.4. Préconisations sur la planarité 58
2.3.5. Préconisations sur les structures de visualisation 60
2.4. Prise en compte de la dimension temporelle 61
2.4.1. Données temporelles 61
2.4.2. Espace temps 62
✓ Visualisation linéaire 63
✓ Forme non uniforme du temps 70
✓ Temps cyclique 73
✓ Discussion 77
2.4.3. Représentation de l'espace temps et des données temporelles 85
2.4.4. Détection des tendances émergentes 91
2.5. Conclusion 91

Chapitre 3. Notre contribution pour les graphes statiques 95
3.1. Introduction 96
3.2. Visualisation de données : méthodes et représentations 96
3.2.1. Approche 96
3.2.2. Espace de représentation des données relationnelles 98
3.2.3. Métriques 98
✓ Métriques associées aux sommets 98
✓ Métriques associées aux arêtes 100
✓ Représentation des métriques 100
3.2.4. Placement des sommets 102
3.2.5. Représentation des liens 103
3.2.6. Algorithmes de représentation de graphe 103
✓ Représentation de graphe 103
✓ Algorithme basé sur les force_directed_placement (FDP) 104
3.2.7. Autres types de graphes 114
✓ Graphes orientés 114
✓ Graphes spécifiques 116
✓ Représentation caméléon 119
3.3. Les fonctionnalités 122
3.3.1. Positionnement de l'utilisateur 122
3.3.2. Identification et analyse de structure de graphe 123

3.3.3. Filtrage 125
3.3.4. *K*-Core 126
3.3.5. Transitivité 128
3.3.6. Retour aux documents 131
3.3.7. Focalisation 132
✓ Elaboration 132
✓ Restitution des résultats 133
3.3.8. Partitionnement de graphes 134
✓ Principe général 134
✓ Markov Clustering intégré à VisuGraph 135
✓ Manipulation du graphe réduit 136
3.4. Conclusion 140

Chapitre 4. Notre contribution pour les graphes dynamiques 143
4.1. Introduction 144
4.2. Spécifications 145
4.3. Espace de représentation des données temporelles 146
4.4. Métriques 146
4.4.1. Métriques associées aux sommets 146
4.4.2. Métriques associées aux arêtes 147
4.4.3. Codage des métriques 147
4.5. Positionnement temporel 148
4.6. Algorithme de représentation de graphe 150
4.7. Analyse de structure temporelle 153
4.8. Morphing de graphe : du graphe global au graphe de période 155
4.8.1. Définition du morphing de graphe 155
4.8.2. Principe 156
✓ Représentation globale 156
✓ Représentation par période 157
4.8.3. Morphing sans transition 162
4.8.4. Morphing avec transition 164
✓ Le point de vue du repère temporel de l'instance considérée 164
✓ Le point de vue de la structure temporelle 164
4.8.5. Discussion sur le morphing de graphe 167
4.9. Les graphes orientés 168
4.10. Fonctionnalités 170
4.10.1. Points de vues 170
✓ Le point de vue global 170
✓ Le point de vue évolutif 170
4.10.2. Filtrage 171
4.10.3. K-Core 172
4.10.4. Transitivité 173
4.10.5. Partitionnement de graphe 173
4.11. Application de notre contribution à des analyses non temporelles 175
4.12. Synthèse 177

Chapitre 5. Expérimentations et évaluation de VisuGraph 179
5.1. Présentation des expérimentations 180
5.2. Architecture et conception du prototype 180
5.3. Evaluation de la représentation graphique 181
5.3.1. Evaluation sur la représentation des données temporelles 181
5.3.2. Evaluation sur l'espace des données 183
5.4. Expérimentation de VisuGraph 184
5.4.1. Méthodologie 184
5.4.2. Technique utilisée 184

5.4.3. Population étudiée ____ 186
5.4.4. Accès à l'expérimentation ____ 187
5.4.5. Résultats ____ 188
✓ Durée d'expérimentation ... 188
✓ Niveau de difficulté par type de données ... 188
✓ Analyse des autres résultats ... 201
✓ Conséquences du sondage ... 202
5.5. Synthèse de l'expérimentation et de la validation de VisuGraph ____ 202

Conclusion et perspectives ____ *205*
6.1. Résumé des contributions ____ 206
6.2. Limites et approches envisagées ____ 208
6.2.1. Volume de données ____ 208
6.2.2. Ergonomie ____ 208
6.3. Perspectives de recherche ____ 208
6.3.1. Données en entrée ____ 208
6.3.2. Ajout de fonctionnalités ____ 209
✓ Classification ... 209
✓ 3 dimensions (3D) ... 210
✓ Amélioration du morphing de graphe ... 210
✓ Granularité temporelle ... 210
✓ Image de synthèse ... 210
✓ Réalisation d'un rapport automatique ... 210

Liste des figures ____ *213*

Liste des tableaux ____ *216*

BIBLIOGRAPHIE ____ *219*

GLOSSAIRE ____ *237*
✓ Veille ... 237
✓ Théorie des graphes ... 239
✓ Informatique ... 240
✓ Statistique ... 241

Annexes ____ *245*
Questionnaire d'évaluation de VisuGraph, outil graphique d'analyse de données relationnelles et évolutives. 246
Graphe symétrique ____ 251
Graphe asymétrique orienté ____ 252
Graphe asymétrique, non orienté ____ 253
Graphe asymétrique, orienté et issu du croisement de trois variables ____ 254
Graphe temporel, orienté ____ 254
Graphe symétrique des dépôts de brevets ____ 256
Graphe symétrique dans le domaine des véhicules hybrides ____ 257

Introduction générale

> *« Si pour obtenir une réponse correcte et complète à une question donnée, une construction requiert un temps de perception plus court qu'une autre construction, on dira qu'elle est plus efficace pour cette question. »* (Bertin, 1970)

Contexte de travail ____ 13
Problématiques ____ 14
Contributions ____ 15
Organisation du mémoire ____ 17

Contexte de travail

Une des caractéristiques du monde actuel est la forte liaison entre les différents secteurs de l'activité humaine d'où la notion de « mondialisation » non seulement de l'économie mais aussi de la finance, de la santé, du développement durable,… et de leurs interactions. L'expression« effet papillon » stipule que chaque évènement, même le plus anodin, peut avoir, à long terme, des conséquences colossales.

Le surendettement dû initialement au dumping économique et dont la crise des prêts « subprime » américains n'est qu'une des composantes s'est transformé en crise globale du crédit puis de la liquidité. Les difficultés financières issues des mondes de l'immobilier et de la banque ont atteint progressivement le milieu boursier puis l'économie mondiale. Cette crise se caractérise par une sévère chute de la consommation et de l'investissement, touchant ainsi l'activité économique, mais aussi le développement durable et la santé, qui sont momentanément mis entre parenthèses. Cela engendre, entre autres, une baisse de la qualité alimentaire, mais aussi une augmentation des nuisances portant atteinte à la santé déjà fragilisée par le stress provoqué par la crise et ses conséquences. Pourtant, un renforcement du développement durable permettrait de relancer intelligemment l'économie et de surcroit aurait un impact direct sur la santé.

Il est indéniable que notre planète constitue un système fermé, que l'économie est un jeu à somme quasiment nulle, d'où la nécessité de modèles intégrant l'ensemble des paramètres stratégiques. Tout sous modèle ne peut donc pas être considéré comme fermé car il possède obligatoirement des interactions avec son complémentaire. D'où la nécessité, au départ de chaque étude stratégique, de prendre en compte l'environnement global.

Moins d'une semaine après son apparition au Mexique, la crainte d'une pandémie de grippe A h1n1 s'est répandue dans le monde entier. En même temps que le virus a muté, pour être aujourd'hui transmissible d'homme à homme, il a passé les frontières, engendrant un renforcement des contrôles dans les aéroports, une restriction des voyages au Mexique et dans certains états américains et une suspension des importations de viande de porc en provenance des régions infectées. Ici encore, la forte interconnexion des activités humaines nous oblige à réagir ensemble, ce qui va dans le sens d'une gouvernance globale.

Toute activité humaine ayant d'importantes conséquences sur notre monde, on observe une très forte implication de l'analyse du relationnel dans les réactions immédiates, la gouvernance et la prospective.

Ainsi, s'informer, communiquer, anticiper est primordial afin de renforcer la solidarité, favoriser la cohésion, l'implication, la productivité, la pérennité en planifiant des actions s'inscrivant dans une démarche à long terme.

Pour déterminer et mener à bien ces actions, il faut pouvoir prendre connaissance de l'information pertinente parfois très rapidement, afin de la rendre disponible au bon moment aux bonnes personnes. En effet, l'information est de plus en plus utilisée comme objet de référence et comme outil d'aide à la décision d'ordre stratégique. Ainsi, l'élaboration et la mise en œuvre de stratégies pensées permettent d'atteindre efficacement des objectifs cohérents et durables.

Il faut rechercher, identifier, récupérer, valider, normaliser, synthétiser des informations, confronter les découvertes de connaissance avec ce que l'on sait déjà et qui n'est pas nécessairement sous forme électronique, pondérer le tout et prendre une décision alors que la vision de l'environnement immédiat est basée sur des données souvent imprécises, incomplètes, évolutives et fortement liées aux autres secteurs d'activité.

Pour appréhender certains types d'informations, la représentation graphique, sous forme de cartes géographiques, d'histogrammes, de diagrammes présente un intérêt non négligeable puisque l'esprit humain peut traiter une plus grande quantité d'information de manière visuelle et donc instinctive. La représentation globale des données permet d'obtenir une vue d'ensemble, avant de se focaliser sur des détails plus précis, de naviguer, d'investiguer et de comprendre. Cette approche relève à la fois de la visualisation scientifique, du datamining, de l'interface homme machine, de l'imagerie et des graphiques. Elle offre à l'utilisateur une représentation claire, lisible, synthétique d'une information, initialement difficile d'accès. Cette représentation est souvent complétée par des méthodes d'analyse de données en permettant une exploration encore plus efficace. Appliquée à des modèles équilibrés entre fidélité et maîtrise, la visualisation de données permet d'analyser le relationnel des diverses activités humaines, via des graphes exploitables suivant différents degrés de granularité : graphe global, graphes réduits, graphes partiels, sous graphes.

Comme une grande part des informations servant de référentiel aux décisions est d'origine textuelle, il nous faut, dans un premier temps, collecter les données brutes, sous forme semi-structurée, à l'aide d'un procédé automatisé. Dans un second temps, les informations contenues dans ces textes sont extraites sous forme d'entités, puis sont prétraitées afin de les rendre cohérentes. Enfin, les entités mises en relation sont représentées graphiquement et peuvent être manipulées de manière interactive par l'utilisateur via différentes vues complémentaires, pour une investigation efficace, notamment lorsqu'elle vient appuyer le travail de l'expert.

La représentation graphique est un excellent vecteur d'analyse des données complexes qui s'inspire d'idées ancrées dans des traditions d'origines diverses, dont la statistique graphique, la cartographie, le graphisme par ordinateur, l'interaction homme-machine, la psychologie cognitive, la sémiotique, le design graphique et l'art graphique (Tufte, 1983); (Tufte, 1990); (Tufte, 1997), (Card et al. 1999), (Herman et al. 2000), (Spence, 2000).

Problématiques

Dans un contexte de veille stratégique, l'information synthétique prend souvent une forme relationnelle : liens entre acteurs du domaine, réseaux sémantiques, alliances, fusions, acquisitions, collaborations, cooccurrences de tous ordres. En se basant sur le constat de la très forte implication du relationnel dans la prospective, nos travaux se situent à l'interface de ces deux domaines, afin d'élaborer des techniques graphiques permettant la compréhension des activités humaines, de leurs interactions mais aussi de leur évolution, dans une perspective décisionnelle.

L'un de nos objectifs est la disponibilité et la manipulation de mécanismes de visualisation pour aider l'utilisateur à faire un choix parmi différentes alternatives, pour atteindre le but qu'il s'est fixé.

Cette approche est indissociable du concept «d'action perception», (Casati et al., 2009)[1] c'est-à-dire l'exploitation des visualisations proposées. La visualisation cumulée de plusieurs périodes conduit souvent à des interprétations erronées. Afin d'identifier et d'analyser les caractéristiques temporelles, il est important de décomposer la représentation en sous graphes de périodes.

[1] http://www.institutnicod.org/detail_c.htm

Ainsi l'utilisateur obtient une vue de l'organisation générale et des propriétés structurelles. Cela permet de mettre en évidence les éléments les plus importants du graphe : ceux qui disparaissent, ceux qui émergent ou encore ceux qui persistent.

Il est possible de pousser plus loin l'analyse en traduisant de façon dynamique leur évolution dans le temps par le biais de l'analogie espace/temps. Il devient alors beaucoup plus facile d'appréhender, non seulement la fonctionnalité des structures implicites découvertes, mais aussi de comprendre leur évolution et donc de détecter les évènements clés et, éventuellement, les stratégies mises en œuvre.

De plus, un des problèmes classiques de la navigation dans des bases d'information de grande taille est le phénomène de désorientation de l'utilisateur (Nielsen, 1990). Nos travaux se basent donc sur les mécanismes cognitifs de l'utilisateur, en particulier le mécanisme de satisfaction. Pour cela, les techniques de représentations proposées doivent fournir à l'utilisateur une compréhension qualitative du contenu de l'information qu'il manipule.

Elles doivent répondre aux différents points suivants :

- Avoir une vue de l'ensemble des connaissances claire et compréhensible;
- percevoir les dynamiques des structures, en identifiant les acteurs intervenants mais surtout les acteurs de la dynamique ;
- permettre à l'utilisateur final de faire des découvertes, proposer des explications ou prendre des décisions ;
- manipuler aussi bien des motifs (clusters, tendances, émergences, anomalies) ou des ensembles d'éléments ou encore des éléments isolés ;
- communiquer efficacement des informations ;
- faciliter la découverte de connaissances grâce à une représentation graphique basée sur des corpus d'informations ;
- permette la surveillance et l'anticipation des évolutions.

Les outils de datamining permettent de construire des modèles de manière plus ou moins interactive avec l'utilisateur. On trouve aussi des produits presse-bouton qui s'adressent à des non-spécialistes. Les produits intermédiaires proposent généralement une certaine interaction avec l'utilisateur tant dans le paramétrage de l'apprentissage que pendant la recherche du modèle. D'autre part, les techniques statistiques requièrent un maniement par des statisticiens professionnels, bien que certains outils commencent à évoluer vers une meilleure convivialité et une assistance à l'utilisateur accrue.

La lisibilité ou la puissance ? Nos travaux portent ainsi sur la conception d'un outil alliant simplicité d'utilisation et précision d'analyse. Notre objectif essentiel est d'offrir un modèle présentant un bon pouvoir de prédiction et une bonne lisibilité des résultats. Il existe souvent un compromis entre clarté du modèle et pouvoir prédictif.

Contributions

Dans notre démarche, nous avons recours à la visualisation pour étudier, dans un contexte décisionnel, des données relationnelles évolutives, le plus souvent issues d'un processus d'extraction de connaissances à partir de corpus textuels volumineux et impliquant fortement l'utilisateur dans chacune des étapes.

Cette thèse a été soutenue par l'équipe des **S**ystèmes d'**I**nformations **G**énéralisés (SIG), plus particulièrement les membres de l'équipe d'**E**xploration et de **V**isualisation d'**I**nformation (EVI) de l'**I**nstitut de **R**echerche en **I**nformatique de **T**oulouse (IRIT) et la **D**élégation **G**énérale de l'**A**rmement (DGA).

Les domaines abordés dans le cadre de nos travaux concernent la fouille de texte, l'extraction de connaissance, la visualisation statique et dynamique de données notamment dans un contexte de veille stratégique. Nous avons recours aux méthodes proposées par le domaine de l'Interaction Homme-Machine (IHM) pour visualiser les données.

Notre démarche cible l'analyse des informations relationnelles évolutives reposant sur des interfaces de visualisation riches et pertinentes et des modes d'interaction adaptés aux tâches de l'utilisateur voulant effectuer la veille d'un domaine spécifique.

Les techniques de représentation graphique doivent tenir compte de deux aspects : la structure propre aux données et des conventions graphiques consensuelles. L'objectif de cette visualisation est de matérialiser des données relationnelles et évolutives ou non sous forme de graphes avec une répartition spatiale des sommets-données et des liens-relations qui est à la fois optimisée.

L'aspect statique de notre modèle de visualisation repose sur un espace de représentation, dans lequel les préceptes de la théorie des graphes sont appliqués. Le recours à des sémiologies spécifiques telles que le choix de formes de représentation, de granularité, de couleurs significatives permet une visualisation plus juste et plus précise de l'ensemble des données.

L'originalité de notre contribution repose sur la conception de graphes orientées ou non, enrichis en informations par typage des sommets et des arcs ou arêtes. Les graphes orientés proposés ne se limitent pas à la direction des arcs. La distinction claire, précise et qualificative des acteurs ou des termes origines ou extrémités enrichissent la représentation. De même, les valeurs des arêtes pour des graphes non orientés ne se limitent pas au quantitatif mais peuvent aussi représenter des caractéristiques qualitatives.

L'utilisateur étant au cœur de nos préoccupations, notre contribution repose sur l'apport de fonctionnalités spécifiques, qui favorisent l'identification et l'analyse de structures de graphes. Nous proposons des algorithmes qui permettent

- de cibler le rôle des données au sein des structures,
- d'analyser leur voisinage, tels que le filtrage, le *k*-core, la transitivité,
- de retourner aux documents sources,
- de partitionner le graphe ou de se focaliser sur des spécificités structurelles.

Une caractéristique majeure des données stratégiques est leur forte évolutivité. Or l'analyse statique ne permet pas d'étudier cette spécificité, d'anticiper les risques encourus, d'identifier les sources à l'origine d'une tendance, d'observer les acteurs ou termes ayant un rôle décisif au cœur des structures évolutives.

La prise en compte de l'aspect temporel au sein de l'analyse permet de situer les événements, les stratégies, les actions dans le passé par reconstruction de la chronologie des données, le présent par orientation temporelle et le futur par anticipation. Pour ce faire, un changement de l'espace de représentation est établi, prenant désormais en compte la dimension temporelle. Notre contribution pour les graphes dynamiques repose sur la proposition d'une stratégie de positionnement des sommets selon leurs caractéristiques temporelles. Pour ce faire, nous proposons d'utiliser des repères temporels, assimilés à des sommets spécifiques, intégrés au dessin de graphe. Dans la métaphore des horloges, les repères sont assimilés aux indicateurs des heures et les sommets sont positionnés selon leur appartenance aux différentes périodes. Cette solution permet de regrouper les sommets dont les caractéristiques temporelles sont semblables. Cette contribution est complétée par la proposition d'un algorithme de représentation de graphes, permettant de favoriser le positionnement temporel des sommets, par augmentation de l'attirance des repères des périodes durant lesquelles l'entité est fortement présente.

Le point majeur de notre contribution pour les données relationnelles et temporelles, est le morphing de graphe.

L'objectif est de faire ressortir les tendances significatives en se basant sur la représentation, dans un premier temps, d'un graphe global toutes périodes confondues puis en réalisant une animation entre les visualisations successives des graphes attachés à chaque période. Ce procédé permet d'identifier des structures ou des événements, de les situer temporellement et surtout une lecture prédictive.

Nous orientons nos travaux sur la réalisation d'un outil, nommé VisuGraph qui place l'utilisateur au centre de nos préoccupations. L'objectif principal de VisuGraph est la facilité d'utilisation et de manipulation.

Organisation du mémoire

Ce mémoire est composé de cinq chapitres, ordonné selon notre démarche d'analyse, comme le montre la Figure 1.

Les deux premiers chapitres décrivent l'existant et son organisation dans le domaine de la veille et de l'extraction de connaissances, ainsi que sur les méthodes de représentation du temps.

Notre contribution porte sur les graphes statiques, chapitre 3 et sur les représentations dynamiques, chapitre 4.

Le dernier chapitre présente les expérimentations effectuées pour valider notre modèle.

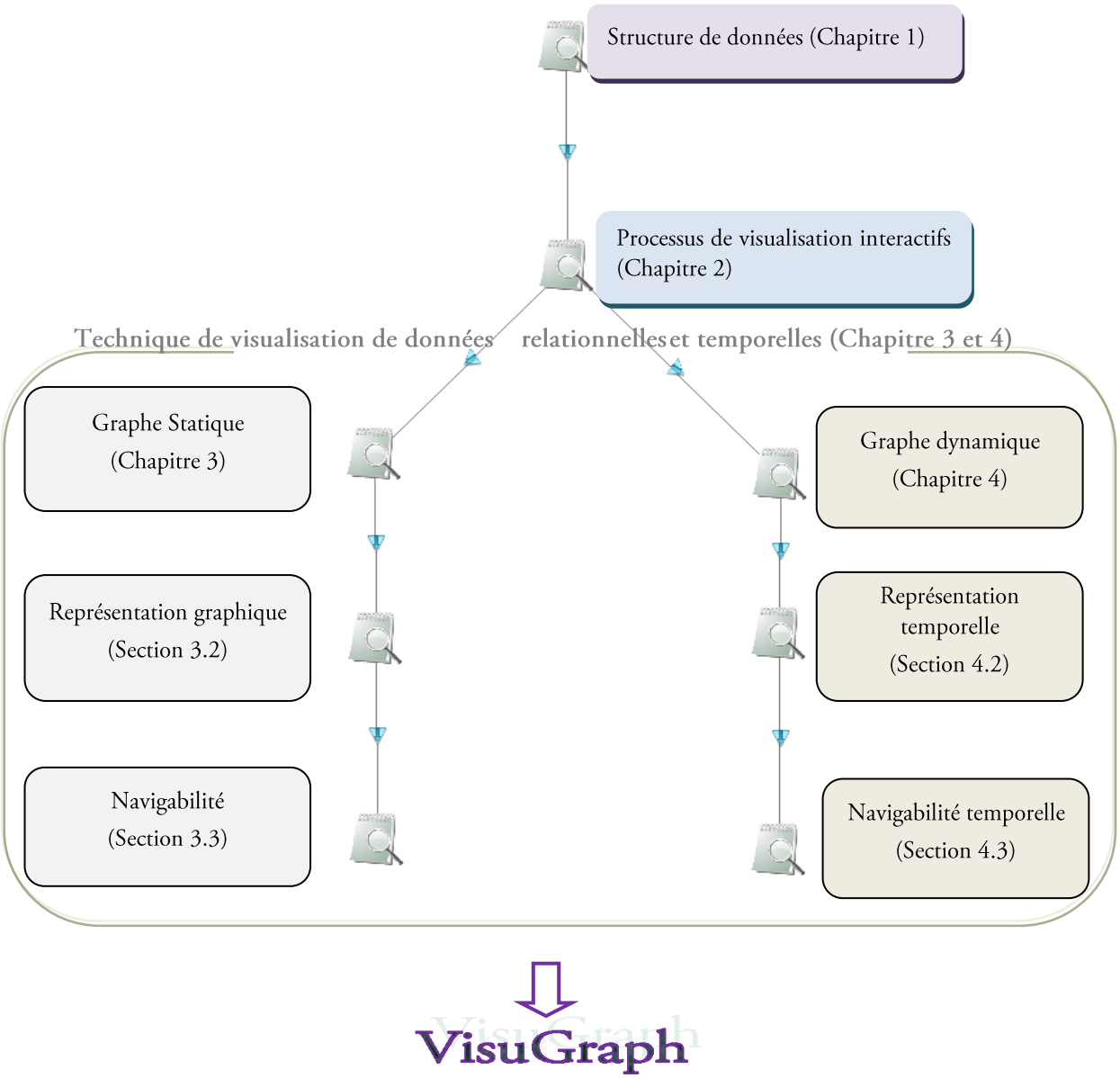


Figure 1. *Notre approche de travail.*

Le premier chapitre présente notre contexte d'application à savoir le domaine de l'intelligence économique et plus particulièrement de la veille stratégique. Pour que cette activité soit efficace, il est indispensable de passer par les phases d'observation et d'analyse de l'environnement scientifique et technologique, puis de diffusion des informations stratégique utiles à la prise de décisions.

Pour cela, le recours à l'extraction de connaissance est indispensable. Nous présentons alors les différentes étapes, de sélection des données cibles selon un besoin déterminé, puis l'extraction et le traitement de l'information.

Pour illustrer chacune de ces approches d'extraction et de connaissances, nous présentons notre méthodologie à travers des exemples concrets obtenus sur la plateforme de veille stratégique Tétralogie et nous mettons en avant tous les pré-requis indispensables à la conception d'un outil efficace permettant l'analyse de ces données récoltées et homogénéisées. Cette plateforme permet la création de matrices de cooccurrences. Ces matrices de croisements sont utilisées entre autres comme données en entrée de notre contribution.

Dans le second chapitre, nous présentons la visualisation de données en détaillant les différents types d'outils de visualisation existants et en particulier sur ceux permettant la représentation de données évolutives. Plusieurs méthodes de visualisation temporelles existent, telles que la représentation *linéaire*, dont la caractéristique majeure est de présenter un axe sur lequel le passé, le présent et le futur sont représentés, *non uniforme* permettant la visualisation de grands volumes de données avec une relation d'ordre ou *circulaire,* permettant de révéler la continuité de façon naturelle et intuitive.

Nous comparons ces techniques et nous effectuons une synthèse afin d'appuyer notre contribution sur les avantages de ces méthodes, en limitant les inconvénients. Pour cela, nous sélectionnons des critères spécifiques, issus de la littérature du domaine. Nous évaluons les outils utilisant ces méthodes et nous réutilisons ces critères ultérieurement, en chapitre 5, pour valider notre approche.

Dans le troisième chapitre, notre premier axe de contribution est développé, concernant la visualisation statique de données relationnelles. L'espace de représentation est établi et le recours à la théorie des graphes permet de mettre en place une méthodologie de visualisation. L'utilisateur étant au cœur de notre problématique, nous présentons l'aspect interactif et les méthodologies d'Interface Homme-Machine utilisées. Afin de faciliter la lisibilité du graphe, des artifices visuels sont utilisés pour coder l'information, ainsi que les liens entre ces dernières. Afin d'améliorer cette représentation, nous proposons des algorithmes de placements dirigés des sommets permettant de limiter les entrecoupements des arêtes. Nous comparons les solutions que nous proposons. Puis, nous présentons les différents graphes possibles, à savoir orientés ou non, avec des liens qualitatifs et/ou quantitatifs. Enfin, nous présentons les fonctionnalités interactives permettant une bonne analyse de structure de graphe.

Dans le quatrième chapitre, notre second axe de contribution concernant les graphes dynamiques est présenté. Basé sur des matrices de cooccurrences divisées en périodes homogènes, toutes les spécificités proposées dans le chapitre précédent sont adaptées au cas temporel, au niveau de la représentation graphique mais aussi des fonctionnalités.

Le point majeur de notre contribution pour les graphes temporels est alors présenté. Il s'agit du morphing de graphe. Après avoir défini cette notion, nous expliquons son principe. Pour chaque période considérée, un repère temporel invisible est placé sur la fenêtre de représentation, telles les heures sur un cadran d'horloge. Puis les données sont placées selon leur appartenance aux différentes périodes, à une distance relative de ces repères. Enfin, nous étudions deux aspects distincts du morphing à savoir la visualisation de l'évolution des données par transition ou non et nous appliquons les fonctionnalités d'analyse de structure aux résultats obtenus. Enfin, nous ouvrons notre contribution à des domaines non évolutifs, permettant de prendre en compte, à la place du temps, une autre dimension.

Dans le cinquième chapitre, nous expérimentons les contributions présentées, en évaluant les méthodes selon les critères précédemment retenus. Puis, une enquête est réalisée auprès d'une population ciblée, constituée d'individus informaticiens, statisticiens, de l'Intelligence Economique et de personnes dont la spécificité n'est pas liée directement aux domaines étudiés dans ce mémoire. Suite à la réalisation d'études par notre équipe, la population interrogée a effectué les mêmes en jugeant le niveau de difficulté de chacune des tâches.

Les résultats sont ensuite comparés à ceux de notre équipe et un bilan est effectué pour cibler les points forts et les points faibles de nos contributions, engendrant une discussion de nos résultats.

Enfin, nous concluons sur l'ensemble des travaux présentés dans ce mémoire. Nous présentons nos perspectives de recherche sur ces axes d'étude, en reprenant les résultats du sondage, mais aussi en évaluant les problématiques engendrées par nos travaux et que nous souhaiterions développer par la suite.

Chapitre 1.
De la découverte de connaissance à son utilisation en veille stratégique

> « *Le savant n'est pas l'homme qui fournit les vraies réponses. C'est celui qui pose les vraies questions* » (Lévy-Strauss, 1963).

1.1. Préambule ____ *20*

1.2. Processus de veille ____ *21*

1.2.1. Intelligence économique et veille ____ 21

1.2.2. Concept de veille stratégique ____ 23

- ✓ Définition ____ 23
- ✓ Objectifs de la veille stratégique ____ 24

1.3. Découverte de connaissance ____ *24*

1.3.1. Définition ____ 24

1.3.2. Information et connaissance ____ 25

- ✓ Qu'est-ce que l'information? ____ 25
- ✓ Qu'est-ce que la connaissance? ____ 26

1.3.3. Présentation des étapes du processus de découverte de connaissance ____ 26

1.3.4. Sélection des données cibles ____ 27

- ✓ Sources formelles ____ 27
- ✓ Sources informelles ____ 28

1.3.5. Prétraitement ____ 29

- ✓ Metadonnées ____ 29
- ✓ Niveaux d'information ____ 30
- ✓ Multi-termes ____ 30
- ✓ Synonymie ____ 31
- ✓ Filtrage ____ 34

1.3.6. Transformation ____ 34

1.3.7. Fouille de données ____ 35

- ✓ Entités textuelles ____ 35
- ✓ Différentes mesures du relationnel ____ 36
- ✓ Matrices et cubes ____ 37

1.3.8. Evaluation du modèle ____ 44

1.3.9. Validation ____ 45

1.3.10. Discussion ____ 45

1.4. Synthèse ____ *47*

1.1. Préambule

Dans la société de l'information et du savoir, la performance organisationnelle passe par la gestion des savoirs et des connaissances. Actuellement, un grand nombre d'informations ponctuelles, indexée, pertinente est organisée, grâce au web sémantique, visant à rendre le contenu des ressources du World Wide Web accessible et utilisable par les programmes et agents logiciels, grâce à un système de métadonnées formelles. De plus, les travaux récents en Recherche d'Information permettent de cibler davantage les ressources pertinentes en ciblant le contexte ou encore en élaborant un profil utilisateur afin de répondre de façon précise aux besoins informationnels.

Au sein des bases de données accessibles se cachent des informations d'importance stratégique mais qui ne sont pas explicitées (Martinez et Guillaume, 1998). En effet, situées au cœur d'un flux important d'informations dont la pertinence est faible, elles sont à priori difficilement décelables.

D'autre part, une information partielle risque de déboucher sur une décision erronée, si des éléments environnementaux venaient à manquer.

L'entreprise n'est plus, depuis la crise, le cœur du problème, il y a des enjeux beaucoup plus fondamentaux, tels que le développement durable, la régulation des marchés, l'imbrication économiques : il y a plus à gagner d'initialiser des partenariats stratégiques que de rentrer en concurrence ouverte. Il y a une mutation de l'intelligence vers une gouvernance globale. D'où une nécessité d'analyser le relationnel dans sa globalité planétaire, et de pouvoir à tout moment zoomer sur une spécificité donnée mais sans perdre de vue le modèle complet.

vis-à-vis des nouveaux enjeux et du nombre d'acteurs impliqués il est important de disposer et surtout de maîtriser l'information afin de pouvoir anticiper, décider et agir rapidement et efficacement.

La veille stratégique devient alors une nécessité dans un monde où l'innovation et la compétitivité qui étaient le moteur de l'économie risquent d'être progressivement remplacées par des choix plus durables sur le plan écologique, financier, sociétal, ...

La maîtrise de l'information stratégique passe par le recours à des outils puissants, permettant de contrôler pleinement des flux de données en perpétuelle expansion. Cette maîtrise du veilleur passe par les étapes « *d'observation et d'analyse de l'environnement scientifique et technologique, suivie de la diffusion bien ciblée aux responsables des informations sélectionnées et traitées utiles à la prise de décisions stratégiques* », selon (Jakobiak, 1995). Cependant, dans la logique de contrôle de l'environnement par le biais des informations, il est essentiel de suivre ces données dans le temps. Jakobiak écrit que « *le suivi systématique de l'évolution technologique des domaines critiques de l'entreprise permet, c'est un des objectifs de la veille technologique, de saisir les opportunités de développement sans perdre de temps* ». La notion de perception du temps est donc primordiale dans tout travail de veille.

Dans ce chapitre, nous définissons, dans un premier temps, le contexte dans lequel se situe notre contribution, à savoir le domaine de l'intelligence économique et plus particulièrement de la veille stratégique. L'approche que nous favorisons consiste à déterminer les quatre principes majeurs constituant le cycle de la veille, comme le montre la première colonne de la figure 2. Nous détaillons plus précisément le concept et les axes de la veille stratégique dans la première partie de ce chapitre.

Puis, nous précisons les étapes intermédiaires entre ces principes de veille, illustrées par la colonne du milieu de la figure 2. Enfin, nous approfondissons chacune de ces approches en les mettant en pratique par la plateforme de veille scientifique et stratégique (Dousset et Benjamaa, 1988), (Dkaki et al., 1995), (Dousset et al., 1995), (Dousset, 1995), (Dousset, 2006), (Dousset, 2009) développée au sein de l'équipe des Systèmes d'Information Généralisés de l'Institut de Recherche en Informatique de Toulouse (IRIT), montré dans la troisième colonne de la figure 2, et nous mettons en avant tous les pré-requis, indispensables à la conception d'un outil efficace (phase d'évaluation, d'interprétation et d'exploitation) permettant l'analyse de ces données récoltées et homogénéisées.

Afin d'exploiter les données dans un contexte de veille stratégique, cette approche via cette plate-forme est composée d'un module de manipulation de corpus permettant d'interfacer les bases de données bibliographiques ou autres sources d'informations. Elle permet de découper les sources en unités pertinentes.

Ainsi, une fois cette étape de découpage réalisée, les données sont alors croisées et les méthodes d'analyse de données sont applicables, telles que l'Analyse en Composantes Principales (ACP), l'Analyse Factorielle des Correspondances (AFC), la Classification Ascendante Hiérarchique (CAH), la Classification par Partitions (CPP). Des méthodes d'analyse de l'évolution (relative et absolue) comme les rotations procustéennes, l'analyse relationnelle des données portant sur les liens (secondaires, ternaires, ...) sont également proposées.

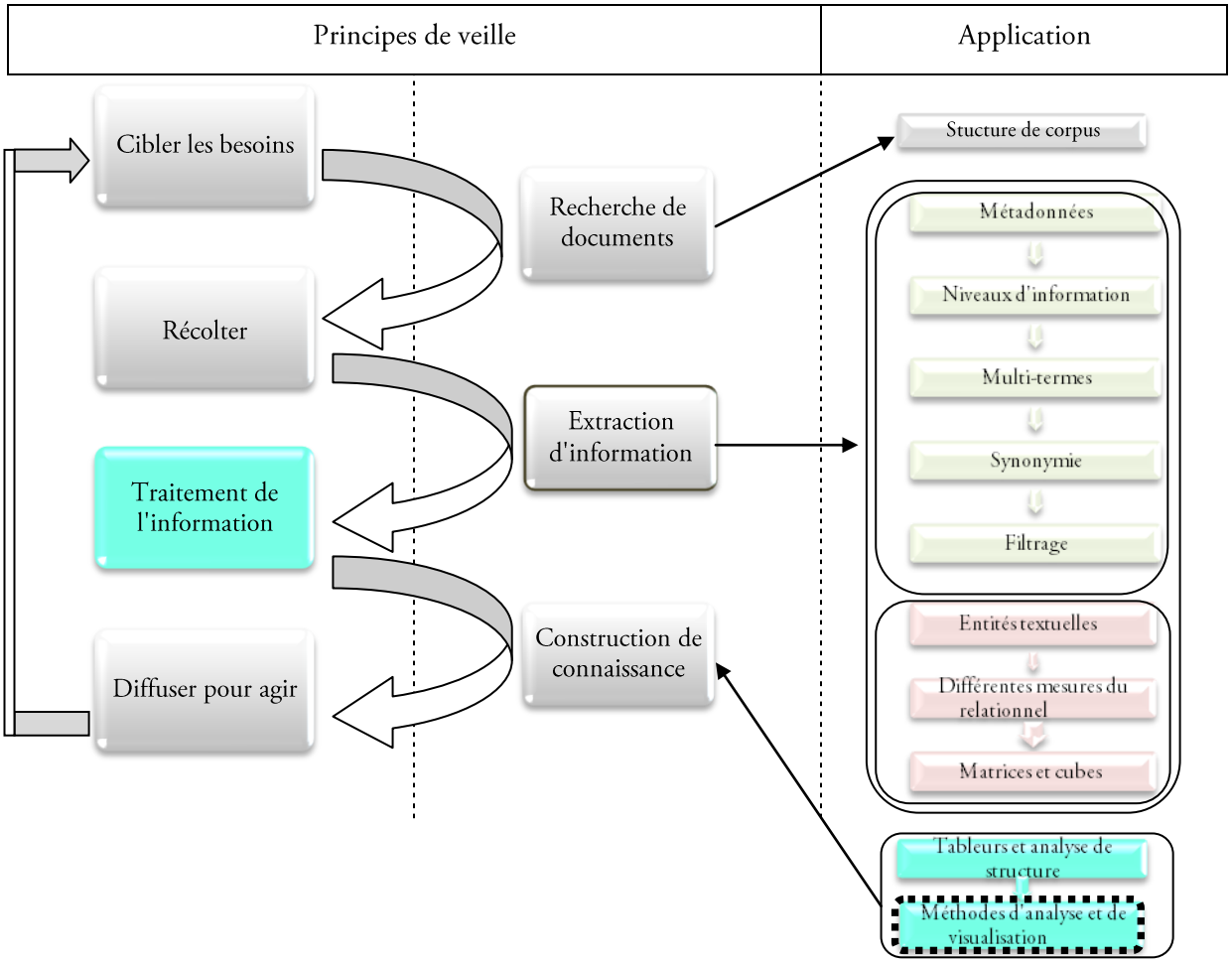


Figure 2. *Principe de veille et découverte de connaissance.*

Nous étudions le besoin de créer un outil supplémentaire permettant la visualisation de données relationnelles, dans un contexte évolutif, que nous présentons dans notre contribution, dans les chapitres 3, 4 et 5.

1.2. Processus de veille

1.2.1. Intelligence économique et veille

"*La compétence est individuelle, l'intelligence est collective.*"

La démarche d'Intelligence Economique (IE) est par nature un processus transverse, et met en réseau les hommes autour d'un besoin d'information, qui va éclairer la décision.

Henri Martre définit l'Intelligence économique comme « *L'ensemble des actions coordonnées de recherche, de traitement et de distribution en vue de son exploitation, de l'information utile aux acteurs économiques (…)* », dans (Martre, 1994). L'IE englobe initialement des notions comme la guerre économique, la sécurité économique, le renseignement, le lobbying, la veille, mais aussi, plus récemment, la gestion des connaissances, les nouvelles technologies de l'information et de la communication, l'analyse stratégique, la prospective, la gouvernance, comme le montre la figure 3.

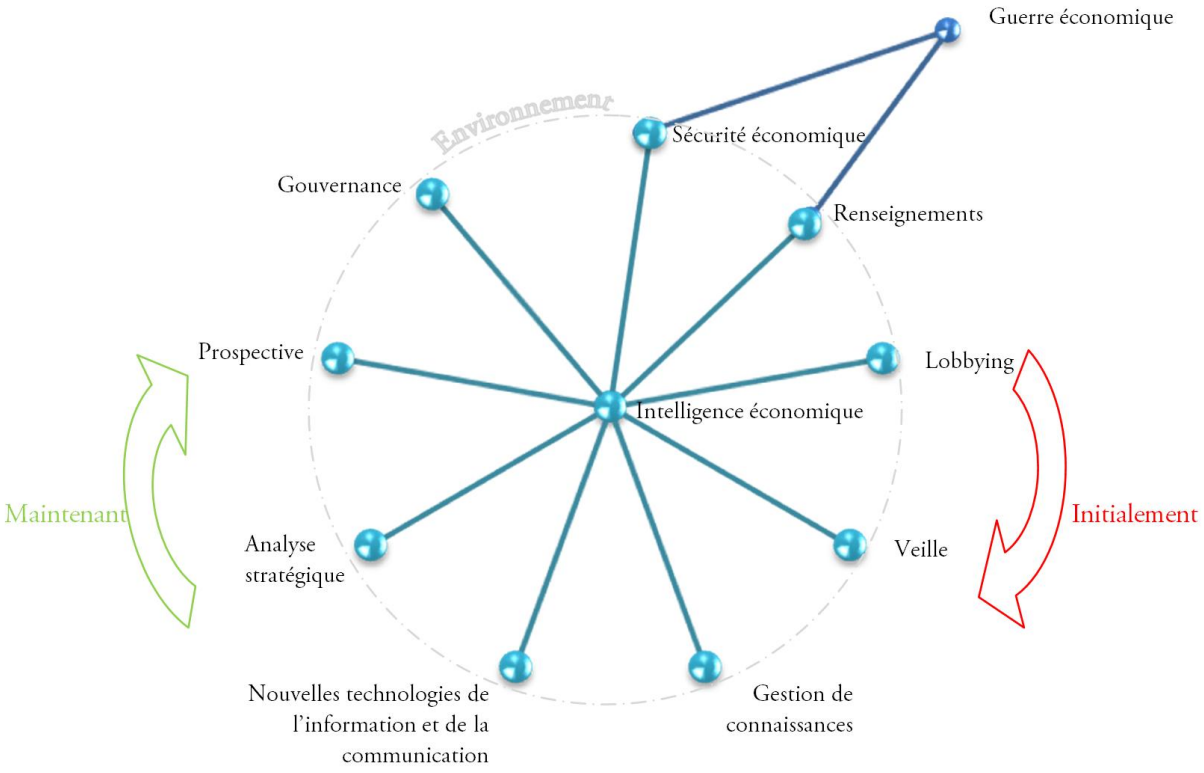


Figure 3. *Le concept d'intelligence économique*

Les concepts de veille et d'intelligence économique sont fortement imbriqués : l'enclenchement du processus d'IE est voulu par le décideur pour couvrir un besoin informationnel précis. Maintenant que tout le processus de veille associé est activé, il continue de fonctionner sans que nécessairement d'autres demandes lui parviennent tels que les abonnements à des flux RSS, à des listes de diffusion, à des alertes mail, à des alertes sur les réseaux,... Le point d'activation de la veille se situe au niveau des sources d'informations, alors que celui de l'IE se situe au niveau du décideur. Les définitions « économiques » de l'IE font souvent appel à des points pouvant influencer les entreprises dans leur action. Selon (Martinet et Ribault, 1989), « *l'entreprise aura besoin d'identifier et de connaître les éléments clés des paramètres qui conditionnent son existence* ». La veille stratégique permet l'observation et l'analyse de l'environnement scientifique, technique, technologique et économique de l'entreprise pour en détecter les menaces et saisir les opportunités de développement (Hussein et al., 2004).

Cette activité de veille met en jeu les observateurs, les experts et les décideurs. Ainsi, l'IE englobe la veille stratégique, contenant les différents types de veille, comme le montre la Figure 4.

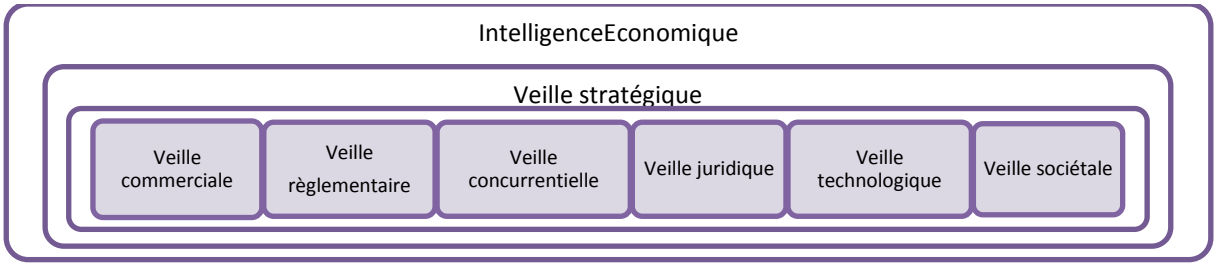


Figure 4. *De l'IE aux différents types de veille.*

La veille stratégique peut se décliner en fonction de ses objectifs en plusieurs types de veille.

La veille commerciale s'intéresse au pouvoir de négociation des clients et des fournisseurs.

La veille réglementaire prend en compte l'évolution des textes de lois, des normes nationales ou internationales, des accords commerciaux, des dépôts de brevets, des nouveaux labels de produits.

La veille concurrentielle permet d'identifier et de surveiller les firmes rivales et prend en compte les menaces que représente l'arrivée de nouveaux acteurs dans le secteur d'activité.

La veille technologique est plus orientée vers le développement et touche donc le monde industriel et commercial. Ses supports traditionnels sont les bases de brevets, les documentations techniques et plaquettes publicitaires, la presse et les sites Internet de l'ensemble des acteurs du domaine étudié.

La veille scientifique s'intéresse essentiellement au monde de la recherche, elle s'appuie pour cela sur les très nombreuses bases bibliographiques disponibles en ligne ou sur CD-Rom et depuis quelques temps sur les innombrables écrits scientifiques présents sur Internet.

1.2.2. Concept de veille stratégique

✓ ***Définition***

La démarche de veille est définie comme une « *attitude organisée d'écoute des signaux provenant de l'environnement de l'entreprise susceptible de mettre en cause ses options stratégiques* », selon les travaux de (Martinet et Ribault, 1989). La veille stratégique est le processus informationnel volontariste visant la recherche des informations à caractère anticipatif concernant l'évolution de l'environnement socio-économique d'un domaine dans le but de se créer des opportunités et de réduire ses risques liés à l'incertitude.

Une autre définition de la veille est développée par l'AFNOR[2] comme une « *activité continue et en grande partie itérative visant à une surveillance active de l'environnement technologique, commerciale, etc., pour anticiper les évolutions* » (Norme XP X50-053). Pour ce faire, l'AFNOR propose un processus de veille type, Figure 5, dans lequel différentes étapes sont définies, de l'élaboration des axes de surveillance aux résultats obtenus de veille.

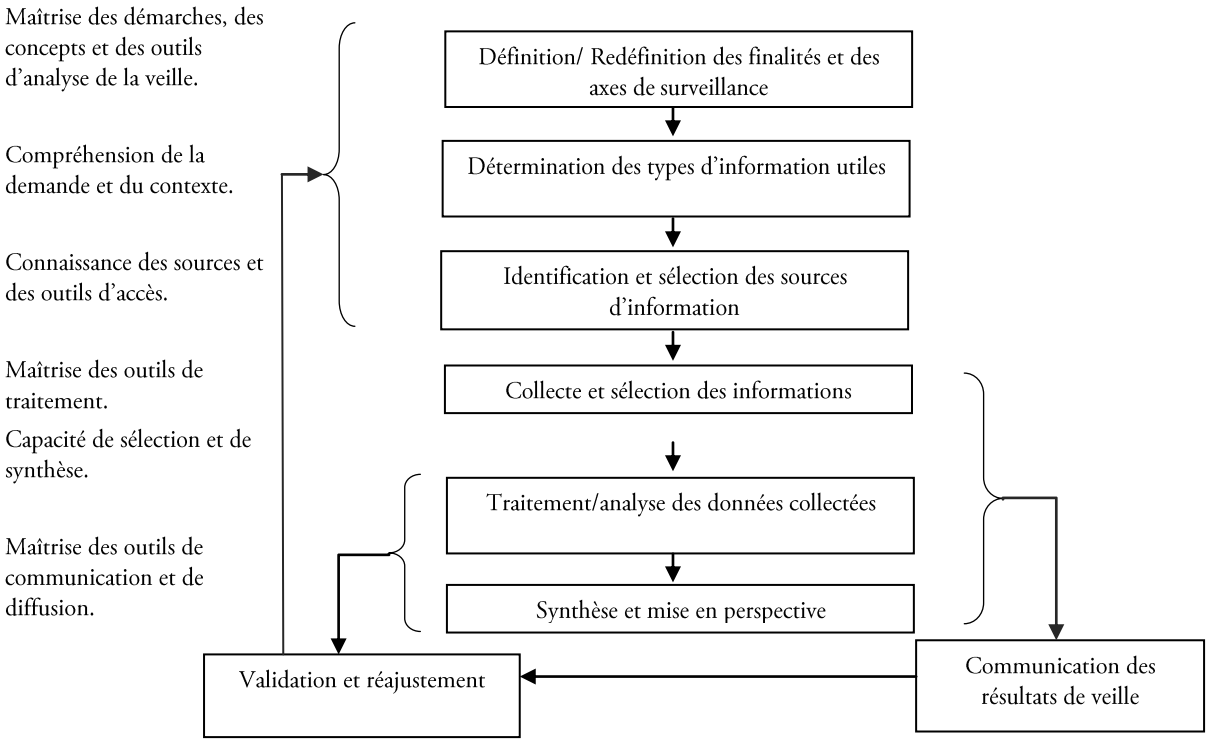


Figure 5. *Processus de veille type adapté de l'Afnor X50-053*

L'enjeu d'une veille scientifique et technologique est donc de détecter les tendances, de surveiller les découvertes, les innovations afin de positionner l'entreprise par rapport à celles-ci.

[2] AFNOR, **A**ssociation **F**rançaise de **Nor**malisation, est un groupe international de services organisé autour de 4 grands domaines de compétences : la normalisation, la certification, l'édition spécialisée et la formation.

Cette dernière peut alors se positionner soit en concurrence, si elle se situe sur le même marché, soit éventuellement en collaboration, si son marché est différent. Ainsi la veille consiste en un processus anticipatif d'observation et d'analyse de l'environnement, suivi de la diffusion ciblée des informations utiles à la prise de décision.

✓ ***Objectifs de la veille stratégique***

Les objectifs de la veille stratégique sont d'avoir une vue d'ensemble compréhensible, de développer, élargir, améliorer mais aussi recentrer le champ d'activité, de surveiller et anticiper les évolutions (technologiques, de marché, ...) pour appréhender les menaces et les opportunités de développement. Elle permet aussi de percevoir les dynamiques des structures, en identifiant les acteurs intervenants dans la structure mais surtout les acteurs de la dynamique, de faire de l'information un outil de développement à haute valeur ajoutée, de favoriser la pro activité plutôt que la réactivité, de prendre des décisions avec une meilleure sécurité, d'être en amont des projets innovants.

1.3. Découverte de connaissance

La veille stratégique cible la prise de décisions qui engage le devenir, l'évolution d'un domaine spécifique, en relation avec les changements de son environnement socio-économique. De ce fait, la découverte de connaissance est indispensable dans une activité de veille.

1.3.1. Définition

Les systèmes d'informations sont actuellement alimentés par une quantité sans cesse croissante de données électroniques de différentes natures. Les chercheurs de l'université de Berkeley ont estimé que le volume d'information a augmenté de 30% chaque année entre 1999 et 2002 (Lyman et Varian, 2003). Une telle masse d'informations est bien trop complexe pour pouvoir être appréhendée par un utilisateur sans une assistance informatique de qualité (Martinez et marchand, 1998), (Martinez et Mouaddib, 1999), (Martinez et Loisant, 2002), (Chrisment, 2007).

Parallèlement, le domaine de l'Extraction de Connaissances à partir de Données (ECD) s'est développé pour répondre à une volonté de découverte de connaissances via ces masses de données.

L'ECD est « un processus non trivial d'identification de connaissances inconnues, valides, potentiellement exploitables et compréhensibles dans les données », selon (Fayyad et al., 1996).

L'intérêt des connaissances extraites est validé en fonction du but de l'application. Seul l'utilisateur peut déterminer la pertinence des résultats obtenus par rapport à ses objectifs.

Deux grands courants de recherche peuvent être distingués : la recherche d'informations spécifiques (RI) et l'extraction de connaissances où l'objectif est d'analyser le contenu d'une collection de documents afin de répondre à un besoin, a priori non défini, d'information. Dans la logique plus traditionnelle de recherche d'information, le moteur de recherche permet l'extraction de documents à partir d'un univers lexical considéré comme non structuré, parfois indexé par des mots-clés; il suffit simplement d'indiquer dans la requête le contenu que l'on souhaite observer dans les documents sélectionnés. La seconde approche, plus récente, consiste en l'extraction de connaissances représentées par un ensemble de traits structuraux caractéristiques de la collection de documents analysée, comme le montre la Figure 6. Néanmoins ces deux types d'approches sont le plus souvent complémentaires et finalement peu dissociables.

Deux dénominations courantes, mais pas tout à fait équivalentes, se rencontrent habituellement dans la littérature anglosaxonne : le Knowledge Discovery in Databases (KDD) et le Data Mining (DM). La différence entre ces deux désignations réside dans le type d'approche utilisée : intelligence artificielle pour le KDD avec utilisation d'heuristiques provenant de l'apprentissage symbolique, statistique pour le DM considéré comme une industrialisation des techniques d'analyse des données.

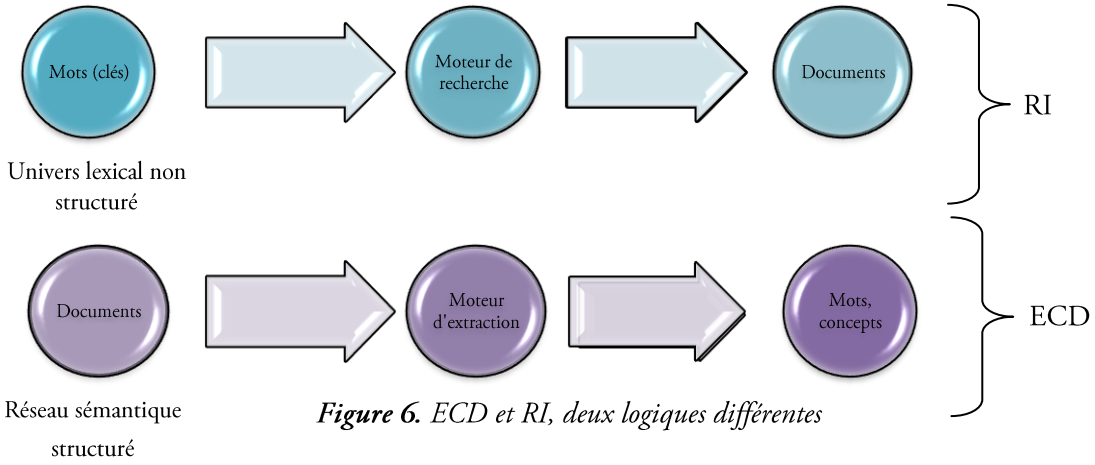


Figure 6. *ECD et RI, deux logiques différentes*

A l'image des chercheurs d'or qui doivent transporter, broyer, trier et filtrer de grandes quantités de terre pour extraire le métal précieux, le Data Mining est l'art d'interpréter intelligemment, à l'aide d'outils informatiques, les informations disponibles pour parvenir à des connaissances opérationnelles. L'objectif est d'acquérir des connaissances enfouies dans des ensembles de texte, d'en extraire des informations utiles et de les analyser. Cette étape se nomme la « fouille ».

Plusieurs types d'analyse de texte existent.

L'analyse lexicale identifie le lexème, unité élémentaire de signification. Elle suppose la prise en compte du contexte ou du domaine.

L'analyse morphologique vise à ramener tous les mots reconnus dans une phrase à leur forme canonique, en séparant les variations grammaticales (pluriels, conjugaisons, flexions,…).

L'analyse syntaxique étudie les relations entre les mots dans une phrase.

L'analyse sémantique permet d'analyser le sens. Elle consiste à associer l'ensemble des éléments linguistiques en une représentation pouvant en corriger le sens.

1.3.2. Information et connaissance

La découverte de connaissance se base sur la définition et sur la différence entre l'information et la connaissance. Dans cette section, nous définissons ces deux principes.

- ✓ ***Qu'est-ce que l'information?***

L'information est « *une connaissance inscrite sous forme écrite, orale ou audiovisuelle sur un "support spatio-temporel" (imprimé, signal électrique, sonore, etc.). L'information comporte un élément de sens. C'est une signification transmise à un être conscient par le moyen d'un message inscrit sur un support* », d'après (LeCoadic, 2004).

L'information portée par des sujets correspond à la connaissance et à la compétence des sujets ou individus. L'information objective est représentée par des textes et les publications scientifiques stockées dans les bases de données. Quand nous parlons d'analyse, le sujet sur lequel elle s'exerce est l'information au sens objectif, celle qui se trouve stockée dans les bases de données. Nous nous intéressons à la formalisation et à la représentation de la connaissance objective, enfouie dans les données en information scientifique et technique, telles que les publications scientifiques, la documentation technique, les brevets,…

Selon (Goria, 2006), il existe une hiérarchie entre les notions de « connaissance », d' « information » et de « données ». D'après la théorie de l'information développée par (Shannon et Weaver, 1975), les données sont susceptibles de devenir des informations lorsqu'elles sont perçues et interprétées par l'individu comme lui apportant des éléments nouveaux. En ce qui concerne le concept d'information, on distingue trois types d'information.

L'information *blanche* est publique ou réservée, elle est issue de banque de données, publications scientifiques, périodiques, plaquettes d'entreprises, entretiens avec des experts de centres techniques, des fournisseurs, des clients, des partenaires... elle est donc libre d'accès et d'exploitation.est aisément et licitement accessible [Norme AFNOR XP X 50-053].

L'information *grise*, est essentiellement réservée, elle se constitue d'informations ayant fait l'objet d 'une appropriation par l'obtention d'un droit privatif : brevets, modèles, droits d 'auteurs... Son exploitation est limitée, soumise à l'autorisation du titulaire.

L'information *noire* est confidentielle, protégées par le secret : secrets de fabrication, secrets commerciaux tels que les études de marché, prévisions de vente, ou relatifs à l'organisation,... Son accès est soumis à des risques de sanctions civiles et pénales telles que le vol, le débauchage, la corruption, … et son exploitation est libre si accès légal, sauf copie servile ou agissements parasitaires.

Seulement l'information blanche et la grise concernent la veille.

✓ ***Qu'est-ce que la connaissance?***

La connaissance est constituée d'éléments permettant de construire de nouveaux faits ou permettant de déterminer de nouvelles actions à entreprendre et non pas comme des éléments descriptifs, (Pitrat, 1990). La connaissance est le résultat d'un assemblage d'informations traitées auquel l'esprit humain a pu assigner un sens (Malhotra, 2000).

C'est une manière de comprendre, de percevoir, elle régit les rapports entre les afférences cognitives de l'individu et le monde extérieur.

La connaissance résulte d'un processus complexe ; elle est utile à un agent donné, bien que la capacité de transformation de données brutes en connaissance constitue une valeur ajoutée incontestable pour l'entreprise. Mais c'est en fait l'individu qui applique son intelligence pour apporter signification et pertinence à l'information, transformant ainsi l'information en connaissance.

1.3.3. Présentation des étapes du processus de découverte de connaissance

Le processus d'Extraction de Connaissances à partir de Bases de Données (ECBD, où encore KDD, en anglais : **K**nowledge **D**iscovery in **D**atabases) que nous présentons maintenant est semi-automatique et itératif, découpé en six parties : la sélection, le prétraitement, la transformation, la fouille de données et l'interprétation/l'évaluation. L'enchaînement des différentes étapes est présenté dans la Figure 7.

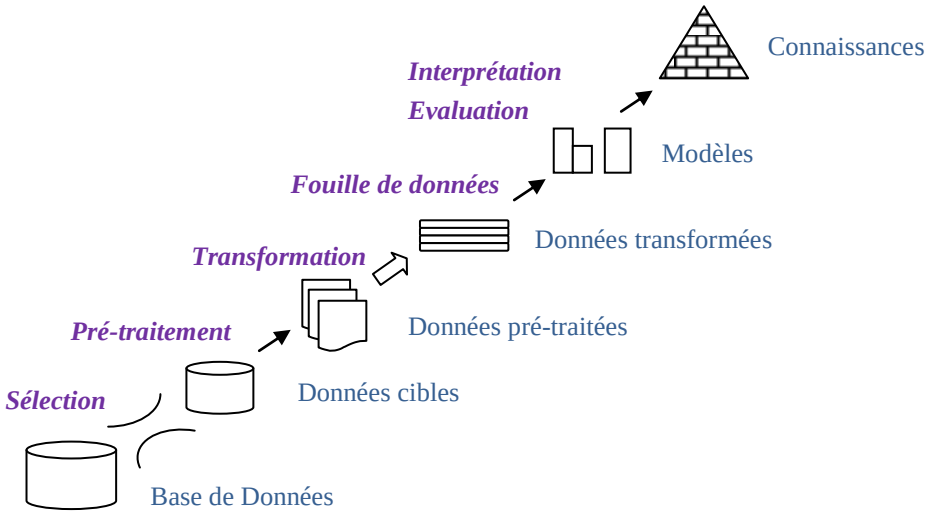


Figure 7. *Les étapes du processus d'Extraction de Connaissances à partir de Bases de Données (ECBD) ((Fayyad et al., 1996), cité par (Toussaint, 2004)).*

En pré requis, il est important de connaître le domaine d'application et de définir les buts de l'application. Dans les sections suivantes, nous développons chacune de ces étapes. Les travaux réalisés par notre équipe permettent d'effectuer chacune de ces étapes. Pour positionner l'avancement des recherches de notre laboratoire, nous illustrons chacune de ces phases en utilisant la plateforme de veille stratégique Tétralogie, développée au sein de l'équipe des Systèmes d'Informations Généralisés – Exploration et Visualisation d'Information.

1.3.4. Sélection des données cibles

Une première phase consiste à l'élaboration d'un corpus ciblé, en fonction de l'objectif d'exploration, qui par la suite est analysé via les méthodes de cette plate-forme. On emploie souvent le terme corpus pour désigner de vastes ensembles de données textuelles semi ou totalement structurés et mis sous forme électronique.

Un corpus est une collection de données langagières qui sont sélectionnées et organisées selon des critères linguistiques explicites pour servir d'échantillon du langage, selon (Habert, 2000).

De plus, un corpus électronique est un corpus qui est encodé de manière standardisée et homogène pour permettre des extractions non limitées à l'avance. L'origine et la provenance des données langagières sont notées.

En effet, la simple existence sur support électronique ne fait pas d'un ensemble de textes un corpus électronique. Encore faut-il que ce document respecte des conventions de représentation, de codage répandues, voire consensuelles, qui permettent la transmission et la réutilisation des données textuelles.

L'étape de sélection permet de se focaliser suivant des critères prédéfinis, sur des données supposées à la fois « interprétables » et à potentiel informatif.

La préparation des données consiste dans un premier temps à les *sélectionner* en accord avec les objectifs que l'on s'impose, en ayant recours aux techniques de recherches d'information, (Maron et Kuhns, 1960), (Saltan, 1970), (Rocchio, 1971). Ce processus (Salton et McGill, 1984) cherche à mettre en correspondance une collection de documents et le besoin de l'utilisateur (Maniez et Grolier, 1991), traduit sous la forme d'une requête (Kleinberg,1999) à travers un système d'information. Ce dernier est composé d'un module d'indexation automatique ou semi-automatique ; d'un module d'appariement document-requête et éventuellement d'un module de reformulation de la requête.

Différents modèles sont utilisés selon les moteurs de recherches utilisés, pour l'appariement entre la requête et le document, tels que le modèle probabilistes (Maron et Kuhns, 1960), booléen (Salton, 1971), (Robertson, 1977), flou (Paice, 1984), connexionniste (Mothe, 1994), (Boughanem et al., 2000), flexible (Sauvagnat, 2005), …

Les données sélectionnées proviennent le plus souvent de bases de production ou d'entrepôts dans lesquels les données sont structurées en champs typés.

Deux types de sources à identifier, sélectionner et collecter existent :

- ✓ ***Sources formelles***

L'information est dite formelle dès lors qu'elle est publiée sur support papier, informatique, microfilm…Elle peut être structurée ou non, mais il s'agit dans tous les cas d'une information directement accessible (sous réserve des contraintes définis par son auteur) et exploitable. Ce type de sources correspond à l'information blanche.

Les sources formelles sont composées principalement de la presse, la télévision, la radio, les livres, banques de donnée et CD-ROM, les brevets, les informations légales, les études réalisée par des prestataires publics ou privés, internet. Ces sources ont l'avantage d'être sûres et assez exhaustives, de faible coût (sauf le cas certaines banques de données telles que Pascal, …), faciles d'accès. Les banques de données sont principalement utilisées, elles peuvent être factuelles, en texte intégral ou encore bibliographiques.

Les banques de données factuelles correspondent à des données brutes bien structurées, telles que la Statistique et les données chiffrées.

Les banques de données en texte intégral concernent les données à dominante littérale (juridique, économique, …). Elles ont l'inconvénient d'être généralement mal structurées.

Les banques de données bibliographiques correspondent à des données sous forme de références bibliographiques.

Dans un contexte de veille stratégique, les bases de données les plus consultées sont à dominante scientifique, technologique, réglementaire et se trouvent sur des bases bibliographiques. Chaque référence fait l'objet d'une structuration complète, composée par un titre, un/des auteur(s), une/des affiliation(s) et l'identification de la source.

Parmi les bases de données les plus intéressantes, nous pouvons citer dans le domaine économique Factiva[3], physique avec Inspec[4], orienté entreprise avec Kompass Europe[5], multidisciplinaire avec Pascal[6], médical avec PubMed[7], …

✓ ***Sources informelles***

L'information informelle est constituée de toutes les informations non formalisées et non disponibles directement. Il est donc nécessaire d'entreprendre des démarches directes auprès des détenteurs supposés de cette information. Ce type de sources correspond à l'information grise. Ces sources peuvent être les expositions et les salons, les fournisseurs, les colloques, les congrès, les clubs: on y échange des informations, on y communique. L'information qui circule alors peut être d'une grande valeur stratégique, les concurrents (portes ouvertes, communication commerciale et financière, publication de journal interne,…), les sources internes de l'entreprise : 80% des informations que recherche un décideur se trouvent dans son entreprise, certains sites web : des sites personnels, des études et recherche menées par un groupe d'étudiants ou de thésard, etc… les réseaux personnels : le cousin, l'ami commercial de chez X, le représentant de Y, le voisin qui travaille chez Z, l'écoute, « par hasard », d'une conversation dans un avion, un train, lors d'un dîner, … dans la limite de la légalité et de la déontologie.

Les données récoltées sont ensuite analysées par rapport aux besoins émis au début du projet. A ce stade du projet, la veille ne sert à rien si le résultat n'est pas diffusé auprès des collaborateurs qui pourront agir en conséquence. Il faut donc que l'information sélectionnée et mise en avant remonte vers les acteurs cibles. Les résultats du traitement des données représentent une base de travail pour les différents services : recherche et développement, commercial, ... Une information est fondamentalement une action en devenir pour qui sait la mettre en perspective, elle procure la capacité à mettre en œuvre des actions en vue d'influer sur l'environnement. Néanmoins, l'information est périssable, sa valeur diminue avec le temps et globalement plus la source est formalisée, plus l'information est obsolète.

Dans notre approche, les données cibles sont sélectionnées (Dkaki et al., 1997) en fonction de l'objectif d'exploration. Comme le décrit (Rousseau-Hans, 1998), l'utilisation d'outils infométriques dans une démarche de veille technologique permet une approche globale de l'information contenue dans un corpus. Ces outils découpent d'abord les données en unités (mots, dates ou chaînes de caractères), puis appliquent des calculs mathématiques et statistiques afin d'obtenir sous forme de graphiques ou de cartes une représentation des unités en fonction de relations ou proximités calculées.

L'utilisateur va effectuer une recherche d'information, en interrogeant des sources identifiées comme pertinentes issues de déchargement de CD/Rom, télédéchargement[8] de bases de données en ligne, aspirateur d'URL (Wisigot, MémoWeb, Teleport pro), aspirateurs de site tels que MémoWeb ou Teleport pro permettant de récupérer l'intégralité ou une partie d'un site, agent de monitorage, tels que Pragtec, Digimind, Kbcrawl. ils permettent de surveiller les changements qui interviennent sur les pages Web ou dans des articles de nouvelles préalablement sélectionnés.

---

[3] Dow jones Factiva. http://factiva.com/ . Base de données de presse et d'informatique économique.

[4] EBSCO Industries. http://support.epnet.com/. Base de données bibliographiques en physique.

[5] http://www.kompass.fr/ip. Base de données sur les entreprises européennes.

[6] INIST (Institut National de l'Information Scientifique et Technique). Base de données multi-disciplinaire.

[7] Base de données bibliographiques, interrogeable par le Mesh (Medical subject heading)

[8] En informatique, le téléchargement est l'opération de transmission d'informations — programmes, données, images, sons, vidéos — d'un ordinateur à un autre via un canal de transmission, en général internet. Wikipédia http://fr.wikipedia.org/wiki/Télédéchargement

Il permet d'automatiser les tâches reliées à la surveillance des informations sur le Web. Enfin, il peut avoir recours à des robots spécialisés permettant la segmentation de requêtes complexes ou la récupération massive de documents depuis les bases de données du web.

Les corpus utilisés sont composés de notices, c'est à dire des documents structurés en champs (Dkaki et al., 2000). Le mot, unité sémantiquement trop pauvre, a été supplanté par la notion de termes que l'on peut associer à un concept dans une ontologie (Chrisment et al., 2006), (Chrisment et al., 2008).

Un champ, l'unité de base, est le contenu informationnel identifié par une balise et une donnée, par exemple auteur, date, adresse, organisme. Un item est le contenant du champ, un terme, c'est-à-dire la donnée.

Il peut être (Mothe, 2000), (Dousset, 2003) :

- *mono-valué* ne pouvant avoir qu'une seule valeur possible telle que la date ou encore la langue. Par exemple PUBLICATION YEAR=2007 ;
- *multi-valué* en ayant plusieurs valeurs, comme par exemple plusieurs noms d'auteurs pour un article coécrit, délimités par des séparateurs;
- *diversifié*, si le champ contient plusieurs valeurs représentant des concepts différents. Pa exemple, SOURCE=2007-11,32p., ce champ peut se décomposer en une date de publication : 2007-11, qui se divise elle-même en année et en mois, une référence : 32p. indiquant le numéro de la page.

1.3.5. Prétraitement

Une première difficulté est de pouvoir relier des données qui parfois sont hétérogènes. Des problèmes de format de données apparaissent et des conversions sont souvent nécessaires. Une deuxième difficulté est la récupération de valeurs manquantes et éliminer des données aberrantes et la phase de nettoyage est certainement de nouveau utile. Par exemple, dans USPTO, CA est lié à LosAngeles, mais correspond aussi au Canada. Donnée mal explicitée, mais l'automate ne le comprend pas et ne le distingue pas. Besoin d'analyser non automatiquement la ressource. La ressource joue un rôle fondamental.

L'objectif du nettoyage de données est d'isoler et structurer l'information. Pour cela, il faut extraire les parties utiles. Il faut ensuite définir et segmenter les éléments d'information, par exemple en segmentant les données en mots (ou mots composés). Il convient alors de traiter la ponctuation, les chiffres, les unités fréquentes, gérer la casse (tout en majuscule ou en minuscule).

Les règles d'extraction définies permettent d'isoler l'information à partir de documents, hétérogènes ou non, dont la localisation est possible par le biais de balises (Chrisment, 1997). Ainsi chaque champ est distinguable et surtout peut être facilement extrait. L'information explicite est directement lisible. Elle est issue de l'environnement d'un domaine et mémorisée dans les banques de données.

Chaque base ayant sa propre structure, il est important de s'adapter à chacune d'entre elles par dérivation du schéma global et par recours à des outils de description des formats : les *meta données* (Dousset, 2003).

✓ ***Metadonnées***

Le terme de métadonnées est utilisé pour définir l'ensemble des informations techniques et descriptives ajoutées aux documents pour mieux les qualifier. Pour que ces données soient utilisables par d'autres, elles doivent s'inscrire dans des modèles largement reconnus par les acteurs du Web.

En effet, chaque source a son format, qui lui-même a son descripteur spécifique (métadonnées de premier niveau). Il faut alors (Dousset, 2003) trouver une technique pour différencier les documents les uns des autres, déterminer les balises des champs sémantiques présents dans la base, leur donner un nom et un sigle standard, définir leur utilité et leur priorité et déterminer d'astucieuses techniques de découpage pour extraire chaque type d'information.

Une collection de corpus aux formats hétérogènes est gérée par un descripteur générique (métadonnées de second niveau). Les données brutes sont associées à des métadonnées qui récapitulent l'ensemble des informations potentiellement exploitables.

Cette base de métadonnées permet d'établir une fonction d'association de champs équivalents issus de bases différentes. L'exemple suivant montre la structure d'extraction relative à des notices issues de la base Pascal.

✓ ***Niveaux d'information***

Le champ analysé peut-être de type :

- acteur : auteurs, inventeurs, journaux, villes, pays, sources …
- sémantique : mots-clés, index, classifications, termes libres, multi-termes…
- temps : années, mois, jours, périodes, date de priorité, date de dépôt, date d'application...

Le champ est alors analysé par dénombrement de l'ensemble de ses items. Le nombre d'occurrences obtenue est absolue et correspond au nombre de fois où l'item apparaît dans le corpus analysé.

✓ ***Multi-termes***

L'extraction d'information permet de passer d'une représentation brute du texte (succession de mots) à une succession d'unités terminologiques. Les différents items sont alors identifiés (titre, mots du résumé) par association de chaque modalité et de sa fréquence dans le corpus. Il est important de souligner qu'une unité terminologique peut être sous la forme de radicaux auxquels il est possible de ramener certains uni-termes par des algorithmes de radicalisation (Lovins, 1968), (Porter, 1980), de (paice, 1996) et de Carry (Paternostre et al., 2002), c'est-à-dire par suppression de tous les affixes d'un mot. Par affixes on entend suffixe (cafe-**tière**), préfixe (**sur**-population) et infixe (rêv-**ass**-er). Il peut aussi être de la forme d'un mot simple ou uni-terme.

Enfin, il peut être un mot composé qu'on nommera alors « multi-terme » (Dousset et Kanoun, 1998). Si nous prenons l'exemple du multi-terme « data mining », la plupart des méthodes traditionnelles dissocient le mot « data » du mot « mining ». L'objectif de la méthode des multi-termes est d'associer «data » à « mining ». Les traitements statistiques (Diday, 2005), (Diday, 2008b) qui seront appliqués par la suite porteront alors sur « data mining» et non pas sur les deux termes distincts. L'utilisation de groupes de termes plutôt que des termes simples a fait l'objet de nombreuses études en recherche d'informations (Fagan, 1987), (Mitra, 1997), (Chevallet et Bruandet, 1997), (Kraaij et Pohlmann, 1998).

La détection des multi-termes s'effectue de la manière suivante. L'utilisateur choisit dans un premier temps les traitements de détection des multi termes qui lui semblent les plus adéquats. Pour chaque champ sémantique (mots-clés, titres, résumés, texte intégral…), un dictionnaire est crée, contenant toutes les valeurs du champ rencontrées. Puis, ces dictionnaires sont fusionnés et les doublons sont supprimés. Tous les mots composés, séparés par des tirets dans les titres, résumés, texte intégral, sont conservés sans leurs acronymes, générant alors un dictionnaire unique de la spécialité. L'utilisateur peut ajouter manuellement d'autres multi-termes à ce dictionnaire.

Un dictionnaire de synonymes est généré pour prendre en compte les acronymes, ainsi que les variations morphologiques (inversion, terminaisons, pluriels,…). Grâce à ce dernier, le système considère le multi-terme et sa variante comme une même et seule entité. Ici aussi, l'utilisateur peut compléter manuellement ce dictionnaire en rajoutant à la main d'autres synonymies de multi-termes.

La détection statistique se base sur la recherche de l'ensemble des multi-termes non détectés, à cause de l'absence de tiret entre les deux termes. Pour cela, nous recherchons de façon statistique quelles sont les expressions qui reviennent suffisamment souvent (au moins deux fois) et qui sont absentes du dictionnaire « conscient » précédent. Cette étape est très importante car elle permet de trouver des concepts clés du domaine nouveaux ou encore non officiellement reconnus.

Suite au choix d'un ou plusieurs traitements de détection de multi termes, un nouveau champ d'indexation est inséré dans chaque document et contient tous les multi-termes du document. Un des problèmes de l'indexation des multi termes est la taille très importante du dictionnaire généré, composé très souvent de plusieurs dizaines de milliers d'entrée. Afin de réduire cette quantité et ainsi de limiter le nombre de croisements sémantiques, les mots vides de la langue sont éliminés, ainsi que les termes qui ne sont présents qu'une fois dans le corpus, nommés

apax, ainsi que ceux qui sont distribués uniformément sur l'ensemble des documents, termes de l'équation de recherche, mots usuels ou trop généraux, …. Une fois cette première sélection effectuée, seuls les termes ayant une forte densité dans certains documents sont conservés.

Cette densité correspond au rapport entre la densité locale d'un terme dans chaque document et la densité globale du terme dans le corpus (Kanoun, 1998). Les termes ayant un rapport important dans au moins deux documents sont qualifiés et conservés, comme l'illustre la Figure 8.

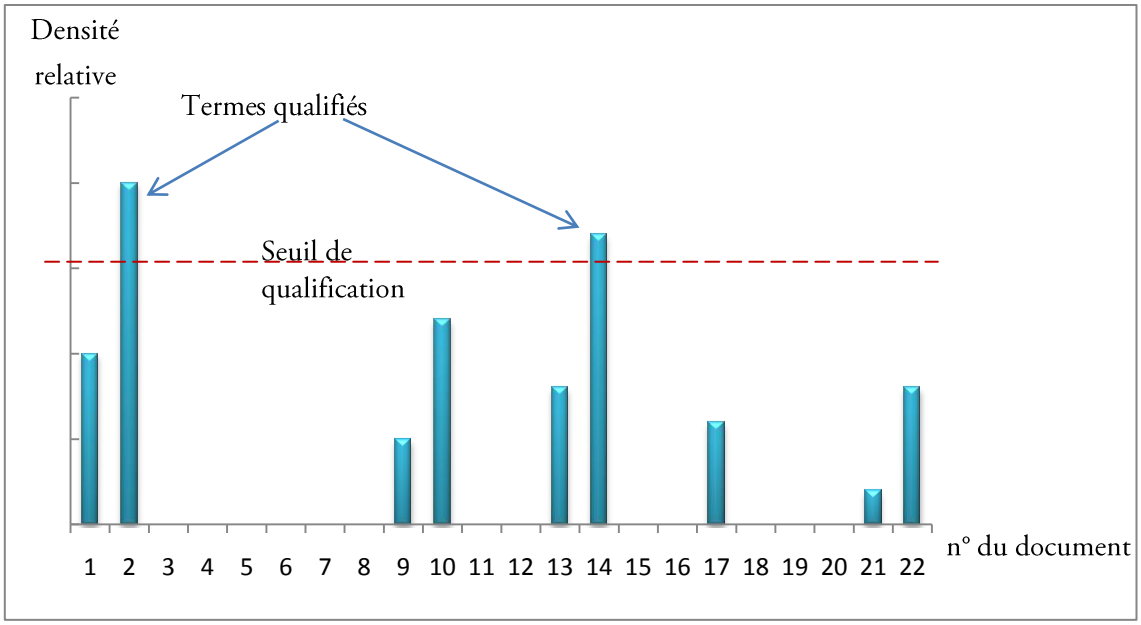


Figure 8. *Qualification des multi-termes à conserver dans le dictionnaire (Roux, 1998).*

Les multi-termes sont plus aptes à désigner des entités émergentes ou des concepts que les mots uniques et constituent alors une meilleure représentation du contenu sémantique utile des documents.

Dans la littérature, différentes approches pour la détection des multi-termes sont utilisées, comme l'utilisation des mesures de similarité, telles que l'information mutuelle ou le coefficient de Dice, ou bien la découverte de combinaisons de mots en fonction de leur régularité d'apparition (Church et Hanks, 1990) ou encore l'utilisation d'une mesure de proximité entre les contextes des termes (Besancon et al., 1999) prennent en considération la distance entre les mots ou l'ordre de leur apparition. D'autres travaux combinent une approche statistique avec une approche linguistique. Ces systèmes construisent des groupes de mots en sélectionnant les candidats à partir de schémas syntaxiques puis en les filtrant à l'aide de méthodes de la statistique (Simoni, 2000). (Grefenstette, 1992) combinent les cooccurrences statistiques des termes et la technique de "tête modifieur" pour extraire des combinaisons de mots relatives à un même contexte. (Hearst, 1992) utilise des patrons syntaxiques pour extraire des syntagmes reflétant une relation sémantique entre les termes alors que (Morin, 1999) se base sur les travaux de Hearst en intégrant des mesures statistiques pour le filtrage. Ce dernier met en évidence des schémas lexico-syntaxiques par l'intermédiaire de phrases qui utilisent des couples de termes liés par la même relation sémantique.

- ✓ ***Synonymie***

La notion de synonymie de termes traitée dans cette partie est relative à un ensemble de connaissances permettant de déterminer des classes d'équivalence 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 concernent la détection de synonymes, la normalisation de données (Jolibois et al., 2000) avec des approches plus ou moins automatisées. Cependant, les premiers résultats de l'étude de l'INIST (Inist, 2005) portant sur des outils d'analyse montre que de nombreux problèmes persistent.

Les techniques de prétraitement que nous proposons permettent 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 module est qu'il est paramétrable et efficace 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. Des conventions sont appliquées sur les termes afin que les données puissent être nettoyées et analysées.

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 variations morphosyntaxiques, …, et de proposer un représentant par classe d'équivalence (le plus fréquent ou généré suivant certains critères). La synonymie traitée met en avant la notion de proximité des termes (d'après une distance d'édition) et un regroupement en classes d'équivalences notamment pour mieux exploiter les champs en texte libre comme le titre ou le résumé d'un article. D'autres classes d'équivalences peuvent regrouper des termes synonymes sémantiquement, et c'est dans ce cas que l'introduction de connaissances extérieures a le plus d'importance (Loubier, 2007).

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 données comme la spécialisation/généralisation de termes géographiques et l'inclusion de dates dans des intervalles temporels : notions primordiales dans le cadre d'études évolutives.

L'objectif est d'automatiser au maximum les prétraitements textuels, tout en permettant à l'utilisateur d'introduire des connaissances spécifiques à son étude.

Ces connaissances sont des règles de radicalisations très simples telles que féminin ou encore pluriel.

Elles peuvent aussi être des règles de nettoyage des données se basant sur un fichier de rejet, à partir duquel les motifs précisés seront supprimés ou alors rendront synonyme à NULL tout terme le contenant. Par exemple, si le fichier de rejet contient le terme « COMMENT », alors toute occurrence à ce dernier sera supprimée ou synonyme à NULL selon le choix de l'utilisateur. Un autre nettoyage se base sur un fichier de transformation qui contient les descriptions de substitution à effectuer pour la recherche de synonymie. Par exemple, il est possible de transformer le caractère « ‘ » en espace « ».

Ces connaissances suivent aussi une convention en ce qui concerne les termes. L'objectif est d'homogénéiser des termes composés (comme par exemple des adresses). Pour cela, les fichiers permettant de décrire la liste des sous-termes valides ainsi que des règles de convention de terme sont conçus. La convention peut également être suivie d'une complétion des données si des règles de complétions sont présentes dans le fichier préalablement généré. Le premier type de fichier crée doit contenir des types et des listes de sous-termes valides regroupés par type. Les sous-termes peuvent être décrits par des chaînes de caractères ou des expressions régulières. Le second fichier sert à décrire l'ordre des types et éventuellement donne des règles de complétion.

Les sous-termes de chaque terme sont réordonnés en fonction de leur appartenance à un type. De plus, si des règles de complétion sont présentes, elles décrivent quels types permettent d'en déduire d'autres.

Des instances de ces règles sont recherchées dans les données et servent à compléter les termes incomplets. Par exemple, soit la règle de complétion suivante : codePostal → ville.

Une autre technique est la comparaison de termes. L'objectif de ce traitement est de déterminer des classes d'équivalences parmi les termes des données et de proposer un représentant pour chaque classe.

Trois algorithmes de comparaisons de sous-termes sont disponibles :

- *comparaison des préfixes* : deux sous-termes sont équivalents si l'un est préfixe de l'autre et que leurs longueurs sont proches (complexité en *O(n)*). Exemple : 'inst' et 'institut' sont équivalents.
- *distance de hamming* : les caractères de même rang sont comparés, et la distance est augmentée s'ils sont différents. (complexité en *O(n)*). Exemple : 'francois' et 'frqncois' sont équivalents.

- *distance de levenshtein* : le calcul de la distance entre deux sous-termes correspond au coût minimal de transformation d'un sous-terme en un autre en utilisant cinq opérations d'édition. Ces opérations sont : inversion, suppression, insertion, double et substitution de caractères. (complexité en *O(n^2)*).

Pour les deux distances, un seuil est utilisé : c'est un nombre réel dont la valeur par défaut est fixée par expérimentation à 0.21. Si la distance entre deux sous-termes est inférieure à ce seuil, les deux sous-termes sont considérés comme équivalents. Il est possible de donner une autre valeur à ce seuil compris entre 0 et 1. Plus ce seuil est faible, moins on autorise de variation entre deux sous-termes pour les déclarer équivalents.

Si le seuil est à 0 alors les deux termes sont équivalents si et seulement s'ils sont égaux, comme le montre le Tableau 1. Si le seuil vaut 1 alors tous les termes de longueur proche sont équivalents.

Une autre technique est la synonymie hiérarchique. L'intérêt et 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...

x		*y*
Toulouse	<	Haute-Garonne
Haute-Garonne	<	Midi-Pyrénées
Midi-Pyrénées	<	France
France	<	Europe
Montpellier	<	Hérault
Hérault	<	Languedoc-Roussillon
Languedoc-Roussillon	<	France
Ariège	<	Midi-Pyrénées

Tableau 1. *Relations d'ordre d'informations géographiques.*

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, comme illustré dans le Tableau 2:

Terme	Synonyme
Toulouse	Midi-Pyrénées
Haute-Garonne	Midi-Pyrénées
Ariège	Midi-Pyrénées
Montpellier	Languedoc-Roussillon
Hérault	Languedoc-Roussillon

Tableau 2. *Génération de synonymie.*

Ainsi les termes Toulouse, Haute-Garonne, Midi-Pyrénées et Ariège sont considérés comme équivalents 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.

- ✓ ***Filtrage***

Afin de faciliter l'analyse des données, il est important de réduire la liste des items d'un champ, en construisant des filtres à partir des instances identifiées. Plusieurs types de filtres peuvent être réalisés.

La troncature des dictionnaires permet de réaliser des filtres adaptés aux traitements. Dans ce cas, les termes les plus fréquents sont conservés dans un fichier texte nommé filtre. Ainsi, lors de l'étude de terme, si ce dernier appartient au fichier texte « filtre » alors il est conservé sinon il est ignoré.

L'analyse de la pertinence du contenu permet d'éliminer des termes présents qu'une fois (apax) ou qui sont équidistribués (trop généraux, trop fréquents, ou appartenant éventuellement à l'équation de recherche) qui ne sont pas des mots à distribution typée c'est à dire concentrés dans peu de documents, en se basant sur la loi de Zipf (Zipf, 1949). Cette dernière prévoit que dans un texte donné, la fréquence d'occurrence $f(n)$ d'un mot est liée à son rang n dans l'ordre des fréquences par une loi du genre $f(n) \times n = K$ où K est une constante.

L'extraction de sous-champs permet d'aboutir à la création de filtre. Les bases de données n'étant souvent que semi-structurées, certains champs peuvent contenir des éléments sémantiques recouvrant plusieurs concepts. Par exemple, le champ « adresse postale » peut contenir les sous-champs « nom de l'organisme », « type d'organisme », « ville », « pays », « code postal », « numéro dans la rue », « nom de la rue » ...

Une autre technique concerne le nettoyage grâce à des filtres négatifs, lorsqu'un champ contient des informations de faible intérêt, comme par exemple les éléments de l'adresse autres que le pays et la ville, traducteurs, éditeurs dans le champ auteur.

La sélection orientée par filtres positifs permet la réduction des items d'un champ, en utilisant des dictionnaires dont la thématique correspond à une partie sémantiquement cohérente du contenu du champ. Par exemple les auteurs dont le nombre de publications est supérieur à n, les dates ou périodes qui sont reconstituées par synonymie….

1.3.6. Transformation

Les données sont transformées sous la forme exigée par l'algorithme d'extraction. Les algorithmes qui permettent de calculer les motifs fermés ou les concepts présents dans les bases de données sont nombreux et anciens, ainsi que les études qui les comparent. La littérature concernant les algorithmes d'extraction de l'intégralité des motifs fermés est donc riche (Ganter, 1984, Guénoche, 1993, Godin et al., 1995, Fu et Mephu Nguifo, 2003). Les travaux de (Kuznetsov et Obiedkov, 2002) en dresse un état de l'art et détaille les composants des algorithmes historiques, selon qu'ils sont incrémentaux, trient les attributs, utilisent du hachage, du partitionnement, des structures d'arbres, etc.

Dans le domaine de l'extraction d'information, on utilise plus spécifiquement des algorithmes issus des méthodes de traitement automatique du langage naturel (TALN) qui extraient les concepts-clés des textes puis représentent graphiquement les interrelations entre ces concepts au sein de la collection de documents. Il en résulte une partition des textes ayant une similarité de contenu. Plusieurs techniques permettent le traitement du langage naturel. Des systèmes basés sur des approches traditionnelles l'analysent au niveau des phrases prises individuellement. L'objectif est alors d'en créer une représentation sémantique sous la forme de relations structurées entre les mots représentatifs de la phrase.

L'étape de transformation sert à définir des représentations et/ou des abstractions des données adaptées à la tâche d'extraction de connaissances.

1.3.7. Fouille de données

La fouille de données est le cœur du processus car elle permet d'extraire de l'information des données. Néanmoins, c'est souvent une étape difficile à mettre en œuvre, coûteuse et dont les résultats doivent être interprétés. Il faut aussi noter qu'en situation réelle, pour l'aide à la décision, une très grande majorité des résultats recherchés s'obtient uniquement par requêtes, par analyse multidimensionnelle (Agrawal, 1997) ou grâce aux outils de visualisation.

Une approche traditionnelle pour découvrir ou expliquer un phénomène est décrite par l'algorithme suivant.

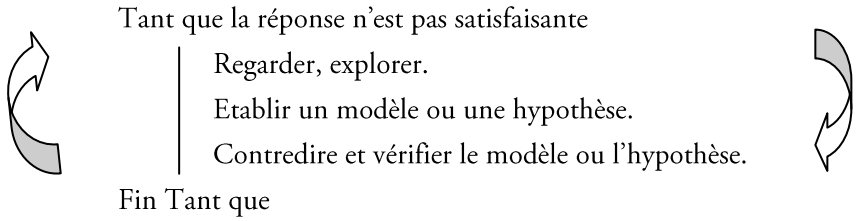


Figure 9. *Approche de la fouille de données.*

La partie d'exploration est traditionnellement réalisée avec des outils de reporting ou d'analyse multidimensionnelle. La réalisation du modèle se base sur une hypothèse ou des idées émises par l'utilisateur. On peut voir les outils de fouille de données comme des procédures qui permettent de faciliter ou encore d'automatiser ce processus.

- ✓ ***Entités textuelles***

Un texte peut être décomposé en plusieurs grains informationnels unitaires appelés entités textuelles.

La population Ω que nous considérons est composée de *n* individus ou objets, associés aux unités textuelles, telles que des champs en texte libre, des paragraphes ou des phrases dont la juxtaposition constitue un corpus.

Le plus grand grain informationnel considéré est le champ en texte libre électronique, numérisé ou non et caractérisé comme un objet porteur d'information ou encore comme « *toute base de connaissance fixée matériellement, susceptible d'être utilisée pour consultation, étude ou preuve* », selon l'Institut International de Coopération Intellectuelle (International Institute for Intellectual Cooperation). Il s'agit d'une agence de la ligue des Nations, en collaboration avec l'Union français des Organismes de Documentation.

En dessous, nous trouvons le paragraphe, qualifié de « segment de texte suivi, dit aussi texte linéaire, compris entre deux alinéas » selon Wikipédia.

Enfin, le plus petit grain informationnel considéré est la phrase, définie comme « *un ensemble autonome, réunissant des unités syntaxiques organisées selon différents réseaux de relations plus ou moins complexes (…). La phrase possède une unité sémantique (ou unité de communication), c'est-à-dire, un contenu transmis par le message (sens, signification...). Ce contenu se dégage du rapport établi entre les signes de la phrase, et dépend du contexte et de la situation du discours* », selon Wikipédia.

Une variable uni modale ne peut prendre qu'une valeur à la fois parmi plusieurs.

Le terme peut être soit un mot soit un groupe de mots employé pour représenter une notion (AFNOR 1987) ou un concept (Hudon, 1994). Cet élément lexical se situe à une granularité inférieure au multi-terme.

Elle peut être de différent type, comme le montre le Tableau 3 :

Type de variable	Exclusivité	Principe	Exemple
Quantitative		la valeur mesurée sur chaque individu représente une quantité. On peut alors calculer un total pour un ensemble d'individus.	Nombre de citations, nombre de lignes du document, nombre de termes du document, degré de pertinence,…
Qualitative		la valeur mesurée sur chaque individu (parfois qualifiée de catégorie ou de modalité) ne représente pas une quantité numérique.	Journal, titre, auteur…
Qualitative ordinale		la valeur mesurée sur chaque individu (parfois qualifiée de catégorie ou de modalité) est numérique.	Année, jour de la semaine, mois…
Qualitative hiérarchique		La valeur mesurée sur chaque individu est ordonnable selon une graduation.	Zones géographiques, inclusions sémantiques, arborescence de fichier,…
Qualitative nominale		La valeur mesurée sur chaque individu est non ordonnée.	Journaux, mots-clés…
Qualitative uni-modale		Une seule modalité de cette variable est requise pour chaque document. Il s'agit d'un champ mono-valué.	Année, revue, langue… Un article doit comporter une seule et unique année, langue et ne doit être publiée que dans une seule revue.
Qualitative multi-modales	exclusive	La valeur mesurée sur chaque individu peut prendre plusieurs modalités différentes qui peuvent apparaître une seule fois chacune dans le même document. Il s'agit d'un champ multi-valué.	Auteurs. Un article peut être signé par plusieurs auteurs, tous différents et n'apparaissant qu'une fois.
	Non-exclusive	Le champ peut comporter plusieurs fois la même modalité. Elle est redondante si la même modalité est répétée. Il s'agit d'un champ multi-valué	Le champ pays peut contenir plusieurs fois le même pays si les auteurs sont concitoyens. Le pays des différents auteurs peut être le même, répété plusieurs fois (la modalité est alors redondante).

Tableau 3. *Les différents types de variables.*

✓ ***Différentes mesures du relationnel***

Selon la nature de l'individu, de la variable étudiée, la mesure du relationnel diffère.

Cas de deux variables quantitatives

Soit X une variable quantitative considérée, supposée à n valeurs $\in R$ notées

$$x_1, \dots, x_i, \dots, x_n$$

Considérons maintenant une seconde variable quantitative Y à n valeurs $\in R$, notées

$$y_1, \dots, y_i, \dots, y_n$$

La mise en évidence de dépendance entre deux variables X et Y permet de réduire l'espace informationnel afin de mieux le maîtriser, par suppression des éléments indépendants. Ainsi, seules les relations les plus significatives stratégiquement sont conservées.

Cas de deux variables qualitatives

Maintenant, considérons deux variables qualitatives observées simultanément sur *n* individus. On suppose que la première, notée *X* possède *n* modalités. L'ensemble de ces valeurs appartiennent à $\{m_1, m_2, \ldots, m_i\}$ avec m_j une modalité considéré.

$$x_1, \ldots, x_i, \ldots, x_n$$

et que la seconde, notée *Y* possède n individus, *s* modalités notées. L'ensemble de ces valeurs appartiennent à $\{m_1, m_2, \ldots, m_s\}$ avec m_k une modalité considéré.

$$y_1, \ldots, y_i, \ldots, y_n$$

Plusieurs mesures de dépendance sont disponibles.

La contingence est issue du **croisement de deux variables uni-modales**. La somme des éléments de la matrice est égale au nombre de documents possédant simultanément les deux variables. Les croisements peuvent être de type « journaux × Années » ou encore « Journaux × Langue ».

La cooccurrence est la présence simultanée de deux unités linguistiques (deux mots par exemple ou deux codes grammaticaux) au sein d'un même contexte linguistique (le champ balisé, le champ textuel, le paragraphe ou la phrase). **Les cooccurrences résultent du croisement de deux variables qualitatives dont au moins l'une n'est pas uni-modale, à modalités multiples, exclusives ou non.**

Un certain nombre de modèles et de coefficients ont été à ce jour proposés : (Lafon, 1984), (Church et Hanks, 1990), (Dunning 1993), (Fung et McKeown, 1997), (Manning et Schütze, 1999), (Véronis, 2003), (Wu et Zhou, 2003), (Véronis, 2004), etc.

La proximité, qui étudie en termes de « distance » deux variables. Pour le texte libre, il est possible de ne prendre en compte que les coïncidences des modalités physiquement proches (à côté, dans la même phrase, à *n* mots de…).

La présence/ absence. Il existe au moins un document du corpus qui contient simultanément les deux modalités.

- ✓ ***Matrices et cubes***

Dans cette partie, nous présentons les matrices, croisant deux entités, et les cubes, prenant en compte trois dimensions, selon le type de données utilisées. Toutes les matrices et cubes proposées peuvent être utilisés en entrée dans nos travaux.

Dans le contexte de nos travaux, nous ciblons le croisement de variables qualitatives. On génère ainsi une matrice dont le nombre de lignes est égal au nombre de modalités de la première variable et le nombre de colonnes à celui de la seconde.

Il existe deux types de matrices :

Matrices symétriques

En algèbre linéaire une matrice symétrique *A* est une matrice carrée. Elle est égale à sa propre matrice transposée. Ainsi *A* est symétrique si : ${}^tA = A$. Intuitivement, les coefficients d'une matrice symétrique sont symétriques par rapport à la diagonale principale. L'ensemble (En théorie des ensembles, un ensemble, désigne intuitivement une collection d'objets, appelés aussi éléments...) des matrices symétriques à coefficients dans un anneau *K* est noté $S_n(K)$. Toute matrice diagonale (En algèbre linéaire, une matrice diagonale est une matrice carrée dont les coefficients en dehors de la diagonale...) est symétrique, puisque tous les coefficients en dehors de la diagonale principale sont nuls. Un théorème (Un théorème est une proposition qui peut être mathématiquement démontrée, c'est-à-dire une assertion qui peut être...) fondamental concernant de telles matrices est le théorème spectral en dimension (Dans le sens commun, la notion de dimension renvoie à la taille ; les dimensions d'une pièce sont sa longueur, sa...) finie, qui énonce que les matrices symétriques dont les coefficients sont des nombres réels sont diagonalisables à l'aide de matrices orthogonales.

Elles sont issues du croisement d'une variable non exclusive avec elle-même (auteurs, pays, villes, citations, brevets cités, mots-clés, multi-termes…). Les croisements effectués permettent de mettre en avant les associations entre les modalités d'une même variable. Ainsi la matrice symétrique, croisant des auteurs permet de révéler leur collaboration, leur stratégie et la formation de leurs équipes de recherches.

Matrices asymétriques

Les matrices asymétriques croisent deux variables différentes, où alors la même variable filtrée par deux jeux différents de modalités c'est-à-dire dont le nombre de modalités diffère. Leur analyse permet de mettre en avant les corrélations croisées entre leurs modalités respectives.
Ainsi le croisement d'une variable avec les documents est fortement utilisé en Recherche d'Information pour les calculs de pertinence, le filtrage de documents… Le croisement d'une variable avec le temps permet de détecter les tendances et les émergences. Les croisements entre des auteurs et des thématiques permettent de révéler les centres d'études les plus importants, les concurrences, les collaborations relatives à un sujet spécifique… Les croisements des variables peuvent s'effectuer de plusieurs façons :

Croisement des variables uni-modales : Matrice de contingence et Matrice de fréquence

A partir de deux variables qualitatives à modalités ayant qu'une valeur choisie parmi n modalités, nous définissons le tableau de contingence croisant les modalités de deux variables.

Ces données sont présentées dans un tableau à double entrée, appelé table de contingence, dans lequel on dispose les modalités de X en lignes et celles de Y en colonnes.

Ce tableau est donc de dimension $r \times s$ et a pour élément générique le nombre $n_{\ell h}$ d'observations conjointes des modalités x_ℓ de X et y_h de Y ; les quantités $n_{\ell h}$ sont appelées les *effectifs conjoints*.

La case à l'intersection de la ligne i et de la colonne j contient le nombre d'individus ayant la modalité i de la première variable et la modalité j de la seconde variable.

Si l'on divise chaque valeur de ce tableau par n, le cardinal de la population, on obtient le *tableau de fréquence*. La somme des éléments de la matrice est égale au nombre d'individus possédant simultanément les deux variables.

Les marginales des lignes et des colonnes représentent la population liée à chaque modalité. Dans le cas de nos travaux, les matrices de contingence peuvent être le résultat du croisement entre journaux et années, ou encore entre premiers auteurs et année. Dans l'exemple illustré dans le Tableau 4, on comptabilise chaque apparition des couples [Nom journal] × [valeur année]. La valeur 480 correspond au nombre total d'individus (ici des documents), par ajout des marginales lignes. Ce nombre correspond aussi au total des marginales colonnes. Elle nous permet d'obtenir l'effectif total représenté dans le tableau. La valeur 218 correspond au nombre d'individus contenant la modalité « Journal of Information Visualisation » ainsi que les modalités « 2007 ou 2008 ou 2009 », le nombre 172 est le nombre total de croisement « journaux » × 2007.

Années / Journaux	2007	2008	2009	Marginale Ligne
Journal of Information Visualisation,	75	87	56	218
Journal of Social Structure	65	28	31	124
Information Design Journal	32	77	29	138
Marginale Colonne	172	192	116	480

Tableau 4. *Matrice de contingence.*

Croisement des variables uni-modales : Tableau de Burt (MARC, 1991)

Ce tableau symétrique croise toutes les variables qualitatives entre elles. Il est composé de tableaux de contingence élémentaires entre toutes les variables prises 2 à 2.

Croisement des variables uni-modales ou à modalités multiples exclusives ou non : Matrice de présence – absence

Le croisement des variables révèle s'il existe au moins un document du corpus qui contient simultanément une modalité de la première variable et une de la seconde. Cette matrice n'est composée que des valeurs zéro et un. Dans le cas du croisement entre auteurs et journaux, nous obtenons directement le nombre de journaux dans lequel chaque auteur a publié, grâce aux marginales lignes, représentant le nombre de colonnes connectées. De même, pour chaque journal, nous pouvons clairement voir le nombre d'auteurs qui ont publié au moins un article dans ce dernier, grâce aux marginales colonnes, représentant le nombre de lignes connectées.

Dans l'exemple du Tableau 5, on récence les auteurs ayant publié dans les différents journaux. En marginale ligne, nous obtenons le nombre d'auteurs ayant publié au moins un article dans le journal correspondant. En marginal colonne, nous obtenons le nombre de publications de l'auteur correspondant.

Auteurs / Journaux	Martin	Dubois	Petit	Marginale Ligne
Journal of Information Visualisation,	1	0	0	1
Journal of Social Structure	1	1	0	2
Information Design Journal	0	0	1	1
Marginale Colonne	2	1	1	4

Tableau 5. *Matrice de présence/absence.*

Matrices de cooccurrences simples

Il existe plusieurs types de matrices de cooccurrences, indiquant la présence simultanée de deux modalités des deux variables qualitatives.

Le croisement des deux variables qualitatives multi modales révèle le nombre de documents dans lesquels on retrouve simultanément les deux modalités. Il est important de remarquer que la cooccurrence simple est identique à la contingence si les deux variables sont uni-modales. En effet la contingence est un cas particulier de la cooccurrence.

Dans l'exemple du Tableau 6, nous croisons à nouveau les journaux et les auteurs et nous obtenons pour chaque auteur le nombre de publications qu'il a signé pour le journal correspondant.

Auteurs / Journaux	Martin	Dubois	Petit	Marginale Ligne
Journal of Information Visualisation,	2	0	0	2
Journal of Social Structure	3	2	0	5
Information Design Journal	0	0	1	1
Marginale Colonne	5	2	1	8

Tableau 6. *Matrice de cooccurrence simple.*

La marginale ligne indique donc le nombre de signatures par journal. Son intérêt est limité puisque nous aurons autant de comptages que de signataires. Un seul article signé par n auteurs sera comptabilisé n fois.

Cooccurrence multiples

Nous nous trouvons dans le cas où au moins une des variables est à modalité multiple et non exclusive. Par exemple, les termes dans une phrase ou les pays dans un champ d'adresses multiples. Si, par exemple, la matrice de cooccurrences multiples croise des pays et des auteurs ; dans le cas de cooccurrences simples, le fait de trouver un auteur et un pays associés dans un même document reviendrait à rajouter +1 aux cooccurrences, que ce croisement apparaisse une ou plusieurs fois. Par contre dans le cas de cooccurrences multiples, on comptabilise ces croisements autant de fois qu'ils sont découverts dans chaque document et non pas une seule fois par document. Ainsi, l'échelle d'étude n'est pas la même. En effet, si nous étudions le terme t_1 dans l'ensemble des documents $D=\{d_1, d_2\}$. Admettons que d_1 soit un article contenant une seule fois simultanément les termes t_1 et t_2 et d_2 soit un livre de plus de 1000 pages, contenant 100 fois t_1 *et 200 fois* t_2. Notons $t_1 \times t_2$ la cooccurrence résultant du croisement entre le terme *t1* et le terme t_2.

Dans le cas de la cooccurrence simple,

$D1 : t_1 \times t2=1$

$D2 : t_1 \times t2= 1$

Par contre dans le cas de cooccurrences multiples :

$D1 : t_1 \times t2=1$

$D2 : t_1 \times t_2 = 100X200=20000$

Ainsi, la cooccurrence multiple est principalement utilisée en Recherche d'Information (Boughanem et Dousset, 2001) afin de faire ressortir les documents les plus pertinents, c'est à dire ceux dans lesquels le terme recherché apparait le plus souvent, de façon significative.

Elle est équivalente à la cooccurrence simple dans le cas de deux variables multimodales à modalités exclusives ou dans le cas du croisement d'une variable multimodale et une variable uni-modale. Elle est identique à la contingence, dans le cas de croisement de deux variables uni-modales.

Cooccurrence pondérée

Nous distinguons plusieurs types de pondération, selon le type de matrice utilisée.

Dans le cas des matrices symétriques de cooccurrence simple, par exemple une matrice « Auteurs × Auteurs », un article coécrit par deux auteurs va être comptabilisé quatre fois. Dans le croisement « Auteur1 × Auteur2 », dans le croisement « Auteur1 × Auteur1 », dans le croisement « Auteur2 × Auteur1 », dans le croisement « Auteur2 × Auteur2 ». Ils comptent donc pour ¼ chacun. Chaque marginale représente alors la part des publications de chaque auteur. Dans l'exemple suivant, les auteurs « Martin » et « Dubois» ont coécrit trois articles, qui seront pondéré à ¼. De même, les auteurs « Petit » et « Dubois » ont coécrit un article, pondéré aussi à ¼. Les valeurs en diagonale de cette matrice indiquent le nombre pondéré d'articles écrits par un auteur spécifique. Les marginales représentent la part des publications pouvant être attribuée individuellement à l'auteur spécifique.

Auteurs / Journaux	Martin	Dubois	Petit	Marginale Ligne
Martin	5+3/4	¾	0	6,5
Dubois	3/4	1	1/4	2
Petit	0	¼	3+1/4	3,5
Marginale Colonne	6,5	2	3,5	12

***Tableau** 7. Matrice symétrique de cooccurrence simple pondérée.*

Ainsi chaque cooccurrence pondérée dans le cas d'une matrice symétrique peut être formalisée de la façon suivante. Pour *m*, le nombre d'auteurs ayant coécrit un article et pour chaque document, si le couple *{Auteur1, Auteur2}* est rencontré, que ce soit une ou plusieurs fois, alors la cooccurrence suivante est ajoutée une seule fois.

$$Cooccurrence_{\{Auteur1X\ Auteu2r\}} = +\frac{1}{m^2}$$

La cooccurrence multiple pondérée, dans le cas d'une matrice symétrique, est formalisée de la manière suivante :

$$Cooccurrence_{\{Pays\times\ Pays\}} = \frac{nb\ occurrences\ \ multiples}{(\sum m)^2}$$

Il s'agit ainsi de prendre en compte le nombre de croisements multiples total du champ, comme vu précédemment et de le rapporter au nombre total de fois où ce champ a été observé.

Dans le cas des matrices asymétriques, la cooccurrence simple pondérée peut être formalisée de la façon suivante. Pour *m*, le nombre d'auteurs ayant coécrit l'article et *n* le nombre de multi-termes correspondant à cet article,

$$Cooccurrence_{\{AuteursX\ multi\text{-}termes\}} = +\frac{1}{m\times n}$$

Ainsi, si dans un article, écrit par 3 auteurs et dans lequel nous trouvons deux multi-termes, cet article va être comptabilisé 6 fois dans les croisements $m_1\times n_1$, $m_2\times n_1$, $m_3\times n_1$, $m_1\times n_2$, $m_2\times n_2$, $m_3\times n_2$.

Dans le cas de matrices asymétrique, la cooccurrence multiple pondérée est comptabilisée de la manière suivante :

$$Cooccurrence\ \{Pays\times\ multi\text{-}termes\} = \frac{n\times m}{\sum n\times\sum m}$$

Pour *m* le nombre de pays et *n* le nombre de multi-termes trouvés dans l'ensemble des documents, prenons l'exemple de deux pays « France » apparait 3 fois et « Italie », 2 fois. Maintenant, considérons deux multi-termes « *Data-mining* » apparait 4 fois et « *Knowledge-Management* », 2 fois. La cooccurrence pondérée apparaîtra de la manière suivante :

Multi-termes / Pays	Data-mining	Knowledge-management	Marginale Ligne
France	$\frac{4\times 3}{(3\times 4)+(2\times 4)+(2\times 3)+(2\times 2)}$	$\frac{6}{30}$	$\frac{18}{30}$
Italie	$\frac{8}{30}$	$\frac{4}{30}$	$\frac{12}{30}$
Marginale Colonne	$\frac{20}{30}$	$\frac{10}{30}$	1

Tableau 8. *Matrice asymétrique de cooccurrence multiple pondérée.*

De manière générale, les matrices de présence/absence, ainsi que de contingence sont fréquemment utilisées par les statisticiens. Les matrices de cooccurrences simples sont principalement utilisées en analyse de données. Enfin, les cooccurrences multiples sont utilisées dans le domaine de la Recherche d'Information.

Les cubes

Nous nommons « cube » les tableaux multiples 3D, issus du croisement de trois variables. Dans le cadre de nos travaux, la troisième variable ciblée correspond au temps, correctement discrétisée en périodes homogènes. Cependant, cette troisième variable peut ne pas être temporelle. Les volumes relationnels du temps doivent rester comparables.

La prise en compte de la dimension temporelle dans l'analyse de l'information permet d'effectuer des prédictions (Mendelzon et Vaisman, 2000), à travers l'étude de l'évolution du vocabulaire utilisé, de l'importance des mots-clés, des différentes associations observées… Le cube est alors matérialisé par des regroupements de documents homogènes en périodes. Le cube peut être sous deux formes distinctes :

une matrice symétrique, dans le cas où il s'agit d'un croisement d'une variable avec elle-même, ainsi qu'avec la variable temps ;

une matrice asymétrique, si elle résulte du croisement entre trois variables différentes, dont la variable temps.

Par ce croisement, il est possible d'analyser les différentes matrices et de comparer de nouvelles sources à une déjà connue et analysée et de situer l'apport d'informations récentes dans une analyse déjà finalisée.

Dans le contexte de nos travaux, nous ciblons plus particulièrement les cubes sur des croisements de type cooccurrences de variables qualitatives. Cependant, cette décomposition temporelle est applicable sur les autres matrices vues dans la section précédente. Dans l'exemple suivant, nous décomposons une matrice symétrique en trois périodes homogènes.

Cette décomposition permet d'observer l'augmentation du nombre de publications au cours du temps de l'auteur « Petit », mais aussi la diminution d'activité de l'auteur « Martin » et la constance de l'auteur « Dubois ». Cette décomposition temporelle permet ainsi d'observer l'évolution des variables mais aussi d'étudier plus minutieusement les collaborations.

1999-2002

Auteurs / Auteurs	Martin	Dubois	Petit	Marginales lignes
Martin	15	2	1	18
Dubois	2	8	4	14
Petit	1	4	7	12
Marginales colonnes	18	14	12	44

2002-2005

Auteurs / Auteurs	Martin	Dubois	Petit	Marginales lignes
Martin	7	0	0	7
Dubois	0	9	6	15
Petit	0	6	13	19
Marginales colonnes	7	15	19	41

2005-2008

Auteurs / Auteurs	Martin	Dubois	Petit	Marginales lignes
Martin	3	0	1	4
Dubois	0	8	5	13
Petit	1	5	14	20
Marginales colonnes	4	13	20	37

Tableau 9. *Décomposition d'une matrice Auteurs* X *Auteurs en trois périodes homogènes.*

Méthodes d'analyse et de visualisation de structure

Dans cette partie, les différentes méthodes d'analyse de base sont présentées pour étudier les structures matricielles.

Les méthodes suivantes permettent d'analyser les entités (Chrisment, 1997), (Mothe, 1998), (Hubert et al., 2001), (Dousset, 2003), (Mothe, 2006).

L'Analyse en Composantes Principales (A.C.P.) s'applique aux données quantitatives et éventuellement aux matrices issues du qualitatif comme celles de contingence et de cooccurrence et notamment dans le cas de l'analyse relationnelle entre acteurs. L'objectif de l'Analyse en Composantes Principales (ACP) est de revenir à un espace de dimension réduite en déformant le moins possible la réalité. Il s'agit donc d'obtenir le résumé le plus pertinent possible de l'ensemble des informations de la matrice. Le nuage des individus (lignes) est représenté dans l'espace des variables (colonnes).

L'ACP engendre la réduction du nombre de caractères permettant des représentations géométriques des individus et des caractères, c'est-à-dire de visualiser les données à n dimensions ($n>3$) dans un espace à p dimensions ($p<n$) à l'aide d'une projection de ces données sur les plans définis par les p dimensions. C'est la matrice des variances-covariances (ou celle des corrélations) qui permet de réaliser ce résumé pertinent, parce qu'on analyse essentiellement la dispersion des données considérées. De cette matrice, on va extraire les vecteurs propres associés aux valeurs propres de plus forts modules de cette matrice pour déterminer les composantes principales de ce modèle optimal.

Comme la nature et la dispersion des variables sont parfois très hétérogènes, une normalisation de celles-ci est alors nécessaire pour obtenir des cartes lisibles et sur lesquelles les variables ont toutes des rôles similaires. C'est le principe de l'Analyse en Composantes Principales Réduite A.C.P.R. Les variables sont alors réduites par normalisation (division par la norme de chaque vecteur colonne), ce qui a tendance à arrondir le nuage et donc à générer des valeurs propres de plus faibles modules. La matrice à diagonaliser est alors celle des corrélations (diagonale unitaire) et non plus la matrice de variance-covariance.

L'Analyse Factorielle des Correspondances (A.F.C.) s'appuie sur la même logique que l'ACP à l'exception qu'elle s'applique à des données qualitatives. La technique de l'AFC est essentiellement utilisée pour de grands tableaux de données toutes comparables entre elles, si possible exprimées toutes dans la même unité. Elle sert à déterminer et à hiérarchiser toutes les dépendances entre les lignes et les colonnes du tableau. Chaque ligne correspond à un profil unitaire, il suffit donc de faire une analyse en composantes principales de ces profils.

Comme les variables sont ici représentées sous forme d'individus complémentaires, une seule carte factorielle suffit pour rendre compte de façon globale des correspondances :

- Entre les individus,
- Entre les variables,
- Entre les individus et les variables.

Dans le cas des ACP, un observatoire graphique des caractéristiques de l'analyse nous permet de connaître la profondeur de l'exploration à entreprendre pour atteindre 80 à 90% de l'information.

L'Analyse Factorielle des Correspondances Multiple (A.F.C.M.) est une autre technique, disponible via la plateforme, qui permet seulement d'étudier des variables multimodales qualitatives. On appelle AFCM des variables ($X_1,\ldots,X_p$) relativement à l'échantillon considéré, l'AFC réalisée soit sur la matrice X soit sur la matrice.

Une autre technique d'analyse disponible est la *classification ascendante hiérarchique* (C.A.H.).

Elle considère initialement toutes les observations comme étant des clusters ne contenant qu'une seule observation (singleton), et leur distance est alors le plus souvent définie comme étant leur distance euclidienne. La première étape consiste donc à réunir dans un cluster deux observations les plus proches. Puis le principe de CAH continue, fusionnant à chaque étape les deux clusters les plus proches au sens de la distance choisie.

Le processus s'arrête quand les deux clusters restant fusionnent dans l'unique cluster contenant toutes les observations. Les méthodes de réalisation de ces classifications sont relatées dans (Dobrowolski, 1964),

(Bouroche, 1989) (Bellot, 2004) entre autres. Cette analyse classique basée sur une matrice de distances est entièrement interactive, permettant, entre autres, le choix du niveau de coupure, l'obtention du détail d'une classe, l'exportation de la classification vers le tableur, les cartes factorielles, ou encore les cartes géostratégiques.

La classification par partitions (C.P.P), autre technique disponible, ne propose pas de hiérarchie de classes imbriquées autorisant parfois plusieurs niveaux de coupure cohérents, mais définit simplement une partition composée d'un nombre maximum de classes défini à l'avance. Pour cela, nous devons préalablement choisir un ou plusieurs initiateurs de classes (éléments représentatifs de chaque classe).

Plusieurs représentations de cartes factorielles sont possibles pour étudier les données (Chrisment et al., 2004).

- La 2D est une représentation traditionnelle, elle s'effectue par sélection des axes à représenter.
- La 3D permet la visualisation simultanée de trois axes d'analyse, avec la possibilité d'effectuer une rotation d'axes et donc d'obtenir une information mieux représentée et plus précise.
- La 4D, pour laquelle la rotation ou encore le changement d'axes à visualiser est possible (Dkaki et al., 1991), (BenAmmar et Dousset, 1999).

Le module de réalisation de cartes géostratégiques permet la génération de visualisations géographiques à partir de matrices (Dousset et Karouach, 2002). Afin de relativiser les résultats, une pondération par le Produit national Brut (PNB), par la population, par la surface ou encore d'autres données externes est possible. Ce système offre à l'utilisateur la possibilité de définir des régions spécifiques d'analyse, sous une autre échelle. Ainsi, il est possible de visualiser les cinq continents ou encore les Amériques, … Enfin, le codage de l'information se traduit par le choix de couleur spécifiques.

Enfin, *l'Analyse Procustéenne* (AP) est une méthode qui permet d'ajuster par rotation, translation et homothétie, un nuage de points sur une configuration cible de points. La plupart du temps, lorsque les données sont contenues dans un plan (deux variables), les résultats d'une telle analyse se résument, à une mesure numérique de l'adéquation entre les deux nuages et à une représentation graphique des deux nuages de points. Il est possible de déduire la trajectoire relative suivie par les différents points. Pour cela, les différents tableaux sont centrés de sorte à faire coïncider leurs centres de gravité et ils sont modifiés par rotation pour minimiser la distance entre les tableaux pris deux à deux. La distance entre les tableaux est définie pour une série de tableaux centrés K^h, $h \in [1 \ldots m]$.

1.3.8. Evaluation du modèle

Il s'agit d'évaluer et valider les connaissances extraites afin de les déployer en vue d'une utilisation définitive. La qualité du modèle obtenu se mesure selon la rapidité de la création, de l'utilisation, de la facilité de compréhension par l'utilisateur, de la bonne qualité des performances, de la fiabilité du modèle. Les performances ne doivent pas se dégrader dans le temps et doit pouvoir évoluer.

Il va de soit qu'aucun modèle n'aura toutes ces qualités. Il n'existe pas une meilleure méthode de fouille. Il faudra faire des compromis selon les besoins dégagés et les caractéristiques connues des outils.

Pour une utilisation optimale, une combinaison de méthodes est recommandée, telles que (Agrawal, 1997), (Chrisment, 1997) :

- la classification, la régression, la prédiction

Il s'agit de trouver une classe ou une valeur selon un ensemble de descriptions. Ce sont des outils très utilisés. Les algorithmes reposent sur des arbres de décision, des réseaux de neurones, la règle de Bayes, les k plus proches voisins.

Il s'agit de rechercher un ensemble de prédicats caractérisant une classe d'objet et qui peut être appliqué à des objets inconnus pour prévoir leur classe d'appartenance (Bahsoun et BeyLagoun, 2006), (Huyn et al., 2007). Par exemple, une banque peut vouloir classer ses clients pour savoir si elle accorde un crédit ou non.

- L'association, sequencing

Il s'agit de trouver des similarités ou des associations. Le sequencing est le terme anglais utilisé pour préciser que l'association se fera dans le temps. Par exemple, *si un étudiant effectue un master 2 Recherche, il y a de forte chances pour qu'il devienne thésard l'année suivante* (sequencing) ou encore *Si le doctorant souhaite effectuer une carrière d'enseignant chercheur et si le poste correspondant à son profil est disponible alors l'individu va certainement postuler* (association).

1.3.9. Validation

Les méthodes de validation dépendent de la nature de la tâche et du problème considéré. Nous distinguons deux modes de validation : par un expert du domaine puis par méthodes statistiques. Il est essentiel que le modèle produit soit compréhensible.

Pour les problèmes d'apprentissage non supervisé (segmentation, association), la validation est du ressort de l'expert. Pour la segmentation, le programme construit des groupes homogènes, un expert peut juger de la pertinence des groupes constitués. La encore, on peut combiner avec une validation statistique sur un problème précis utilisant cette segmentation. Pour la recherche des règles d'association, c'est l'expert du domaine qui jugera de la pertinence des règles. En effet, s'il fournit des règles porteuses d'information, l'algorithme produit également des règles triviales et sans intérêt.

1.3.10. Discussion

Les techniques d'analyse, précédemment présentées, permettent d'obtenir des renseignements très importants sur nos données. Les méthodes d'analyses se basent sur la notion de distance et non pas sur l'aspect relationnel des données. Les résultats peuvent alors être incomplets. Pour illustrer ce problème, nous ciblons le cas de l'AFC qui permet de visualiser et d'analyser les écarts des distributions conditionnelles à la distribution marginale, pour les lignes comme pour les colonnes. Deux points-profils d'une même variable suffisamment proches représenteront deux modalités ayant des distributions similaires suivant les modalités de l'autre variable. Cependant, ce type de visualisation ne permet pas d'étudier l'aspect relationnel des données.

En effet, si nous prenons l'exemple, issu du croisement entre entreprises, nous obtenons l'AFC de la Figure 10.

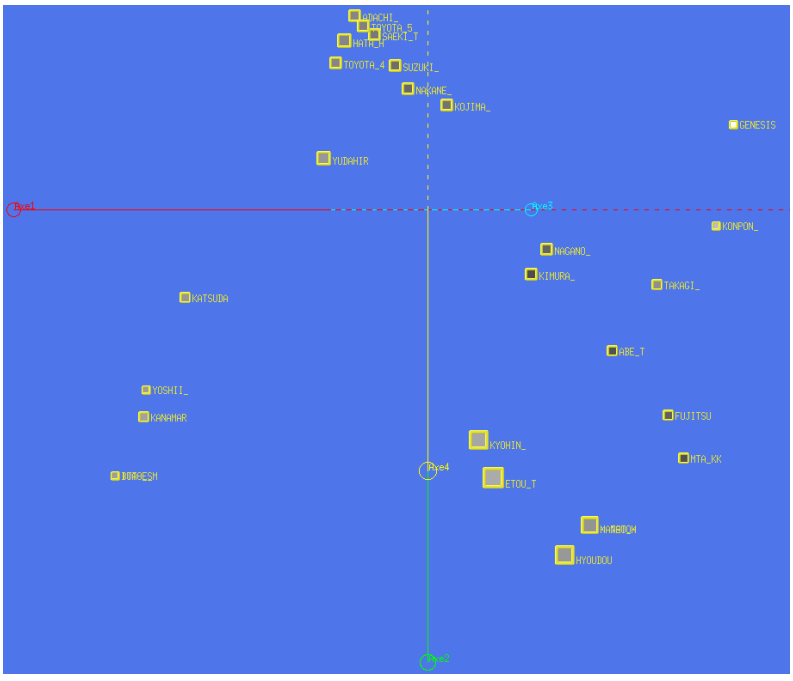


Figure 10. *AFC basée sur des croisements d'entreprises.*

En visualisant cette dernière, l'interprétation graphique des données nous mène à distinguer plusieurs regroupements d'entreprises.

Si nous nous intéressons particulièrement aux entreprises « Nagano » et « Kimura », zoomés dans la Figure 10, la proximité des données pourrait s'expliquer selon deux hypothèses :

- Ces données peuvent avoir des distributions similaires et être liées.
- Ces données se situent à cet emplacement indépendamment l'une de l'autre, suite à leur liaison avec les autres données telles que « *Tagaki* », « *Kojima* », « *Abe_t* », …

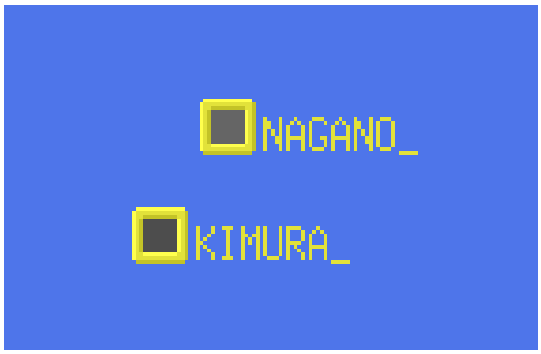


Figure 11. *Zoom de deux données de l'AFC sur les croisements d'entreprise.*

En observant la matrice de croisements à l'origine de ce graphe, nous constatons qu'en effet, ces données ne sont pas liées et que leur proximité est du à leur liaisons respectives avec les autres données. La Figure 12 est un extrait de la matrice de croisement entre les différentes entreprises.

	ABE_T	ADACHI_	KATSUDA	KIMURA_	KOJIMA_	KONPON_	KYOHIN_	MATSUOK	MTA_KK	NAGANO_
abe_t	1			1						
adachi_		1								
bito_s										
etou_t							1	1		
fujitsu						3				
genesis						1				
hata_h		5			1					
hyoudou							1	1		
iwase_m										
kanai_h							1	1		
kanamar										
katsuda			1							
kimura_	1			1	8					
kojima_				8	1					
konpon_						1				
kyohin_							1	1		
matsuok								1		
mta_kk									1	
nagano_										1
nakane_		3			5					1
saeki_t		2			1					
suzuki_		1			5					
takagi_	1			4		1		1	1	1
toyota_4		4	2		1					
toyota_5		1			1					
yoshii_			4							
yudahir							1			

Figure 12. *Extrait de la matrice de cooccurrences entre des entreprises.*

Nous pouvons observer que le croisement entre « Nagano » et « Kimura » est nul, reflétant l'absence de liaison entre ces deux modalités. La proximité révélée par l'AFC aurait donc pu nous mener à une erreur d'interprétation. Si nous prenons un autre exemple, basé sur « *Nakane* » *et* « *Kojima* », nous voyons que ces données sont très proches sur l'AFC. Si nous consultons la matrice de croisement, ces données sont effectivement

croisées et leur liaison est importante. Les outils proposés jusque là se basent aussi sur une notion de distance entre les données non jointes et non pas en terme de liaison. Il est alors difficile d'étudier les données relationnelles.

Pour remédier à cela, une solution est de proposer un module complémentaire de visualisation, permettant de révéler les différentes liaisons, afin de cibler l'étude vers une analyse relationnelle. Un module de visualisation de graphe répond à cette attente (Karouach, 2003), (Gay et Dousset, 2006); les nœuds représentent les données et les arcs sont assimilés aux différentes liaisons. Il est alors possible d'étudier les différentes collaborations, alliances mais aussi la formation de groupes d'acteurs du domaine analysé. Notre contribution se base sur l'extension de ce module graphique et plus particulièrement sur la prise en compte de la dimension temporelle dans ce type de représentation. Nous développons ce procédé dans les chapitres 3 et 4.

1.4. Synthèse

La maîtrise de l'information, d'un point de vue qualité et manipulation, permet de renforcer la décision et de susciter l'action. Le travail du veilleur se situe au niveau de la recherche d'information, du traitement des données mais aussi de l'analyse. Pour ce faire, le Data-Mining rassemble les techniques et les outils informatiques permettant d'acquérir des connaissances à partir de grands corpus textuels pour en extraire l'information utile.

Ainsi les différentes étapes du processus d'extraction et d'analyse de la connaissance se basent sur

- La collecte de l'information ;
- Le pré-traitement de l'information : l'information structurée ou semi-structurée est décrite afin d'en extraire une représentation adaptée à son analyse ; les données sont croisées sous forme de matrices spécifiques au type de variables étudiées. Le tableau suivant synthétise les différents cas.

Type de variable	Uni-modale	Multi-modale à modalités exclusives	Multi-modales à modalités non exlusives
Uni-modale	- Contingence - Présence/ absence *Statistique*	- Cooccurrences simples - Présence/ absence	- Cooccurrences simples - Cooccurrences multiples - Présence/ absence
Multi-modales à modalités exclusives	- Cooccurrences simples - Présence/ absence *Data-Mining*	- Cooccurrences simples - Présence/ absence	- Cooccurrences simples - Cooccurrences multiples - Présence/ absence
Multi-modales à modalités non exclusives	- Cooccurrences simples - Cooccurrences multiples - Présence/ absence	- Cooccurrences simples - Cooccurrences multiples - Présence/ absence	- Cooccurrences simples - Cooccurrences multiples - Présence/ absence - Pondération ***Recherche d'Information***

Tableau 10. *Type de matrice(s) possible(s) et domaine d'application (en italique) selon le type des variables croisées.*

- L'exploration de l'information : de multiples techniques sont basées en particulier sur des méthodes d'analyse de données, ils peuvent coopérer pour aboutir aux résultats finaux ou au contraire être utilisés en parallèle pour conforter les résultats obtenus.
- La visualisation : l'information cachée découverte dans les masses d'informations (corrélations, clusters,…) est visualisée sous forme graphique. Ces outils sont puissants pour analyser des données en terme de distance mais l'analyse relationnelle n'est pas privilégiée. En effet, l'interprétation des résultats passe souvent par une analyse de proximité entre les données représentées.

La veille s'effectue extrayant les informations enfouies dans une masse de données complexes. Cette complexité doit être gérée en décomposant l'espace à étudier, et en l'organisant par niveaux de détails dans un espace fragmenté. Il est indispensable de pouvoir naviguer dans cet espace et d'obtenir des informations détaillées sur les

données analysées. L'étude statique de l'information ne suffisant plus, la dimension temporelle doit impérativement être prise en compte pour suivre l'évolution de toute brique d'information.

La typologie de l'information se définit ainsi de la manière suivante, comme illustré dans la Figure 13:

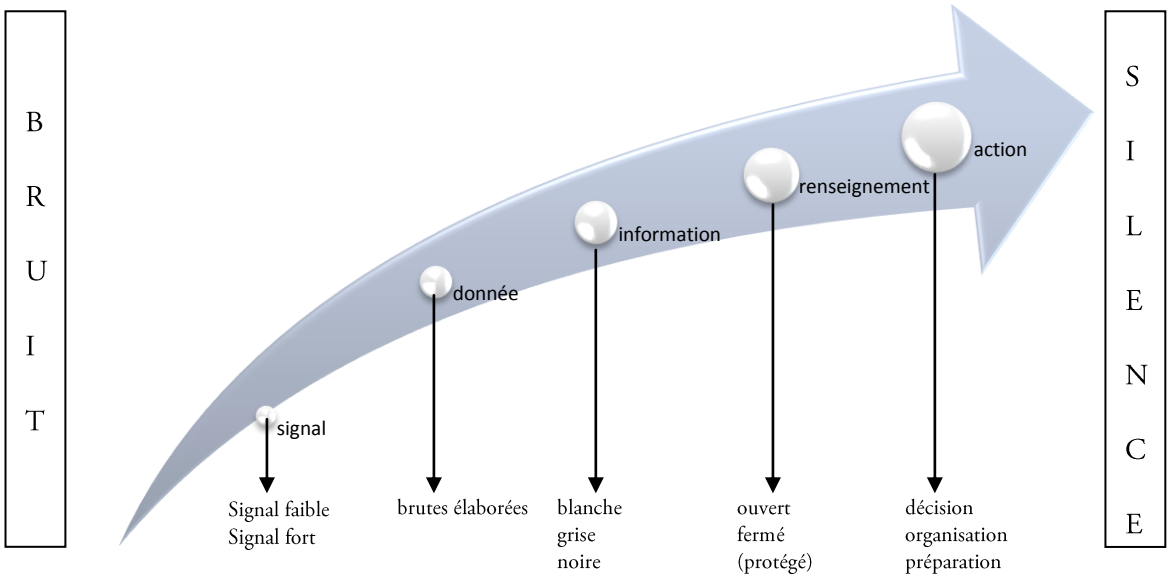


Figure 13. *Typologie de l'information*

Dans le cadre de nos travaux, nous proposons de compléter ces techniques d'analyse de données, par extension d'un module complémentaire de visualisation de données relationnelles par le biais de graphes, que nous présenterons dans les chapitres 3 et 4. Les liens entre les données sont alors visibles et l'étude des relations (telles que les collaborations, les alliances, les formations d'équipes de recherche…) peut alors compléter l'analyse élaborée par les techniques présentées dans ce chapitre. L'ajout de ce module de graphes a pour objectif d'offrir des techniques d'analyse de données relationnelles s'appuyant sur des interfaces de visualisation et de manipulation interactive, présentées dans le chapitre 2.

Chapitre 2.
Visualisation de données : vers le temporel

> *« Si familier, si intime, si essentiel que soit le temps à notre existence, il s'échappe comme du sable dès que nous essayons de le saisir par la pensée. Le passé n'est plus, l'avenir n'est pas encore, et le présent n'est qu'un passage fugitif entre ces deux néants. Le temps n'est pas « ce dans quoi » apparaissent les choses, mais plutôt la manière d'apparaître de toute réalité. (...). Le temps est la présence fuyante de l'âme aux choses. »* (Farago, 2004)

2.1. Introduction ____ 50
2.1.1. Spécifications ____ 50
2.1.2. Présentation du domaine ____ 50
2.1.3. Enjeux de la visualisation de données ____ 51
2.1.4. Définitions ____ 51
2.2. Domaines d'application ____ 52
2.2.1. Réseaux sociaux ____ 52
✓ Réseaux d'acteurs ... 53
✓ Réseaux de citation ... 53
✓ Réseaux de co-auteurs ... 53
✓ Réseaux de contacts ... 53
✓ Réseaux d'appels téléphoniques ... 53
2.2.2. Réseaux sémantiques ____ 53
✓ Graphes conceptuels ... 53
✓ Ontologies ... 54
2.2.3. Autres réseaux ____ 54
✓ Réseaux de neurones ... 54
✓ Réseaux trophiques ... 54
✓ Réseaux d'interactions entre protéines ... 54
✓ Réseaux de distribution électrique ... 55
✓ Réseaux informatiques ... 55
✓ Réseaux de transports ... 55
2.3. Principes de la visualisation de graphes ____ 55
2.3.1. Complexité de la visualisation ____ 55
2.3.2. Sémiologie graphique ____ 56
2.3.3. Problèmes d'ergonomie ____ 57
2.3.4. Préconisations sur la planarité ____ 58
2.3.5. Préconisations sur les structures de visualisation ____ 60
2.4. Prise en compte de la dimension temporelle ____ 61
2.4.1. Données temporelles ____ 61
2.4.2. Espace temps ____ 62
✓ Visualisation linéaire ... 63
✓ Forme non uniforme du temps ... 70
✓ Temps cyclique ... 73
✓ Discussion ... 77
2.4.3. Représentation de l'espace temps et des données temporelles ____ 85
2.4.4. Détection des tendances émergentes ____ 91
2.5. Conclusion ____ 91

2.1. Introduction

2.1.1. Spécifications

Dans ce chapitre, nous allons aborder la visualisation de données, à savoir sa définition et ses enjeux. Nous présentons le domaine d'application, section 2.2, puis nous approfondissons en proposant un état de l'art sur les différents types d'outils de visualisation existant, section 2.3 et en particulier sur ceux permettant la représentation de données évolutives, section 2.4, ce qui est un élément essentiel de notre contribution.

La représentation graphique permet de fournir au lecteur un maximum de renseignements synthétiques, qui ne sont que très rarement explicités dans les données brutes. Ce type de visualisation est un excellent vecteur d'analyse des données complexes, (Tufte, 1983,1990, 1997).

Par ailleurs, le but de cette représentation est d'exploiter les caractéristiques du système visuel humain pour faciliter la manipulation et l'interprétation de données informatiques variées. Les travaux de (Tufte, 1983) et (Bertin, 1977) ont montré comment exploiter, de façon intuitive ou ad hoc, ces caractéristiques de perception globale. Les travaux en perception visuelle ont montré que l'être humain a une perception d'abord globale d'une scène, avant de porter son attention aux détails (Myers, 2000).

La visualisation d'information cherche à exploiter ces mêmes caractéristiques de façon plus systématique en projetant les données dans un espace de représentation via un algorithme spécifique de conception de graphe.

Par définition, la visualisation de données est caractérisée comme «l'utilisation informatisée de représentations visuelles interactives de données abstraites, de manière à amplifier la cognition. La visualisation de l'information est un champ de recherche en informatique qui se consacre à la création d'interfaces visuelles riches aidant l'utilisateur d'un système à comprendre et à naviguer au travers d'espaces informationnels complexes » (Polanco, 2002), (Fekete et Lecolinet, 2006).

Par exemple, on peut se poser la question : Existe-t-il des regroupements caractéristiques dans ce réseau ? La visualisation graphique peut nous donner une vue sur l'organisation des données ou en faire apparaître les propriétés structurelles pour définir si un élément spécifique est important dans le réseau. Ces tâches d'investigation seraient très difficiles, voire impossibles, en basant seulement l'analyse sur le texte brut, en particulier pour les grands volumes de données.

2.1.2. Présentation du domaine

Le phénomène de la visualisation de l'information est une réaction récente due à la quantité de données à laquelle nous avons accès et qui progresse à un rythme qui ne cesse d'accélérer. La visualisation de l'information (ou l'infovis) concerne la création d'outils qui exploitent le système visuel humain pour aider les utilisateurs à explorer ou à expliquer ces données.

De nombreux travaux ont tenté d'approcher certaines formalisations des représentations disponibles pour restituer les processus spatio-temporels (Langran 1993), (Gayte et al., 1997). L'analyse de l'évolution d'informations relationnelles est principalement basée sur la visualisation de graphes dynamiques. De nombreux chercheurs ont développé des systèmes de visualisation de réseaux, (Di Battista et al., 1999), en prenant en compte une cartographie des connectivités liées à Internet, les réseaux d'appels téléphoniques, les réseaux de citation ainsi que la visualisation progressive des domaines évolutifs de connaissances.

Depuis une quinzaine d'années, sous l'impulsion de chercheurs comme Ben Shneiderman (Card et al., 1999), Robert Spence (Spence, 2000), Colin Ware (Ware, 2000) ou (Frank et al., 2001), la visualisation d'information est devenue un axe de recherche à part entière. Géographes et cartographes se sont également intéressés, plus récemment, à ces questions (Josselin et Fabrikant., 2003), (Paque, 2007).

De nombreuses techniques de visualisation de données temporelles ont été proposées, à ce jour, dans diverses applications, tels que l'étude de données personnelles, telles que celles contenues dans un dossier médical (Plaisant et al., 1996), de données hydrométriques (Kramer et Jozsa, 1998), de données géographiques (MacEachren et al., 1998), de données cliniques (Shahar et Cheng, 1999).

De même, elles ont été mise en place à différentes fins tels que la recherche des tendances significatives (Havre et al., 2000), l'exploration des traces de programmes (Renieris et Reiss, 1999, la représentation des abstractions temporelles (Shahar et cheng, 2000) ou la visualisation des règles d'associations temporelles (Rainsford et Roddick, 2000).

« L'interaction avec une représentation visuelle de données soigneusement conçue peut nous aider à former des modèles mentaux qui nous permettent d'exécuter des tâches spécifiques plus efficacement » (Munzer, 2003).

Pour nous, qui nous intéressons à l'évolution des thématiques de recherche, la visualisation de l'information, selon (Kapusova, 2004) et (Fekete et Lecolinet, 2005), combine des aspects de la visualisation scientifique, des Interfaces Homme-Machine, de la fouille de données, de l'imagerie et des graphiques (Fong et al., 2006), (Chittaro, 2006), (Compieta et al., 2007).

2.1.3. Enjeux de la visualisation de données

De nombreuses techniques de visualisation de l'information ont été proposées ces quinze dernières années comme en témoigne l'ouvrage récapitulatif de (Card *et al.*, 1999), proposant d'identifier 6 principales causes d'amplification de la cognition par la visualisation :

- la réduction des ressources cognitives mobilisées par l'utilisateur pour traiter et analyser les informations, l'interaction élevée avec l'utilisateur, perception menée en parallèle, facilité d'accès à une grande quantité d'information… ;
- la simplification de la recherche d'information, en visualisant beaucoup de données dans un petit espace et en regroupant des données, par exemple, par critère ;
- l'augmentation des possibilités de détection de structures, relations entre données, regroupements significatifs, positions stratégiques, centralité, …;
- l'inférence à l'aide de la perception visuelle. Certains problèmes paraissent plus simple ou même évidents à l'aide d'une représentation visuelle, …;
- la surveillance des événements, changements de structures, apparition ou mouvement dans les motifs, regroupements, …;
- la manipulation des données par la navigation interactive.

Ainsi, la visualisation doit permettre à l'utilisateur final de faire des découvertes, proposer des explications ou prendre des décisions (Poulet et Kuntz, 2006). Ces actions peuvent se faire aussi bien sur des motifs tels que les clusters, tendances, émergences, anomalies ou sur des ensembles d'éléments ou encore sur des éléments isolés. Les technologies de la visualisation permettent de communiquer efficacement des informations via des cartes cognitives et facilitent la découverte de connaissances grâce à une représentation graphique issue de l'analyse d'un corpus d'informations par des cartes sémantiques (Balmisse, 2005).

2.1.4. Définitions

La visualisation de données se base sur la notion de graphe, comme un objet mathématique défini par deux ensembles : les sommets et les arêtes. L'ensemble des sommets, noté *V*, correspond à l'ensemble des nœuds d'un réseau. Les arêtes, dont l'ensemble sera noté *A,* décrivent les relations entre les sommets du graphe. Ainsi deux sommets *u* et *v* de *X* seront reliés par une arête si la paire (*u; v*) appartient à l'ensemble *A*.

Un graphe $G = (X,A)$ est constitué d'un ensemble X ($|X| = N$) de sommets et d'un ensemble A d'arêtes. Une arête est une paire orientée ou non de sommets du graphe.

Sauf si c'est explicitement spécifié, nous supposerons G non orienté, c'est-à-dire qu'il n'y a pas de distinction entre (u, v) et (v, u) pour u et v dans X. G est considéré comme simple, c'est-à-dire qu'il n'y a pas de boucle (v,v) dans A et il existe au plus un lien entre deux nœuds. D'autre part, nous traitons aussi, dans ce qui suit, les graphes bipartis.

Un graphe biparti est un graphe dont l'ensemble des nœuds peut être divisé en deux ensembles disjoints *U* et *V* tel que chaque arête a un sommet en *U* et un sommet en *V*. Un graphe biparti permet notamment de représenter une relation binaire entre un ensemble *U* et un ensemble *V*. Si chaque nœud de *U* est relié à chaque nœud de *V,* alors le graphe biparti est complet.

Considérons les définitions suivantes :

Le nombre de sommets est appelé *ordre* du graphe. Le nombre de nœuds de *G* est noté $n = |X|$ et $m = |A|$ est son nombre de liens.

Lorsque $a = \{u, v\} \in A$, on dit que *a* est l'arête de *G* d'extrémités *u* et *v*, ou que *a* joint *u* et *v,* ou que *a* passe par *u* et *v*. Les sommets *u* et *v* sont dits adjacents dans *G*. Graphiquement, les sommets peuvent être représentés par des points et l'arête *a = (u, v)* par un trait reliant *u* à *v*.

Le *degré* de *u*, noté *d(u)*, est le nombre d'arêtes incidentes à *u*, c'est-à-dire contenant *u*. Lorsque *d(u) = 0*, on dit que le sommet *u* est isolé, lorsque *u = 1*, il est dit pendant.

Etant donné un nœud *v* dans *X,* son voisinage est noté *N(v)*, c'est-à-dire l'ensemble des nœuds liés à lui : $N(v) = \{u \in X, (v, u) \in A\}$. $d(v) = |N(v)|$ correspond au degré de *v*, c'est-à-dire son nombre de voisins.

Un graphe ne possédant pas de boucle, ni d'arêtes parallèles (deux arêtes distinctes joignant la même paire de sommets) est appelé *graphe simple* ou *1-graphe*.

Un *p-graphe*, appelé aussi graphe généralisé est un graphe pour lequel il n'existe jamais plus de *p* arêtes de la forme *(u, u)*.

Dans un graphe, un chemin entre deux nœuds est une suite de liens qui connecte ces deux sommets. La longueur du chemin correspond au nombre de liens de cette suite. La longueur d'un plus court chemin entre deux nœuds *u* et *v* est appelé distance entre *u* et *v* et sera notée *d (u, v)*. Le diamètre du graphe est la distance maximale définie par $max_{u,v \in V}\ d\ (u,v)$.

Si aucun des sommets composant la séquence n'apparait plus d'une fois, le chemin est dit *élémentaire.* Si aucune des arêtes composant la séquence n'apparait plus d'une fois, le chemin est dit *chemin simple.*

Un graphe est *connexe* si l'on peut atteindre n'importe quel sommet à partir d'un sommet quelconque en parcourant différentes arêtes. De manière plus formelle, un graphe *G* est connexe s'il existe au moins une chaîne entre une paire quelconque de sommets de *G*. Les ensembles de nœuds connexes maximaux d'un graphe sont appelés ses composantes connexes.

Un point d'articulation d'un graphe est un sommet dont la suppression augmente le nombre de composantes connexes.

Un ensemble d'articulation $A \subset X$ d'un graphe connexe *G* est un ensemble de sommets tels que le sous-graphe *G'* déduit de *G* par suppression des sommets de *A*, ne soit plus connexe.

Un graphe est *planaire* lorsqu'il admet une représentation sur un plan telle que deux arêtes ne se rencontrent pas en dehors de leurs extrémités.

Une distance est une application qui formalise l'idée intuitive de distance, c'est-à-dire la longueur qui sépare deux sommets d'un graphe. Nous considérons, dans tout ce qui suit, la distance euclidienne entre deux sommets de coordonnées (x_i, y_i).

$$\sqrt{\sum_{i=1}^{n} |x_i - y_i|^2} \qquad [1]$$

2.2. Domaines d'application

2.2.1. Réseaux sociaux

Les graphes sont utilisés couramment, entre autres, en sciences sociales pour modéliser des interactions entre individus. Un graphe fortement connexe, sans boucle et ayant plus d'un sommet, est appelé un réseau.

Un réseau social est constitué d'un ensemble d'individus et de relations qui les unissent. Les sommets représentent alors les entités ou individus, et les arêtes ou arcs une relation ou interaction entre eux. Par exemple la relation de connaître quelqu'un ou d'avoir collaboré avec cette personne.

- ✓ ***Réseaux d'acteurs***

Le réseau des acteurs est un graphe dont les sommets sont des acteurs. Deux de ces derniers sont reliés s'ils ont joué ensemble dans un film (Barabasi et Albert, 1999), (Amaral et al., 2000).

- ✓ ***Réseaux de citation***

Le graphe des citations est un graphe dont les sommets sont des publications scientifiques. Deux de ces dernières *u* et *v* sont liées par l'arc *(u,v)* si la publication *u* a cité en références la publication *v*.

Dans (Redner, 1998), des études ont été réalisées sur un graphe de 783 339 sommets issu du catalogue de l'*Institute for Scientific Information* et un graphe de 24 296 sommets issus des publications de la revue *Physical Review D*.

- ✓ ***Réseaux de co-auteurs***

Le graphe de co-auteurs est un graphe dont les sommets sont des auteurs scientifiques et deux auteurs sont reliés s'ils ont une publication commune (Newman et al., 1998), (Newman, 2001), (Barabasi et al., 2002).

- ✓ ***Réseaux de contacts***

Le graphe de connaissance est un graphe dont les sommets sont des personnes. Deux de ces dernières sont liées si elles se connaissent. Cette méthode est basée sur une propriété des réseaux sociaux qui sont des petits mondes (notion introduite dans les années soixante par Milgram (Milgram, 1967). Cette propriété correspond au fait que pour aller d'une personne à une autre dans le réseau, il suffit en moyenne de six intermédiaires. Formellement, cette définition est traduite par une mesure sur les sommets du réseau.

- ✓ ***Réseaux d'appels téléphoniques***

Le graphes des appels téléphoniques est un graphe orienté dont les sommets représentent des numéros de téléphones et un arc signale qu'un numéro a au moins une fois appelé un autre (Aiello et al., 2000).

2.2.2. Réseaux sémantiques

Un concept n'existe que par l'ensemble des relations dans lesquelles il intervient. Le graphe des relations entre les concepts constitue le réseau sémantique. On fait remonter l'idée des réseaux sémantiques à (Quillian, 1968) qui a proposé un système de représentation dans lequel les concepts acquerraient leur signification par les liens qui les unissaient à d'autres concepts.

Le réseau sémantique est représenté sous la forme d'un graphe étiqueté et orienté. Un arc lie un nœud de départ à un nœud d'arrivée. Chaque nœud peut être relié par un ou plusieurs arcs. Les inférences possibles dépendent de la nature des liens.

- ✓ ***Graphes conceptuels***

Le modèle des graphes conceptuels se base sur la représentation de connaissances du type réseaux sémantiques qui a donné lieu à un certain nombre de travaux depuis son introduction par Sowa (Sowa, 1984). L'une des particularités de ce modèle est de permettre de représenter des connaissances sous forme graphique.

Un graphe conceptuel est un graphe biparti étiqueté, les deux classes de sommets étant étiquetées respectivement par des noms de « concepts » et des noms de « relations conceptuelles » entre ces concepts. Une telle

représentation graphique des connaissances permet à des utilisateurs de comprendre, créer ou modifier directement des objets de ce type, de façon beaucoup plus simple (en comparaison avec une représentation sous forme de formules logiques). Les graphes conceptuels ont été utilisés dans les systèmes d'information pour la représentation de requêtes et de documents dans (Guarino 1999).

- ✓ ***Ontologies***

« An ontology defines the basic terms and relations comprising the vocabulary of a topic area, as well as the rules for combining terms and relations to define extensions to the vocabulary", selon (Neches et al., 1991) . Suite à cette première élaboration, la définition communément admise d'une ontologie est énoncée par (Gruber, 1993) comme « la spécification explicite d'une conceptualisation". Cette définition a également été récemment précisée pour devenir : « *Définir une ontologie pour la représentation des connaissances, c'est définir pour un domaine et un problème donnés, la signature fonctionnelle et relationnelle d'un langage formel de représentation et la sémantique associée* » (Bachimont, 2000).

Une ontologie est donc une structure de graphe qui regroupe un ensemble de concepts décrivant complètement un domaine. Ces concepts sont liés les uns aux autres par des relations taxonomiques (hiérarchisation des concepts) d'une part, et sémantiques d'autre part.

Il existe plusieurs types d'ontologies et ses applications sont diverses :

L'*ontologie informatique* permet de représenter précisément un corpus de connaissances sous une forme utilisable par une machine. Elle représente un ensemble structuré de concepts. Les concepts sont organisés dans un graphe dont les relations peuvent être des relations sémantiques et/ou des relations de composition ou encore d'héritage au sens objet.

L'*ontologie de l'information* contribue à organiser et clarifier les idées des collaborateurs sur un projet en exposant le schéma global du système avec tous ses liens et ses raisonnements. Ce type d'ontologie est utilisé dans un projet dans le but de réduire les incompréhensions et les quiproquos.

L'ontologie du domaine est utilisée pour représenter un domaine (les composants informatiques, l'immobilier, le droit, la génétique …) sous forme de base de connaissances. Elle présente les concepts-clés, les attributs, les instances relatifs au domaine.

2.2.3. Autres réseaux

- ✓ ***Réseaux de neurones***

Les réseaux de neurones sont représentés par un graphe dont les sommets correspondent aux cellules, interconnectées entre elles. Chaque point de connexion (appelé coefficient ou poids) entre deux cellules joue le rôle d'une synapse, c'est l'élément principal d'interaction entre les neurones (McCulloch et pitts, 1943), (Hopfielf, 1982), (Minsky et Papert, 1988), (Dreyfus et al., 2004).

- ✓ ***Réseaux trophiques***

Les réseaux trophiques sont représentés par un graphe dans lequel un sommet est une espèce animale ou végétale.Deux espèces sont reliées si une des deux espèces est un prédateur de l'autre (Schindler et Scheuerell 2002).

- ✓ ***Réseaux d'interactions entre protéines***

Le réseau d'interactions entre protéines est modélisé par un graphe dans lequel les sommets sont des protéines. Deux de ces dernières sont reliées si elles réagissent l'une avec l'autre (Jeong et al., 2000).

✓ ***Réseaux de distribution électrique***

Le réseaux de distribution électrique est un graphe dont les sommets sont des stations (générateurs, transformateurs) électriques et deux stations sont liées si une ligne de haute tension existe entre les deux stations (Watts et Strogatz, 1998), (Amaral et al., 2000). Plusieurs applications sont utilisées dans les réseaux d'interactions : routage efficace sur Internet, algorithmes de recherche d'information efficaces sur le Web, ...

Une connaissance approfondie et exacte des réseaux d'interactions permet de concevoir des applications mieux adaptées et plus performantes. En effet, l'étude des réseaux d'interactions en tant que graphe a permis, sur le Web, de développer des algorithmes très efficaces tels que l'algorithme PageRank utilisé par de nombreux moteurs de recherche.

✓ ***Réseaux informatiques***

Un réseau informatique caractérise un ensemble d'ordinateurs (représentés par des nœuds), reliés entre eux grâce à des lignes physiques (matérialisés par des liens entre les sommets) et échangeant des informations sous forme de données numériques.

✓ ***Réseaux de transports***

Un réseau de transports est défini comme un ensemble d'infrastructures et de disposition permettant de transporter des personnes ou des biens entre plusieurs zones géographiques.

Ils peuvent être de type :

- *aériens*. Fréquemment matérialisé par une carte géographique, les nœuds du graphe sont assimilés aux différents aéroports et les liens correspondent aux voies aériennes en service.
- *routier*. Matérialisés aussi par une carte géographique, les différentes villes sont représentées, dans un graphe de transport, par des sommets et les liens correspondent aux axes routiers, qui peuvent être de type différents (autoroute, route, sentier, voie ferrée…).
- *maritimes*. Les graphes sont composés de nœuds représentant les ports et les liens entre ces derniers constituent les voies navigables.
- *ferroviaires /métro*. Les différentes stations sont reliées par les voies fréquentées par le train/métro.

2.3. Principes de la visualisation de graphes

La visualisation des structures complexes est une composante essentielle des outils utilisés dans de nombreuses applications en sciences. Les graphes sont utilisés pour visualiser des informations modélisées sous forme d'objets connectés.

Par conséquent, la représentation des données doit être facile à lire et à comprendre (DiBattista et al., 1999).

2.3.1. Complexité de la visualisation

La plupart des algorithmes de placement de graphes ont une complexité telle qu'ils ne sont pas adaptés à la représentation de très grands graphes. Il convient donc d'étudier les différents types de graphes, les problèmes rencontrés quand à leur représentation et leur complexité (Tamassia, 1997).

Une idée de la complexité des grandes classes d'algorithmes est illustrée dans le Tableau 11 :

Classe de graphes	Problème	Complexité
Graphes généraux	Minimisation des intersections	NP-difficiles
Graphes généraux	Détermination du sous-graphe planaire maximal	NP-difficiles
Graphes généraux	Test de planarité	O(n)
Graphes orientés généraux	Test de planarité	$O(n^2)$
Graphes orientés généraux à source unique	Test de planarité	O(n)
Graphes planaires	Dessin en segment de droite à longueur constante	NP-difficiles
Graphes planaires	Dessin en segment de droite avec une résolution angulaire maximale	NP-difficiles
Graphes planaires	Dessin planaire en segment de droite sur une grille en une surface en $O(n^2)$ et une résolution angulaire en $O(1/n^2)$	O(n)
Graphes planaires	Dessin planaire en polylignes avec une surface en $O(n^2)$ et une résolution angulaire en $O(1/d)$	O(n)
Graphes planaires de degré 3	Dessin orthogonal minimisant le nombre de segment par arête avec une surface en $O(n^2)$	$O(n^5 \log n)$
Graphes planaires de degré 4	Dessin orthogonal avec une surface en $O(n^2)$ et un nombre de segments par arête en $O(n)$	O(n)

Tableau 11. *Problèmes et complexités temporelles des graphes (Tamassia, 1997).*

2.3.2. Sémiologie graphique

La compréhension de la visualisation graphique repose sur des règles de construction de la symbolique, c'est la sémiologie (étude des signes et de leur signification), elle repose également sur une utilisation codifiée des écritures et sur des principes esthétiques généraux.

Bertin est considéré comme l'initiateur en terme de cartographie d'information (Bertin, 1970). Il s'intéresse à la construction de visualisation par des symboles graphiques. La sémiologie graphique repose sur la signification des dessins, le choix des légendes, des symboles, des icônes, une méthodologie pour faire passer un message visuel.

D'après Bertin, le lecteur de la carte perçoit six variations sensibles attachées aux symboles de la carte. Il les appelle variables visuelles, « *composantes du système d'expression* », synthétisées dans le **Tableau 12**.

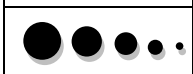
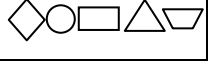
Taille	Valeur	Grain	Couleur	Orientation	Forme
					

Tableau 12. *Sémiologie graphique selon (Bertin, 1970).*

La variation de taille permet de traduire parfaitement les variations quantitatives. La taille n'est pas forcément assujettie à la dimension de l'objet qu'elle représente, mais selon l'objectif de la représentation, à l'importance que l'on désire attribuer au message.

La variation de valeur d'une couleur est une variation d'intensité lumineuse du plus sombre au plus clair, ou inversement ; elle traduit une relation d'ordre et des différences relations quantitatives. Cependant, notre capacité à reconnaître est bien plus limitée que notre aptitude à apprécier. La sensibilité différentielle de l'œil à l'énergie lumineuse n'est pas directement proportionnelle à l'intensité du flux. L'appréciation des dégradés est plus faible dans les couleurs claires que dans les foncées. Notre sensibilité chromatique différentielle n'est pas uniforme, non plus, tout au long du spectre. Il en résulte qu'en cartographie on estime qu'en fonction des couleurs le nombre de paliers (longueur de la variable) sera de : 6 du Blanc au Noir, 5 pour les Violets et Rouges, 4 pour les Bleus et Orangés, 3 pour les Verts, 2 ou 3 pour les Jaunes.

Les grains constitutifs des trames combinent déjà plusieurs variables telles que les formes, la taille et traduisent une relation d'ordre et des différences relatives (relation quantitative). Cette sémiologie permet de positionner un signe par rapport aux deux axes du graphique; elle exprime les différences et gagne en efficacité en combinant les variables de grains et de valeur.

Les couleurs traduisent des différences et peuvent les ordonner entre elles par le biais de dégradés; elles sont de plus chargées de significations culturelles et psychologiques. Bien que notre œil soit capable d'apprécier quelques milliers de nuances et que l'artiste puisse se permettre une infinité de coloris, la palette du cartographe sera réduite à ce que l'utilisateur est capable de différencier et surtout de mémoriser en fonction du contenu de la carte (Cosmin Porumbel et al, 2007), (Cosmin Porumbel et al, 2009). Une vingtaine de couleurs différentes semble être la limite de la variable.

L'orientation permet de positionner un signe par rapport aux deux axes du graphique; elle exprime les différences et gagne en efficacité en combinant les variables de grains et de valeur. Dans une image complexe, l'œil ne peut discerner sans erreur que les quatre directions principales : les deux axes de la carte et deux obliques opposées. Il n'est pas raisonnable d'infliger au lecteur de multiples et subtiles différences d'orientation qu'il devra vérifier en légende. Quatre sera donc la longueur de cette variable.

Les formes expriment relativement bien l'identité de l'objet à représenter et donc, par relation, les différences; qu'il s'agisse de pictogrammes ou de formes fondamentales (le carré, le cercle, etc.). Leur lisibilité est souvent plus grande que celle des dessins réalistes. La forme peut être figurative (pictogramme), évocatrice (idéogramme), ou purement symbolique. Sa configuration peut être géométrique ou quelconque. La création de formes n'a pour limite que l'imagination du créateur, on dit que cette variable a une longueur infinie.

2.3.3. Problèmes d'ergonomie

Quand la conception d'un graphe n'est soumis à aucune convention graphique, elle doit respecter des critères esthétiques, tels que :

- minimiser le nombre d'intersections,
- minimiser la surface occupée par le dessin,
- minimiser la longueur totale des arcs,
- minimiser la longueur maximale d'un arc,
- minimiser la variance de la longueur des arcs,
- minimiser le nombre total d'inflexions des arcs,
- minimiser le nombre maximal d'inflexions par arc,
- minimiser la variance du nombre d'inflexions des arcs,
- maximiser l'angle minimal entre deux arcs issus du même sommet,
- minimiser le rapport entre le segment le plus long et le segment le plus court,
- afficher des symétries.

Les travaux de (Tufte, 1983) reposent sur d'autres principes d'excellence graphique qui consistent à concevoir efficacement une présentation de données intéressantes, à communiquer des idées complexes de manière claire, précise et efficiente[9], à donner au lecteur le plus grand nombre d'idées dans le plus court intervalle de temps possible et avec le moins d'encre possible dans le plus petit espace possible.

Tufte propose aussi d'autres principes réunis sous le titre de principes d'intégrité graphique (Tufte, 2001). La représentation graphique de valeurs numériques est physiquement mesurable sur le graphique.

L'étiquetage clair, détaillé et complet permet de réduire les déformations graphiques et les ambiguïtés. Les explications du graphique doivent se trouver en dehors du graphique. Les événements important doivent, en revanche, être signalés sur le graphique.

Seules les variations dues aux données sont mises en évidence.

Le nombre de dimensions utilisées pour représenter une donnée ne doit pas excéder le nombre de dimensions de cette même donnée.

Ces principes, inspirés par de multiples observations, peuvent être intéressants en tant que ligne de conduite à tenir lors de la conception de représentations visuelles de données (Kuntz et al., 2006).

2.3.4. Préconisations sur la planarité

Les données à visualiser sont souvent hétérogènes, les graphes possèdent rarement de bonnes propriétés, en particulier pour la planarité. Pour cela, il est nécessaire de disposer d'algorithmes ouvert à une classe de graphe générale, sans propriétés particulières, tels que les algorithmes « Force-Directed-Placement »(FDP).

Pour cela, on distingue deux caractéristiques :

- le modèle basé sur une analogie physique (ressorts, particules, forces, ...);
- l'algorithme simulant la dynamique du modèle permettant d'approcher une configuration «stable» du système physique.

Certains de ces algorithmes ont été proposés pour réaliser de grandes visualisations graphiques (Hachul et Jünger, 2005). Les algorithmes de Placement Dirigés par des Forces sont aussi appelés « à ressort » ou encore « modèles physiques ».

Les travaux portant sur ce domaine reposent sur une méthode de visualisation des graphes qui présente ces derniers comme des systèmes physiques (Tutte, 1963), puis (Eades 1984) où les sommets sont considérés comme des anneaux et les arêtes comme des ressorts reliant les anneaux 2 à 2. Le graphe est initialement positionné de manière aléatoire dans l'espace de tracé. Le système évolue librement de manière à aboutir à un équilibre, où la somme des forces exercées par les élastiques est minimale. Plus deux sommets sont fortement liés, plus ils s'attirent et inversement.

De même, plus le lien entre les deux sommets est faible, voire inexistant, plus ils se repoussent. Ces algorithmes itératifs permettent le déplacement des sommets de manière à ce que leur proximité soit significative de forte liaison, dans le cas où deux sommets sont joints par une arête. Ainsi, le nombre d'entrecoupements d'arêtes est limité et le graphe devient alors plus clair à étudier.

Dans la Figure 14, le principe des algorithmes FDP est illustré. La première figure à gauche représente un graphe non orienté dont la valeur des liens est affichée. Le second représente ce même graphe, mais pour lequel les liens ont été assimilés à des ressorts et les sommets à des anneaux. Dans la troisième figure, les ressorts sont contractés ou relâchés selon l'importance du lien. La dernière figure illustre l'apport de cet algorithme, permettant de visualiser le graphe sous sa forme la plus planaire et la plus significative

Dans (Kamada et Kawai, 1989), il s'agit d'une évolution de l'algorithme d'Eades, intégrant les interactions entre sommets non voisins, par le biais des courts chemins reliant ces sommets. Cet algorithme est basé sur la résolution d'équations différentielles.

[9] l'efficience est le rapport entre l'effort cognitif du lecteur et l'apport informationnel qu'il en tire.

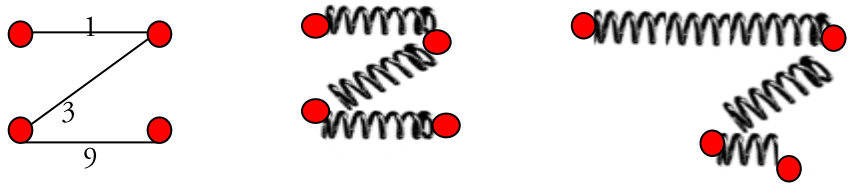


Figure 14. *Principe de l'algorithme de ressorts.*

Kamada et Kawai considèrent le dessin d'un graphe comme un processus de réduction de l'énergie totale d'un système de ressorts reliant des anneaux. En minimisant la somme de la compression et de la tension sur tous les ressorts, les anneaux devront être plus proches de leurs distances idéales les uns des autres.

Les travaux de (Groves et al., 1990) permettent d'optimiser un graphe en prenant en compte un critère d'esthétisme paramétrable en fonction des objectifs à atteindre.

Les travaux de(Fruchterman et Reingold, 1991) aboutissent à l'élaboration d'un algorithme de répulsion électrostatique entre tous les sommets. Une attraction est appliquée entre les nœuds liés. Cet algorithme est basé sur la notion de température de représentation, qui diminue lors du déplacement des nœuds. Le déplacement maximal possible d'un nœud est restreint par une valeur maximale. Cette valeur maximale est réduite après chaque itération pour profiter du fait que la qualité de la représentation d'un graphe devient meilleure après chaque étape. Le déplacement maximal possible devient de plus en plus petit. Plusieurs types de forces peuvent être utilisés : la force électrique et la règle de Hooke.

Les auteurs ont proposé aussi une variante de la méthode, "grid-variant", pour accélérer la vitesse d'exécution de la méthode. Dans cette variante, les auteurs appliquent une grille qui divise l'espace d'affichage en plusieurs carrés.

Chaque nœud est placé dans un carré et à chaque itération, les forces répulsives ne sont calculées qu'entre les nœuds résidant dans des carrés voisins. Le résultat ainsi obtenu est équivalent à celui qu'on obtient en appliquant les forces répulsives entre tous les nœuds.

Dans (Davidson et Harel, 1996), il s'agit d'un algorithme de recuit simulé, l'objectif étant ici de minimiser le niveau d'énergie globale d'un système.

Dans (Frick et al. 1991), chaque nœud du graphe cherche sa position d'énergie minimale en rapport avec ses voisins. Cette méthode présente l'avantage de favoriser certaines caractéristiques esthétiques d'un graphe (Eades et Lin 2000), et surtout de préserver la carte mentale de l'utilisateur (Eades et al., 1991). Ces algorithmes permettent de dessiner des graphes sans propriété particulière.

L'algorithme (Harel et Koren, 2002) se base sur le problème du k-centre.

Dans (Walshaw, 2003), les sommets sont regroupés formant des clusters. Ces derniers sont utilisés pour former un nouveau graphe. Ce procédé est répété jusqu'à ce que la taille du graphe tombe au-dessous d'un certain seuil. L'algorithme de placement des sommets est alors appliqué sur ce graphe grossier. Une fois les positions définies, un retour au graphe initial est établi.

Dans (Koren et al., 2003) l'algorithme se base sur une fonction simplifiée calculant l'énergie du système, permettant ainsi un traitement mathématique plus robuste.

Dans (Harrel et Koren, 2004), la méthode comporte deux phases: tout d'abord intégrer le graphique dans une très haute dimension et ensuite le projeter dans un plan 2D.

Tous ces algorithmes diffèrent non seulement par leurs techniques d'optimisation, mais aussi dans leurs modèles implicites de l'espace de représentation. Plusieurs logiciels ont recours à ces algorithmes. Parmi les plus utilisés, on distingue Pajek (Batagelj et Mrvar, 1998), GraphVis (AT&T, 2003), NetDraw (Borgatti 2002), et d'autres (Krackhardt et al., 1994); (Himsolt, 1995).

Mis à part sa simplicité de mise en œuvre, l'approche FDP présente le grand avantage d'apporter une forte plasticité au graphe. Une déformation de la topologie du graphe en un point (ajout ou suppression d'un nœud, ajout ou suppression d'une relation, etc.) n'induit pas une reconstruction du dessin du graphe dans sa totalité. Cette idée de persistence est présente chez les auteurs qui visent à maintenir une carte mentale chez l'utilisateur.

"*When the graphe is redrawn it is important to preserve the look (the user's mental map) of the original drawing as much as possible, so the user does not need to spend a lot of time relearning the graph*" (Bridgeman et Tamassia, 2000).

Les méthodes de forces étant des méthodes itératives, il est nécessaire de considérer les critères d'arrêt de ces méthodes. Le critère d'arrêt le plus simple est basé sur le nombre d'itérations à réaliser. Ce nombre peut être simplement fixé de manière arbitraire. Dans la réalité, ce nombre reste difficile à choisir (une valeur commune pour tous les cas de graphes). Un autre critère repose sur des notions comme l'énergie totale ou la température totale : l'arrêt survient quand ces quantités deviennent nulles ou restent constantes.

2.3.5. Préconisations sur les structures de visualisation

Il existe un grand nombre de techniques de visualisation de l'information qui ont été développées pendant la dernière décennie pour permettre l'exploration de grands ensembles de données.

Plusieurs typologies ont été proposées :

Les travaux de Ben Shneiderman (Shneiderman, 1996) s'appuient sur la nature des données, unidimensionnelles, bidimensionnelles, tridimensionnelles, multidimensionnelles, temporelles, hiérarchiques, réseaux.

La typologie présentée par Daniel Keim (Keim, 2002) s'appuie sur trois critères : les techniques de visualisation, le type de données et le type d'interaction.

Enfin, celle de Keith Andrews (Andrews, 2002) qui s'appuie également sur les types de données.

Pour des raisons d'ordre didactique, les outils de visualisation sont répertoriés selon la catégorisation de Keith Andrews qui prend en compte les données de type vectoriel que Ben Shneiderman ne mentionne pas dans sa typologie.

Keim propose une classification (Keim, 2002), illustré dans la Figure 15, selon trois critères : les données à visualiser, la technique de visualisation, la technique d'interaction et de déformation.

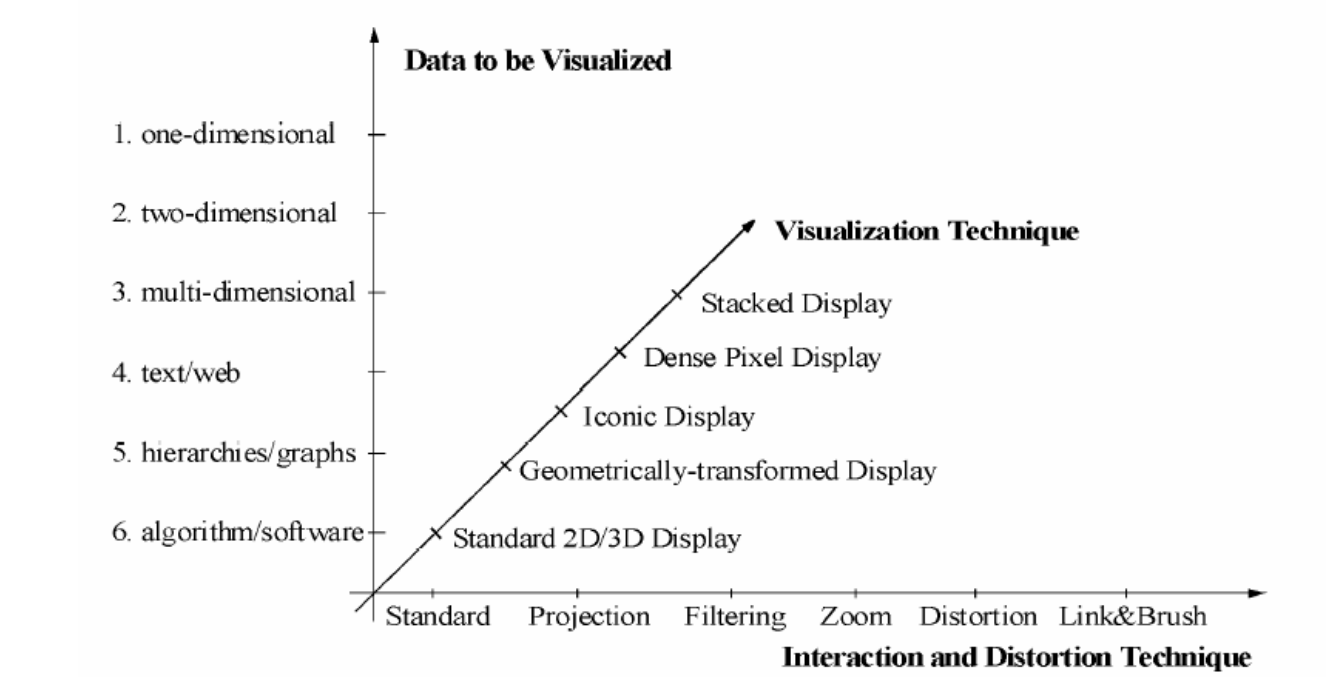


Figure 15. *Classification de (Keim, 2002).*

Les données unidimensionnelles ont habituellement une dimension dense, comme par exemple, les données temporelles de ThemeRiver (Nowell et al., 2001).

Les données bidimensionnelles ont deux dimensions distinctes, les axes *X* et *Y* permettent de montrer des données bidimensionnelles, comme par exemple les données géographiques où les deux dimensions sont longitude et latitude. Polaris (Tang et al., 2001), MGV (Abello et al., 2001).

Beaucoup d'ensembles de données se composent de plus de trois attributs et donc, ils ne permettent pas une visualisation simple à deux ou trois dimensions. Ces données sont qualifiées comme multidimensionnelles (ou multivariables). Elles peuvent être par exemple des tables des bases de données relationnelles qui ont souvent des dizaines et même jusqu'à des centaines de colonnes (ou attributs). Polaris (Tang et al., 2001), Scalable Framework (Lopez et al., 2001).

Tous les types de données ne peuvent être décrits en termes de dimensionnalité. A l'âge du web, un type important de données est le texte et l'hypertexte ainsi que les contenus des pages web multimédia. Dans la plupart des cas, une transformation de données en des vecteurs descriptifs est d'abord nécessaire avant que des techniques de visualisation puissent être utilisées, comme par exemple les articles de nouvelles, les documents web.

Les enregistrements ont souvent un certain lien avec d'autres informations. Des graphiques sont largement répandus pour représenter de telles interdépendances, telles que les hierarchies. Un graphique se compose de l'ensemble d'objets, appelé des nœuds, et les connexions entre ces objets, appelés les bords, comme par exemple les corrélations des e-mails entre des personnes, leur comportement d'achats, la structure des fichiers du disque dur, les hyperliens du web.

Pour les algorithmes et logiciels, le but de la visualisation est de soutenir le développement de logiciel en aidant à comprendre des algorithmes, par exemple en montrant le flux de l'information dans un programme, pour améliorer la compréhension du code écrit en représentant la structure des milliers de lignes du code source comme des graphiques, et pour soutenir le programmeur en corrigeant le code, c'est à dire en visualisant des erreurs.

La technique de visualisation utilisée peut être classifiée dans des affichages:

- standards en 2D/3D (standard 2D/3D display), comme par exemple les diagrammes à barres, *x-y* plot, graphiques linéaires…
- géométriquement transformés (Geometrically transformed Displays). Ces techniques visent à trouver des transformations « intéressantes » des ensembles de données multidimensionnels. Les techniques rencontrées peuvent être de type Scatterplots (nuage de points), Landscapes, Projection Pursuit, Prosection Views, Hyperslice, Parallel Coordinates…
- basés sur des icônes (Icon-based Displays). L'idée est de tracer les valeurs d'attribut d'une donnée multidimensionnelle sous forme d'une icône. Comme exemple, nous pouvons citer Chernoff faces, Stick figures, Shape-Coding, Color-icons, TileBars, …
- denses en pixel (dense pixel displays). L'idée de base de cette technique est de tracer chaque valeur d'une dimension en un pixel coloré et de grouper les pixels appartenant à chaque dimension en des zones adjacents. Puisqu'en général cette technique emploie un pixel par valeur de donnée, elle permet de visualiser la plus grande quantité de données possibles sur le même dispositif.
- empilés (stacked displays). Ces techniques sont conçues pour présenter des données divisées d'une manière hiérarchique.

2.4. Prise en compte de la dimension temporelle

2.4.1. Données temporelles

Une donnée temporelle est caractérisée comme une association entre une dimension valeurs temporelles et une dimension valeurs structurelles (Singh et al., 2006), (Gao et revesz, 2006). Cette dernière correspond aux valeurs de la donnée. La dimension temporelle correspond aux valeurs temporelles celles de la donnée (Ankerst et al, 2008), (Graham et al., 2008), (Castellani et al, 2008).

La suite d'observations $(y_t, t \in T)$ d'une variable y à différentes dates test appelée série temporelle. Habituellement, T est dénombrable, de sorte que $t=1,\ldots T$. La dimension temporelle est ici importante car il s'agit de l'analyse d'une chronique historique : des variations d'une même variable au cours du temps, afin de pouvoir comprendre la dynamique. La périodicité de la série n'importe en revanche pas : il peut s'agir de mesures quotidiennes, mensuelles, trimestrielles, annuelles… voire même sans périodicité. La fonction première pour laquelle il est intéressant d'observer l'historique d'une variable vise à en découvrir certaines régularités afin de pouvoir établir une prévision. Il s'agit ici de supposer que les mêmes causes produisent les mêmes effets. Avec une analyse fine, il est même possible d'établir des prévisions "robustes" vis-à-vis de ruptures brusques et de changements non anticipables. Ainsi, une série temporelle n'existe que dans la mesure où le caratère mesuré est invariant dans le temps.

Deux types d'unités temporelles sont généralement considérés dans les représentations symboliques: l'unité temporelle qui dure, appelée intervalle et celle qui ne dure pas, l'instant.

L'*instant* est l'entité temporelle sans durée ou ponctuelle, par analogie avec un point sur une droite. Un instant peut être représenté numériquement par une date. Représenter les instants de manière symbolique consiste à les identifier et à les mettre en relation.

L'*intervalle* est l'entité temporelle qui dure. Par analogie à la droite, il peut être assimilé à un segment. C'est ainsi que l'intervalle sera représenté de manière numérique, soit par une date de début et une date de fin, soit par une date de début et une durée.

A une unité temporelle t_i, la donnée temporelle a une valeur y_i. L'ensemble des données temporelles sont donc issues d'une fonction du temps, sous la forme :

$$D = \{(t_1,y_1),(t_2,y_2),\ldots,(t_n,y_n)\} \text{ avec } y_i=f(t_i)$$

Différentes caractéristiques peuvent être extraites de ces données temporelles. La modélisation markovienne suppose qu'un évènement X survenu à l'instant t peut être prédit en fonction des k évènements précédents $\{X_{t-1}, X_{t-2}, \ldots, X_{t-k}\}$, encore appelés retards. Un individu peut alors être représenté comme étant composé de k variables exogènes, les retards et d'une variable endogène, l'évènement à l'instant t, caractérisant alors le système comme autorégressif. Cette notion est importante dans un contexte d'analyse de données relationnelles évolutives et principalement dans la détection des signaux faibles.

2.4.2. Espace temps

Le temps n'ayant qu'une dimension, sa représentation est assez pauvre par rapport à celle de l'espace. Il est défini selon des caractéristiques spécifiques telles que le passé, le présent, le futur, selon une relation d'ordre (Keogh, 2005).

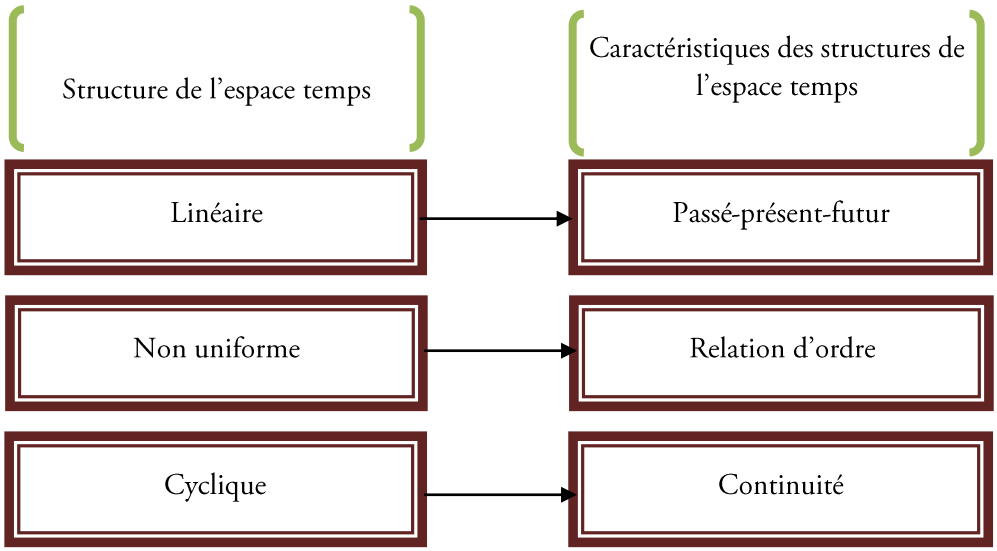


Figure 16. *Structure et caractéristiques de l'espace temps.*

En se basant sur les travaux de (Daassi, 2003), nous retenons trois principales représentations du temps, sous forme *linéaire* ouverte (time line), mettant l'accent sur les évènements passés, présents et futurs, sous forme *non uniforme*, mettant en avant une relation d'ordre des évènements ou encore sous forme d'une ligne fermée circulaire, c'est-à-dire *cyclique*, en raison de l'aspect répétitif de certains évènements tel que, par exemple, une journée où les heures seront graduellement les mêmes et qui constitue une certaine continuité, comme le montre la Figure 16.

✓ ***Visualisation linéaire***

La visualisation linéaire est la méthode la plus répandue pour représenter le temps, dominant notre conception de la durée.

Ce type de représentation utilise un axe horizontal dont la partie gauche correspond généralement au passé, le centre donne une vision du présent et la partie droite est assimilée au futur. La *timeline*, encore appelée la ligne de temps ou la chronologie est l'outil qui permet de représenter des données en les situant sur ce qui pourrait être qualifié comme la version numérique de la frise chronologique.

Les timelines sont le plus souvent utilisées quand on doit placer des événements parallèles par rapport à un axe temporel absolu. Elles sont utilisées dans des domaines comme le traitement du son, le montage vidéo, la gestion de projets, ... Dans la plupart des systèmes, l'auteur peut placer directement les icônes sur l'axe temporel. De plus la granularité du timeline peut être adaptée aux besoins.

Parmi les outils permettant l'analyse de données évolutives, nous pouvons citer timeline (Morris et al., 2003), qui permet la construction, l'exploration, et l'interprétation des lignes de temps afin d'identifier et de visualiser les citations d'articles de recherches sur des thématiques précises sur plusieurs périodes. Le graphe se base sur un repère dont les abscisses représentent le temps et les ordonnées sont les domaines de recherche, classées selon leurs similarités, comme le montre la figure 17. Ainsi chaque sommet du graphe, dont la taille est relative à la valeur de sa métrique correspond au nombre de citations, à une date et à une activité de recherche particulière.

Dans l'exemple de la Figure 17, les papiers sont représentés par des cercles dont la taille est proportionnelle au nombre total de fois où ils ont été cités. Les cercles remplis sont des papiers qui ont été cités huit fois ou plus au cours des 12 derniers mois.

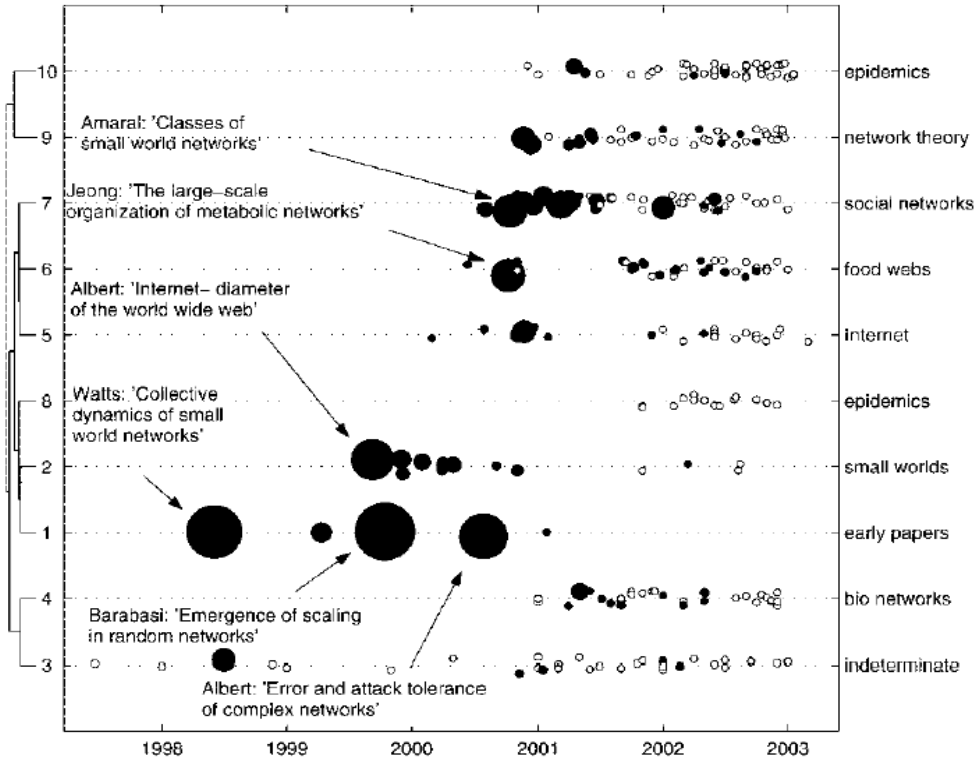


Figure 17. *TimeLine sur des articles de recherche (Morris et al., 2003).*

Le système Starfield (Jog et Schneiderman, 1995), (Cailleteau et Plaisant, 1999) de la figure 17, développé pour visualiser des informations relatives à des films telles que la date d'apparition, popularité, liste d'acteurs et actrices, etc., utilise une ligne de temps. Le graphe de cette figure est réalisé à partir d'une base de données constituée de deux mille films. Starfield se base sur la représentation d'un film par un rectangle coloré selon la catégorie du film telle que drame, comédie, …, situé par rapport à l'axe des ordonnées selon sa popularité et à l'axe des abscisses représentant le temps en années.

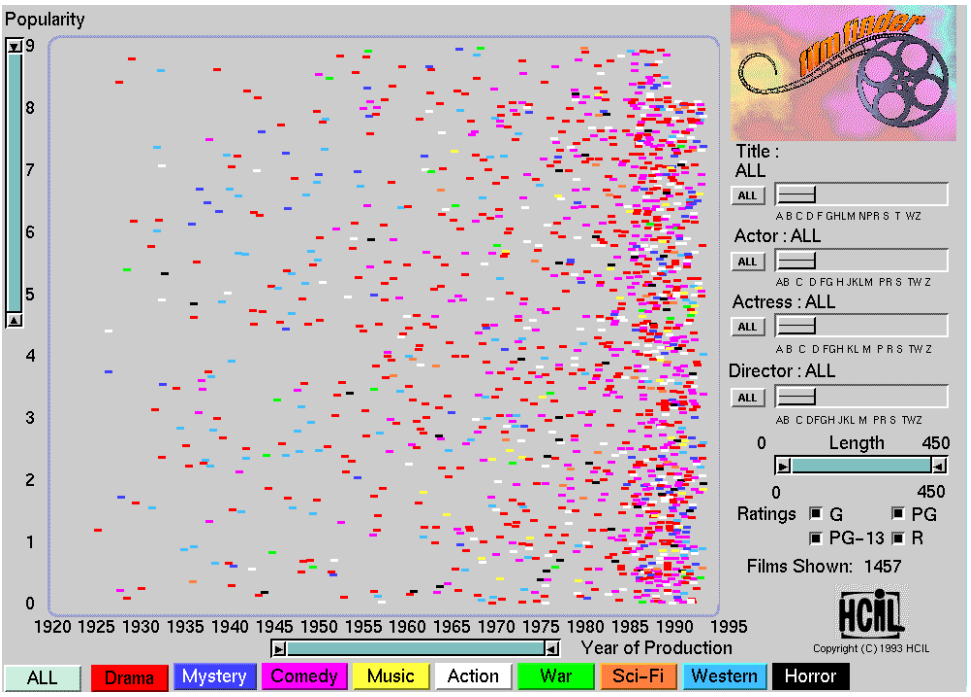


Figure 18. *Starfiel, système dédié à la visualisation d'une base de données de films.*

Une approche pour concevoir des techniques animées par images (*frame-based animation*) (Hansen, 2001) consiste à créer des images individuelles et à les afficher rapidement dans un ordre chronologique, ce qui crée une illusion de mouvement ou de changement. Une taxonomie détaillée des techniques d'animation est présentée dans (Thalmann et Thalmann, 1994).

L'outil ThemeRiver (Havre et al., 1999) de la Figure 19 utilise la métaphore d'une rivière pour représenter une animation des valeurs temporelles : le temps passe comme le flux d'une rivière.

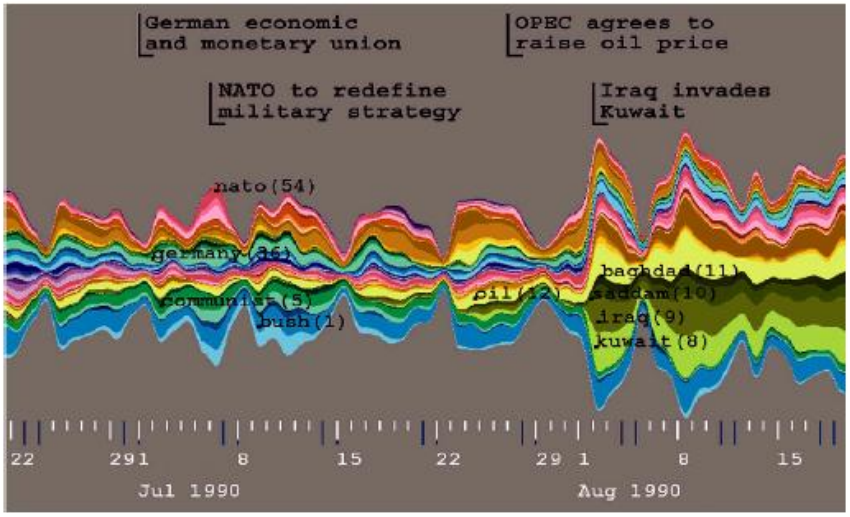


Figure 19. *ThemeRiver (Havre et al., 1999), (Havre et al., 2000).*

L'animation dans la technique ThemeRiver est ainsi horizontale alors qu'elle est en profondeur pour le système TimeMap Viewer. Ce dernier (Johnson et Wilson, 2002), présenté Figure 20, est développé en Borland Delphi. Cet outil repose sur un système d'informations géographiques (SIG) pour la cartographie de données spatio-

temporelles en deux dimensions. TimeMap Viewer est principalement utilisé pour des applications archéologiques. Le temps est représenté en abscisse et peut être zoomé pour obtenir une précision en jours, en mois, en années ou encore la période totale des données affichées.

Chaque image correspond à une période spécifique et l'animation est utilisée pour observer les changements au cours du temps. Il est alors possible de revenir sur une image correspondant à une période antérieure *(i-1)* ou inversement de revenir sur une vision actuelle à l'instant *i*.

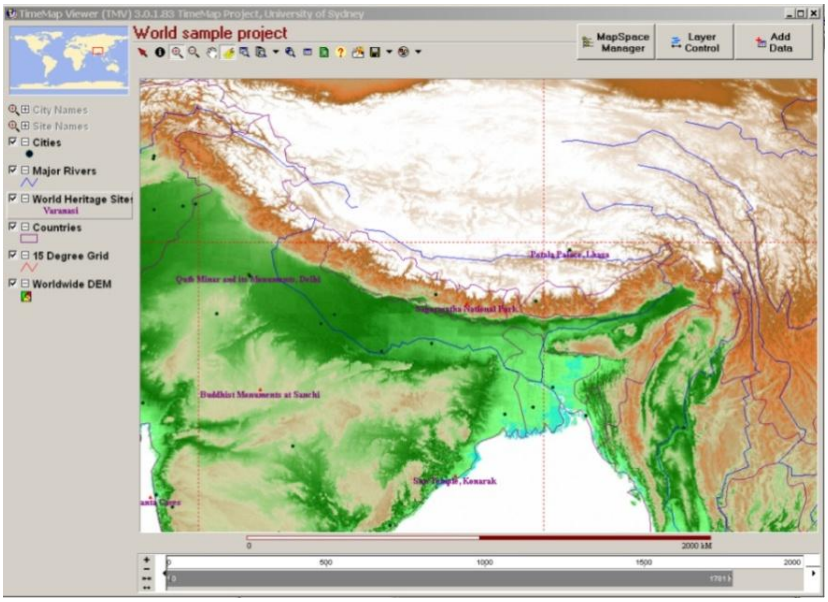


Figure 20. *System TimeMapViewer (Johnson et Wilson, 2002).*

Nous pouvons aussi citer le système OITL (Bui et al., 2001), qui représente le temps par une dimension dans un espace final en deux dimensions sous la forme d'un navigateur référencé par l'étiquette ligne temps. Le système OITL, présenté en Figure 21, permet de visualiser l'historique médical d'un patient : les images radio, les différents rapports médicaux, etc.

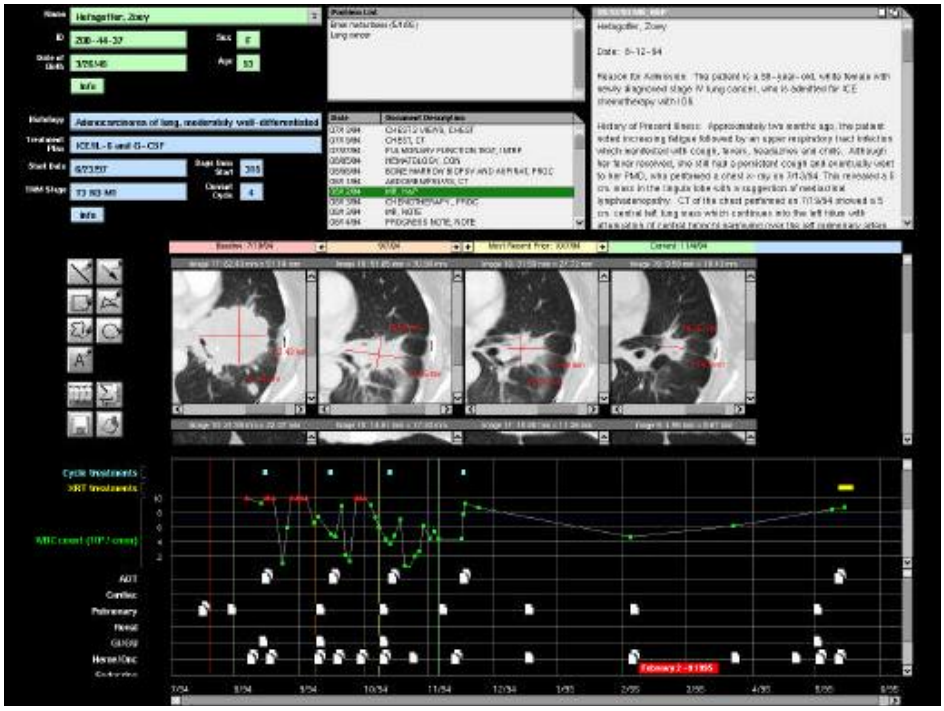


Figure 21. *Système OITL utilisant le principe du timeline (Bui et al., 2001).*

L'avantage majeur de cette approche est la facilité avec laquelle on visualise les événements qui se déroulent en parallèle. De plus, la représentation est très intuitive ce qui lui permet d'être facilement compréhensible par le lecteur.

De même, nous pouvons citer l'outil LifeStreams pour des documents (Freeman, 1997) ou LifeLines pour des données médicales personnelles (Plaisant et al. 1998).

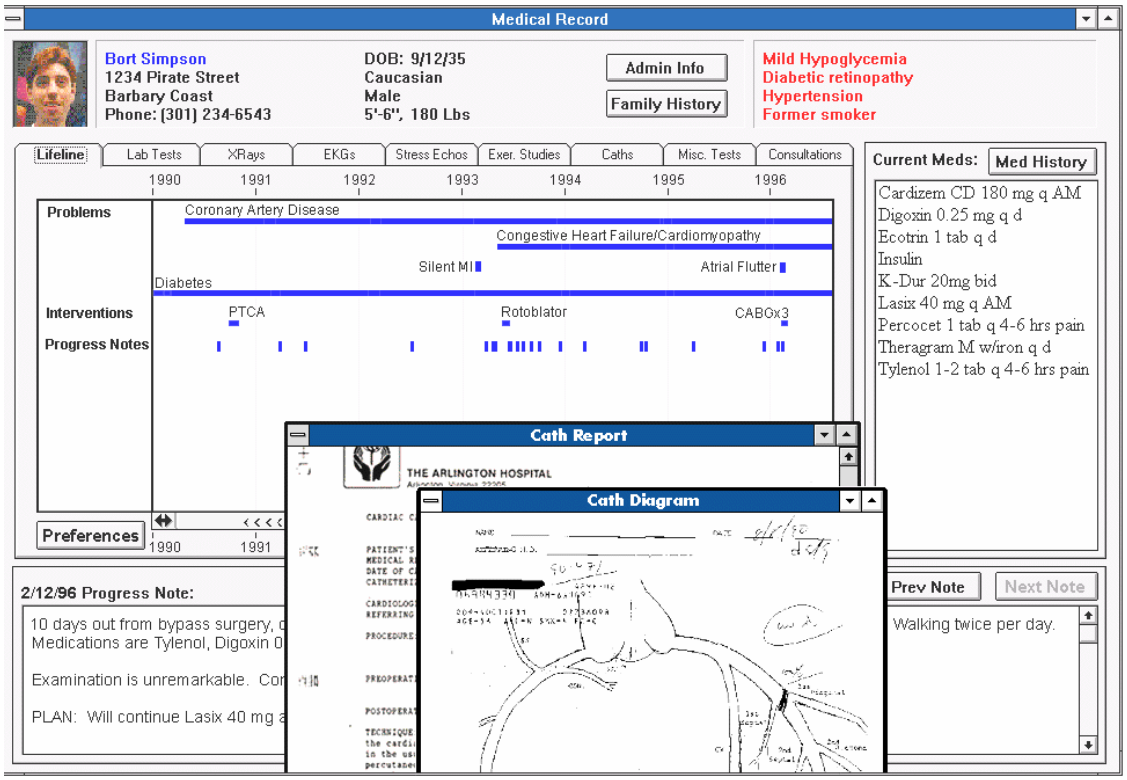


Figure 22. *LifeLines (Plaisant et al. 1998).*

La technique illustrée dans la Figure 23, développée par (Kullberg, 1995), représente le temps par une dimension dans un espace final en trois dimensions. Cette technique attribue un des axes au temps, et un autre aux photographies correspondantes.

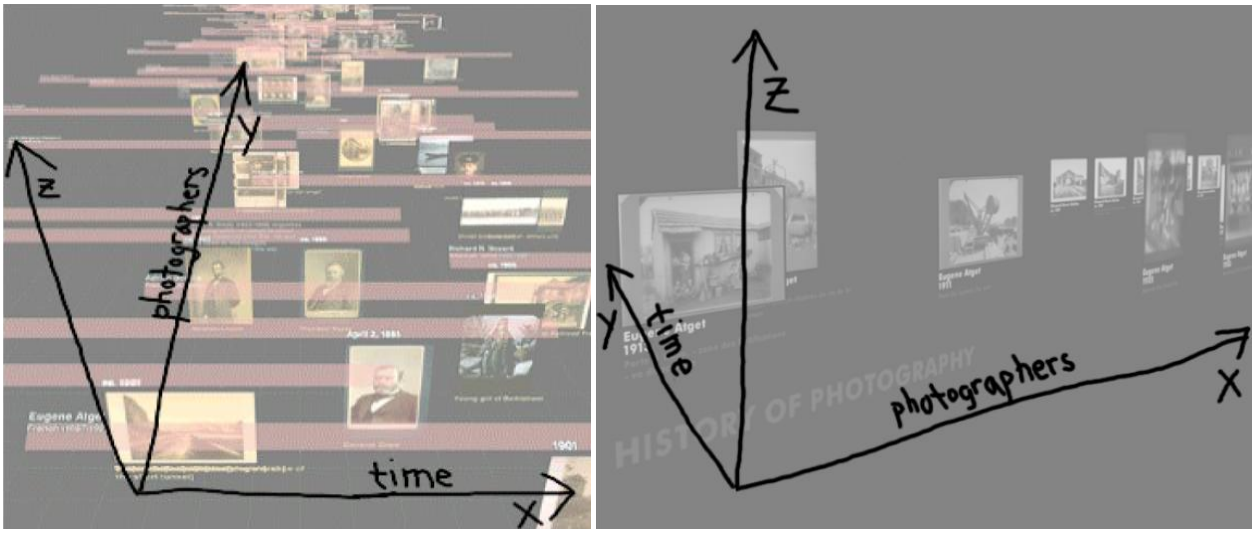


Figure 23. *Timelines dynamiques pour la visualisation d'historique de photographies (Kullberg, 1995).*

Dans la Figure 24, l'espace temps est défini par deux dimensions, l'une pour représenter les heures et l'autre pour représenter les jours, dans un espace final en trois dimensions. La représentation du temps par deux dimensions se base sur un tableau à deux entrées, contenant pour chaque croisement de l'heure et du jour, une mesure quantitative. L'ensemble de ces dernières sont traduites en objets graphiques tels que leurs hauteurs et leurs colorations sont proportionnelles aux valeurs structurelles correspondantes.

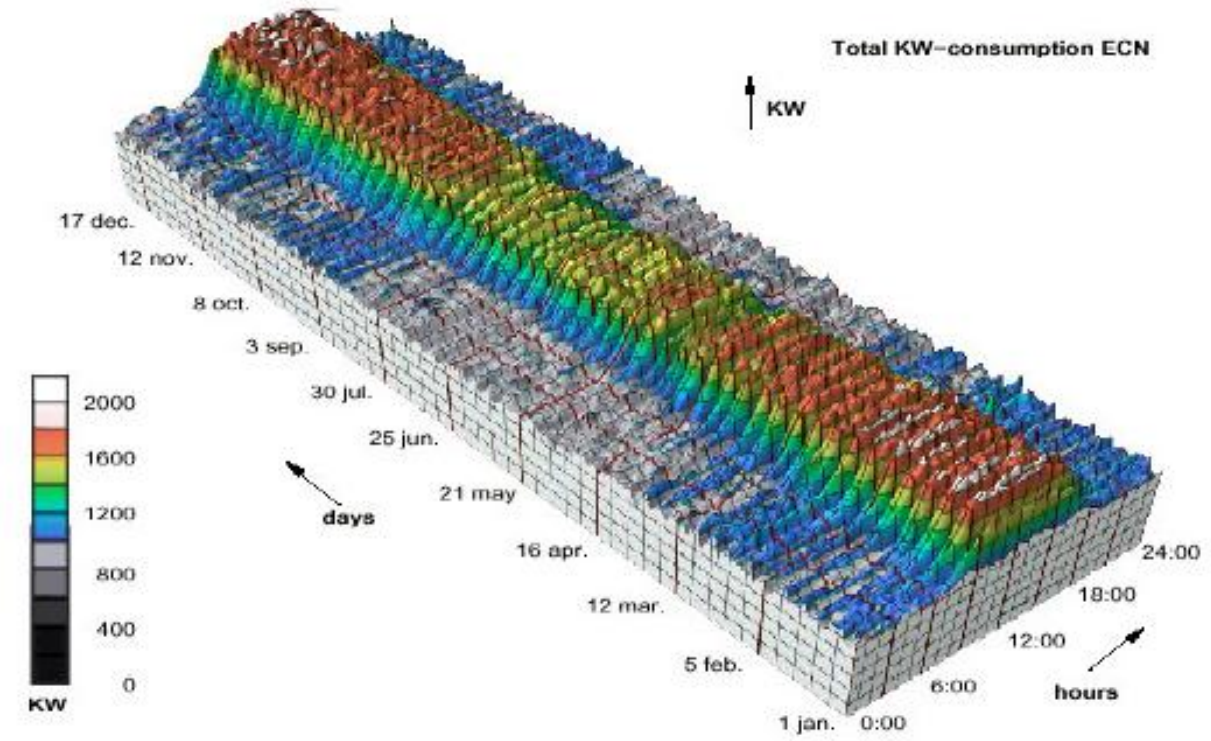


Figure 24. Deux axes pour représenter le temps (Wijk et Selow, 1999).

De nombreux outils, développé par des entreprises à des fins commerciales et non dans un but de recherche, sont disponibles sur internet. Ils se basent généralement sur le même principe de représentation des données temporelles. Ces dernières sont représentées sous forme d'une icône, et sont positionnées selon un axe horizontal. Parmi ces outils, nous pouvons citer l'outil AllOfMe[10], qui permet de créer une ligne de temps personnelle à partir de ses propres photos, vidéos, documents. Chaque icône symbolise un évènement spécifique, positionné selon l'axe horizontal du temps. Il est possible de changer la granularité du temps ; un zoom sur les semaines, les mois, les années est possible. De plus, comme le montre la Figure 25, la ligne horizontale représente le temps est comparable avec des évènements extérieurs, recensés dans une base de données.

Figure 25. Timeline conçue à partir de l'outil AllOfMe.

[10] http://www.allofme.com/

Par exemple, une étude effectuée sur l'évolution d'une entreprise commerciale peut être à l'origine d'une timeline. Cette dernière pourra être mise en parallèle d'une autre, par exemple, la timeline portant sur la conjoncture sociale et les évènements historiques.

Nous pouvons citer aussi l'outil BeeDocs'Timeline[11] qui est spécifique aux utilisateurs de Macintosh, cet outil associe un texte descriptif à chacune des images situées par rapport à l'axe horizontal du temps.

L'avantage est qu'initialement, chaque donnée temporelle est visualisée sous la forme d'une icône avec un petit libellé, comme le montre la Figure 26. Un zoom est possible en cliquant sur une icône pour obtenir un descriptif plus complet de la donnée visualisée.

Figure 26. *Timeline obtenue à partir de Bee Doc's TimeLine.*

Dans l'outil Dandelife[12], illustré en Figure 27, développé en entreprise, chaque évènement est représenté par un lien.

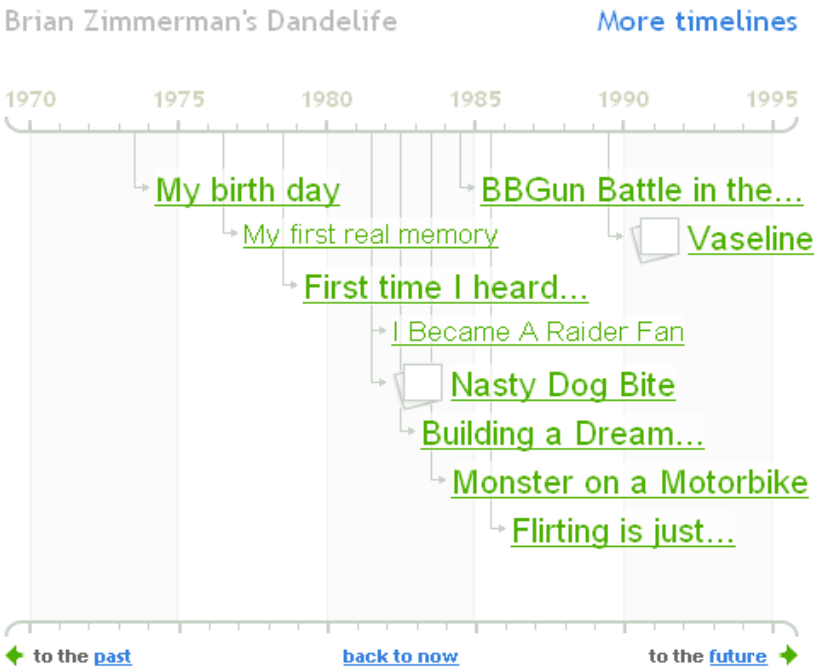


Figure 27. *Dandelife : un réseau social basé sur la biographie.*

[11] http://www.beedocuments.com/index.php
[12] http://dandelife.com/

L'apport de cet outil, comparé aux autres est la possibilité de cliquer sur un lien, pour revenir sur un contenu textuel relatif à l'évènement. Cette timeline permet de se déplacer en circulant dans le passé, en visualisant le présent ou en ciblant le futur.

Dans le même principe, l'outil Mnemograph[13] présenté en Figure 28, permet de concevoir une timeline classique. Pour chaque évènement, un double click sur l'intitulé et une fiche de renseignements complémentaires apparait pour donner davantage d'information. En parallèle, une seconde fenêtre, placée en dessous du time line, permet d'afficher des informations complémentaires, de type images, vidéos… situé par rapport à l'axe horizontal du temps. Ainsi, pour une date donnée, il est possible d'obtenir une information textuelle mais aussi plus visuelle. Cet outil offre aussi la possibilité de changer la granularité temporelle et de modifier l'échelle en semaines, mois ou années.

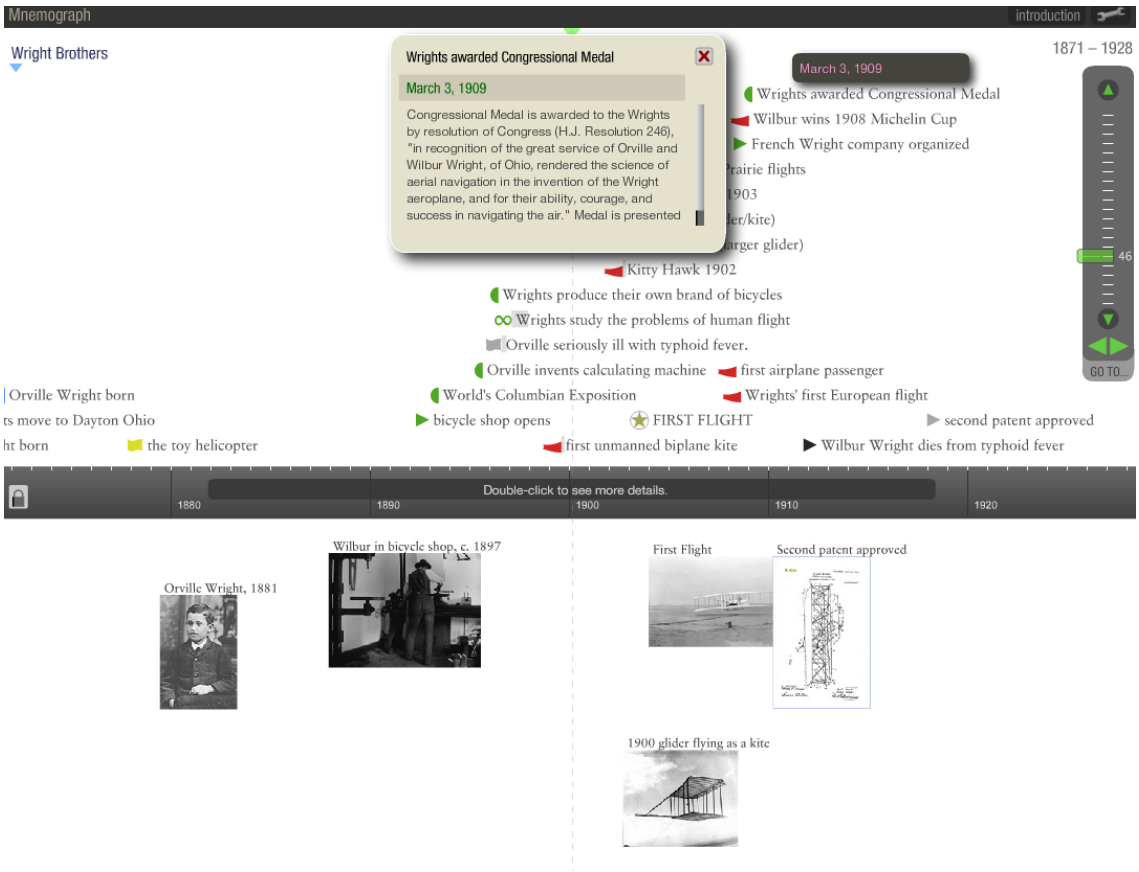


Figure 28. *Timeline obtenue avec l'outil Mnemograph.*

Sur le même principe, nous pouvons citer d'autres tels que Dipity[14], Simile[15], TimeLineCreator[16], TimeLineIndex[17], CircaVie[18], Timetoast[19], Viygo[20], Xtimeline[21], sachant que cette liste est non exhaustive.

Nous ne présentons pas en détail ces outils car ils reposent sur les mêmes principes énoncés précédemment et n'apportent pas un avantage primordial et spécifique, par rapport à ceux présentés. Ils se basent sur la représentation du temps sous forme de timeline et sur le placement d'icônes ou de libellés pour caractériser des évènements à des dates précises.

---

[13] https://mnemograph.com/login.php
[14] http://www.dipity.com/
[15] http://simile.mit.edu/timeline/
[16] http://timeline.cer.jhu.edu/support.htm#FAQ
[17] http://www.timelineindex.com/content/insertMain.php
[18] http://www.circavie.com/
[19] http://www.timetoast.com/
[20] http://www.viygo.com/
[21] http://www.xtimeline.com/

✓ ***Forme non uniforme du temps***

L'espace d'une représentation graphique étant par nature restreint et le nombre d'informations à visualiser toujours plus important, il est très difficile d'afficher toutes les informations avec un maximum de détails. Les opérations de contrôle de point de vue (comme les changements de zoom) permettent de résoudre en partie ce problème. Cependant, lors de changements de point de vue, l'utilisateur est vite déstabilisé. Cette problématique est à l'origine des techniques de visualisations non uniformes.

Les différentes formes non uniformes du temps ont pour objectif de remédier au problème du manque d'espace par rapport à la quantité d'information à cartographier. Dans un contexte temporel, le principe de cette technique est d'afficher les données temporelles du graphe avec un niveau de détails variable en fonction de l'intérêt que leur porte l'utilisateur.

Dans les zones de focalisation, les structures visuelles sont représentées de manière optimale et pour le reste de la carte, elles sont visualisées sous forme « dégradée », plus « abstraite », mais néanmoins riche en information (Leung & Apperley, 1994); (Sarkar & Brown, 1992). Le résultat de ce type de visualisation est une vue dite non uniforme, obtenue en appliquant une fonction nommée « fonction de transformation » (Leung & Apperley, 1994).

L'illustration de la figure 29 présente l'application d'une fonction de transformation à une structure visuelle de l'espace représenté.

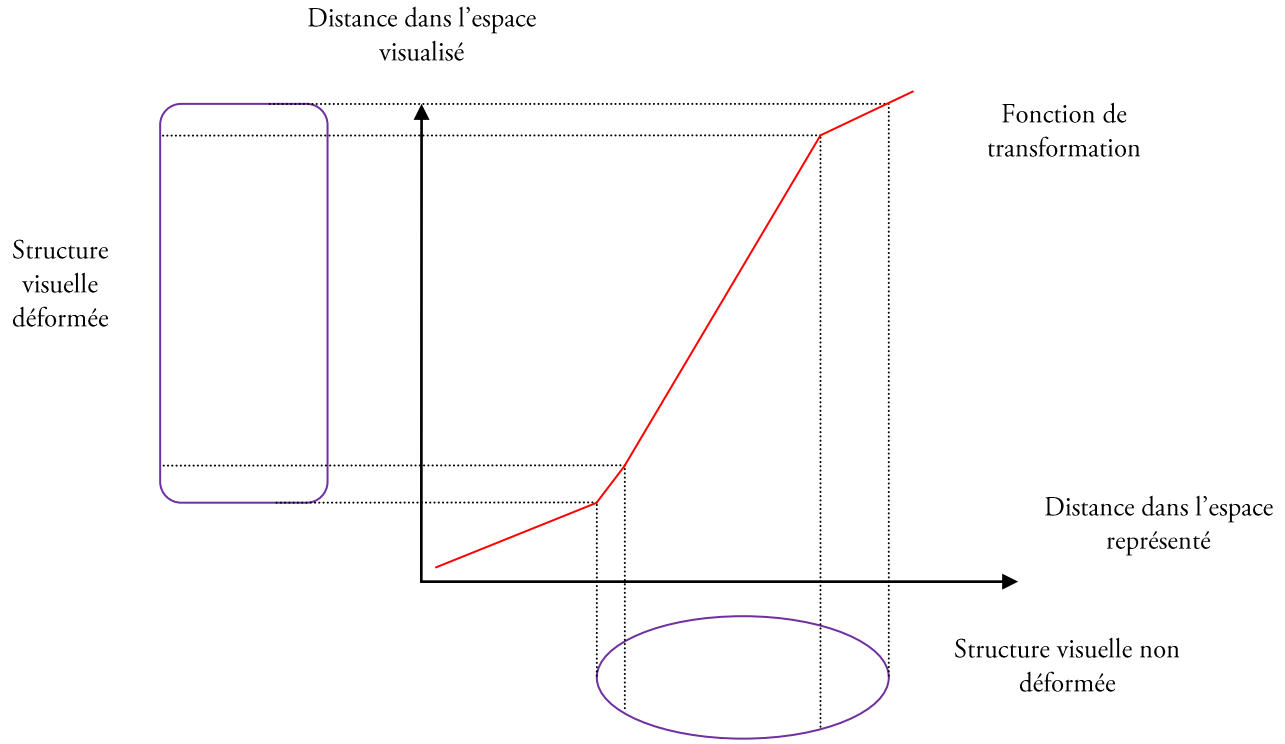


Figure 29. *Vues non uniformes : la fonction de transformation permet de déformer les structures visuelles (Leung et Apperley, 1994).*

La dérivée de la fonction de changement nommée « fonction de magnification » fournit le profil de la transformation, amplification ou réduction, et plus particulièrement, un facteur de magnification (correspondant à la « force » de la déformation).

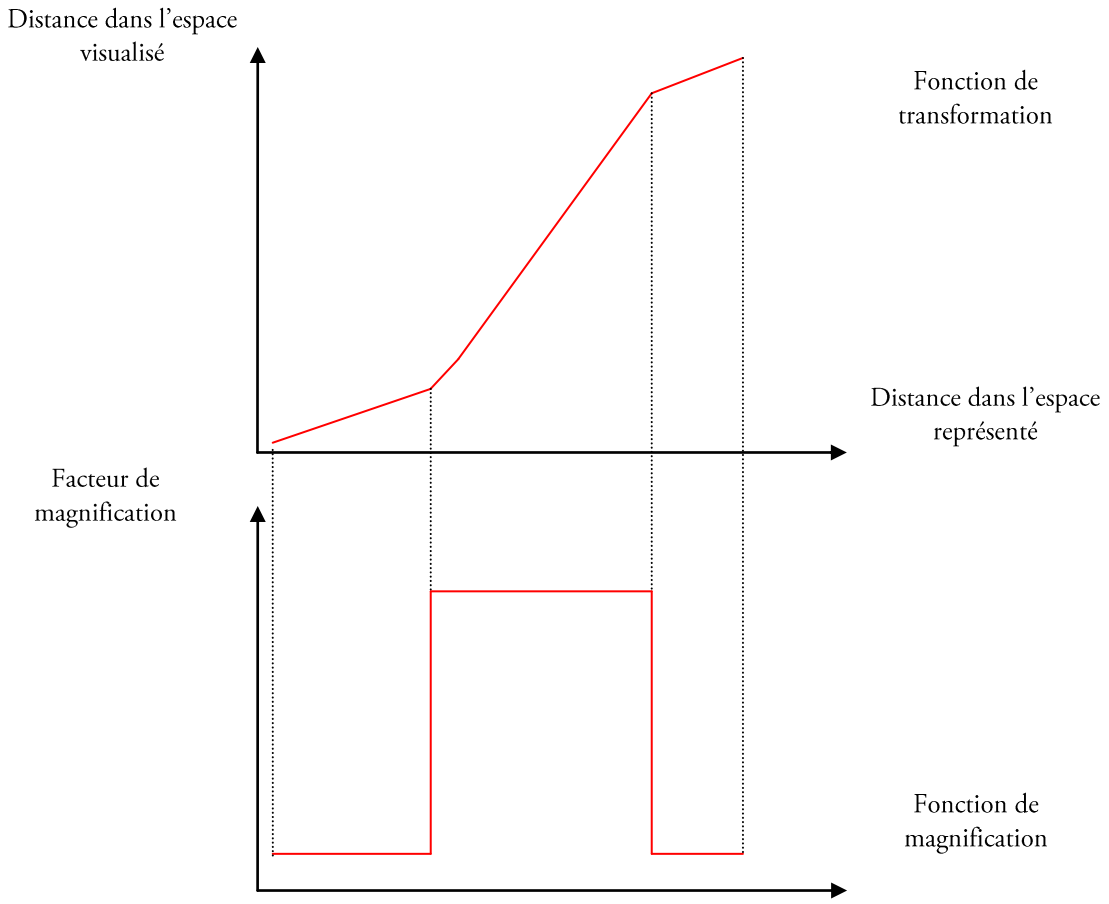


Figure 30. *Vues non uniformes : la fonction et le facteur de magnification (Leung et Apperley, 1994).*

Les murs fuyants constituent une évolution de l'affichage bifocal ; nous pouvons citer l'outil Perspective Wall (Mackinlay et al., 1991), présenté dans la Figure 31, dont le principe réside dans le fait que les données sont présentées chronologiquement sur plusieurs panneaux (trois en général). La partie centrale de la carte n'est pas transformée alors que sur les côtés, elle l'est pour donner une impression de perspective. Les deux panneaux latéraux ont une fonction de magnification « réductrice » proportionnelle à la distance au premier plan de la carte.

Dans la Figure 31, ce qui est en avant correspond à une période donnée où notre attention se focalise (en général le présent), la partie gauche à ce qui est antérieur, le passé, à cette période et la partie droite à ce qui est postérieur, le futur. Cette technique conçue pour gérer de grand espace de données, utilise une déformation de l'espace en un mur à trois faces :

- une face avant pour visualiser les données d'une manière non déformée et détaillée, représentant un écoulement du temps normal et régulier ;
- deux faces en perspectives pour visualiser une grande quantité de données d'une manière déformée et moins détaillée, représentant un avancement plus ou moins rapide dans le temps.

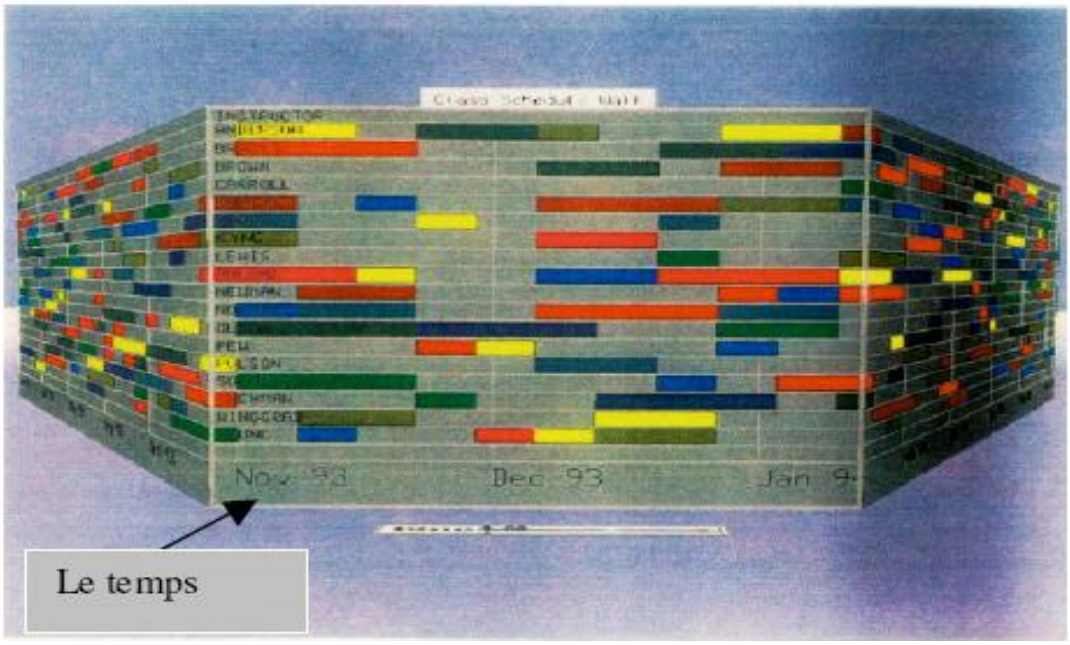


Figure 31. *Mur en perspective (Mackinlay et al., 1991).*

Dans le système *Document Lens* (Robertson et Mackinlay, 1993), illustré en Figure 32, les documents sont affichés sous la forme de petites images représentant des documents et sont disposés sur une grille dans l'ordre de la lecture. Ainsi dans la vue d'ensemble les différents documents ne sont pas lisibles mais leur aspect visuel miniature constitue un indice qui peut permettre de se déplacer directement vers le document recherché. Lorsque l'utilisateur pointe vers l'un des documents de la grille, une déformation de la grille permet de rendre ledit document lisible, tout en gardant le contexte des autres documents. Avec cette technique, le parcours de la liste de documents peut être plus rapide, car l'affichage du contexte aide à l'orientation et aux déplacements directs. La proportion du visible par rapport à l'invisible est également améliorée car un grand nombre d'informations peut être visualisé simultanément. Par contre, la surface nécessaire à l'affichage de ces présentations reste particulièrement importante.

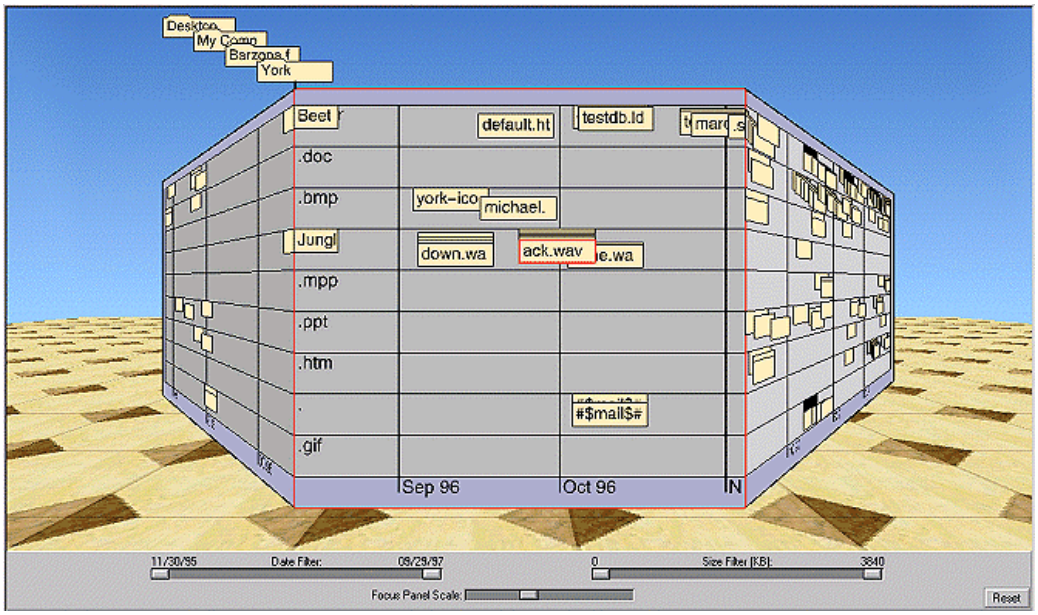


Figure 32. *Mur fuyant (Robertson et Mackinlay, 1993).*

✓ ***Temps cyclique***

Le temps cyclique se base sur le constat du temps engendrant la répétition de certains événements. La science nous montre que la nature fonctionne dans des cycles : cycles de la reproduction, cycles biologiques, cycles des climats etc. Un cycle suppose une évolution circulaire et non pas linéaire.

Le principe général d'une représentation en cycle ou en spirale est de représenter un axe d'une façon continue, qui commence à partir d'un point d'origine, autour duquel il tourne en spirale et progresse. Cet axe peut être vu comme une succession de formes circulaires. Les données sont par la suite représentées sur ces partitions circulaires.

Basé sur cette vision, des outils offrent une visualisation cyclique du temps. Le système illustré en Figure 33 (Mackinlay et al., 1994) utilise une spirale pour la visualisation d'un calendrier, en représentant chaque donnée du calendrier sous forme d'icônes, situées sur la spirale.

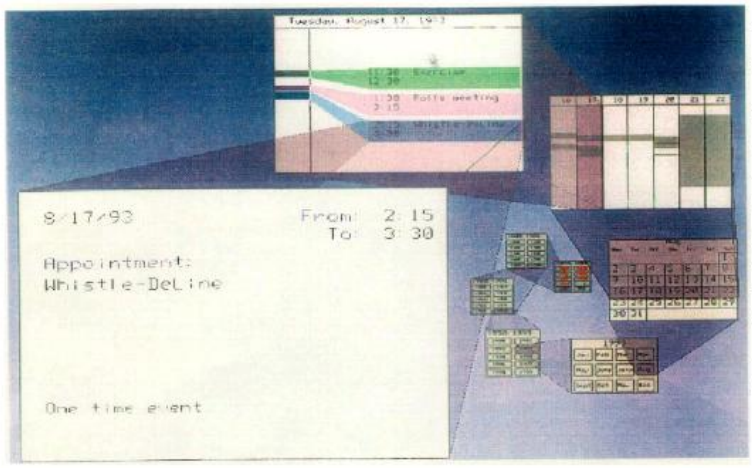


Figure 33. *Recours à une spirale pour visualiser un calendrier (Mackinlay, 1994).*

Cette technique permet de visualiser des données temporelles sous la forme d'un calendrier. Plusieurs vues des données temporelles sont disponibles, toutes connectées entre-elles selon une forme de spirale.

Dans (Tominski et al., 2003), la visualisation est basée sur des tranches en 3D, représentant des données de séries temporelles multi variées. Ces axes sont liés par un axe central, symbolisant le temps.

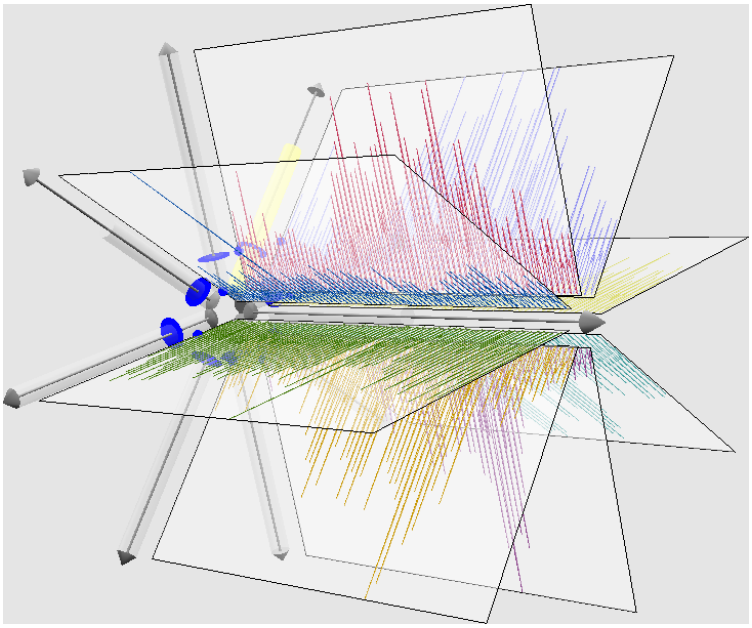


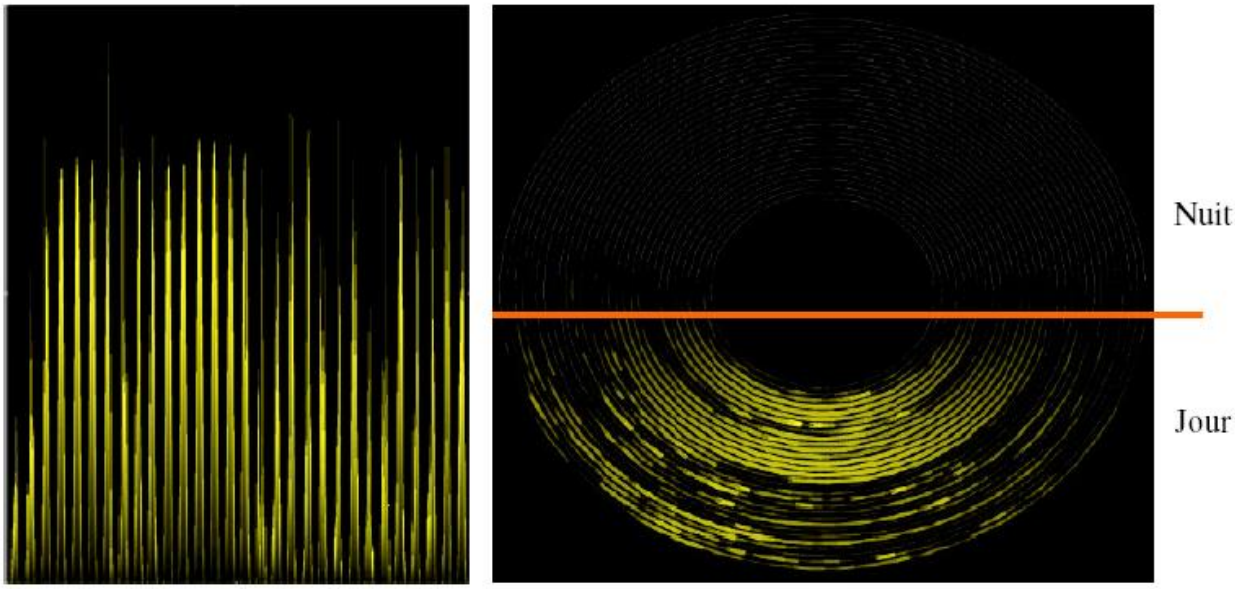
Figure 34. *Visualisation 3D représentant huit periodes différentes, basées sur un axe temps central (Tominski et al., 2003).*

Les travaux de (Hewagamage et. al., 1998) se basent sur la 2D et la 3D pour visualiser les spirales spatiotemporelles basées sur la notion d'événements.

La notion de spirale est décrite selon des coordonnées polaires, sous la forme $r = f(\varphi)$, où f est une fonction monotone. Dans les travaux de (Weber, 2001), une spirale est décrite par :

$$r = f(\varphi),\ \frac{df}{d\varphi} > 0, \varphi \in R^{+} \quad \begin{cases} r = \alpha\varphi \ tel\ que \ \ r = \sqrt{x^2 + y^2} \\ \varphi = \tan^{-1}\left(\frac{y}{x}\right) \\ x = r\cos\varphi \\ y = r sin\varphi \end{cases}$$

La représentation du temps peut aussi s'effectuer sous forme de spirale (Weber, 2001) : la visualisation concerne l'intensité solaire. La représentation en spirale, visible en Figure 35a, fait apparaître des cycles et facilite l'observation des événements. Par exemple, il est plus facile de voir les périodes de nuages dans la représentation en spirale de la Figure 35b que dans la représentation linéaire de la Figure 35a. Les périodes de nuages correspondent aux zones de couleur foncée au niveau des arcs de la spirale représentant les jours, distinguée par l'étiquette « Jour » dans la Figure 35b.



a. Représentation linéaire b. Représentation spirale

Figure 35. *Visualisation de l'intensité solaire dans le temps (Weber, 2001).*

La Figure 36 présente une visualisation des traces d'un programme dans le temps d'une spirale (Renieris et Reiss, 1999). Cet outil permet de visualiser et d'explorer l'exécution de la trace d'un programme pour aider les agents de maintenance dans la compréhension des programmes. Les données sont collectées et représentées sous la forme d'une trace, attachée sous forme linéaire et sous forme d'une spirale. La spirale fournit une image plus dense que des méthodes linéaires.

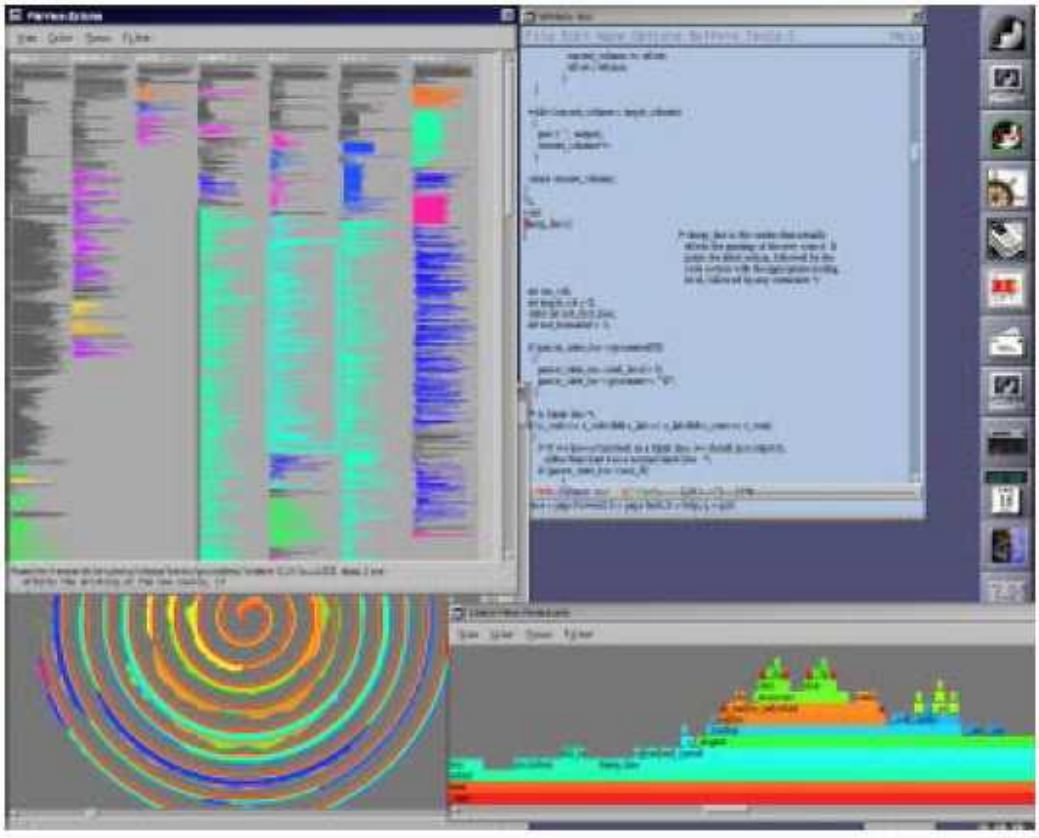


Figure 36. *Représentation d'une spirale (Renieris et Reiss, 1999).*

La technique SpiraClock (Dragicevic et al., 2002) de la Figure 37 est un exemple de cet ensemble de techniques où la représentation de la dimension temporelle définit la structure de l'espace perceptible par l'utilisateur. Cette technique utilise la métaphore d'une montre pour représenter le temps sous la forme en spirale. Elle visualise des distances temporelles à des événements au fur et à mesure que ces derniers s'approchent. Dans la Figure 37, l'horloge indique 12:11. Les différents évènements sont représentés par des couleurs différentes. L'événement en bleu commence à 12:15 et se termine à 12:22. L'évènement représenté en rouge commence à 12:55 et se termine à 13:40: chaque révolution de la spirale est une heure.

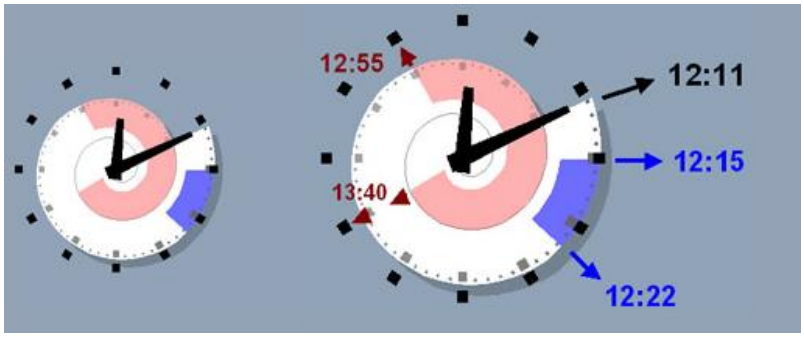


Figure 37. *Technique SpiraClock (Dragicevic et al., 2002).*

Une autre approche, proposée dans la littérature est la « *technique des cercles concentriques» (TCC)* (Daassi et al., 2000). Elle consiste en un ensemble de cercles de même centre. Chaque cercle correspond à une année dans la Figure 38 et chaque rayon représenté sur les cercles correspond à un mois dans l'exemple de la Figure 38. Les valeurs relatives à cette période sont représentées par des rectangles positionnés sur les cercles. La *TCC* a été conçue pour manipuler une ou deux données temporelles dont les valeurs structurelles sont quantitatives. Une couleur et une position sont attribuées aux rectangles correspondants aux valeurs. Une ligne temps, qui correspond aux périodes visualisées dans l'espace de référence permet de naviguer selon le temps.

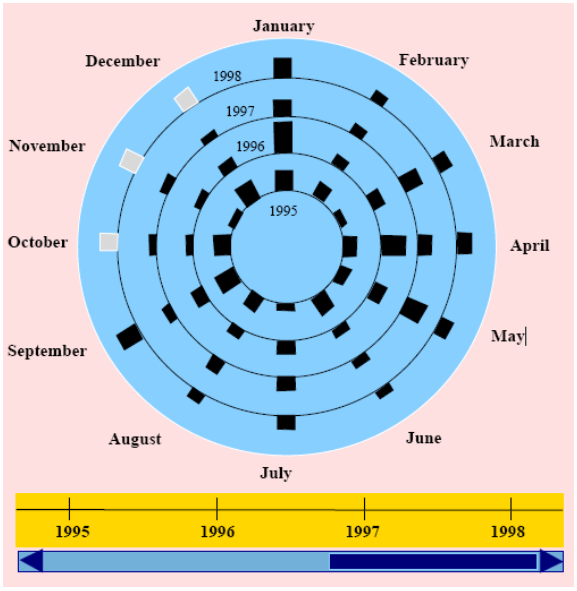


Figure 38. *Technique des cercles concentriques (CCT) (Daassi et al., 2000).*

La représentation en spirale (Carlis et Konston, 1998) de la Figure 38 met en avant le fait que les données sont périodiques. Les valeurs de l'unité grossière sont tout d'abord représentées sous la forme d'une spirale puis les valeurs de l'unité fine sont représentées par des rayons au-dessus de la spirale.

Chaque partition circulaire de la spirale correspond à une année, unité grossière, et chaque rayon correspond à un *mois* de l'année, unité fine. Les valeurs des données observées à l'unité d'observation *mois* sont ensuite représentées par des éléments graphiques et sont placées directement à l'intersection de l'axe temps (la spirale) et du rayon correspondant. Nous soulignons l'existence d'une indépendance visuelle entre la ligne temps et les objets graphiques comme le disque dans la Figure 39, traduisant les valeurs structurelles. En effet, le processus de visualisation de la dimension structurelle peut être remplacé par un autre processus tout en conservant celui de la dimension temporelle. Ainsi, d'autres transformations visuelles peuvent être utilisées pour traduire les valeurs des données en signes graphiques.

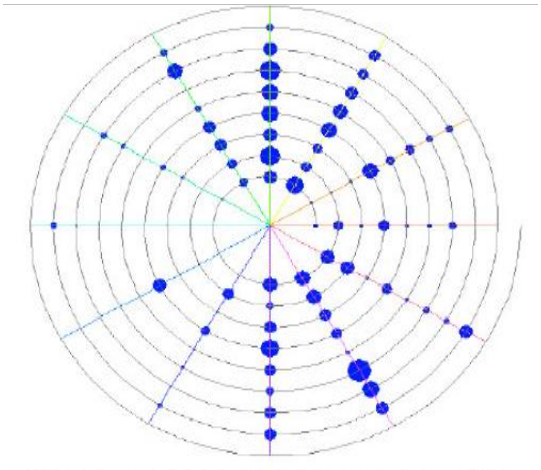


Figure 39. *Spiral indentée, utilisant l'axe temps comme support pour représenter les données temporelles (Carlis et Konston, 1998).*

Un autre outil, nommé Spiral Graph (weber et al., 2001) permet de visualiser, de comparer et d'analyser des données cycliques, comme le montre la Figure 40. L'axe du temps *Y* est représenté par une spirale. Les données de départ sont placées au centre de la spirale tandis que les données de fin d'observation sont placées sur la partie externe. La période est donnée par un tour complet de la spirale. La valeur d'observation peut être donnée par une modification de la couleur ou de l'épaisseur du trait. En général, cette méthode s'applique pour une seule variable par axe de temps. Dans la Figure 40, nous étudions la présence des employés d'une entreprise durant sept jours consécutifs.

Nous avons le 1er janvier comme début de l'observation au centre de la spirale et le 23 décembre comme fin de l'observation à la périphérie de la spirale. Un tour complet de spirale représente une semaine ; le lundi se trouvant en bas à droite et le vendredi en au haut à gauche. Les parties plus foncées correspondent à la journée du lundi au vendredi. L'alternance entre une partie foncée et une partie claire provient du contraste entre la nuit et le jour. En journée, nombreux sont les employés présents dans les locaux et inversement la nuit et le week-end.

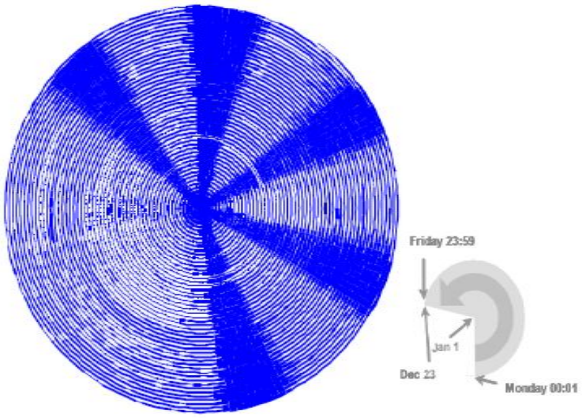


Figure 40. *Représentation cyclique des données avec l'outil Spiral Graph. (weber et al., 2001).*

✓ ***Discussion***

Les trois formes de représentation du temps, linéaire, non uniforme ou encore cyclique, étudiées précédemment facilitent grandement l'étude de données temporelles. Nous comparons les trois formes de visualisation proposées précédemment, ainsi que les outils cités. Nous sélectionnons plusieurs critères, cités dans ce chapitre pour qualifier les outils présentés précédemment. Il en résulte le Tableau 13.

Critères retenus	Caractéristique du critère
Qualité de la représentation d'un point de vue général, au niveau de l'efficience, de la clarté et de la précision.	Qualité-représentation
Recours à une sémiologie particulière pour la représentation de la donnée : couleur, forme, grain… facilitant l'interprétation du graphe et les différences entre données temporelles (comme l'appartenance à des catégories différentes).	Sémiologie-donnée
Facilité de comparaison entre les données temporelles au niveau quantitatif.	Comparaison-quantitative
Minimisation de la surface de représentation des données temporelles.	Minimisation-surface
Etiquetage clair et complet des données.	Etiquetage
Accessibilité à l'information complète : retour aux sources ou obtention de complément d'informations sur la donnée.	Complément-info
Visualisation de grands volumes de données.	Volume-données
Changement d'échelle de la notion de temps. Possibilité de visualiser les données de manière globale ou plus précise : granularité temporelle.	Navigabilité-temps
Représentation de la donnée temporelle sous une seule forme.	Représentation-unique
Détection des tendances ou des comportements périodiques dans un ensemble de données temporelles.	Détection-tendance

Tableau 13. *Critères choisis pour comparer les outils de visualisation de données temporelles, basés sur les trois formes du temps étudiées.*

Basée sur les préconisation issues de la littérature (weber et al., 2001), pour visualiser des données temporelles et permettre de les analyser et de les comparer, la représentation de l'ensemble de l'information doit se baser sur une seule technique de visualisation appropriée pour les données nominales, ordinales, et quantitatives.

De plus, la visualisation doit utiliser des icônes révélant l'évolution de la donnée, rendre accessible l'information complète, visualiser de grands volumes de données. La représentation a pour but de permettre la lecture comparative des données temporelles sous une vision globale, toutes périodes confondues, mais également permettre la comparaison de plusieurs cycles individuels dans un ensemble de données. Enfin, elle doit permettre la détection de comportements périodiques et les différentes tendances dans un ensemble de données temporelles.

A partir de ces critères, nous élaborons la matrice de présence/absence suivante. La présence d'un 1 signifie que l'outil répond spécifiquement au critère correspondant.

	Outils \ Critères	Qualité-représentation	Sémiologie-donnée	Comparaison-quantitative	Minimisation-surface	Etiquetage	Complément-info	Volume-données	Navigabilité-temps	Représentation-unique	Détection-tendance
Temps linéaire	Timeline	1		1	1			1		1	1
	Starfield		1					1		1	1
	Frame-based animation							1		1	1
	ThemeRiver	1	1	1				1	1	1	1
	TimeMap viewer		1	1			1		1	1	
	OITL	1	1		1		1			1	1
	LifeStreams	1		1	1	1	1		1	1	
	LifeLines	1		1	1	1	1		1	1	
	TimeLine dynamique									1	
	Wijk et selow,1999	1	1	1	1					1	1
	AllofMe	1				1			1	1	1
	Bee Docs'Timeline	1				1	1			1	
	Dandelife	1				1	1		1	1	
	Mnemograph	1	1			1	1		1	1	
	Dipity	1				1	1		1	1	
	Simile	1				1			1	1	
	TimeLine Creator	1	1			1	1		1		
	TimeLineIndex	1				1	1			1	
	CircaVie	1				1	1		1	1	
	Timetoast Timelines	1			1	1	1	1		1	
	Viygo	1			1	1	1		1	1	
Temps non uniforme	xtimeline	1				1	1		1	1	
	Perspective Wall		1		1			1		1	1
	Document Lens				1	1				1	1
Temps cyclique	Spiral Calendar Visualizer	1	1			1			1		
	(Tominski et al., 2003)	1	1	1	1			1		1	1
	Hewagamage	1	1	1	1	1		1		1	
	(Weber, 2001)		1	1	1			1		1	1
	Renieris et Reiss	1	1	1	1		1	1		1	1
	SpiraClock	1	1		1				1	1	1
	TCC	1	1	1	1	1			1	1	1
	(Carlis et Konston, 1998)	1	1	1	1					1	1
	SpiralGraph	1	1		1			1		1	1

Tableau 14. *Matrice de présence/ absence des critères de qualité des techniques de visualisation de données temporelles.*

A partir de cette matrice asymétrique, le graphe correspondant est tracé, Figure 41. Les critères sont représentés par des cercles de couleur claire et les différents outils sont visualisés en utilisant une couleur plus foncée. La présence d'un lien entre une caractéristique montre que l'outil possède la caractéristique.

Les outils situés à proximité d'une caractéristique sont souvent liés à cette dernière, révélant ainsi leurs propriétés spécifiques. A partir de ce graphe, il est possible d'extraire des widgets[22], visualisant un sommet ainsi que ses voisins directs. Dans notre perspective de caractérisation, un widget est conçu par critère. Les widgets des dix critères sont visualisés dans les figures 41et 42.

Cette fonctionnalité permet ainsi de résumer visuellement l'information, c'est-à-dire à partir d'un critère spécifique, nous pouvons déterminer quels outils répondent à cette propriété.

Les numéros des widgets correspondent aux critères suivant :

N° Widget	Critères
1	Minimisation_surface
2	Détection_tendance
3	Représentation_unique
4	Volume_données
5	Sémiologie_donnée
6	Navigabilité_temps
7	Complément_info
8	Etiquetage
9	Comparaison_quatitative
10	Qualité_représentation

Tableau 15. *Correspondance entre widgets et critères.*

A partir du graphe des critères et des outils, nous traçons l'Analyse Factorielle des Correspondances (AFC), Figure 43. Les types de représentations du temps se distingue par trois couleurs différentes sur l'AFC obtenue. L'analyse du graphe se base sur la notion de proximité : plus un outil est proche d'un critère, plus il est considéré comme ayant cette propriété.

[22] Le Widget est un module interactif qui s'intègre sur le poste utilisateur. Petite application spécialisée pour une tâche et permettant d'accéder rapidement à des informations (visualisation ciblée) ou des fonctions utilisées fréquemment (transitivité) dans une fenêtre de petite taille.

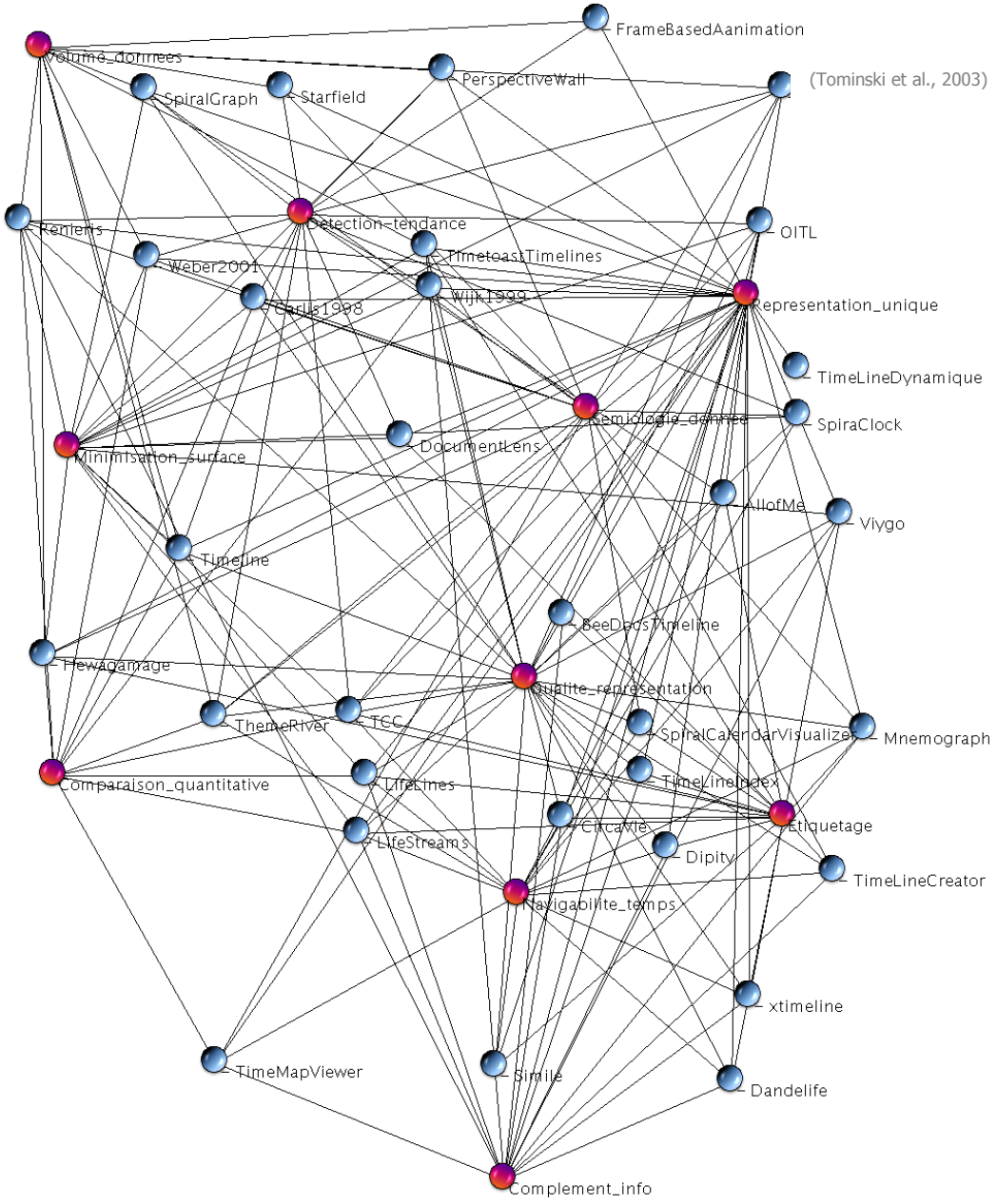


Figure 41. *Graphe des critères et des outils de visualisation de données temporelles, effectué sous VisuGraph (Loubier et Dousset, 2008)*

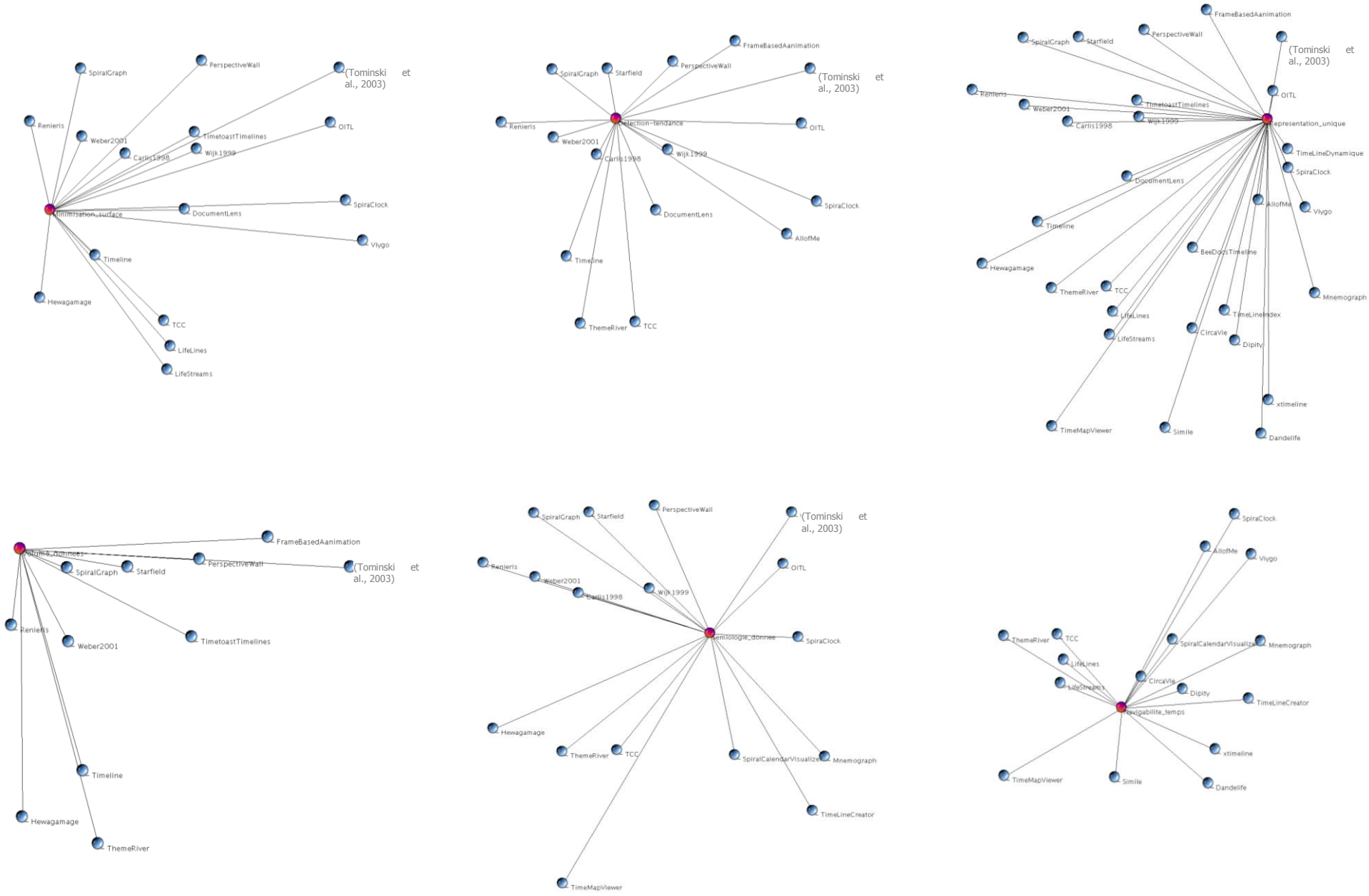


Figure 42. *Widgets des dix critères de comparaison, obtenus sous VisuGraph (Loubier et Dousset, 2008).*

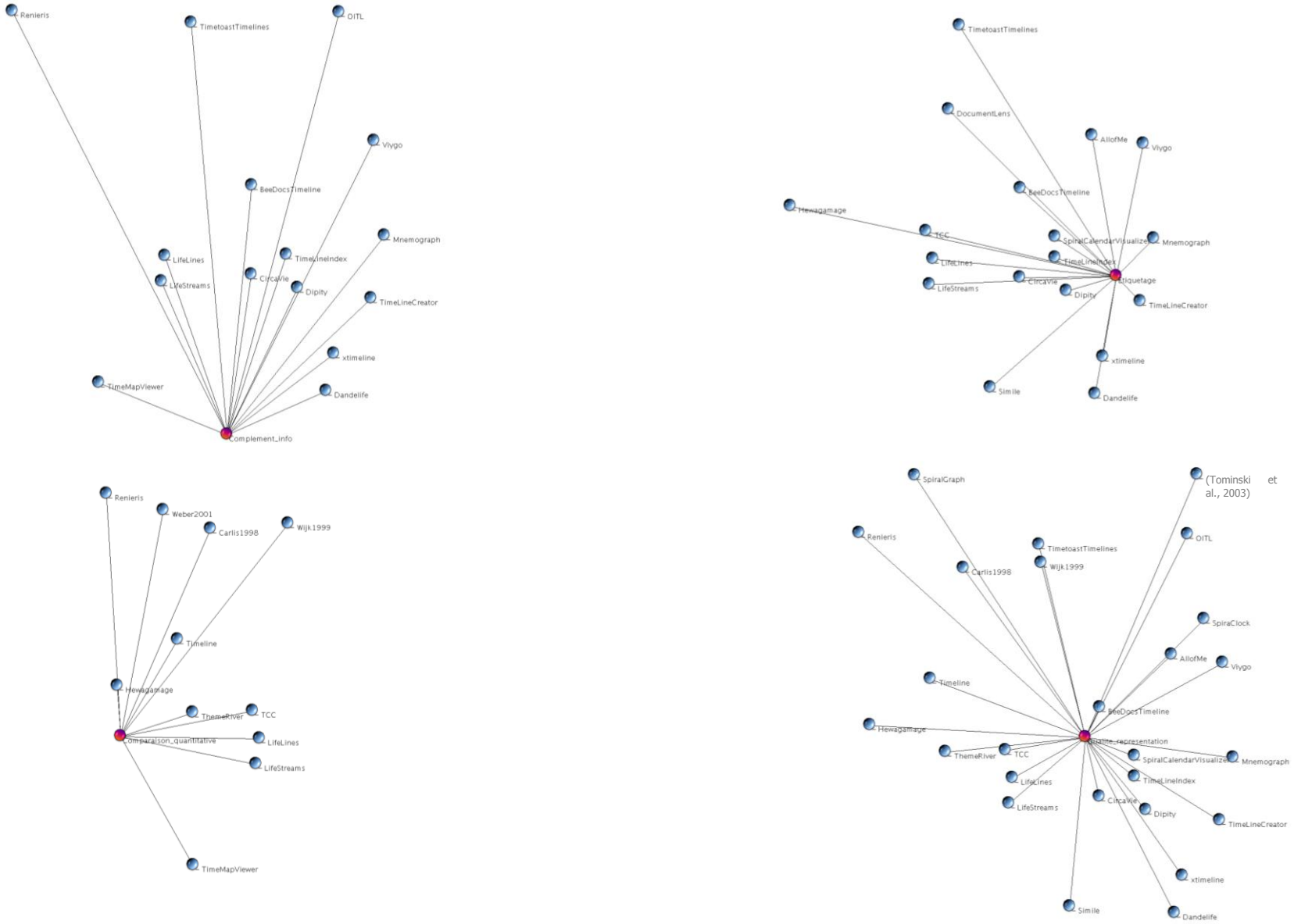


Figure 43. *Widgets des dix critères de comparaison, obtenus sous VisuGraph (Loubier et Dousset, 2008).*

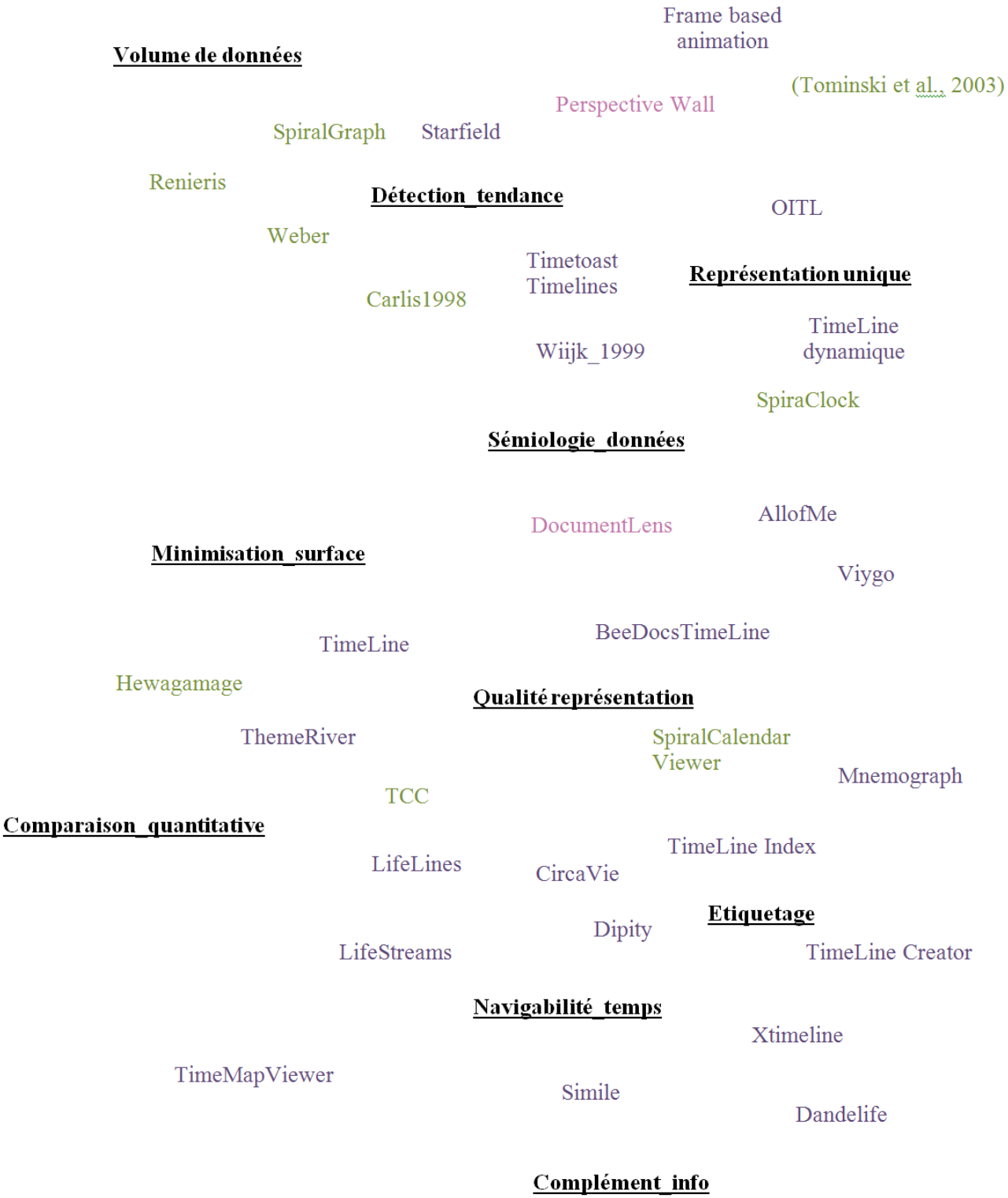


Figure 44. *AFC des critères et des outils de visualisation de données temporelles.Les critères sont en gras et soulignés.*

Pour comparer les différents outils mais surtout les techniques de représentation des données temporelle, citées précédemment, nous analysons les résultats graphiques obtenus dans la Figure 44. Nous distinguons ainsi trois catégories, les outils se basant sur la représentation linéaire que nous regroupons sous le titre « *Les timelines* », ceux sur « *la visualisation non uniforme*» et ceux sur la représentation du « *temps circulaire* ».

Les timelines

L'avantage de la représentation linéaire du temps provient de la nature intuitive de ce mode de représentation, de la contrôlabilité de la granularité de l'information représentée ainsi que la possibilité de visualiser explicitement des événements qui se déroulent en parallèle. Dans le graphe précédent, les outils se basant sur des timelines sont représentés en violet.

Nous pouvons constater que les timelines sont en général très clair, leur manipulation est simple et l'effort intellectuel fourni par l'utilisateur pour décoder l'information est relativement faible. De plus, dans les outils analysés, les données temporelles sont souvent étiquetées clairement, permettant ainsi de mieux les distinguer et il est possible d'accéder à un complément d'information, que ce soit par un retour aux sources ou par un texte informatif. De manière générale, ces outils permettent de zoomer sur des périodes spécifiques ; ainsi il est possible d'obtenir un graphe sur l'ensemble de la période, mais aussi d'obtenir une granularité temporelle de l'ordre de l'année, du mois, de la semaine, du jour et même parfois de l'heure.

Les inconvénients de ce procédé sont le fait que la donnée est souvent représentée sous forme textuelle, sous un intitulé linéaire, ou par une image. Il est alors difficile de comparer des valeurs quantitatives, ce qui réduit la capacité de ces outils à détecter les tendances majeurs ainsi que les signaux faibles. Si les données temporelles sont catégorisées, il est difficile de les distinguer par l'absence de sémiologie dans la représentation des données. Le problème de la ligne de temps est principalement du aussi à la non limite de la surface de représentation. Le présent est mis en avant, ainsi que le passé et le futur proche. Pour des périodes éloignées, l'étude précise, de manière linéaire, est plus délicate. Ainsi, la préconisation graphique portant sur la minimisation de la surface de représentation n'est pas la plus respectée pour les timelines. Le volume de données est aussi limité, puisque pour chaque granularité du temps en abscisse, doivent correspondre toutes les données temporelles adéquates. Ainsi, plus le nombre de ces dernières à une date t est important, plus le graphe est surchargé et plus vite il devient illisible.

La visualisation non uniforme

Cette visualisation des données temporelles permet de visualiser une grande quantité d'informations. Elle minimise la surface de représentation, permettant ainsi à l'utilisateur de cibler directement les données. Au niveau de la détection des données émergentes, des signaux faibles ou alors de la perception des différentes tendances, l'accès et la comparaison de données temporelles sur de longues périodes est possible. Cette visualisation correspond à la distorsion interactive de (Keim, 2002), (Keim et al., 2005) vue précédemment, où la zone temporelle d'intérêt est rendue visible suite à une déformation locale comme si on déplaçait une grosse loupe sur cette dernière. Les autres zones sont toujours visibles mais avec un moindre niveau de détail. Au niveau de la sémiologie, une seule représentation des données est disponibles ce qui facilite l'exploration visuelle. De manière générale, il s'agit de rectangles utilisant un code couleur pour les qualifier ou alors contenant le libellé de la donnée. De manière générale, la dimension de la forme de l'icône traduit la persistance des données ; plus un rectangle est étendu, plus la donnée est persistante et inversement. L'observation des données est focalisée sur le présent. Le passé et le futur sont présentés de manière moins détaillée. De plus, il y a un suivi temporel des événements et on peut avoir une meilleure résolution sur les périodes et les événements d'intérêt tout en gardant une vue d'ensemble.

L'inconvénient de cette représentation des données temporelles est l'impossibilité d'étudier l'aspect relationnel entre les données, à savoir leurs liaisons. Il est donc difficile d'étudier les différentes collaborations, les alliances, les liens entre les données.

De plus, les données sont comparables d'un point de vue temporel, il est possible de déterminer si une donnée est plus persistante qu'une autre, mais il est difficile de les comparer sur d'autres critères, tels que le voisinage du sommet à une date précise.

D'autre part, l'échelle du temps étant fixe, la navigabilité dans le temps est réduite car le changement de granularité est difficile. En effet, si l'échelle temporelle se base sur une granularité du temps telle que l'année, il est difficile de visualiser les données par semaines, jours, heures… Le focus étant sur le présent, une étude minutieuse du passé sur un même graphe est laborieuse. Il faut savoir localiser la zone temporelle intéressante et situer un événement particulier ou une donnée particulière dans le temps.

Le temps circulaire

Le temps circulaire possède une propriété de représentation naturelle et intuitive, l'homme a souvent tendance à assimiler le temps à une horloge. Ainsi cette représentation ne demande généralement pas de grands efforts pour comprendre et analyser un graphe représenté sous cette forme; une des préconisations de cette représentation étant le sens des aiguilles d'une montre pour la chronologie des évènements. La qualité de représentation est une caractéristique forte de ce type d'outils, qui utilise très souvent une sémiologie permettant de mieux cibler les données. De manière générale, la forme est utilisée pour traduire une valorisation numérique. Que ce soit sous la forme unitaire de cercle ou de rectangle, plus la valeur quantitative de la donnée temporelle est importante, plus la forme de représentation l'est en conséquence. Ainsi, il est possible de comparer les données une à une. De plus, le temps circulaire permet la minimisation de la surface de représentation, que ce soit sous la forme d'un cercle ou d'une spirale. Les données sont placées chronologiquement, permettant ainsi de voir leur évolution au cours du temps et de détecter les tendances, à savoir, quels sont les éléments majeurs dominant des données moins importantes.

Les inconvénients de cette représentation sont le manque d'informations visuelles sur la donnée. En effet, le volume de données visualisables reste limité et l'étiquetage est souvent très restreint, voire absent, à cause de la minimisation de l'espace de visualisation, pour ne pas surcharger le graphe. Enfin, il est difficilement possible de changer l'échelle du temps. Certains outils proposent plusieurs granularités temporelles au sein d'un même graphe, par exemple TCC (Daassi, 2003), vu précédemment, mais aucun changement d'échelle interactif n'est possible.

Ainsi, nous constatons que ces trois formes de représentation du temps proposent de nombreux avantages mais aussi des inconvénients. Notre objectif est de réaliser un outil combinant ces méthodes de visualisation, afin de n'en extraire que les parties avantageuses et qui permettent de répondre positivement à l'ensemble des critères de qualité que nous avons sélectionnés pour comparer ces outils. Cependant, la conception d'un bon outil de visualisation de données relationnelles et temporelles ne dépend pas seulement de la représentation de ces dernières mais aussi de celle de l'espace temps, à savoir le nombre de dimensions utilisées.

2.4.3. Représentation de l'espace temps et des données temporelles

L'ensemble des données temporelles étudiées peuvent être représentées sur un seul et même espace, afin de faciliter leurs comparaisons et leurs analyses. Une grande majorité des outils de visualisation de données temporelles ont recours à ce procédé. L'utilisation de deux dimensions est caractérisée par un premier axe représentant le point de vue temporel et le deuxième axe est utilisé pour combiner l'aspect temporel à un autre point de vue. Dans l'espace à trois dimensions, la première est mobilisée pour représenter le temps et les deux restantes de l'espace de construction sont utilisées pour la combinaison avec plusieurs autres points de vue.

Dans cette partie, nous ne nous intéressons pas à la forme de représentation des données temporelles, vues dans le paragraphe précédent, mais plutôt aux espaces de visualisation dans lesquels les données sont représentées.

L'application TimeSearcher (Hochheiser, 2002a), (Hochheiser, 2002b) illustré dans la Figure 45 est un exemple où toutes les dimensions structurelles des données sont visualisées dans un même espace graphique et par rapport à un seul espace temps. Par exemple, dans cette figure, les courbes sont tracées dans le même repère.

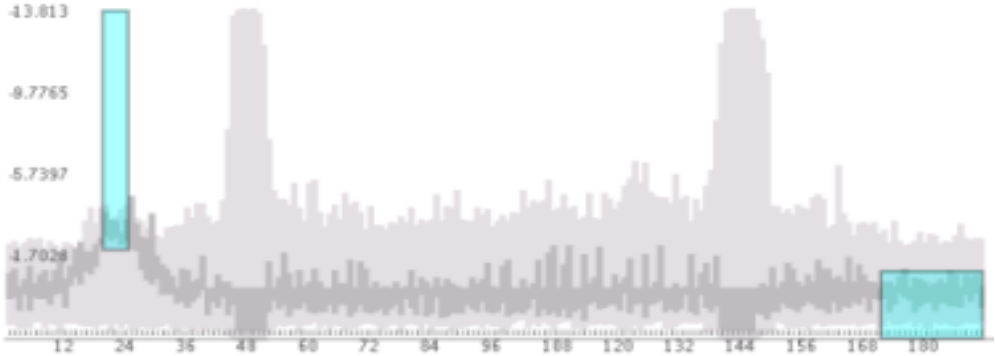


Figure 45. *Timesearcher (Hochheiser, 2002a),(Hochheiser, 2002b).*

L'outil CiteSpace (Chen, 2004), visualise les réseaux de co-citation les plus importants, à partir d'articles publiés dans le domaine considéré. Un raccord des différentes périodes considérées, par tranche de temps, des réseaux de Co-citation permet une visualisation panoramique. Le dessin de graphe inclut la prise en compte du placement stratégique des sommets du graphe, tels qu'ils ont été étudiés par (Fruchterman et Reingold, 1991), puis par (Tamassia et al., 1988). Dans la Figure 46, cela se traduit par la distinction de branches encerclées par un trait épais pour la période 1993-1995 et une mise en évidence de la période 1999-2000, encerclée par un trait fin.

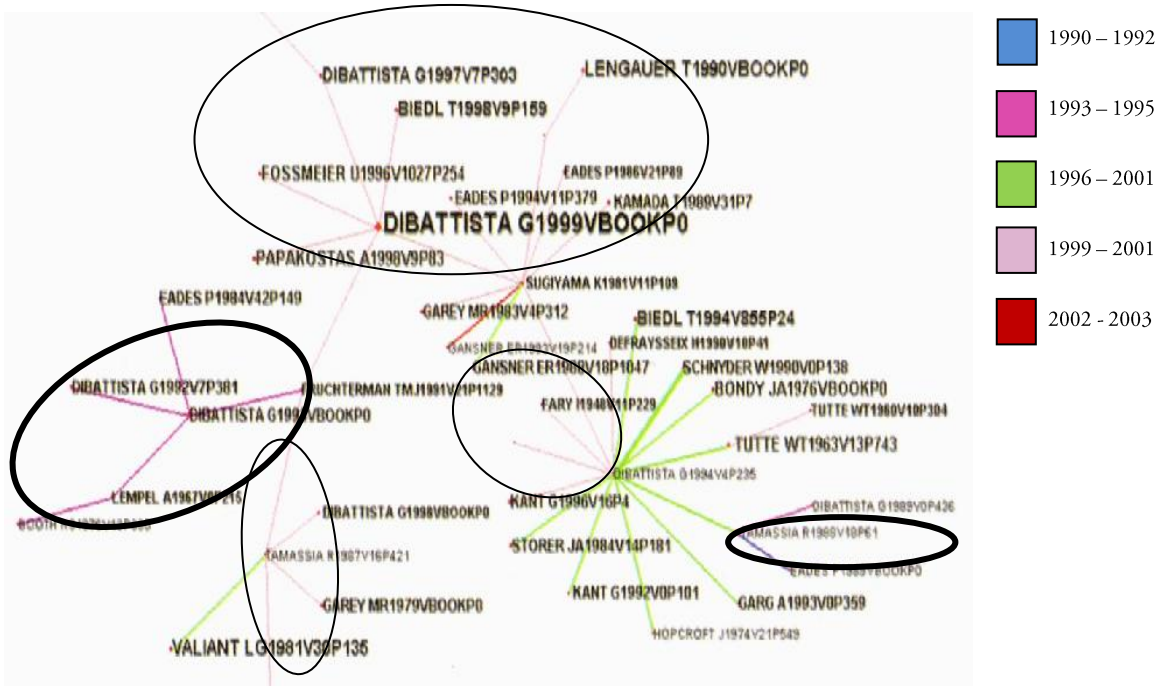


Figure 46. *Graphe des co-citations d'auteurs scientifiques (Chen, 2004).*

Dans la Figure 47, l'outil Tulip (Auber, 2001) utilise une représentation cyclique de l'axe temps ainsi que la couleur, pour représenter plusieurs dimensions structurelles. L'utilisation d'un seul espace temps pour représenter plusieurs dimensions structurelles est donc indépendante des perceptions linéaires, cycliques, logarithmiques du temps.

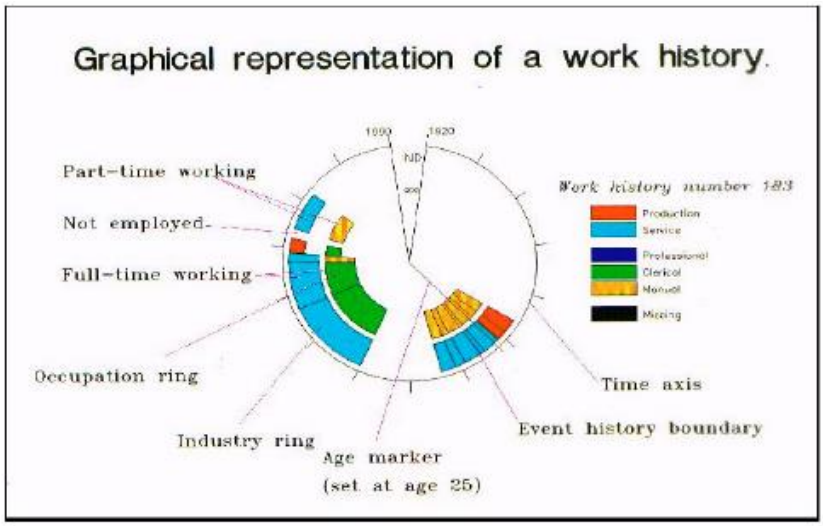


Figure 47. *Forme cyclique du temps, représentant plusieurs dimensions structurelles par rapport à un seul espace temps.*

Aussi, indépendamment du nombre de dimensions de l'espace final de représentation, une seule dimension temps peut être utilisée pour représenter plusieurs dimensions structurelles. Dans les quatre exemples précédents, le temps est représenté dans un espace en deux dimensions.

La technique Lexis Pencils[23] présentée en Figure 48 utilise la forme d'un stylo pour représenter le temps dans un espace en trois dimensions. L'axe temps est représenté par l'axe défini par le stylo même. Ce dernier est représenté par différentes facettes et chacune est utilisée pour représenter une dimension structurelle. Les valeurs sont représentées par des variables rétiniennes comme la couleur et la texture. L'exemple présenté dans la Figure 48 est une visualisation de l'historique de la vie d'un couple. Le temps varie de la gauche vers la droite commençant à la date du mariage. La face en haut du stylo représente l'historique du travail de la femme, celle au milieu représente l'historique du travail de l'homme, et la face en bas représente l'âge des enfants. Les valeurs des données sont traduites en couleurs.

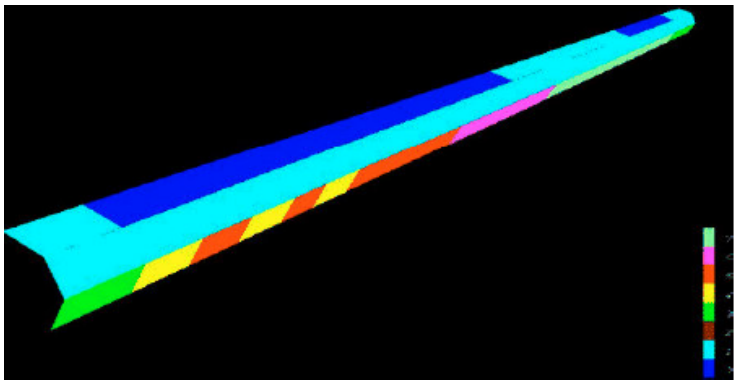


Figure 48. *Technique Lexil Pencils (Brian, 2003).*

La visualisation des dimensions structurelles sur les facettes externes du stylo avec la technique Lexis Pencils nécessite que l'utilisateur tourne le stylo dans l'espace de représentation pour observer une donnée particulière. Il est ainsi difficile de comparer, par exemple, les données entre elles. Pour la facette du bas, chaque intensité de couleur correspond à l'absence, puis la présence d'enfant selon leur tranche d'âge ...

La technique Data Tube (Ankerst, 2000) de la Figure 49 visualise les données au niveau des facettes internes d'un tube. Ce dernier représente l'axe temps. Cette technique permet ainsi d'observer l'évolution de toutes les données en même temps. Cette figure présente une visualisation de 50 données de Janvier 1974 à Avril 1995.

Avec la technique Data Tube, chaque valeur est représentée par un segment dont la couleur est définie dans un niveau de gris proportionnellement à la valeur.

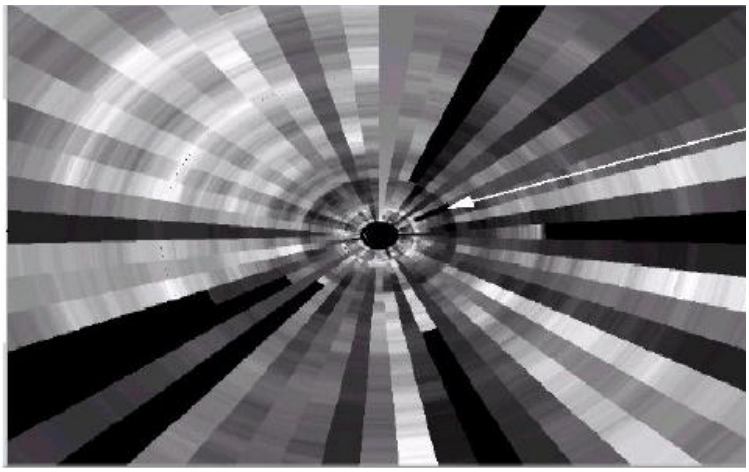


Figure 49. *Data Tube permettant de visualiser un grand nombre de données. (Ankerst, 2000), (Ankerst, 2001).*

[23] http://www.agocg.ac.uk/reports/visual/casestud/francis/compar_1.htm

D'autres techniques de visualisation n'utilisent pas de représentation explicite d'un espace temps. L'approche utilisée consiste à décomposer le temps en éléments : instants ou intervalles, puis à traduire les valeurs structurelles de chaque élément temporel en une représentation visuelle, et enfin à placer les représentations selon l'ordre de succession des éléments temporels.

La distinction entre les moments temporels se fait par une distinction des représentations des dimensions structurelles.

Le système Browser[24] présenté en Figure 50, visualise chaque dimension structurelle dans un espace séparé. Les représentations des dimensions structurelles ne sont pas superposées, et elles sont placées par rapport à un seul axe temps.

Figure 50. *Plusieurs données visualisées sur un axe de temps unique (Browser, 2003).*

Un système pour la visualisation de l'évolution de réseaux en 3D peut aussi s'effectuer sous forme de couches dont chacune représente le réseau pour une tranche de temps donnée (Brandes et Corman, 2003). Les sommets, correspondants à une entité, restent dans des positions semblables d'une couche à une autre, afin de conserver la carte mentale, comme le montre la Figure 51.

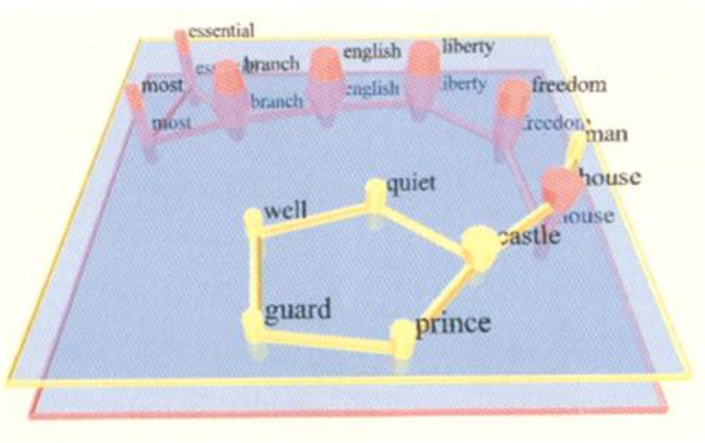


Figure 51. *Représentation 3D d'un réseau évolutif : (Brandes et Corman, 2003).*

[24] http://www.ncea.org.au/

Le système *TGRIP* (Erten et al., 2004) permet l'analyse visuelle de l'évolution de collaborations entre chercheurs d'un domaine donné, comme extension de *GRIP*. *TGRIP* produit une série de représentations 2D, visible en Figure 52, une pour chaque période, en fixant tous les sommets communs à chaque période. Les sommets et les arêtes du graphe étudié possèdent un poids calculé en fonction de la structure du graphe. Ainsi, chaque sommet a une taille relative à son poids. Le poids d'une arête est utilisé pour calculer la force d'attraction entre les sommets lors du dessin de graphe.

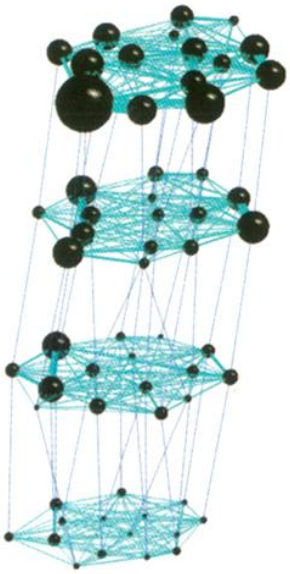


Figure 52. *Représentation 2D de l'évolution de collaborations par TGRIP (Erten et al., 2004).*

L'approche proposée par (Chen et Carr, 1999) visualisé en Figure 53 consiste à visualiser séparément les réseaux pour chaque période. Les différentes périodes sont visualisées séparément.

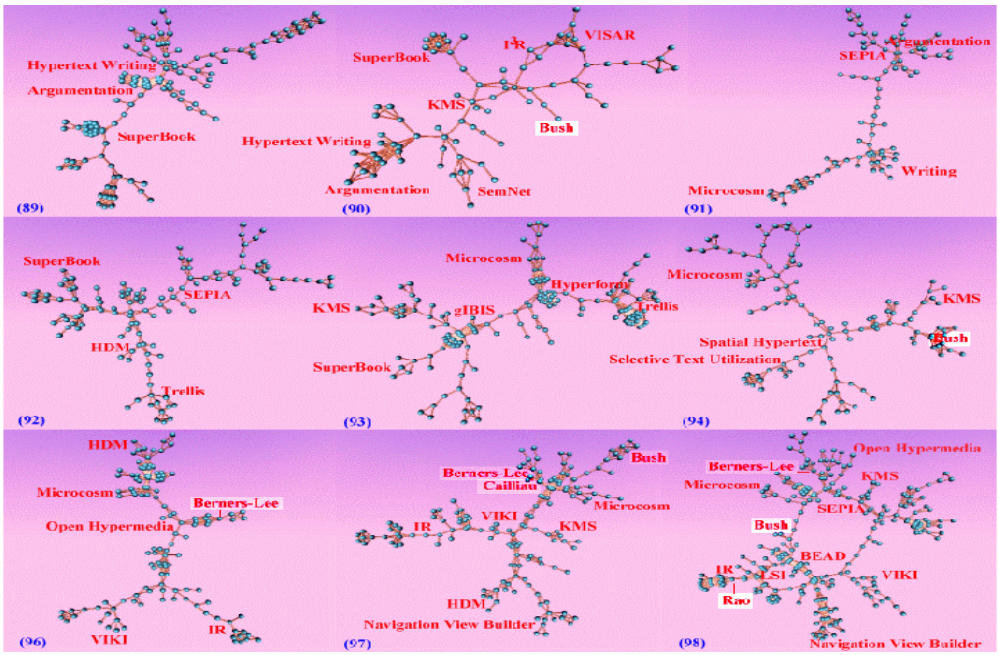


Figure 53. *Représentation séparée des réseaux évolutifs (Chen et Carr, 1999).*

Nous comparons ces techniques de représentation de l'espace temporel et des données temporelles, dans le Tableau 16.

Critères / Travaux	Vue globale	Vue spécifique	Surcharge	Facilité interprétation	Analyse structure	Données relationnelles
(Chen et Carr, 1999)	-	++	-	+	++	+
AsbruView (Kosara et Miksch, 1999)	++	+-	-	+	++	-
Data Tube (Ankerst, 2000), (Ankerst, 2001)	+	--	+	-	-	-
Tulip (Auber, 2001)	+	-	-	-	--	+
TimeSearcher (Hochheiser, 2002a)	++	--	+	+	++	-
Lexil Pencils (Brian, 2003)	+	-	-	--	--	-
(Browser, 2003)	--	+-	-	++	++	-
(Brandes et Corman, 2003)	++	+	+	++	-	+
CiteSpace (Chen, 2004)	++	+	+	++	-	+
TGRIP (Erten et al., 2004)	--	++	+	++	-	+

Légende

++ répond totalement au critère

+ répond assez bien au critère.

+- répond partiellement au critère.

- ne répond pas au critère.

-- ne répond pas du tout au critère.

Tableau 16. *Comparaison des travaux sur la représentation de l'espace temps et des données temporelles.*

Les critères retenus sont :

- une vue globale, c'est-à-dire une perception efficace de l'ensemble des périodes étudiées, notée « Vue globale »;
- une vue spécifique, permettant d'analyser unitairement chacune des données et de la quantifier à tout moment, notée « Vue spécifique » ;
- « Surcharge », une surcharge du graphe, le rendant difficile à analyser;
- «Facilité interprétation », une facilité d'interprétation du graphe;
- « Analyse structure », une comparaison des changements de structures au cours du temps.
- une application permettant la comparaison claire de « Données relationnelles ».

D'après ces résultats synthétisés dans le Tableau 16, on constate de manière générale que si la vision globale est favorisée, c'est au détriment de la vue spécifique et inversement. La surcharge des graphes est caractérisée par la facilité à analyser les alliances mais aussi l'évolution des données, elles-mêmes. La facilité d'interprétation provient de l'effort de cognition que doit fournir l'utilisateur pour comprendre la stratégie de visualisation évolutive du graphe. L'analyse de structure révèle si la détection rapide des changements de structure est possible tels que le changement d'alliance, la disparition d'acteur, l'émergence d'acteur…. Enfin, une partie de ces outils n'est pas adaptée pour les données relationnelles.

D'un point de vue de la représentation de l'espace temps, la décomposition du temps en périodes puis les représentations de ces dernières selon l'ordre de succession des éléments temporels facilite grandement la détection des changements de structure des données, d'un point de vue global mais aussi individuel.

2.4.4. Détection des tendances émergentes

La veille consiste à être attentif aux évolutions. Comme nous l'avons vu dans le premier chapitre, il est nécessaire de pouvoir prendre des décisions à temps afin d'être à même de réagir face à un environnement. Le captage et le traitement du signal est un enjeu extrêmement important dans le devenir du domaine étudié.

Un «signe d'alerte précoce » est une information dont l'interprétation suggère qu'un événement susceptible d'être important pour l'avenir d'une firme pourrait s'amorcer. Il est probable que plus un signe d'alerte est anticipatif, plus il est un signe de faible intensité d'où l'expression « signal faible » utilisée par (Ansoff, 1975). Les *informations anticipatives* visent la connaissance de l'environnement concurrentiel (Porter, 1982) mais aussi la surveillance de l'ensemble des acteurs influents en transaction directe ou indirecte avec l'entreprise (Martinet, 1984). Plus largement, les informations d'anticipation visent la connaissance des acteurs actuels et potentiels de l'entreprise (Lesca, 1986) que ce soit des clients, des concurrents, des fournisseurs ou divers prescripteurs de changement en général.

La prise en compte de la dimension temporelle dans la visualisation de données dites « *relationnelles* » par la présence possible de liaisons entre elles, permet d'analyser les informations anticipatives. Dans un contexte de veille stratégique au sein d'une entreprise, ces signaux s'amplifient avec le temps mais en contrepartie, cette dernière dispose d'un délai moindre pour réagir. Le signal faible est un produit informatif qui se conserve difficilement et qui a une durée de vie limitée à l'annonce de l'événement qu'il porte. Chaque signal acquiert une signification propre, liée à l'interprétation qu'en fait le dirigeant sur la base de ses connaissances, mais aussi par rapport à différentes hypothèses ou interrogations. Un processus « d'ouverture des possibles » est proposé (Piaget, 1970) où l'individu doit se poser des questions sur l'existence de sa réalité et construire d'autres réalités possibles. L'exploitation des signaux faibles s'inscrit dans une vision interprétative de l'environnement au sens où l'entend (Koenig, 1996).

La détection, via des graphes, des signaux faibles est particulièrement importante lorsque l'expert cherche à anticiper les ruptures susceptibles de se produire dans son environnement économique, technologique, social, etc...

Par exemple, la disparition brutale d'un sous domaine, d'une équipe, d'un acteur majeur peut être une information stratégique, à l'origine d'importants changements dans les axes de décision comme la réorientation de thématique, le changement d'alliance ou tout simplement l'arrêt d'une collaboration.

Ainsi la visualisation de données temporelles a pour but d'étudier l'évolution relative des points les uns par rapport aux autres, afin de connaître la typologie de leur dynamique et de répondre à des questions (Dousset, 2003) comme :

- Cet élément évolue-t-il plus vite ou moins vite que la moyenne ?
- Evolue-t-il en sens inverse ?
- Se déplace-t-il dans le nuage de points?
- Se rapproche t-il d'un autre élément ?
- Quelles sont ses occurrences anormales ?
- Voit-on apparaître une stratégie cohérente ?

2.5. Conclusion

La visualisation scientifique, la visualisation d'information est un domaine plus récent que la recherche d'information. L'augmentation de la puissance de calcul et de la qualité des interfaces graphiques a permis d'explorer et de développer des techniques interactives.

Dans ce chapitre, nous avons présenté le domaine de la visualisation de données relationnelles et évolutives. Il s'agit de représenter dans un espace physique sous la forme de graphiques une information souvent abstraite (Hinum et al., 2005). Cette information peut comprendre des données, des processus, des relations ou des concepts.

Il s'agit de fournir à l'utilisateur une compréhension qualitative du contenu de l'information. Cette visualisation doit permettre à l'utilisateur final de faire des découvertes, de proposer des explications ou de prendre des décisions (Meliker et al., 2005). Ces actions peuvent se faire aussi bien sur des motifs (clusters, tendances, émergences, anomalies) ou sur des ensembles d'éléments ou encore sur des éléments isolés. La visualisation de données temporelles doit permettre de communiquer efficacement des informations et faciliter la découverte de connaissances au travers d'une représentation graphique via des cartes cognitives issue de l'analyse d'un corpus d'informations.

Nous avons aussi étudié les différents domaines d'application de ces visualisations, en ciblant les réseaux sociaux, sémantiques ou encore d'interactions.

Suite à cela, nous avons vu les principes de base pour réaliser une bonne représentation graphique, à savoir les règles d'esthétique et les algorithmes utilisés permettant de les suivre, les principes sémiologiques, en particulier dans un contexte temporel.

Puis, nous avons proposé une taxonomie de techniques de visualisation des données temporelles. Nous avons ainsi ciblé les trois principales représentations du temps, à savoir la représentation linéaire, cyclique ou encore non uniforme.

Le recours la taxonomie de Schneiderman permet de mieux appréhender le recours à une visualisation spécifique, selon la nature de la donnée, comme le montre le Tableau 17.

Nous avons étudié la représentation de l'espace temps et des données temporelles. Plusieurs exemples d'outils disponibles ont été cités et illustrés. L'ensemble des données temporelles étudiées peuvent être représentées sur une ou plusieurs dimensions, selon les objectifs fixés par l'utilisateur.

Type de données	Description
1D	Données organisées de manière linéaire ou séquentielle : les lignes de code d'un programme, un texte, d'une liste de noms, etc. Chaque entité possède des attributs qui sont représentés visuellement : il peut s'agir de la taille d'un fichier, de l'auteur d'un texte, de la date de modification des lignes d'un code source, etc.
2D	Données dont la localisation ou la géométrie dans le plan est primordiale, comme par exemple les Systèmes d'Information Géographique ou les outils de mise en page.
3D	Données dont la localisation ou la géométrie dans l'espace prévaut sur tout autre attribut. Ces données sont celles manipulées dans des domaines tels que, apr exemple, la chimie (visualisation de la structure des molécules), la médecine, la mécanique ou encore l'architecture.
Temporelles	Données ayant une existence dans le temps. Elles se distinguent des données linéairement structurées par la nature des entités qui composent l'ensemble de données. Chaque entité est en effet caractérisée par un début, une fin et par conséquent une durée. Deux entités peuvent alors se chevaucher ou se recouvrir dans le temps. L'édition de scénarios temporels ou la visualisation d'historiques médicaux sont deux exemples représentatifs des travaux effectués dans ce domaine.
Multidimensionnelles	Données dont le caractère spatial n'est pas dominant et dont le nombre *n* d'attributs est élevé ($n > 3$). Cette catégorie se distingue des ensembles de données dont la structure est temporelle, linéaire, bidimensionnelle ou tridimensionnelle par l'absence de structure dominante (le temps, le plan, l'espace, etc.) au sein des *n* dimensions. On trouve par exemple dans cette catégorie les données statistiques, les collections de documents, ou le contenu de bases de données.
Hiérarchiques	Les données possèdent un lien vers une unique entité parente : système de fichiers, généalogie, etc.
Relationnelles	Les données sont liées entres elles par un lien quelconque mais explicite et forment un graphe dont la structure n'est pas arborescente : documents hypertextes, réseaux informatiques, etc.

Tableau 17. *Description des types de données identifiés dans la taxonomie de B.Schneiderman*

Enfin, nous avons énoncé les liens entre l'extraction de connaissance, vue dans le premier chapitre, dans un contexte de veille stratégique et la visualisation de données relationnelles et temporelles. En effet, l'un des objectifs de la veille est d'anticiper et de détecter les signaux d'alerte. Après avoir défini ces principes, nous

constatons que le recours à des graphes dynamiques permet de distinguer les structures émergentes, les changements d'alliances et les acteurs importants.

En nous appuyant sur ces principes de visualisation de données temporelles, notre objectif est la conception d'un outil de graphe répondant aux précognitions citées dans ce chapitre. Dans un contexte de veille stratégique, l'outil est destiné à l'analyse de structure et à la détection d'informations significatives et révélatrices de changement. Nous présenterons l'outil développé dans les chapitres suivants.

Chapitre 3.
Notre contribution pour les graphes statiques

> *« Comment trouver un diamant dans un tas de charbon sans se salir les mains ? »*(SAS Enterprise Miner)

3.1. Introduction ____ 96
3.2. Visualisation de données : méthodes et représentations ____ 96
3.2.1. Approche ____ 96
3.2.2. Espace de représentation des données relationnelles ____ 98
3.2.3. Métriques ____ 98
✓ Métriques associées aux sommets ... 98
✓ Métriques associées aux arêtes ... 100
✓ Représentation des métriques ... 100
3.2.4. Placement des sommets ____ 102
3.2.5. Représentation des liens ____ 103
3.2.6. Algorithmes de représentation de graphe ____ 103
✓ Représentation de graphe ... 103
✓ Algorithme basé sur les force_directed_placement (FDP) ... 104
3.2.7. Autres types de graphes ____ 114
✓ Graphes orientés ... 114
✓ Graphes spécifiques ... 116
✓ Représentation caméléon ... 119
3.3. Les fonctionnalités ____ 122
3.3.1. Positionnement de l'utilisateur ____ 122
3.3.2. Identification et analyse de structure de graphe ____ 123
3.3.3. Filtrage ____ 125
3.3.4. *K*-Core ____ 126
3.3.5. Transitivité ____ 128
3.3.6. Retour aux documents ____ 131
3.3.7. Focalisation ____ 132
✓ Elaboration ... 132
✓ Restitution des résultats ... 133
3.3.8. Partitionnement de graphes ____ 134
✓ Principe général ... 134
✓ Markov Clustering intégré à VisuGraph ... 135
✓ Manipulation du graphe réduit ... 136
3.4. Conclusion ____ 140

3.1. Introduction

L'objectif de la visualisation est de faciliter la recherche des structures, caractéristiques, motifs, tendances, anomalies et des relations entre les individus (Grinstein et Ward, 2002). La visualisation apporte une valeur ajoutée souvent due à un cheminement informel et non démontrable. Les travaux de (Fayyad et al., 2002) caractérisent cet apport comme étant l'augmentation des capacités de perception humaine, en fournissant une vision perspicace des données (Marwah et al., 2009), (Chen et al., 2009).

La base de toute proposition d'outil de visualisation de données relationnelles suppose de s'intéresser aux trois aspects fondamentaux suivants :

- la nature des données représentées,
- la manière dont les composantes du graphe sont exploitées pour transcrire ces données,
- la perception de ces composantes par l'utilisateur.

Tout l'art de la conception de graphe consiste alors à passer de l'espace des informations à une représentation visuelle qui traduise, grâce aux composantes utilisées, l'information originelle. Ce qui nous intéresse est bien l'étape de traduction ou de transcription de l'information vers un espace de représentation visuel. De plus, la distinction doit être faite entre la *visualisation* qui se rapporte au processus qui conduit à une représentation graphique et la *représentation graphique interactive d'informations* qui a trait aux moyens d'interactions qui utilisent une représentation graphique des informations. Le rôle de l'utilisateur dans les outils de visualisation de données est un sujet de préoccupation majeure (Grinstein, 1996), (Fayyad, 2002), (Kuntz, 2003).

En fouille visuelle de données, l'interaction matérialise la boucle de rétroaction entre l'utilisateur et le support visuel (Keim et Kriegel, 1994). L'objet de la visualisation n'est pas simplement limité à la production de représentations graphiques prédéfinies, non modifiables par l'utilisateur. En effet, un critère important d'un bon outil de visualisation de données est la possibilité, de le contrôler et le maîtriser pleinement afin de comprendre l'espace des informations ou d'interagir avec celui-ci (Cross et al., 20009). La visualisation rejoint sur ce point les préoccupations qui sont du domaine de l'*interaction* homme-machine.

Dans ce chapitre, nous abordons l'aspect statique de notre proposition de visualisation de données relationnelles, nommé VisuGraph (Karouach et Dousset, 2003), (Dousset et Karouach, 2005), (Loubier et Dousset, 2007c), module de graphe de la plate-forme de veille stratégique Tétralogie.

Dans un premier temps, nous présentons, en section 3.2, les méthodes et les représentations de la visualisation statique de données relationnelles qui nous intéressent. Au sein de cette section, nous développons notre approche, puis nous définissons l'espace de représentation considéré. La notion et l'élaboration de métriques est proposée au travers des sommets et des arêtes, permettant l'élaboration du graphe, par le placement des sommets, étudié, la représentation des liens et le recours à des algorithmes de représentation de graphe. Les différents types de graphes, accessibles via VisuGraph sont présentés.

Notre contribution repose également sur l'apport de fonctionnalités spécifiques que nous développons en section 3.3. Pour cela, nous considérons le positionnement de l'utilisateur, comme le point central de notre outil. Parmi les fonctionnalités accessibles, nous favorisons et facilitons par nos méthodes l'identification et l'analyse de structure de graphe et nous proposons des algorithmes facilitant la caractéristique des données, ainsi que leur voisinage, tels que le filtrage, le *k*-core, la transitivité, le retour aux documents, la focalisation, la présentation caméléon, ainsi que le partitionnement de graphe. Enfin, nous concluons sur cette contribution pour les graphes statiques.

3.2. Visualisation de données : méthodes et représentations

3.2.1. Approche

Comme nous l'avons vu dans le chapitre 2, les techniques de visualisation, et en particulier l'outil que nous proposons, viennent en aval des étapes de traitement automatiques.

Suite à l'application d'algorithmes de découverte des structures, elles permettent de représenter les résultats sous des formes intelligibles facilitant leur interprétation (Loubier et Dousset, 2006), (Loubier et Bahsoun, 2007).

La communication avec l'utilisateur de l'information manipulée et traitée par le système est au cœur de nos préoccupations dans ce contexte de proposition d'outil d'aide à l'analyse de données relationnelles. Notre approche est illustrée en Figure 54.

Pour les graphes statiques, elle repose sur deux axes majeurs :

- une représentation des données, dans un espace défini, en utilisant la notion de métrique pour caractériser les composants du graphe, représentés et placés spécifiquement ;
- une possibilité pour l'utilisateur de naviguer librement, à travers des méthodes d'exploration de graphe.

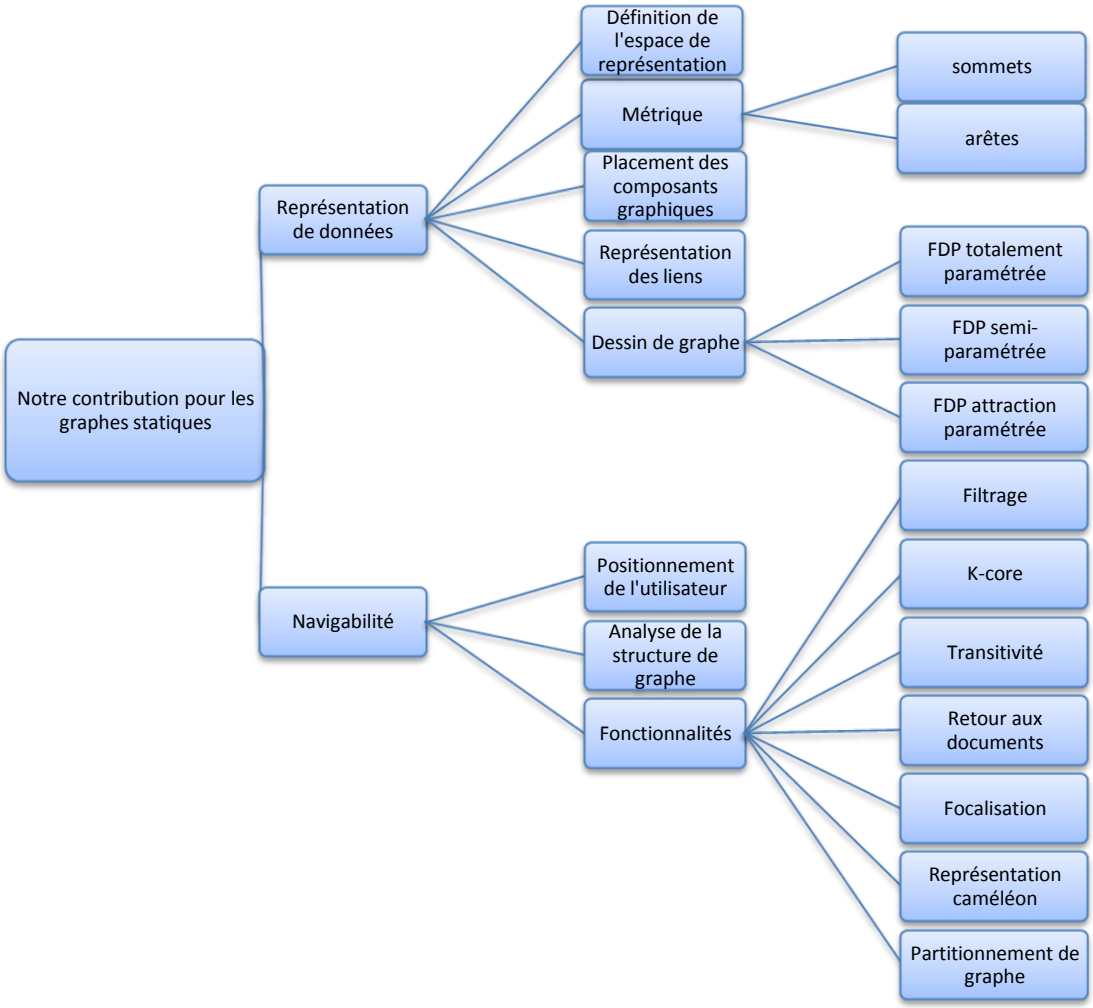


Figure 54. *Notre approche pour les graphes statiques.*

Dans tout ce qui suit, nous utilisons les notations suivantes.

Un graphe simple G est un couple formé de deux ensembles :

$X = \{x_1,x_2,\ldots,x_n\}$ dont les éléments sont appelés sommets ou encore nœuds, n étant fini ;

$A = \{a_1,a_2,\ldots,a_m\}$, partie de l'ensemble $\mathcal{P}_2(X)$ des parties à deux éléments de X, dont les composants sont appelés arêtes. Lorsque $a = \{x,y\} \in A$, on dit que a est l'arête de G d'extrémités x et y, ou que a joint x et y, ou encore a passe par x et y. Les sommets x et y sont dit adjacents dans G. Dans le cas des graphes orientés, si $a = (x,y)$ est un arc du graphe G, x est l'extrémité initiale de a, ou encore appelée *extrémité initiale* et y est l'*extrémité terminale* de a, ou bien *origine* et *destination*. L'arc a part de x et arrive à y.

3.2.2. Espace de représentation des données relationnelles

L'espace de représentation que nous considérons dans VisuGraph est caractérisé par l'espace géométrique et les variables liées aux formes et aspects des éléments graphiques visualisés, que nous nommerons variables visuelles. L'espace géométrique considéré est le plan. Nous élaborons l'espace de représentation selon la définition suivante.

Définition : L'élaboration d'un espace de représentation des données et de variables visuelles repose sur la prise en compte du point de vue de l'utilisateur, en lui proposant un graphe en fonction des informations dont il a besoin pour mener à bien sa tâche.

L'utilisateur doit pouvoir utiliser aisément l'outil. Cependant entre expert et novice, les attentes diffèrent sur les apports de l'outil.

L'expert, spécialiste du domaine étudié, a besoin d'apprentissage pour contrôler le comportement de la machine tandis que le novice peut considérer cette mise en main comme un obstacle puisqu'il n'est pas en mesure d'en évaluer la portée. Ainsi, au-delà du type de la considération de la nature des données, il est important de cibler les attentes de l'utilisateur afin d'y répondre le mieux possible.

Dans le cadre de nos travaux, il s'agit principalement d'étudier les relations entre les données ainsi que la structure proposée, dans un contexte concurrentiel, stratégique. La structure induite correspond à une visualisation graphique spécifique réalisée dans un espace de représentation adaptée. Pour cela, le choix du point de vue sur l'ensemble de données consiste à les organiser dans une structure particulière pouvant être qualifiées comme pertinentes.

Pour définir l'espace de représentation, considérons l'espace des informations E défini par l'ensemble des données $d_1, \ldots, d_n$ contenues dans le corpus D et par l'ensemble des éléments graphiques G qui représentent ces données, à savoir les sommets et les liens.

$$E = \begin{cases} G = (X,A) & [2] \\ D=\{d_1, .., d_n\} \text{ avec } d_i = <a_1, \ldots, a_m>, a_i \in G & [3] \end{cases}$$

Au niveau du choix des dimensions de l'espace de visualisation, le graphe est construit en 2D. Les sommets sont caractérisés par des coordonnées x et y, correspondant aux abscisses et aux ordonnées. Les données considérées sont relationnelles, leur structure est non hiérarchique. L'espace de données caractérisé permet à l'utilisateur de rechercher des relations, des particularités, des motifs ou tout autre phénomène remarquable qui lie les données entre elles. La représentation doit permettre à l'utilisateur de découvrir cette structure. L'objectif de l'utilisateur n'est pas seulement d'obtenir une vue d'ensemble de la répartition des données mais surtout de trouver les particularités liées à cette répartition.

3.2.3. Métriques

Les graphes servent à modéliser des structures relationnelles comportant un ensemble d'entités et des relations liant ces entités entre elles.

Afin de faciliter la compréhension des graphes, le concept de métrique permet la comparaison des différents éléments d'un graphe par affectation de valeur des différentes entités (sommets, arêtes) composant ce dernier.

Définition : Les travaux de Melancon mènent au concept de *nœud métrique* comme une quantité numérique associée aux nœuds et aux arêtes du graphe (Melancon et al., 1999).

Nous ciblons, ici, la métrique basée sur le contenu, c'est-à-dire sur les valeurs des données (Auray et al., 2001). Dans VisuGraph, les métriques sont associées aux sommets mais aussi aux arêtes (Loubier et Dousset, 2008b).

- ✓ ***Métriques associées aux sommets***

Soit A la matrice de cooccurrences du graphe G quelconque de n sommets. La fonction f qui associe à un sommet X_i, la valeur m_i :

$$f(x_i) = m_i \quad [4]$$

C'est une métrique structurale, définie sur $\mathcal{S}_m$, puisqu'elle tient compte de la structure du graphe, par le taux d'information qui circule au niveau de chaque sommet. Dans le cas d'une matrice asymétrique, cette valeur de métrique correspond à la somme des liens entre le sommet X_i et tous les autres sommets, c'est-à-dire la somme des cooccurrences. Dans le cas d'une matrice symétrique, il s'agit de la valeur, lue dans la matrice, du croisement de X_i avec lui-même, c'est-à-dire la somme des auto-occurrences. Par exemple, dans le cas d'une matrice symétrique croisant des auteurs, afin de connaitre leur nombre de publications, la valeur de la métrique affectée à un sommet est égale au total des articles coécrits avec d'autres individus et des documents publiés seul. Autre exemple, dans le cas d'une matrice asymétrique croisant des auteurs et des journaux, la valeur de la métrique attribuée à un auteur est égale à la somme des croisements entre cet individu et chacun des journaux.

$\mathcal{S}_m = <x_i, m_i>$
pour $i \in \{1,\ldots,n\}$

Et m_i la métrique attribuée au sommet X_i, définie par

$$f(X_i) = \{<x_i, m_i>\}$$

Cette métrique est très intéressante puisqu'elle indique clairement les sommets les plus importants, les plus centraux au niveau degré, qui sont souvent à la base, par exemple, d'alliances.

Si nous prenons l'exemple de la matrice de la Figure 55, croisant des articles écrits par des auteurs, les valeurs de métriques des sommets sont indiquées en ligne et en colonne. Ces valeurs correspondent au croisement de l'*auteur n* avec lui-même, permettant de prendre en compte sa valeur individuelle mais aussi celles de ses relations, c'est-à-dire avec les autres individus. Par exemple, l'auteur 5 a coécrit 1 article avec l'auteur 3, 4 avec l'auteur 4 et il en a écrit 1 tout seul, ce qui lui vaut une valeur de métrique égale à 6. *f(Auteur5) = {<Auteur5,6>}*. Cependant, une autre interprétation possible est qu'il en a écrit 2 tout seul, 3 avec l'auteur 4 et 1 en commun avec les auteurs 3 et 4. Les différentes interprétations de la matrice montrent les limites de ces dernières.

Auteur *n* / Auteur *n*	Auteur 1	Auteur 2	Auteur 3	Auteur 4	Auteur 5	Valeur de la métrique
Auteur 1	1	1	0	0	0	1
Auteur 2	1	2	1	0	0	2
Auteur 3	0	1	3	1	1	3
Auteur 4	0	0	1	5	4	5
Auteur 5	0	0	1	4	6	6
Valeur de la métrique	1	2	3	4	6	

Figure 55. *Valeur de métrique pour chacun des sommets, dans un cas symétrique.*

Dans l'exemple de la Figure 56, cas asymétriques, des auteurs sont croisés avec des journaux. Les valeurs des métriques des auteurs et des journaux sont, respectivement, en ligne et en colonne.

Journal *j* / Auteur *n*	Journal 1	Journal 2	Journal 3	Journal 4	Journal 5	Valeur de la métrique
Auteur 1	0	0	1	1	0	1
Auteur 2	1	0	1	0	0	2
Auteur 3	1	1	0	1	0	3
Auteur 4	2	2	1	0	0	5
Auteur 5	1	2	1	1	1	6
Valeur de la métrique	5	5	4	3	1	

Figure 56. *Valeur de métrique dans un cas asymétrique.*

✓ ***Métriques associées aux arêtes***

Le principe de métrique, vu précédemment, est applicable aux arêtes du graphe. Chaque paire de sommets adjacents est liée par une arête, la pondération de cette dernière se traduit par la valeur $m_{<xj,xk>}$,métrique associée à l'arête $a_{<xj,xk>}$ (Loubier et Dousset, 2008). Ainsi, si deux sommets x_j et x_k sont liés, la valeur de la métrique de l'arête les joignant est égale à la valeur du croisement entre x_j et x_k, dans la matrice de cooccurrences associée. Dans l'exemple de la Figure 55, les auteurs 4 et 5 sont joints par une arête dont la métrique a pour valeur 4.

La structure de métrique associée aux arêtes, pour un graphe composé de n est caractérisée par:

$$\mathcal{S}_e = \{<a_{<xj,xk>}, m_{<xj,xk>}>\}$$

pour $j, k \in \{1,\ldots,n\}$;

Et $m_{<xj,xk>}$ la métrique attribuée à l'arête $a_{<xj,xk>}$ définie par $f(a_{<xj,xk>})=\{a_{<xj,xk>}, m_{<xj,xk>}\}$

✓ ***Représentation des métriques***

L'image captée par l'œil forme une première représentation qui ne retient de l'image perçue que les informations élémentaires, encore nommées primitives visuelles, telles que les contours, les angles, les couleurs ou les contrastes.

Pour des graphes représentant des données aux valeurs très hétérogènes, il est essentiel de normaliser les valeurs des métriques m_i en les divisant par leur maximum m_m

$$\mu_i = \frac{m_i}{m_m}, \mu_i \in [0,1]. \quad [5]$$

Selon la sémiologie étudiée dans le chapitre 2, la valeur de la métrique est codée par le biais d'artifices visuels spécifiques (Bertin, 1970) tels que :

Les formes

La forme peut-être géométrique ou figurative et sa variation est infinie. De plus, elle est associative mais elle n'est ni ordonnée, ni quantitative, ni sélective.

Il est important de limiter la diversité dans le choix des formes afin d'éviter de surcharger la visualisation mais surtout afin de limiter l'effort cognitif de l'utilisateur pour étudier chacun des signes. L'étude d'un ensemble de signes ne peut s'effectuer qu'en les étudiant tous successivement. En effet, on ne peut regrouper d'un seul coup d'œil tous les signes d'une même forme car la lecture de cette dernière nécessite de ne regarder qu'un signe à la fois. La représentation des nœuds s'effectue par des icônes (cercles, barres…).

La forme utilisée par défaut est le cercle, facilitant l'interprétation du graphe. La lisibilité du cercle est souvent plus grande que celle des dessins réalistes. L'apport d'une forme est également due à sa simplicité (Bonnet, 1989). Les cercles expriment relativement bien l'identité de la donnée à représenter et donc, par relation, les différences. Plusieurs formes différentes peuvent être utilisées dans les graphes conçus sous VisuGraph, comme par exemple la visualisation de sommets sous forme de barres ou encore le recours à des cercles de diamètre fixe mais d'intensité de couleur variable (Karouach, 2003).

La variation de taille

Initialement, les sommets sont représentés sous forme de cercle dont le rayon est relatif à la valeur de la métrique du sommet. Ainsi, les nœuds les plus importants se distinguent clairement par leur taille plus grande que les autres. La perception du changement de taille entre différentes données est plus simple que celle des couleurs. La variation du diamètre des cercles permet ainsi de traduire parfaitement les différences quantitatives des métriques des données. Cependant, dans l'outil, nous combinons ces deux critères afin de faciliter davantage les comparaisons des différences de valeurs.

D'un point de vue ergonomique, le diamètre du cercle $\mathcal{D}(x_i)$ du sommet x_i vaut :

$$\mathcal{D}(x_i) = \sqrt{\mu_i}$$

Ainsi, $\mathcal{D}(x_i)$ ne peut dépasser 1 et permet, d'un point de vue ergonomique, d'homogénéiser la représentation des sommets.

Les couleurs

Pour distinguer les différentes composantes, des couleurs distinctes sont utilisées. Ainsi, les sommets en colonne dans la matrice de cooccurrences asymétrique sont représentés d'une certaine couleur et ceux en ligne sont pigmentés autrement pour bien les distinguer. Comme les formes, les couleurs traduisent des différences mais ne peuvent cependant les ordonner entre elles.

L'influence du fond sur la forme agit sur la perception de l'image. Ainsi, dans VisuGraph, l'utilisateur choisit le fond lui facilitant l'interprétation des structures de données, à savoir un fond blanc, noir ou encore d'une autre couleur.

La variation de valeur

Des dégradés sont utilisés pour représenter les valeurs des métriques, que ce soit pour les sommets ou encore pour les arêtes. La variation de valeur d'une couleur d'arête est un changement d'intensité lumineuse du plus sombre au plus clair, ou inversement ; elle traduit une relation d'ordre entre les différentes valeurs de métrique. Elle est obtenue par des variations de valeurs, modifiant les tons du gris, utilisé pour représenter les arêtes. La valeur est donc dissociative car elle entraîne une variation de la visibilité.

A partir des valeurs μ_i, nous définissons un spectre de nuances de couleurs adapté à la distribution de ces valeurs.

Pour coder l'intensité de la couleur à partir des valeurs métriques relatives à chaque sommet, nous utilisons le modèle défini par la famille de fonctions non linéaires, illustré dans la Figure 57. Le paramètre n joue le rôle d'amplificateur de l'intensité dans le cas de faible valeur métrique. x est la valeur de métrique visualisée dans la matrice de cooccurrences, comme nous l'avons vu dans la Figure 55.

$$f_n(x) = \frac{(n+1).x}{n.x+1} \quad [6]$$

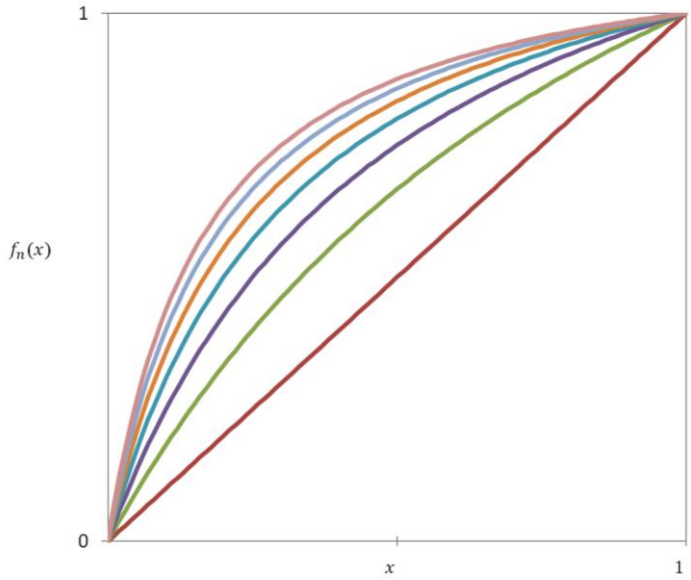


Figure 57. *Prise en compte de la non linéarité de la fonction de codage.*

Les mêmes fonctions sont utilisées pour la coloration des arêtes afin d'identifier les liens forts et faibles dans le graphe (Karouach, 2003). Ces variations sont ordonnées mais elles ne sont pas quantitatives puisque les différents niveaux de couleurs utilisés, que ce soit pour la coloration des arêtes ou encore des sommets, peuvent être classés mais il est impossible de chiffrer la différence entre deux valeurs (Loubier et Dousset, 2006b).

3.2.4. Placement des sommets

La position individuelle des sommets, sans prendre en compte l'aspect structural constitué avec leur voisinage, n'est pas significative. Dans un contexte non évolutif, elle ne traduit pas la valeur d'attribut des données, mais celle relative aux liens entre les sommets. Ainsi, le positionnement des sommets est exclusivement calculé de manière à satisfaire un certain nombre de critères esthétiques ou pratique de construction de graphe, tels que la minimisation des croisements d'arêtes, l'optimisation de la surface de représentation,…. La visualisation graphique des données nécessite l'attribution de coordonnées x et y pour chaque nœud visualisé dans l'espace de représentation.

Le placement d'un ensemble X d'éléments dans un espace à n dimensions peut être défini par une fonction :

$$\lambda : V \rightarrow \mathbb{R}^n. \qquad [7]$$

On notera x_i la $i^{ème}$ coordonnée du sommet v tel que

$$\lambda(V) = (x_1, x_2, \ldots, x_n) \qquad [8]$$

Dans le cas d'un graphe statique représenté par VisuGraph, les sommets sont initialement placés de manière circulaire, comme le montre la Figure 58.

Dans le cas de graphe biparti, les sommets sont placés sur deux cercles concentriques. Les sommets qui correspondent aux lignes de la matrice sont situés sur le cercle extérieur et ceux associés aux colonnes sont sur le cercle intérieur. Cette représentation par défaut permet de visualiser sans grand effort cognitif, la répartition des deux ensembles de sommets, dans le cas biparti (Loubier et Dousset, 2007).

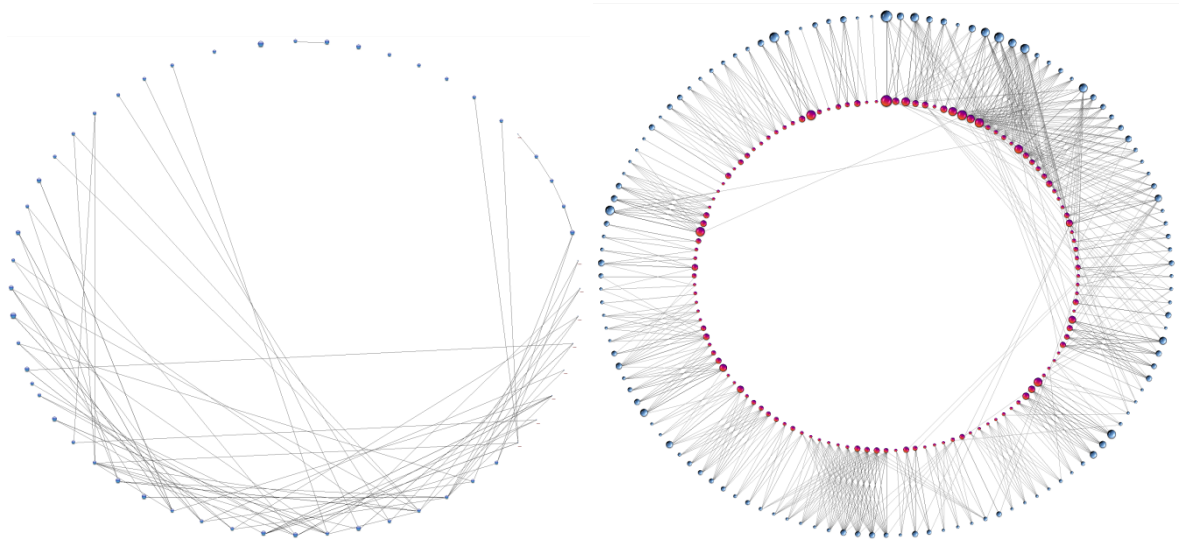


Figure 58. *Graphe basé sur une matrice symétrique (à gauche) et sur une matrice asymétrique (à droite).*

3.2.5. Représentation des liens

Dans le cas du dessin de graphes, les liens entre deux nœuds, symbolisant les données relationnelles, sont matérialisés par des arêtes. Les arêtes, appartenant à l'ensemble A, sont représentées, en utilisant une fonction de routage. Cette dernière correspond à l'ensemble des points par lesquels passe une arête. Le routage d'une arête $a \in A$ est une séquence ordonnée de points de l'espace de représentation (de cardinal k plus grand ou égal à deux).

Caractérisons un point par ses deux coordonnées,

$$p_i = (x_i, y_i) \text{ pour } i \in \{1, \ldots, k\}$$
$$R(a) = \{p_1, p_2, p_3, \ldots, p_k \in \mathbb{R}^k, k \geq 2\} \quad [9]$$

Dans VisuGraph, le tracé des arêtes se base principalement sur deux points, correspondant aux coordonnées des deux sommets à joindre.

Dans le cas d'un graphe orienté, les liens sont représentés de façon similaire, c'est-à-dire basés sur deux points principaux, l'un de départ et l'autre d'arrivée. L'usage de plusieurs points au sein du routage est primordial dans le cas où plusieurs liens se superposent et où il devient nécessaire de courber légèrement l'un par rapport à l'autre. Par exemple, si nous étudions les ventes de licences entre deux entreprises e_1 et e_2. Il se peut qu'e_1 ait vendu une licence à e_2 mais aussi que l'entreprise nommée e_1 en ait acheté une à e_2. Dans ce cas là, un graphe orienté est utilisé afin de distinguer l'achat des ventes, par le sens de la flèche. Le problème est alors la superposition de l'arc achat et de celui de la vente. Afin de les distinguer, deux routages distincts sont élaborés pour chaque lien, permettant ainsi une représentation différentiable de chaque flèche.

3.2.6. Algorithmes de représentation de graphe

✓ ***Représentation de graphe***

La notion de *représentation de graphe* regroupe l'ensemble des techniques permettant d'élaborer une visualisation des données dans le plan de façon à en faciliter la lecture.

Représenter un graphe de manière optimale consiste à disposer ses sommets et ses arêtes selon les règles d'esthétique, à homogénéiser les longueurs des arêtes, à limiter les entrecoupements des liens, à distribuer les nœuds de façon rationnelle sur la surface de représentation, comme nous avons pu l'étudier dans le chapitre 2.

Le but de la visualisation d'information est d'exploiter les caractéristiques du système visuel humain pour faciliter la manipulation et l'interprétation de données informatiques variées.

Les travaux en perception visuelle ont montré que l'être humain a une perception d'abord globale d'une scène, avant de porter son attention aux détails (Myers, 2000). Cette caractéristique est à la base de nombreuses illusions visuelles. Les travaux de (Bertin, 1977) et (Tufte, 1983) ont montré comment exploiter, de façon intuitive ou ad hoc, ces caractéristiques de perception globale. La visualisation d'information cherche à exploiter ces mêmes caractéristiques de façon plus systématique.

La représentation de graphe étant destinée à explorer des données (Loubier et Dousset, 2007b), les tâches rencontrées en visualisation d'information sont liées à la Recherche d'Information au sens large :

- exploration rapide d'ensembles d'informations inconnues ;
- mise en évidence de relations et de structures dans les informations ;
- mise en évidence de chemins d'accès à des informations pertinentes ;
- classification interactive des informations.

La qualité de la représentation est primordiale pour l'appropriation de la visualisation par l'utilisateur (Purchase, 2000). Il faut cependant noter que l'efficacité d'une technique de visualisation est difficile voire impossible à évaluer de façon absolue.

On retient quatre concepts de base (Kuntz, 2003) :

- la convention de tracé qui spécifie les règles géométriques de lecture du tracé, comme le tracé des arêtes sous forme de droite ou encore de courbe ;
- les contraintes de support et de l'œil humain qui imposent des écarts minimums entre les sommets pour éviter leur superposition ou encore le chevauchement des libellés ;
- les critères esthétiques qui facilitent la lisibilité comme vus dans le chapitre précédent ;

✓ ***Algorithme basé sur les force_directed_placement (FDP)***

Afin d'améliorer la représentation de graphe et d'obtenir une visualisation la plus planaire possible, c'est-à-dire minimisant le nombre d'entrecoupements d'arêtes, nous nous basons sur l'analogie « *arc = ressort* ». Notre modèle s'inspire des travaux présentés dans le chapitre 2. Le système, ainsi considéré, engendre des forces entre les sommets, ce qui provoque naturellement des déplacements de ces derniers. La notion d'attraction entre les sommets s'effectue par leur rapprochement pour ceux fortement liés et la répulsion s'établit par éloignement des nœuds. La condition d'arrêt initialement proposée par (Fruchterman et Reingold, 1991) pour un tel système est un nombre maximum d'itérations selon l'évolution du graphe dans le temps. L'utilisateur laisse les forces agir jusqu'à ce qu'il obtienne satisfaction des résultats visuels.

Dans notre proposition nous prenons en compte plusieurs paramètres, à savoir :

- Le dosage, par l'utilisateur, de l'attraction et de la répulsion,
- La distance minimale entre les deux sommets,
- L'aire de représentation du graphe dans la fenêtre de représentation.

Dans un premier temps, nous proposons un algorithme général (Karouach, 2003), (Loubier et al., 2007) permettant un meilleur rendu pour la représentation graphique, quelque soit le type de données (temporelles ou non).

La force d'attraction entre deux sommets u et v est donnée par :

$$f_a(u,v) = \frac{\beta \times d_{uv}^{\alpha_a}}{K} \qquad [10]$$

β est une constante. d_{uv} est la distance entre u et v dans le dessin. α_a sert à augmenter/diminuer l'attraction entre deux sommets.

Le facteur *K* est calculé en fonction de l'aire du dessin et du nombre de sommets du graphe et permet de s'assurer du non dépassement par les sommets, des bords de la fenêtre de représentation. Pour cela, *L* représente la longueur de la fenêtre, *l* la largeur et *N* correspond au nombre de sommets visibles du graphe.

$$K = \sqrt{\frac{L \times l}{N}} \qquad [11]$$

Si les sommets *u* et *v* ne sont pas reliés par une arête alors $f_a(u, v) = 0$.

La force de répulsion entre deux sommets *u* et *v* est définie par :

$$f_r(u, v) = \frac{\alpha_r \times K^2}{d_{uv}^c} \qquad [12]$$

α_r sert à augmenter etdiminuer la répulsion entre deux sommets *u* et *v* ; *c* est, dans ce cas là, une constante.

Afin d'obtenir une forte interactivité entre le système et l'utilisateur, et permettre à ce dernier de contrôler pleinement sa représentation graphique, l'attraction ou/et la répulsion entre les sommets peuvent manuellement être modifiées (Loubier, 2009).

Pour cela, des sliders[25] sont mis à disposition de l'utilisateur, dans le menu, permettant l'augmentation ou la diminution de ces deux types de forces. Le système dispose ainsi d'un slider spécifique aux forces d'attractions et un pour les forces de répulsion. Chacun des sliders est composé de dix graduations et la valeur d'initialisation est par défaut à 5.

Dans l'outil VisuGraph, nous proposons trois variantes de l'algorithme FDP, chacune permettant l'obtention de résultats spécifiques. Les valeurs attribuées aux différentes variables sont issues de multiples tests pour lesquels, l'accent à été mis sur la qualité du dessin de graphe obtenu. Pour illustrer et comparer chacune de ces propositions d'algorithme FDP, nous nous basons sur un même graphe, sur lequel sont appliquées successivement les trois méthodes présentées ci-dessous. Les résultats, des Figure 59, Figure 60, et Figure 61 montrent les différences de représentation.

L'algorithme FDP totalement paramétré

La particularité de cette proposition repose sur la décomposition en deux temps de l'algorithme. Dans un premier temps, la force de répulsion entre tous les sommets est calculée. Dans un second temps, toutes les attractions sont prises en compte, pour toute paire de sommets liés.

Pour éviter toute superposition des sommets, une vérification est effectuée afin d'attribuer des coordonnées uniques à chaque sommet, suite au déplacement provoqué par l'application des forces.

Dans ce premier algorithme de FDP, les paramètres sont étudiés pour obtenir des résultats pertinents.

Ainsi, pour le calcul de la force d'attraction,

β est une constant, initialisée à 1 ;

$d_{uv}^{\alpha_e}$ est la distance entre *u* et *v*, où α_a correspond à la valeur du slider divisée par deux, permettant d'interagir sur l'attraction.

Pour le calcul de la force de répulsion,

α_r correspond à la valeur du slider divisé par 6, permettant d'interagir sur la répulsion ;

[25] Règle graduée, dont le seuil initialement fixé peut être changé.

c est une constante, dans ce cas là initialisée à 1.

```
Pour tout sommet u {
        Si u est visible
        Alors {
                Calcul de la distance d(u,v) ;
                pour tout sommet v du graphe {
                        Calcul des forces répulsives entre u et v ;
                }
        }
}
Pour toutes les arêtes a du graphe {
        Calcul de la force d'attraction entre les deux extrémités u,v de a ;
}
/** Vérification de la non superposition des sommets par comparaison des coordonnnées **/
Pour tout sommet u {
        Pour tout sommet v {
                Si (xu,yu) == (xv, xy)
                alors changer les coordonnées de v ;
        }
}
```

Bases de l'algorithme FDP totalement paramétré

Cet algorithme est appliqué sur un graphe biparti et le résultat obtenu est visible sur la Figure 59. Les sommets s'attirent particulièrement selon leur ressemblance typologique. En effet, on constate que les sommets appartenant au même ensemble de données, de valeurs de métrique relativement proches et appartenant à la même structure connexe par une liaison entre eux, s'attirent très fortement. Cependant deux sommets de même structure connexe, n'appartenant pas au même ensemble de données s'attireront moins que ceux de même nature.

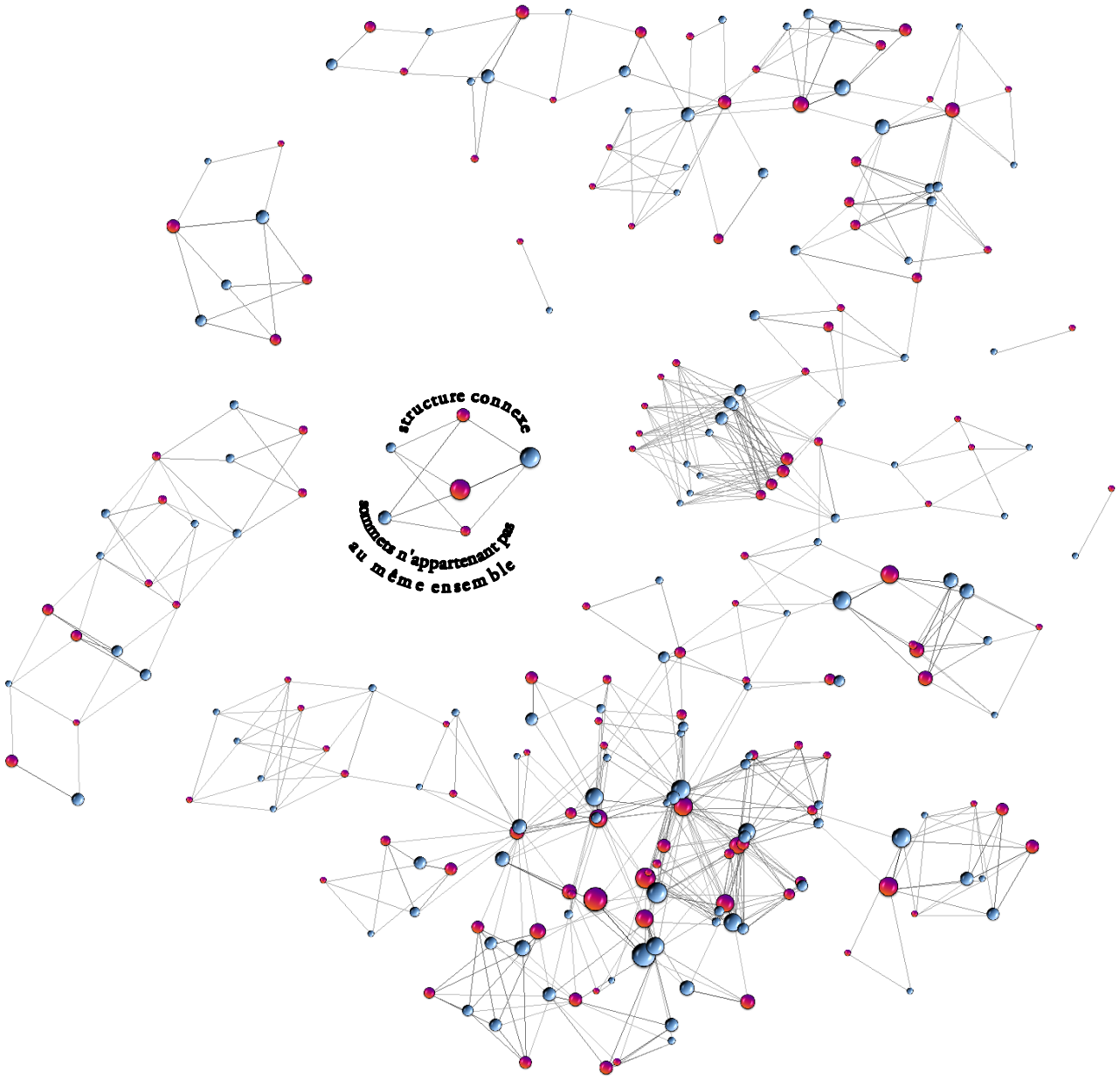


Figure 59. *Graphe sur lequel l'algorithme FDP totalement paramétré est appliqué.*

L'algorithme FDP semi paramétré

Cette autre version de l'algorithme FDP, calcule dans un même temps l'attraction et la répulsion entre deux sommets *u* et *v*.

Dans ce second algorithme FDP, les paramètres sont étudiés pour obtenir des résultats pertinents. De nombreux tests on été effectués et les valeurs initiales des paramètres dépendent fortement du volume de données représentées mais aussi de la complexité des croisements. Les résultats que nous proposons permettent d'obtenir des résultats corrects pour une grande majorité des cas relatif à la taille et à la complexité du graphe. Cependant, suivant la spécificité de la visualisation, nous ne pouvons nier l'obtention parfaite d'un résultat clair et précis sans l'intervention de l'utilisateur dans les valeurs de sliders.

```
Pour tout sommet u {
    Si u est visible
    Alors {
        Calcul de la distance d(u,v) ;
        pour tout sommet v {
            Calcul force répulsion entre u et v ;
            Calcul de la force d'attraction entre u et v ;
        }
    }
}
/** Vérification de la non superposition des sommets par comparaison des coordonnnées **/
Pour tout sommet u {
    Pour tout sommet v {
        Si (xu,yu) == (xv, xy)
        alors changer les coordonnées de v ;
    }
}
```

Algorithme FDP semi paramétré

Ainsi, pour le calcul de la force d'attraction,

β est une constante, initialisée à 1,5 ;

$d_{uv}^{\alpha_a}$ est la distance entre u et v, où α_a correspond à la valeur du slider divisée par deux, permettant d'interagir sur l'attraction.

Pour le calcul de la force de répulsion,

c n'est pas une constante dans ce cas précis et correspond à la valeur du slider permettant d'interagir sur la répulsion ;

α_r est une constante, initialisée à 2.

Cet algorithme a été appliqué au graphe de la Figure 59 et les résultats sont visibles sur la Figure 60. Bien que de manière très générale, la structure globale du graphe reste similaire, on constate qu'au niveau des structures connexes, les données sont davantage réparties et l'attirance s'effectue quelque soit l'appartenance à un ensemble de données, contrairement au cas précédent. Ainsi l'étude locale, au sein des structures connexes est plus claire.

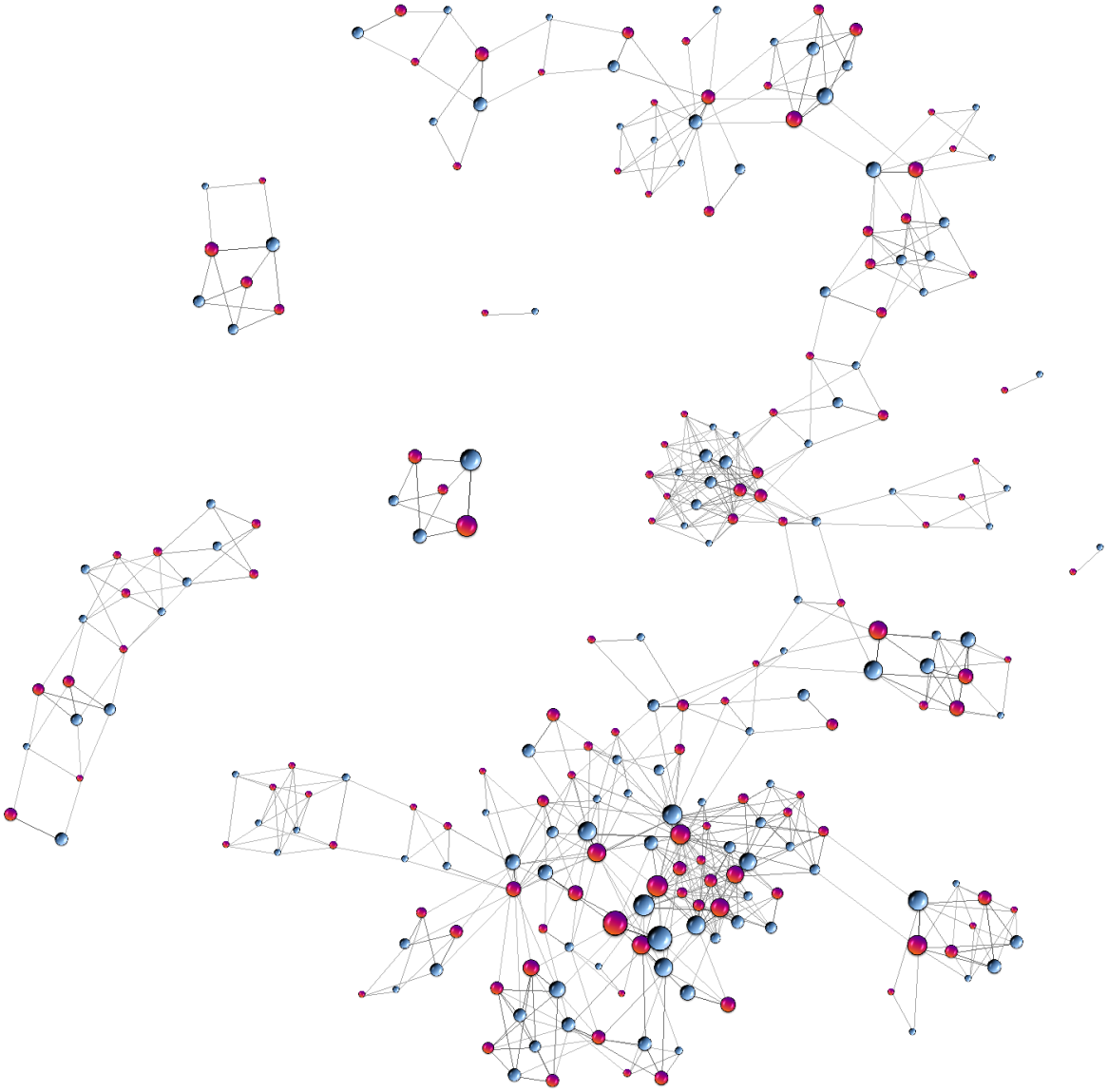


Figure 60. *Graphe sur lequel est appliqué l'algorithme FDP semi paramétré.*

L'algorithme FDP basé sur le paramétrage de l'attraction uniquement

Pour cette nouvelle proposition d'algorithme, seule la force d'attraction est paramétrable. En effet, les deux expérimentations précédentes ont permis de constater que la variation, de manière interactive, d'un paramètre a un impact sur les deux types de forces. Si seul un paramètre d'attraction est modifiable par l'utilisateur, les conséquences sur la répulsion seront réduites. L'objectif de cette proposition est de favoriser l'attraction, de fixer la répulsion et de comparer l'efficacité des résultats, à ceux obtenus avec les deux autres algorithmes.

Dans ce contexte, les forces de répulsion et d'attraction sont appliquées en deux temps distincts. Lorsque la force de répulsion reste stable, sans que l'utilisateur intervienne, la force d'attraction reste ajustable.

```
Pour tout sommet u {
        Si u est visible
        Alors {
                Calcul de la distance d(u,v) ;
                Pour tout sommet v {
                        Calcul des forces répulsives ;
                }
                Pour tous les voisins v de u{
                        Calcul des forces d'attraction entre u et v ;
                }
        }
 }

 /* Vérification de la non superposition des sommets par comparaison des coordonnnées */
 Pour tout sommet u {
        Pour tout sommet v {
                Si (xu,yu) == (xv, xy)
                alors changer les coordonnées de v ;
        }
 }
```

Algorithme FDP basé sur la paramétrage de l'attraction uniquement.

Dans ce troisième algorithme de FDP, les paramètres sont étudiés pour obtenir des résultats pertinents.

Ainsi, pour le calcul de la force d'attraction :

β est une constante, initialisée à 1,5 ;

$d_{uv}^{\alpha_a}$ est la distance entre u et v,

α_a correspond à la valeur du slider, permettant d'interagir sur l'attraction.

Pour le calcul de la force de répulsion,

c est une constante initialisée à 1,5 ;

α_r est une constante, initialisée à 2.

Suite à l'application de cet algorithme, les résultats sont visibles sur la Figure 61. On constate que le graphe obtenu est à mi chemin entre les deux représentations des Figure 59 et 60. En effet, la comparaison des sommets similaires est facilitée par leur proximité, tout en conservant une visualisation de la structure globale, mais aussi locale au niveau des parties connexes. Deux sommets du même ensemble vont s'attirer modérément, s'ils appartiennent au même regroupement connexe.

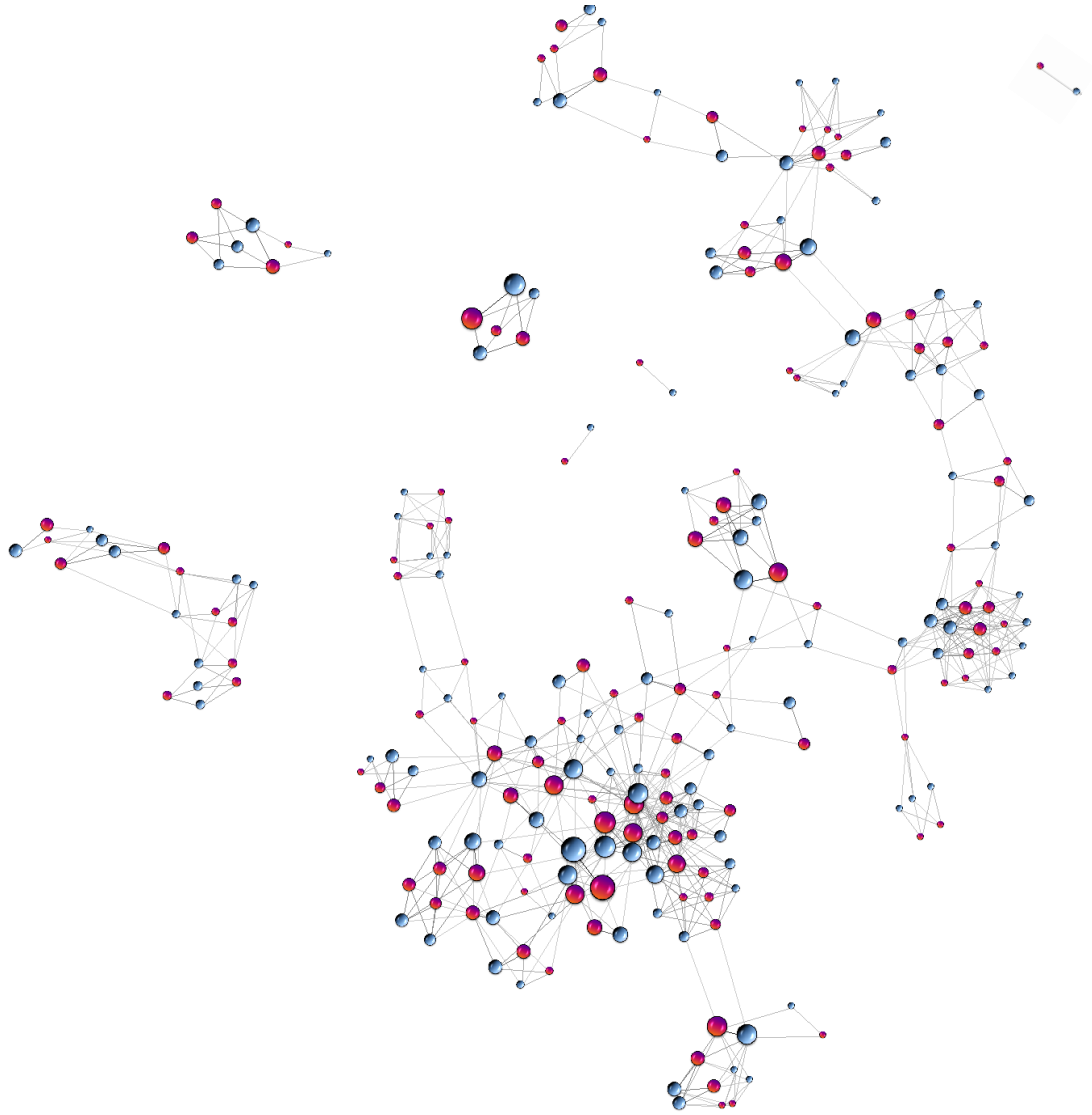


Figure 61. *Graphe sur lequel l'algorithme FDP est appliqué, basé sur le paramétrage de l'attraction.*

La notion de la température globale du système est introduite pour limiter le déplacement excessif des sommets. Cette température est totalement contrôlable par l'utilisateur, via une règle graduée en paramètre. Par changement de valeur de la température, les déplacements des sommets sont plus ou moins importants. La température initiale doit être de forte valeur, afin de permettre le déplacement rapide des sommets et surtout le décroisement des arêtes. Une fois le graphe amélioré, d'un point de vue de la lisibilité, une baisse de la température permet de stabiliser la visualisation.

Nos expérimentations nous mènent à préconiser un ordonnancement spécifique, en trois étapes, pour obtenir un meilleur résultat visuel, en ce qui concerne le paramétrage de ces trois forces (Loubier et Dousset, 2008).

Etape 1 : Appliquer une très forte valeur d'attraction, via le slider spécifique, jusqu'à obtenir un regroupement concentré des données, permettant de distinguer la structure globale du graphe. Mettre la température au maximum via le slider, afin de permettre le déplacement rapide et efficace des sommets.

Etape 2 : Réduire cette force d'attraction et augmenter la répulsion, via les deux sliders, afin d'obtenir un graphe lisible. Réduire la température pour éviter un mouvement trop brutal des sommets.

Etape 3 : Ajuster sensiblement les trois sliders, en baissant la température, jusqu'à obtention d'un résultat satisfaisant.

Ces principes sont illustrés dans la Figure 62, décomposant les différentes étapes du dessin de graphe.

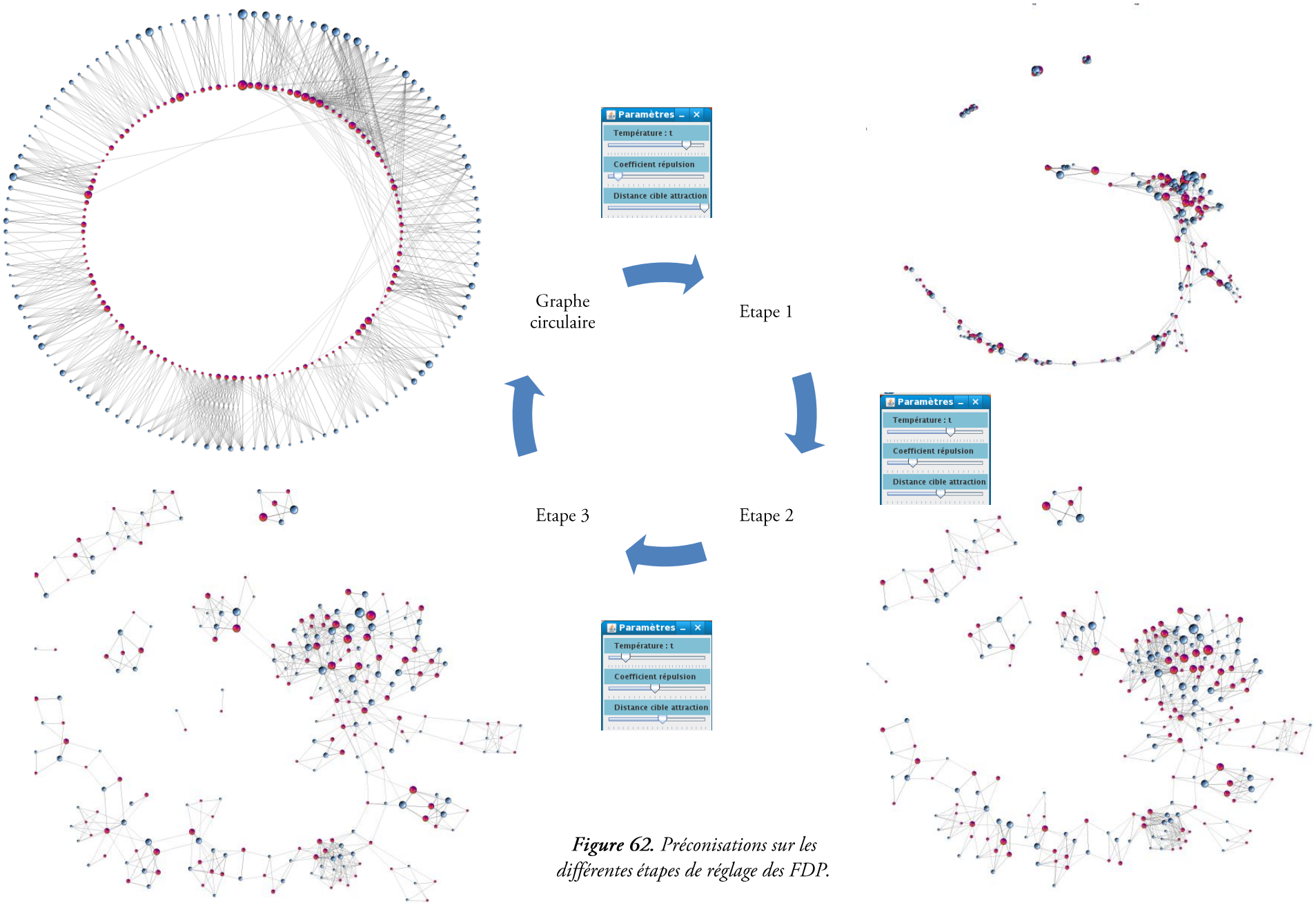


Figure 62. *Préconisations sur les différentes étapes de réglage des FDP.*

Pour synthétiser les résultats obtenus, un bilan comparant ces algorithmes FDP est présenté dans le Tableau 18. Nous ne prenons pas en compte les critères telles que la facilité de lecture du graphe, la limitation de croisement des arêtes et de proximité des sommets fortement liés puisque nous considérons que ces objectifs sont les résultats de base pour les trois types de FDP.

Type de FDP	Caractéristiques	Avantages spécifiques	Inconvénients spécifiques
totalement paramétrées	Calcul de toutes les répulsions dans un premier temps puis des attractions ; $\beta = 1$; $\alpha_r = 6$; c = 1.	Les sommets similaires sont fortement attirés, favorisant leur catégorisation ; Analyse structurale basée sur l'ensemble des sommets reliés ; Distinction au niveau structural entre les différents types d'entités ; Espace de représentation restreint ; Rapidité pour obtenir un graphe lisible.	Superposition de certains sommets, trop similaires ; Lisibilité de l'architecture de moins bonne qualité ; Proximité des liens ne révélant pas totalement leur valeur qualitative ; Difficultés à distinguer les noyaux des structures.
semi paramétrées	Calcul successif des répulsions et attractions ; $\beta = 1{,}5$; c = valeur du slider permettant d'interagir sur la répulsion ; $\alpha_r = 2$.	Distinction des sommets semblables ; Lisibilité des différentes structures améliorées ; Proximité des liens caractérisant vraiment leur valeur quantitative ; Minimisation des superpositions.	Nécessité de quelques retouches manuelles. Certains sommets n'arrivent pas à franchir certains « cols » ; Difficultés de lisibilité si nombre d'arêtes important.
basées sur la paramétrage de l'attraction uniquement	Calcul de toutes les répulsions dans un premier temps puis des attractions ; $\beta = 1{,}5$; c = 1,5 ; $\alpha_r = 2$.	Lisibilité des différentes structures améliorée ; Proximité des liens caractérisant vraiment leur valeur quantitative ; Distinction au niveau structural entre les différents types d'entités ; Les sommets similaires sont fortement attirés, favorisant leur catégorisation.	Nécessité de quelques retouches manuelles. Certains sommets n'arrivent pas à franchir certains « cols » ; Pas de répulsion, minimisant l'interactivité et défavorisant le contrôle des distances au sein des structures.

Tableau 18. *Comparatif des algorithmes FDP proposés.*

Chacun de ces algorithmes répond à un besoin particulier et dans des cas spécifiques chacun a ses avantages et ses inconvénients.

Pour illustrer ces propos, prenons l'exemple de l'étude d'un corpus contenant tous les articles reportant les travaux d'un laboratoire sur une année et plus particulièrement les articles de chercheurs ainsi que leurs thématiques.

Dans le cas où l'analyse porte sur la similitude des acteurs d'une même équipe, collaborant et travaillant sur des sujets similaires, le premier algorithme est préférable. En effet, la présence de caractéristiques communes se traduit par une forte proximité. Dans le cas où l'étude sur ce même type de données porte sur la découverte des différentes équipes et des acteurs dynamiques, le deuxième algorithme est plus favorable. Enfin, si l'utilisateur s'intéresse davantage aux thématiques abordées et aux différents individus traitant de ces sujets, le troisième algorithme est préférable, à mi chemin entre les deux précédents.

La force d'attraction entre les individus peut être qualifiée comme une cohésion entre les membres d'un ensemble pour que celui-ci conserve sa raison d'être, et atteigne les objectifs fixés. La notion de cohésion se précise dans les travaux de (Wasserman et Faust, 1994) où on peut lire qu'un sous-groupe cohésif est un sous-ensemble d'acteurs entre lesquels les relations sont plus fortes, fréquentes, directes ou intenses, que celles qui existent entre ces acteurs et les autres. Selon les relations étudiées, telles que l'affinité, le voisinage, les collaborations, la parenté, etc.…, on peut caractériser la force d'un lien.

3.2.7. Autres types de graphes

✓ ***Graphes orientés***

Nous qualifions une relation orientée entre deux données relationnelles la transmission, de l'une à l'autre. On nomme « arcs » la liaison orientée entre deux sommets.

Un graphe orienté est constitué d'un ensemble fini non vide de sommets X et d'un ensemble de arcs A.

$$A \subset X \times X = \{(x_i, x_j) | x_i, x_j \in X\} \quad [13]$$

A tout graphe orienté $G = (X, A)$, on associe le graphe simple (X, B) où $(x,y) \in B \Leftrightarrow ((x, y) \in A \text{ ou } (y,x) \in A))$

Ce type de graphe exprime la direction en précisant les sources et les destinataires.

Si $a = (x_i, x_j) \in A$, on dit que a est l'arc de x_i à x_j. Si $x_i = x_j$, on dit que a est la *boucle* en x_i. L'*ordre* du graphe orienté G est par définition le cardinal $|X|$ de X et sa *taille* est le cardinal $|A|$ de A.

Les arcs d'un graphe orienté constituent une relation sur l'ensemble de ses sommets.

$$X = \{x_1, x_2, \ldots, x_n\} \text{ et } card(X)=|X|=n \quad [14]$$

$$A = \{a_1, a_2, \ldots, a_m\} \text{ et } card(A)=|A|=m \quad [15]$$

Un chemin c est une séquence d'arcs tous parcourus dans le même sens. Pour qu'un chemin relie deux sommets, un déplacement continu suivant une séquence d'arcs doit être possible. x_1 est l'extrémité initiale ou origine de c et x_j son extrémité finale. La longueur $l(c)$ du chemin est égale au nombre d'arc qu'il comporte.

Par convention, nous appelons *prédécesseur* le sommet x_i, à partir duquel l'arc est tracé, et *successeur* le sommet d'arrivée x_j. Dans l'exemple de la Figure 63, x_1 est le *prédécesseur* et x_2, le *successeur*.

Figure 63. *Prédécesseur et successeur.*

A l'origine d'un graphe orienté, sous VisuGraph, se trouve une matrice croisant les prédécesseurs et les successeurs. Le graphe obtenu permet alors de distinguer clairement qui est à l'origine de qui. Cependant, il est possible d'ajouter des informations à un graphe orienté afin d'enrichir son potentiel informatif.

Par exemple, intéressons nous au cas d'un service de veille d'une entreprise fournissant des licences et cherchant à étudier les forces actuelles du marché dans son domaine. Un graphe orienté croisant les entreprises qui vendent les licences et celles les achetant dans le domaine pharmaceutique est très utile pour le veilleur.

Cependant, ce dernier, peut vouloir disposer sans grand effort cognitif du domaine médical spécifique des entreprises afin de mieux cibler le domaine étudié. Pour cela, VisuGraph permet de compléter le graphe orienté par analyse d'une matrice asymétrique croisant toutes les entreprises, qu'elles soient prédécesseurs ou successeurs avec les pays.

Pour illustrer ce principe, les figures 64 et 65 représentent les croisements des entreprises vendant des licences informatiques de logiciels et celles les achetant. Les deux matrices utilisées pour réaliser le graphe orienté et le résultat obtenu avec la légende permettent de distinguer les états des sommets assimilés aux entreprises. La première matrice permet de générer les sommets, selon leur valeur de métrique et la seconde, permet de générer les liens orientés en considérant les sommets en lignes comme les prédécesseurs et ceux en colonne comme les successeurs.

Nous nommons $P = \{P_1,\ldots P_n\}$, l'ensemble des prédécesseurs et $S = \{S_1,\ldots S_m\}$, l'ensemble des successeurs. Les prédécesseurs correspondent aux entreprises vendant et les successeurs, à celle les achetant.

	S_1	S_2	S_3	S_4	S_5	S_6	…	S_m
P_1	2	2	0	0	0	0	…	0
P_2	0	4	0	2	0	0	…	0
P_3	0	0	1	0	1	1	…	0
…	…	…	…	…	…	…	…	…
P_n	0	0	0	0	0	0	…	0

Figure 64. *Matrice des prédécesseurs X successeurs.*

	USA	UK	Germany	Japan	France	Switzer	…
P_1	1	0	0	0	0	0	0
…	…	…	…	…	…	…	…
P_n							
S_1	0	1	0	0	0	0	0
…	…	…	…	…	…	…	…
S_m	0	0	0	1	0	0	0

Figure 65. *Matrice des {predecesseurs U successeurs} X Pays.*

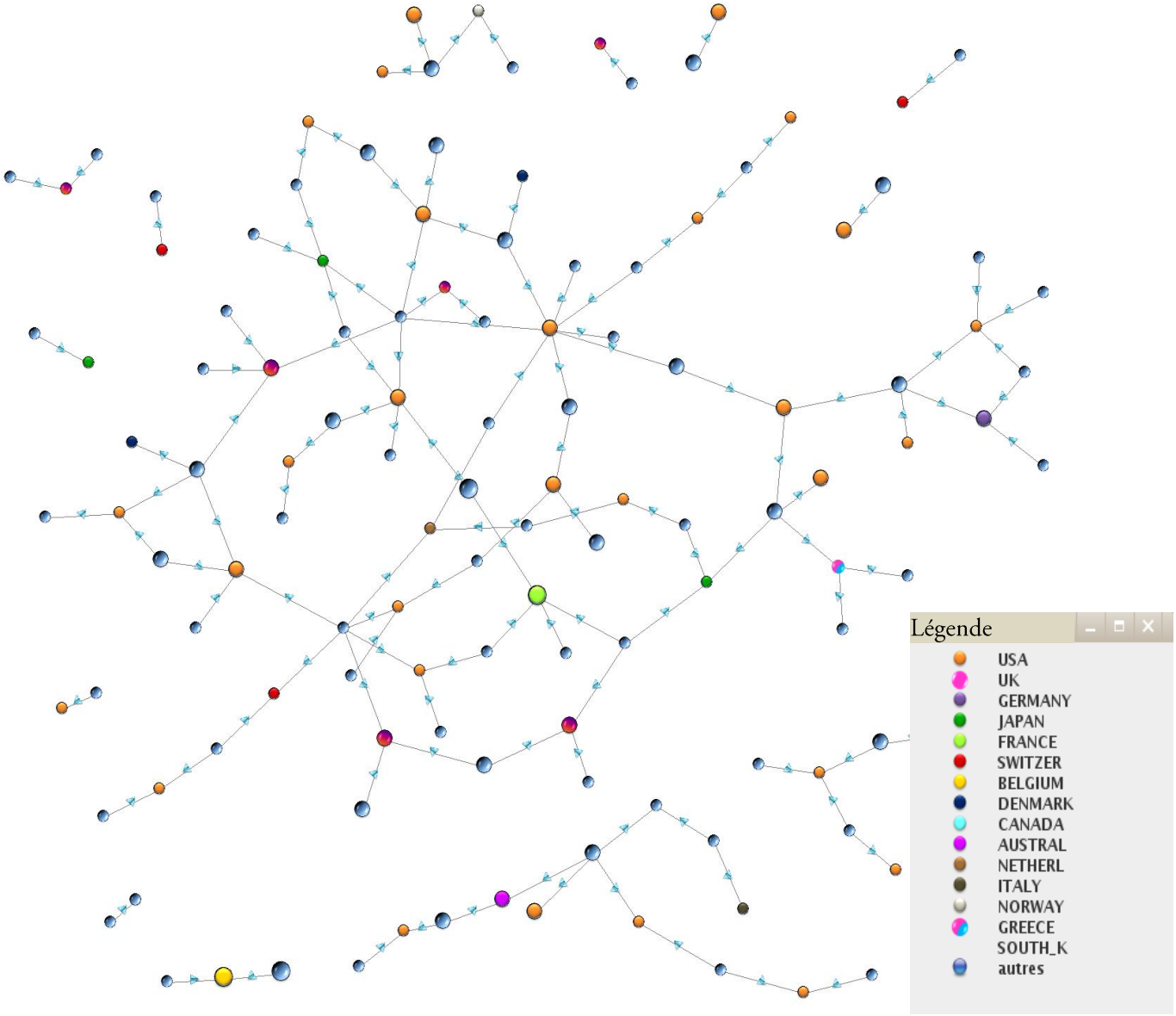


Figure 66. *Graphe orienté biparti.*

✓ ***Graphes spécifiques***

Les graphes proposés par une grande majorité d'outil de visualisation sont soit basés sur le croisement d'une même population, cas des matrices symétriques, soit sur le croisement de deux populations distingues, cas asymétrique. Nous nous intéressons à l'étude et la représentation simultanée des deux types de matrices.

Dans l'outil VisuGraph, nous proposons une nouvelle approche permettant de prendre en compte les liens quantitatifs mais aussi qualitatifs. Pour ce faire, le graphe initial est représenté à partir d'une matrice de croisement symétrique, selon le même mode opératoire que celui décrit dans ce chapitre. A ce stade là, les sommets sont assimilés, comme le montre l'exemple de la Figure 67 à des entités appartenant à la population $X = \{e_1, e_2, e_3, \ldots, e_{10}\}$ et les liens sont quantifiés par les valeurs de croisement de la matrice.

e_n / e_n	e_1	e_2	e_3	e_4	e_5	e_6	e_7	e_8	e_9	e_{10}
e_1	3	1	0	0	0	1	0	0	0	1
e_2	1	3	0	1	0	0	1	0	0	0
e_3	0	0	1	0	1	0	0	0	0	0
e_4	0	1	0	5	0	1	1	1	1	0
e_5	0	0	1	0	2	0	0	0	0	0
e_6	1	0	0	1	0	4	0	0	1	1
e_7	0	1	0	1	0	0	2	0	0	0
e_8	0	0	0	1	0	0	0	2	1	0
e_9	0	0	0	1	0	1	0	1	3	0
e_{10}	1	0	0	0	0	1	0	0	0	1

Figure 67. *Matrice symétrique croisant les individus de X.*

Nous complétons ce graphe, en lisant les informations contenues dans une matrice croisant les entités de la matrice précédente, avec une autre population. Dans le cas de notre exemple, nous croisons les entités avec des pays appartenant à l'ensemble

$$Pa = \{USA, Allemagne, France, Italie, Angleterre, Espagne, Grèce\}, \quad [16]$$

Ainsi, sur un même graphe, il est facile de distinguer à quelles nations sont associées les entités appartenant à *X*.

pays / e_n	e_1	e_2	e_3	e_4	e_5	e_6	e_7	e_8	e_9	e_{10}
USA	1	0	0	0	0	1	0	0	0	1
Allemagne	1	1	0	1	0	0	1	0	0	0
France	0	2	1	0	1	0	0	0	0	0
Italie	0	0	0	5	0	1	1	1	1	0
Angleterre	0	0	1	0	2	0	0	0	0	0
Espagne	0	0	0	0	0	0	0	2	1	0
Grèce	0	0	0	1	0	0	0	0	0	0

Figure 68. *Matrice asymétrique croisant les individus de E et de P.*

La valeur de la métrique des arêtes est par défaut une valeur quantitative, comme le montrent ces deux matrices des Figure 67 et Figure 68. Cependant, une relation entre deux individus peut être de nature qualitative et selon le contexte d'analyse, cette information est primordiale. Nous proposons de compléter cette visualisation en prenant en compte une troisième matrice permettant d'obtenir la nature qualitative des liens.

Nous caractérisons la structure de représentation du graphe résultant $\mathcal{S}_p$, basée sur les trois types de matrices précédemment citées, comme suit :

$$\mathcal{S}_p = \bigcup \{<x_i, a_{q<xi,xj>}, x_j>\}$$

Avec

x_i, x_j l'ensemble des sommets Pour x_i, $x_j \in X$

a_q l'ensemble des arêtes caractérisant qualitativement et quantitativement le lien. Chacune se décompose ainsi:

$$a_{q\ <xi,xj>} = \{<x_i, x_j, valeur\_quantitative, valeur\_qualitative>\}$$

Cette matrice symétrique croise les entités de X, caractérisant

- par une chaîne de caractères la présence et la nature d'un lien,
- par une chaîne vide l'absence de lien.

Les résultats obtenus sont restitués dans la matrice suivante.

e_n / e_n	e_1	e_2	e_3	e_4	e_5	e_6	e_7	e_8	e_9	e_{10}
e_1		Visualisation				Visualisation				RechercheI
e_2	Visualisation			Synonymie		KnowledgeM				
e_3					Collecte					
e_4		Synonymie				Filtrage	KnowledgeM	Collecte	KnowledgeM	
e_5			Collecte							
e_6	Visualisation			Filtrage					Croisements	Collecte
e_7		KnowledgeM		KnowledgeM						
e_8				Collecte					Visualisation	
e_9				KnowledgeM		Croisements		Visualisation		
e_{10}	RechercheI					Collecte				

Figure 69. *Matrice symétrique dont les croisement sont de nature qualitative.*

Au niveau de la visualisation, la représentation des matrices, Figure 70 s'effectue par le recours à deux symboles de couleurs distinctes caractérisant l'appartenance à X ou à P.

Nous considérons l'ensemble des valeurs qualitatives des liens L, constitué des valeurs suivantes :

$$L = \{Visualisation, RechercheI, KnowledgeM, Synonymie, Filtrage, Collecte, Croisements\} \quad [17]$$

A chacune des valeurs de L est attribuée une couleur spécifique permettant l'identification visuelle rapide et efficace de ces dernières. Suite à la lecture des trois matrices, les liens sont caractérisés par une ou deux informations, à savoir la valeur quantitative et la valeur qualitative. Le graphe final est restitué dans la Figure 69. Cet exemple est applicable aussi aux graphes orientés. Dans ce cas là, la matrice croisant les entités de X est symétrique au niveau de la dimension, c'est-à-dire le même nombre de lignes et de colonnes, mais asymétrique au niveau du contenu, les intitulés des entrées ne sont pas forcément les mêmes en vertical et en horizontal. La valeur des croisements révèle dans ce cas les sommets prédécesseurs et les successeurs.

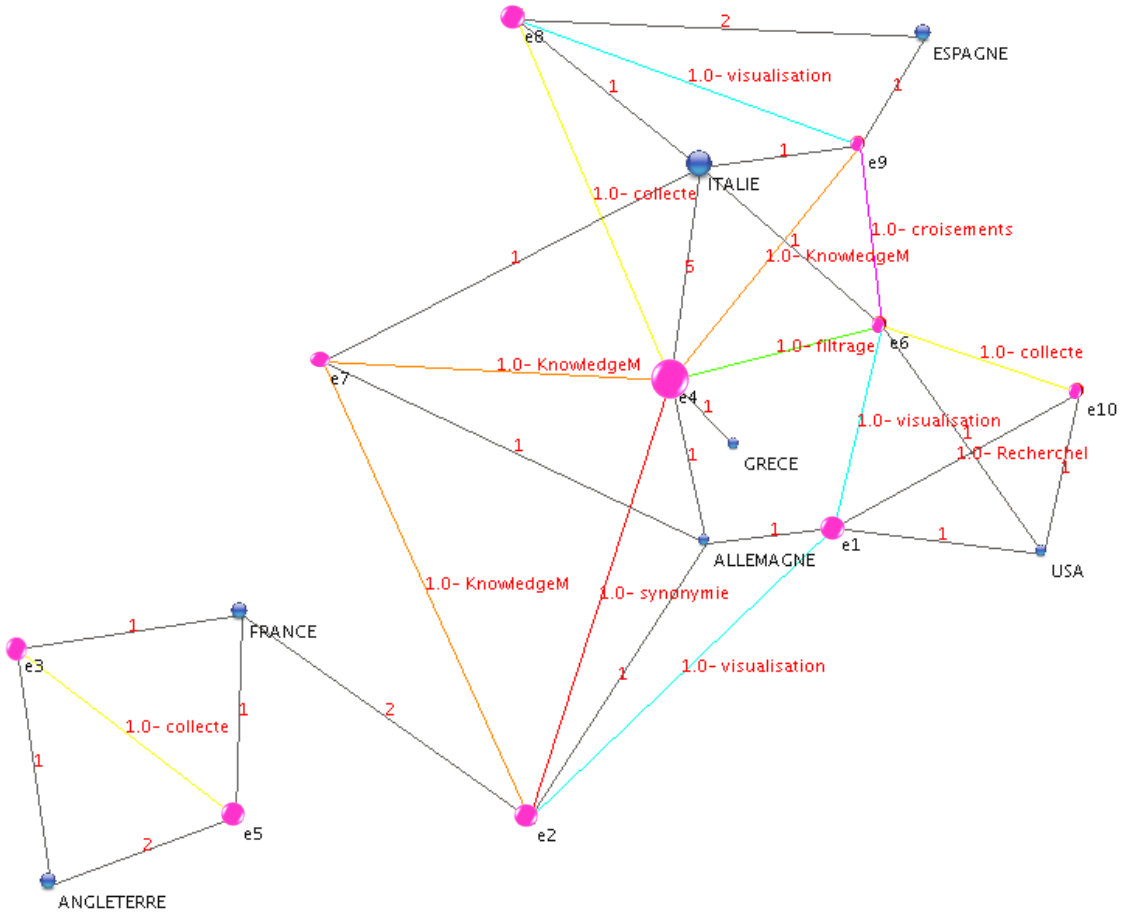


Figure 70. *Graphe basés sur des matrices symétrique et asymétrique dont les liens sont valués qualitativement et quantitativement.*

De nombreuses combinaisons de graphes sont abordables dans VisuGraph, permettant à l'utilisateur de choisir sa représentation selon son besoin. Pour ce faire, un menu en entrée du programme demande à l'utilisateur de préciser le type de visualisation qu'il souhaite, à savoir, un graphe orienté ou non, le nombre de matrices utilisées et les fonctionnalités voulues, telles que la coloration des arêtes qualitatives.

✓ ***Représentation caméléon***

Nous avons vu dans la section 3.2.6, la disponibilité au sein de VisuGraph, d'algorithmes FDP, permettant de mieux analyser le voisinage des sommets. Dans le cas des graphes orientés, il est possible de réaliser une attraction autour des sommets prédécesseurs, en les fixant et en déplaçant les successeurs autour de ces derniers, et inversement. Nous nommons cette fonctionnalité, la *représentation caméléon*, par assimilation à cet animal qui reste immobile et attrape ses proies à distances avec sa langue. Le caméléon est matérialisé par le nœud prédécesseur et les proies sont assimilées aux sommets successeurs.

L'algorithme utilisé reste le même que celui étudié dans la partie précédente à la différence que le déplacement final ne concerne que les sommets mobiles, c'est-à-dire appartenant à l'ensemble, prédécesseur ou successeur choisi. Notons S l'ensemble des sommets « caméléons », qui resteront immobiles.

```
Pour tout sommet u{
    Si u est visible
    Alors {
        Calcul de la distance d(u,v) ;
        Pour tout sommet v{
            Calcul des forces répulsives ;
        }
        Pour tous les voisins v de u{
            Calcul des forces d'attraction entre u et v ;
            Si (!u ∈ S) alors changement_coordonnées(u) ;
        }
    }
}

 /* Vérification de la non superposition des sommets par comparaison des coordonnnées */
Pour tout sommet u{
    Pour tout sommet v{
        Si ((xu,yu) == (xv, xy) && (!u ∈ S))
        alors changer les coordonnées de v ;
    }
}
```

Principes de l'algorithme Caméléon.

Par cette fonctionnalité, il est facile de caractériser les prédécesseurs. Dans le cas de notre exemple de la Figure 70, des entreprises sont croisées et appartiennent à une des deux catégories : les prédécesseurs vendent des licences aux successeurs.

Cette fonctionnalité permet de distinguer clairement les sociétés achetant dans des compagnies différentes, dans le cas où un même sommet possède plusieurs liens vers des sommets différents. Dans le cas inverse, si une compagnie ne possède qu'un seul fournisseur, le lien est unique et l'algorithme caméléon favorise la proximité entre les deux compagnies, à savoir celle achetant et celle vendant.

Au-delà d'une grande diversité de visualisations proposées, VisuGraph est complété par de nombreuses fonctionnalités permettant l'exploration des structures des graphes.

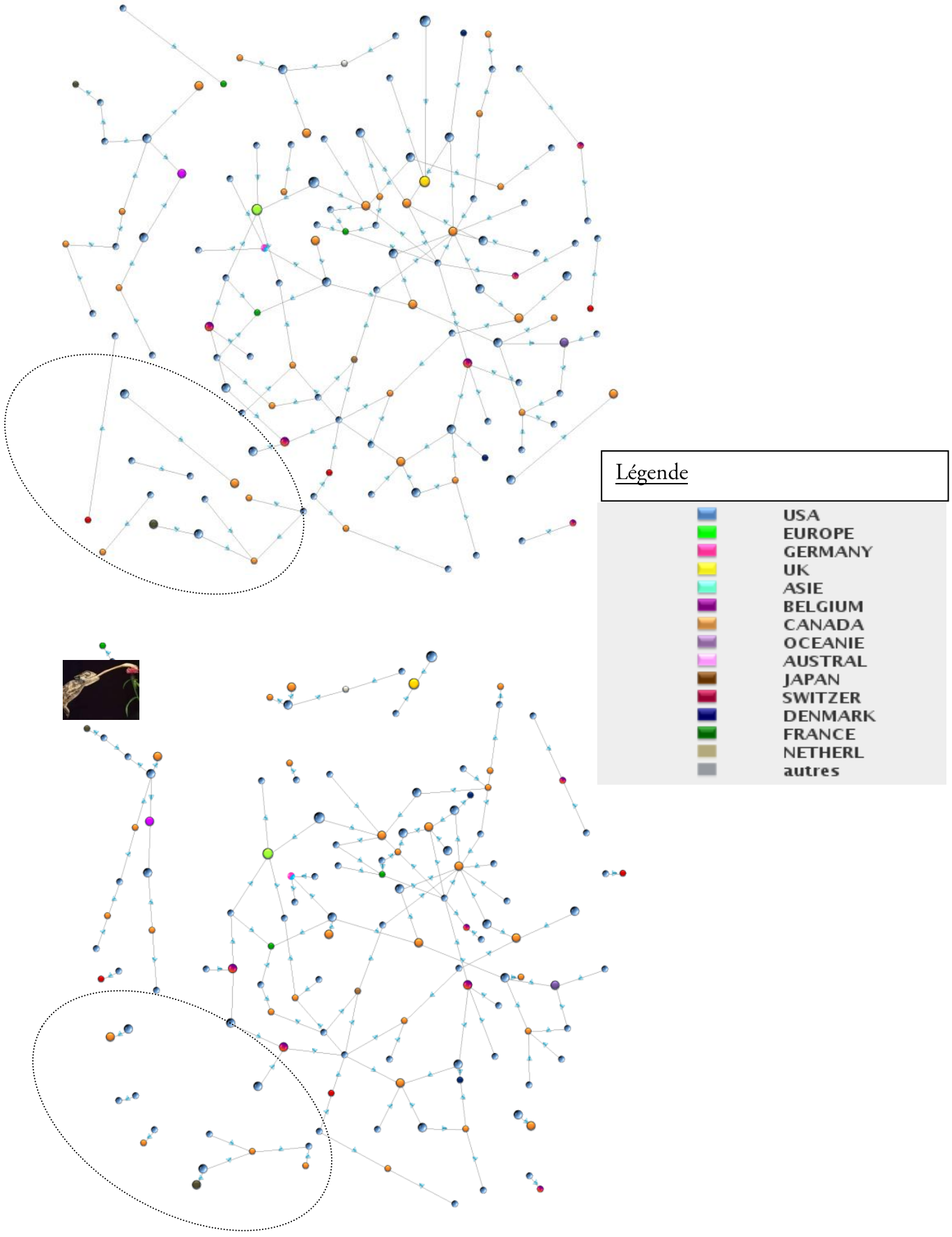


Figure 71. *Application de l'algorithme Caméléon sur le graphe orienté.*

3.3. Les fonctionnalités

Les différentes fonctionnalités proposées dans l'outil VisuGraph sont présentées dans cette section. Il s'agit du filtrage, du K-core, de la transitivité, du retour aux documents, de la focalisation, ainsi que de la représentation caméléon, permettant le rapprochement des sommets successeurs de leurs prédécesseurs, dans le cas des graphes orientés et enfin du partitionnement de graphe (Loubier et Dousset, 2008c).

3.3.1. Positionnement de l'utilisateur

La communication à l'utilisateur de l'information est au cœur de nos préoccupations à travers l'outil VisuGraph.

Pour mieux comprendre les enjeux de la visualisation, d'un point de vue de l'interactivité entre l'homme et le système, il est important de définir ce concept.

Définition : Chacune des deux entités en présence, l'homme et la machine, est un « système » possédant sa propre logique de fonctionnement et qui communique par le biais d'un dispositif appelé « interface » (Durand, 1979).

Dans le cas de l'interface homme-ordinateur, la connexion a lieu entre l'image du système, c'est-à-dire sa manifestation externe et les organes sensorimoteurs de l'utilisateur. La traduction s'effectue entre les formalismes du système et ceux de l'utilisateur. Du côté système, une traduction est assurée entre les représentations internes adaptées au traitement du problème et la représentation externe qui participe à la définition de l'image. Une opération analogue se produit du côté de l'utilisateur entre la représentation mentale de la situation perçue ou à atteindre et les actions physiques à entreprendre.

Deux classes de modèles d'usagers se distinguent :

- Les modèles quantitatifs et empiriques : leur but est de modéliser le comportement externe de l'usager ;
- Les modèles analytiques et cognitifs : leur but est de modéliser le comportement interne de l'usager : connaissances, processus cognitifs etc...

Dans l'outil VisuGraph, nous nous intéressons aux besoins informationnels de l'usager. Nous recherchons donc les facteurs qui déterminent ou conditionnent ces besoins (Loubier et Carbonnel, 2007b), (Loubier et al., 2009).

La tâche de l'utilisateur est l'acquisition de connaissances pour une prise de décision ultérieure (Tamine et al., 2007).

« *L'aspect d'interactivité au sein du système engendre la démarche de l'utilisateur-décideur dans la stratégie de navigation dans un environnement complexe, qui se situe dans un continuum entre la requête (searching) qui correspond à un but précis et bien défini, et le butinage (browsing) qui correspond à un besoin mal exprimé initialement qui conduit l'utilisateur à poursuivre son chemin jusqu'à une certaine satisfaction* » (Kuntz, 2003).

Comme pour certains systèmes interactifs d'aide à la décision, la procédure ne s'arrête donc pas nécessairement avec un test de convergence, mais s'arrête parce que l'utilisateur a le sentiment d'avoir obtenu suffisamment d'informations utiles pour son problème (Vincke, 1992). L'activité, la connaissance du monde, les processus cognitifs sont autant de facteurs à prendre en compte (Kaur Padda et al., 2009).

L'outil VisuGraph repose sur la visualisation interactive des graphes, incluant l'utilisateur-décideur dans la fouille de données. La visualisation proposée est manipulable par l'utilisateur afin qu'il comprenne l'espace des informations et qu'il communique avec le système. L'interaction a lieu à travers la visualisation qui tient lieu d'interface entre l'utilisateur et les données.

L'identification du besoin d'information nous amène à nous pencher sur une modélisation de l'usager lors d'une recherche d'information. Les travaux de (Daniels, 1986) prennent en considération le fait que la modélisation de l'usager, dans un contexte plus général, est constituée de cinq sous-fonctions:

USER: le statut de l'usager

UGOAL: détermine les buts de l'usager

KNOW: les états de connaissances de l'usager dans un domaine

IRS: la familiarité de l'usager avec le système documentaire

BACK: l'expérience de l'usager.

Ces cinq sous fonctions appartiennent toutes aux systèmes cognitifs de l'usager. Nous avons tenu compte de ces aspects dans la validation de notre outil Visugraph que nous détaillerons dans le chapitre 5.

Afin que l'utilisateur soit totalement maître de la représentation graphique des données relationnelles, il est au cœur de toutes les décisions d'application et de paramétrage de toutes les fonctionnalités.

Pour ce faire, nous avons établi plusieurs règles telles que :

- Le menu en deux parties : une première expose les différentes fonctionnalités, que l'utilisateur activera selon ses besoins et une autre permet de régler ces dernières via des sliders spécifiques.
- L'affichage d'indications dans la console accompagnant la fenêtre de visualisation.
- Le choix de l'utilisateur de la couleur du fond de la visualisation, des formes symbolisant les données, accentuation de l'intensité des liens et/ou de la police, ou encore de la taille de cette dernière.
- La mise en place d'un onglet d'aide avec un manuel utilisateur sous forme de document texte, illustré.
- Le déplacement manuel des sommets, via la souris.

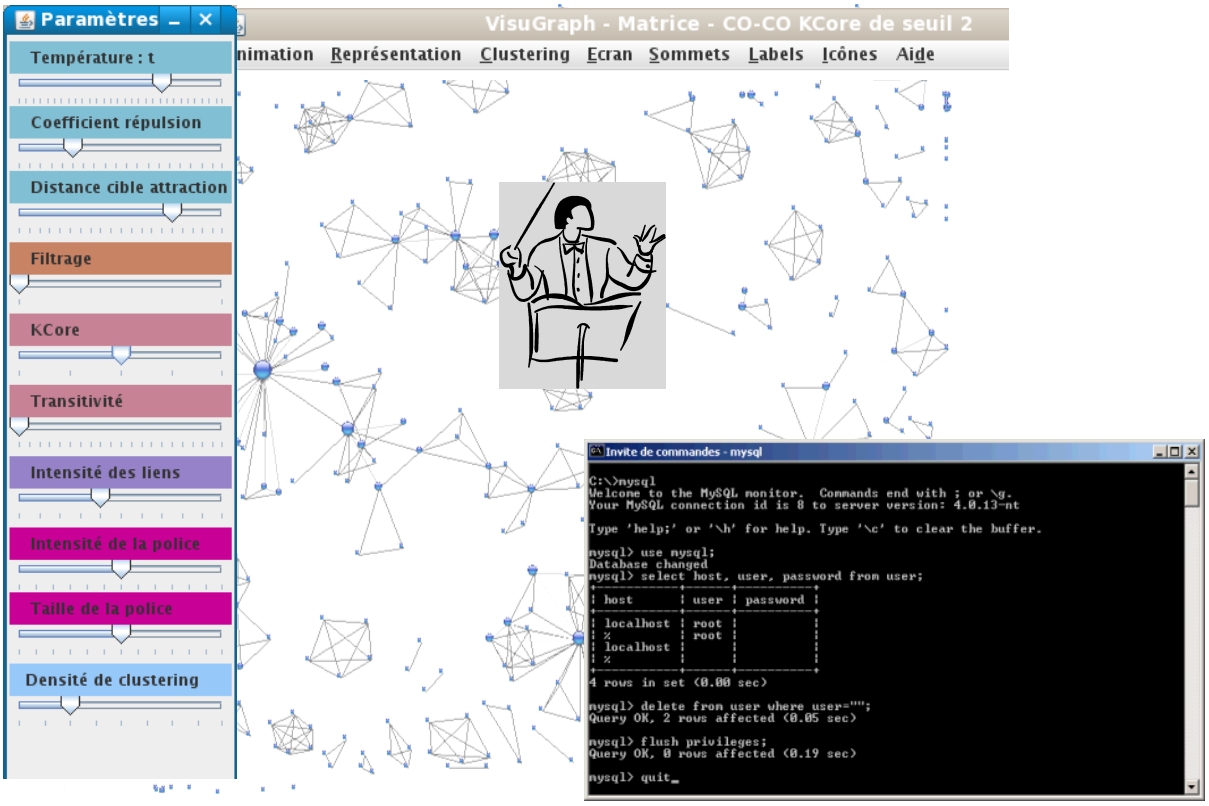


Figure 72. *L'utilisateur, contrôlant la visualisation des données relationnelles.*

3.3.2. Identification et analyse de structure de graphe

La visualisation et l'étude de graphe consistent à analyser le réseau formé par ces entités et la combinaison de leurs relations, afin de comprendre la façon dont la structure contraint les comportements individuels tout en faisant émerger des interactions. L'analyse structurale tente de trouver les régularités de comportement. Plus un individu est proche des autres, plus il est susceptible d'avoir d'informations (Leavitt, 1951), d'accéder à un plus haut statut social (Katz, 1953), d'avoir du pouvoir (Coleman, 1973), de l'influence (Bavelas, 1950 ; Friedkin, 1991), du prestige (Burt, 1992). Un graphe se caractérise d'abord, très simplement, par son ordre, c'est-à-dire par le nombre de ses sommets indiqué dans la console.

Le concept dominant de l'analyse structurale est celui de système, c'est-à-dire qu'il s'agit de rechercher les formes structurelles du système (Eve, 2002). Pour étudier la proximité des sommets, plusieurs critères sont alors utilisables.

- La connexité consiste à repérer des groupes dont les membres sont liés de façon directe ou indirecte ;
- la cohésion s'appuie plutôt sur la densité des relations dans le groupe ;
- l'équivalence introduit un autre point de vue en permettant de rassembler les individus en fonction de leur similitude.

On peut aussi vouloir caractériser chaque acteur d'après sa position dans le graphe, par exemple selon sa centralité. Les études qui utilisent ces notions relèvent de la théorie des graphes (Wasserman et Faust, 1994). Si l'on dispose seulement de données décrivant les réseaux personnels d'un échantillon d'individus, généralement choisis pour être représentatifs d'une population plus large. Il n'est pas impossible de tester l'influence de certaines caractéristiques structurales sur le problème traité.

Par exemple, une question sur la fréquence des relations permet de départager approximativement entre liens forts et liens faibles, distinction dont les travaux de (Granovetter, 1973) ont montré l'importance.

Le travail de description consiste à inventorier la diversité des régimes d'action et des entités mises en relation dans le réseau. Ainsi, les individus les plus centraux dans un graphe occupent des positions privilégiées dans les échanges, notamment par rapport à ceux qui sont situés plus à la périphérie.

Une relation, et plus globalement des relations formant un réseau, peuvent non seulement apporter une information, que l'expert est totalement libre de prendre en compte ou d'ignorer, mais peuvent aussi appuyer celle-ci par une force qui pousse l'expert qui la subit à accepter cette information comme pertinente et donc à modifier son système cognitif. On parle ainsi d'influence lorsque la relation entre deux acteurs comporte à la fois la circulation d'un contenu et l'exercice d'une force visant à imposer ce contenu à l'un des acteurs. Ces processus d'influence, qui déterminent notamment l'adoption des innovations, ont été examinés dès les premiers pas de l'analyse des graphes. Les travaux de Coleman Katz et Menzel (1966) sur la diffusion d'un nouveau médicament ont constitué un point de départ pour une longue tradition de recherche dont les travaux de (Valente, 1995) synthétisent les apports récents. Dans le domaine politique, le réseau peut aussi exercer une influence (Knocke 1990), (Nieuwbeerta et Flap, 2000).

L'approche de (Freeman, 1979) repose sur trois définitions pour la centralité :

La *centralité de degré*, notée C_D, considère le degré d'un sommet, c'est-à-dire le nombre de connexions directes irriguant ce dernier. Un individu est d'autant plus central qu'il est directement lié à un grand nombre d'autres sommets. La centralité de degré mesure localement la capacité d'un individu à communiquer, indépendamment de la centralité des individus auxquels il est directement lié.

L'acteur le plus central est le plus actif, d'un point de vue relationnel. La centralité est donnée par la formule suivante, pour n le nombre de sommets et $d(x_i)$ le degré du sommet i :

$$C_D(x_i) = \frac{d(x_i)}{n-1} \quad [18]$$

La centralité de proximité, notée C_P, donne un point de vue plus global à la centralité puisqu'elle considère la proximité d'un individu avec tous les autres, sur une notion de distance. $d(x_i, x_j)$ étant la distance entre deux acteurs mesurée en nombre minimal de liens. Dans le cas des graphes non orientés, la centralité de proximité est formalisée par la formule suivante,.

$$C_P(x_i) = \frac{n-1}{\sum_{j=1}^{n} d(x_i, x_j)} \quad [19]$$

La centralité d'intermédiarité, notée C_I, défend l'idée qu'un individu peut être faiblement connecté aux autres et même relativement éloigné, mais servir d'intermédiaire dans bon nombre d'échanges entre les autres membres du groupe (Ford et Fulkerson, 1956), (Freeman et al., 1991). Plus il sert ou peut servir d'intermédiaire pour tous les membres, plus il est en position de contrôler la communication ou d'être indépendant des autres pour communiquer.

Un tel individu peut influencer le groupe plus facilement en filtrant ou distordant les informations qui y circulent. Ainsi, pour deux sommets non adjacents t et j qui communiquent si le sommet i se trouve sur leur chemin de communication, alors i est un acteur important. Sa position lui permet également d'assurer la coordination du groupe.

Soit p_{jt} le nombre des chemins les plus courts entre j et t, $p_{jt}(x_i)$ le nombre de chemins les plus courts entre j et t passant par i.

$$C_I(x_i) = \sum_{j=1}^{n} \sum_{j<t}^{n} \frac{p_{jt}(x_i)}{p_{jt}} \quad [20]$$

Les centralités selon la proximité ou l'intermédiarité mesurent la capacité d'un individu à contrôler cette communication qui ne dépend pas forcément du nombre de ses liens avec ses voisins, mais de son rapport à l'ensemble des membres du réseau : rapport de proximité ou d'intermédiarité. Une forte centralisation de connexion est l'indice d'une communication active tandis qu'une forte centralisation de proximité ou d'intermédiarité traduit un petit nombre d'acteurs contrôlent cette communication.

Un graphe, se caractérise non seulement par sa *densité*, mais aussi par sa *connexité*. On mesure la densité d'un graphe par le rapport entre le nombre d'*arcs* de ce graphe et le nombre d'arcs que comporte le *graphe complet* ayant le même nombre de *sommets*. Si un graphe n'est pas connexe, ses parties qui le sont, sont appelées ses *composantes connexes*. Un graphe qui n'est pas connexe, peut être extrêmement dense, par exemple s'il est constitué d'une clique importante et de quelques sommets.

L'analyse de la structure d'un graphe sert à la construction d'indicateurs prometteurs pour caractériser la dynamique d'un ensemble de données relationnelles (Ghalamallah et al., 2007), (Ghalamallah et al., 2008), (Ghalamallah et al., 2008b), (Guenec et al., 2008). Elle sert à caractériser individuellement les données représentées qui assurent ou au contraire réduisent la cohésion globale.

Lorsque le graphe est trop complexe, le recours au masquage des éléments par filtrage facilite l'exploration de la structure.

3.3.3. Filtrage

Une première méthode pour analyser plus simplement un graphe est le filtrage. Filtrer un graphe revient à filtrer ses sommets ou ses arêtes selon certains critères. Ces derniers sont basés sur les propriétés quantitatives ou qualitatives des sommets ou arêtes (Henry 1992), (Huang et al., 2005). Un nettoyage préalable du graphe, basé sur une technique de filtrage appropriée (Huang et al. 2005) permet de révéler une structure et des motifs intéressants. La difficulté est de proposer un filtre qui révèle des caractéristiques sans pour autant dénaturer le graphe.

Dans VisuGraph, le filtrage dynamique, basé sur les valeurs de la métrique utilisée, consiste à ne conserver que les sommets et les arêtes du graphe associés aux valeurs supérieures ou égales à un seuil. La dynamique du filtrage est un concept fondamental de la visualisation de l'information. Grâce au filtrage, l'utilisateur peut contrôler le volume des contenus à afficher pour se concentrer sur ce qui l'intéresse. Cette procédure fait apparaître les sommets les plus représentatifs, ainsi que les composantes importantes de la structure. Dans notre cas, le filtrage s'effectue par masquage des arêtes ayant une valeur de métrique inférieure au seuil fixé par l'utilisateur. De façon générale, on peut effectuer le filtrage d'arêtes de valeur de métrique inférieur à k de façon à ne garder que les arêtes de forte valeur.

Ainsi, la fonction de filtrage de liens se formalise de la façon suivante.

$$f_F(G, Filtre=n)=G^F$$

où G est un graphe, selon la définition initialement définie dans laquelle $G = (X, A)$ avec X l'ensemble des sommets et A l'ensemble des liens.

G^F est le graphe résultant, c'est-à-dire dont les liens sont filtrés.

n est la valeur du seuil du filtre, c'est-à-dire la valeur minimale que devra prendre la valeur quantitative de $a_{<xi,xj>}$, l'arête comprise entre les sommet x_i et x_j, dans le graphe filtré G^F.

Ainsi, le filtrage d'un graphe permet d'obtenir un sous-graphe, comportant des sommets isolés, qu'il est possible de masquer.

Le filtrage permet aisément de détecter la nature des liens entre les acteurs du graphe. Ainsi, un graphe composé majoritairement de liens unitaires peut donner des indications riches sur les échanges entre les données. Si nous prenons l'exemple illustré en Figure 73, il s'agit d'un graphe des auteurs ayant coécrit sur le domaine de la veille. Le premier graphe, à gauche, est réalisé sans filtrage ; le second est filtré par l'utilisateur, via le slider de filtrage, à un seuil égal à 1, ce qui signifie que seules les liaisons symbolisant un nombre de co-signatures supérieur à 1 sont conservées. Par ce filtrage, on distingue clairement que la majorité des auteurs ayant co-écrits ne se sont associés que pour un unique article, tandisque dans le graphe de droite, seuls les auteurs liés par plusieurs publications.

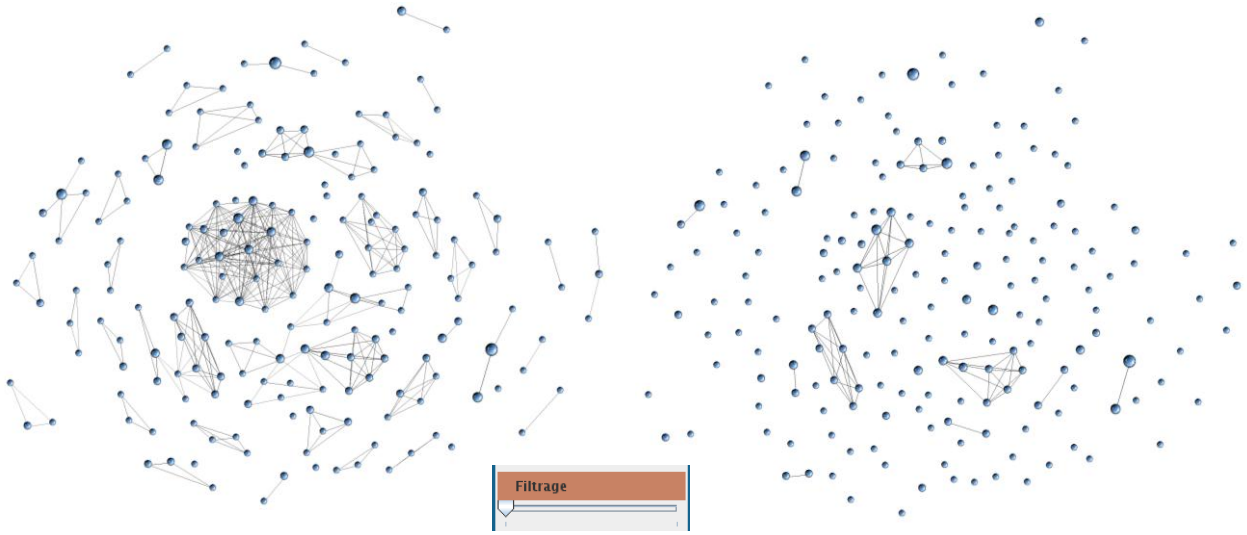


Figure 73. *Filtrage des liens du graphe de co-signatures d'articles scientifiques.*

Le filtrage permet de conserver les liens les plus forts, selon l'importance du seuil fixé via le slider, il est important de disposer de méthodes complémentaires permettant de ne sélectionner que les sommets ayant le plus grand nombre de voisins, c'est-à-dire les nœuds les plus liés. Le filtrage n'étudie que les valeurs des liens en masquant les plus faibles mais ne permet pas de distinguer les nœuds au centre de la structure de graphe. Dans la Figure 74, la classe connexe la plus grande est divisée lors de l'application du filtrage, prenant en compte la valeur quantitative du lien mais pas le nombre de liaison. En effet, un sommet peut être très important, du fait qu'il soit joint à d'autres par des liens de faible valeur. Le filtrage ne permet pas la détection de ces sommets. Pour cela, le *k-core* vient compléter le filtrage.

3.3.4. *K*-Core

Une autre fonctionnalité permettant l'étude de la structure d'un graphe est le *k*-core. Cette décomposition (Batagelj et Zaversnik, 2002), consiste à identifier des sous ensembles particuliers du graphe appelés *k-core*..

Un *k-core* est défini comme suit :

> Un sous-graphe $S = G(C, A|C)$ induit par l'ensemble $C \subseteq X$ est un *k-core* ou un *core* d'ordre k si et seulement si $\forall v \in C: degre_H(v) \geq k$, et S est un sous ensemble maximal avec cette propriété.
>
> Notons que le *k-core* est unique (Alvarez et al., 2005).
>
> Un nœud a un *coreness c*, s'il appartient au *core* d'ordre *c* et s'il n'appartient pas au *core* d'ordre $(c + 1)$.
>
> Un ensemble connexe de *coreness* c_E forme un cluster, ou encore une communauté, au sens de (Alvarez et al., 2005).
>
> Le *k-core* est obtenu par élagage récursif des nœuds qui ont un degré plus petit que *k*. Le graphe restant ne contient que des sommets de degré ≥ *k*.

Dans la Figure 74, les sommets qui n'ont qu'un seul voisin correspondent à un *coreness* de 1. Si nous nous intéressons au *coreness 2*, les sommets appartenant au *1-core* sont masqués. Nous obtenons alors les sommets qui ont au moins deux sommets voisins. Si nous ciblons les sommets correspondant au *3-core*, nous élaguons les sommets des précédents *coreness* et nous obtenons les sommets qui ont plus de trois voisins.

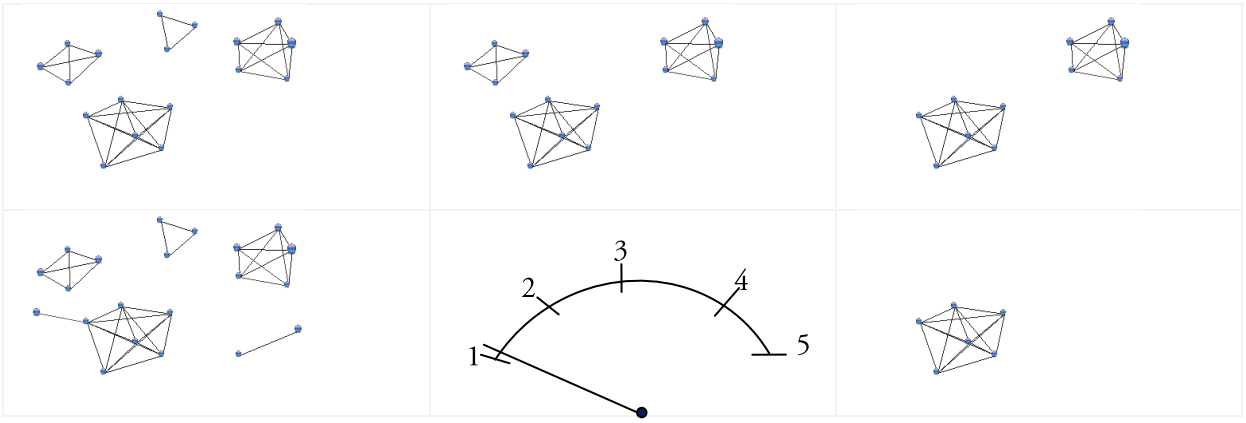


Figure 74. *Décomposition en k-core d'un graphe. Figure réalisée sous VisuGraph (en bas à gauche 1-core, en haut à droite 2-core, au milieu 3-core, en haut à droite 4-core, en bas à droite 5-core).*

Appliqué à VisuGraph, le *k-core* est calculé à partir d'un seuil fixé par l'utilisateur (Loubier, 2009b). Plus ce seuil augmente plus le *coreness* est élevé. Le *k*-core permet de cibler le cœur du graphe, au détriment de sa périphérie.

Nommons f_c cette fonction :

$$f_c(G, seuil=n)=G^k$$

G est le graphe de base, selon la définition $G = (X, A)$ où X est l'ensemble des sommets et A l'ensemble des liens.

n est le seuil permettant de calculer le *k-core* et permettant de caractériser le *coreness*.

$$n = \{0,\ldots,c\}.$$

c est une valeur calculée dès le premier appel de la fonction. Elle correspond au k maximal pouvant être atteint.

G^k est le graphe obtenu pour le seuil fixé, composé de sommets ayant tous au moins n voisins.

On peut décrire la *périphérie* d'un ensemble connexe de nœuds comme la partie externe de cet ensemble.

Les *cœurs du graphe* sont constitués de l'ensemble des nœuds et arêtes restant après avoir atteint le *k-core* maximal. Il constitue le noyau du graphe, à savoir les acteurs les plus importants et par conséquent les plus influents. Dans un contexte concurrentiel, il s'agit des données clés à l'origine de la communication entre toutes les autres. La suppression d'un des sommets du cœur du graphe provoque la déstabilisation de la structure.

Il devient plus aisé de répondre aux concepts généraux proposés par (LeMoigne, 1984) pour définir un système à savoir.

Quelque chose (n'importe quoi, présumé identifiable)

Qui **dans** quelque chose (environnement)

Pour quelque chose (finalité ou projet)

Fait quelque chose (activité, fonctionnement)

Par quelque chose (structure=forme stable)

Qui se **transforme** dans le temps (évolution)

L'analyse du graphe, via des techniques précises telles que les *k-cores* permettent de comprendre l'objet dans son ensemble.

Cependant, une fois une structure détectée, il est important de détailler sa composition afin d'obtenir les principaux acteurs. Pour cela, la transitivité est une fonctionnalité permettant l'exploration des voisins d'un sommet.

3.3.5. Transitivité

La transitivité s'exprime assez justement par *« les amis de mes amis sont mes amis »*.

Dans VisuGraph, cet algorithme s'applique à partir d'un sommet sélectionné par l'utilisateur et par la fixation d'un seuil à l'aide d'un curseur.

Nommons f_t la fonction de transitivité, appliquée à un graphe $G = (X,A)$.

$$f_t(G, x_i, seuil=n)=G'$$

x_i est le sommet initialement choisi, appartenant à X.

n est le seuil fixé pour le calcul de la transitivé. $n = \{1,\ldots,r\}$. r est une variable fixée dès le premier appel de la fonction. Il s'agit du nombre de pas permettant d'obtenir le voisin le plus éloigné du sommet x_i.

L'analyse de graphe est le moyen d'analyser des structures et de s'interroger sur leurs rôles (Mercklé, 2004). Au-delà de la méthodologie (Lazéga, 1998), il s'agit de comprendre en quel sens une structure contraint concrètement des comportements, tout en résultant des interactions (Degenne et Forsé, 2004) entre les éléments qui la constituent.

Dans l'algorithme suivant, S est le sommet initialement choisi, sur lequel est calculée la transitivité. G correspond au graphe $G = (X, A)$. n est le seuil de transitivité fixé par l'utilisateur. *num(SV)* indique le nombre minimum de sommets intermédiaires entre le nœud initialement choisi et *SV*.

```
calculeTransit(G, S){
        int cpt=0;
        getVoisins(S, n);
        Pour tous les voisins directs de S
        {
                SV = Voisin(S);
                if (num(SV)==0){
                        cpt=S_init.num;
                        getVoisins(SV,n);
                        Pour tous les voisins directs de SV
                        {
                                T = Voisins(SV);
                                if ((num(T) < cpt) && (num(T) !=0)) cpt = num(T);
                        }
                        num(SV)=cpt+1;
                        if (num(SV)>init_max) init_max=sv.num;
                        if (NombreVoisibs(SV)>1)calculeTransit(G,SV);
                }
        }
}
```

Algorithme de calcul de la transitivé d'un graphe.

L'étude d'un sommet particulier et de ses relations avec d'autres sommets par le biais de ses connexions permet de définir son rôle au sein de la structure. La donnée visualisée apparaît comme un acteur majeur du domaine.

Si des sommets *A* et *C* sont liés au sommet *B*, ils détiennent peut-être des caractéristiques communes, mais aussi des comportements proches ou « compatibles ».

Le voisinage d'un acteur, ou réseau égo-centré, est l'instrument majeur permettant d'observer les formes de rapprochement que l'individu opère entre des relations, des ressources et des références différentes. En se centrant sur l'analyse des réseaux égo-centrés, le chercheur peut restituer la diversité des relations et préserver le caractère local de l'espace dans lequel elles se développent.

« Nous ne pouvons pas comprendre les interactions d'un groupe donné d'individus si nous ne les considérons pas à la lumière de l'ensemble des liens que chaque acteur entretient en dehors de l'espace commun » (Gribaudi, 1998).

Dans un premier temps, l'analyse du graphe global où toutes les données sont représentées permet de distinguer les tendances générales, à savoir :

- les positions remarquables et stratégiques,
- la présence d'une ou plusieurs grosse(s) structure(s),
- la détection de sommets isolés,
- la détection d'acteur important dont la valeur de la métrique est supérieure à la moyenne.

Dans la Figure 75, il s'agit du graphe des auteurs ayant proposé et présenté un article lors du congrès de Veille Stratégique, Scientifique et Technique (VSST). Ce graphe repose sur la totalité des congrès. Les auteurs sont représentés sous forme de sommet et la coécriture est symbolisée par les liens entre les sommets.

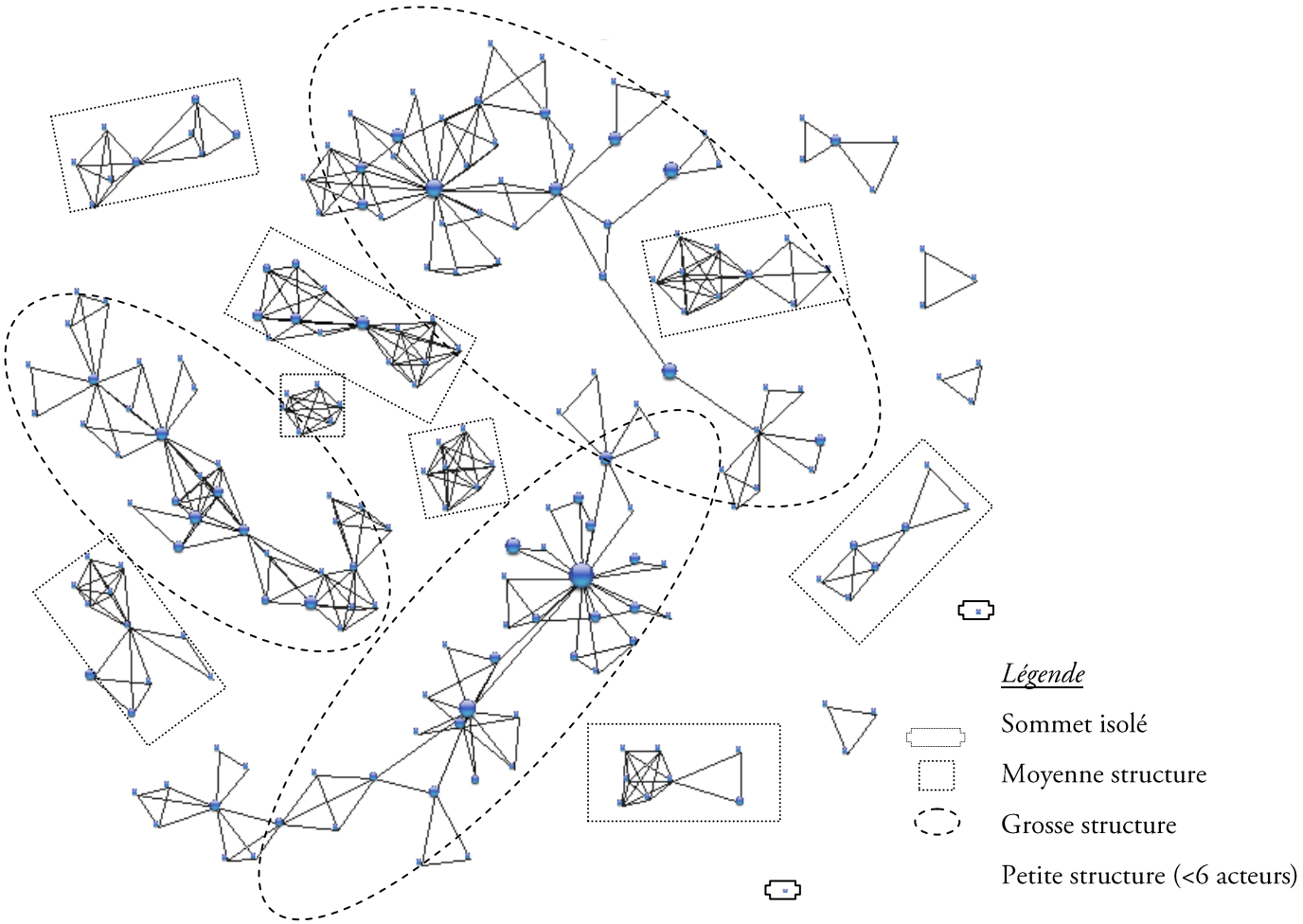


Figure 75. *Décomposition de la structure d'un graphe.*

Ce graphe, de la figure 75, fait ressorti très clairement :

- trois très grandes équipes de recherche ;
- des équipes de taille moyenne ;
- des auteurs uniques ;
- des articles écris en petit comité, mettant en avant que deux ou trois auteurs.

Nous caractérisons l'ensemble des sommets de la structure en plusieurs catégories :

- Les sommets individuels, caractérisés par une forte valeur de métrique, symbolisant son importance dans le domaine étudié.
- Les sommets isolés de faibles valeurs, n'appartenant à aucune structure.
- Les sommets appartenant à une structure et se trouvant à l'extrémité de la structure. Ces éléments sont caractérisés comme membre d'une équipe mais pas comme leader. Ces sommets sont caractérisés comme communiquant peu avec le reste du graphe, avec un nombre de liens faible, même si ces derniers peuvent avoir une valeur de métrique forte. L'étude de leurs transitivités est intéressante puisqu'elle permet de reconstituer l'ensemble de l'équipe et de connaitre le nombre d'intermédiaire entre deux extrémités de la structure. La Figure 76 illustre ce principe.

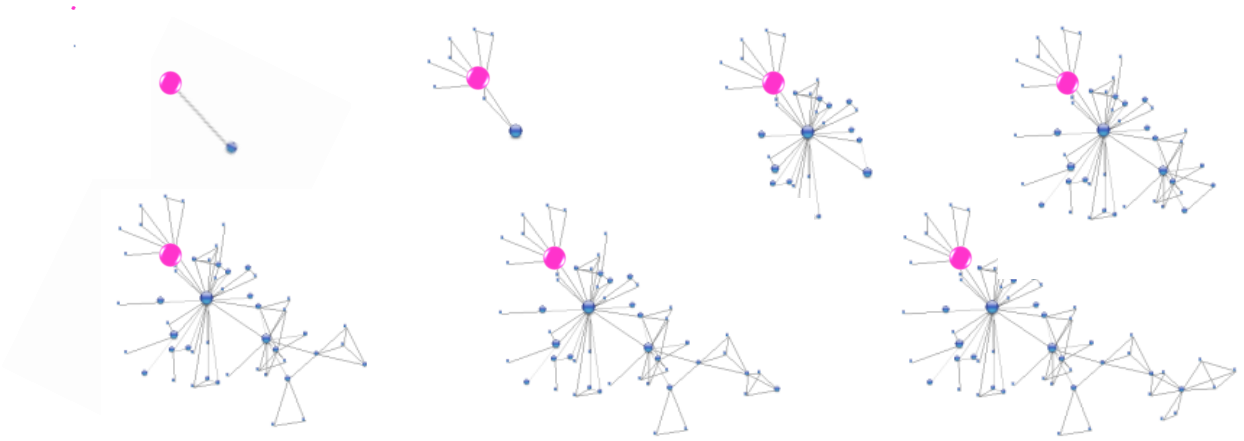


Figure 76. *Calcul des seuils de transitivité d'un sommet en « fin de structure », représenté en rose pour la distinguer des autres sommets.*

Dans cet exemple, la structure est reconstruite par transitivité en sept étapes. Le premier seuil révèle une liaison qu'avec un seul sommet. Puis, par voisinages successifs la structure complète peut être analysée.
Les sommets au centre de la structure sont caractérisés par de nombreux liens avec les autres membres. Leurs suppressions entrainent la rupture en deux. La transitivité permet d'une part d'obtenir l'équipe entière, tout en étudiant le nombre prédominant de liens avec les autres acteurs. Burt nomme ces rotules des « trous sociaux » (Burt, 1992), c'est-à-dire la théorie selon laquelle deux acteurs ne peuvent communiquer entre eux que par l'intermédiaire d'un troisième acteur, qui occupe ainsi une position avantageuse.
Le graphe de la Figure 77 illustre le principe de ce sommet qui, dès le premier seuil, est lié à de nombreux autres. Par les différents seuils de transitivité, on constate que toute la structure repose sur ce sommet, puisque c'est autour de lui que viennent se lier les autres éléments.

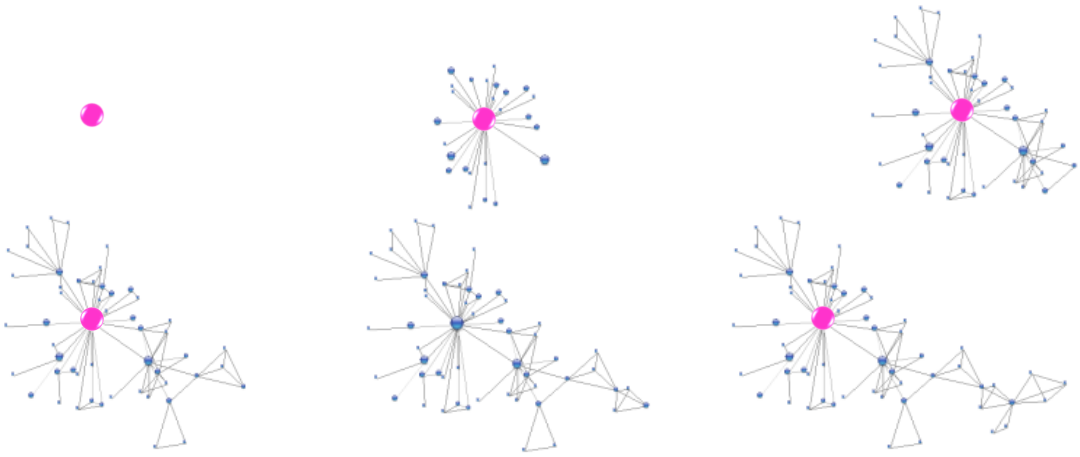


***Figure** 77. Calcul des seuils de transitivité d'une rotule, représentée en rose pour la distinguer des autres sommets.*

La transitivité de seuil 1 diffère de l'affichage des voisins directs. En effet, pour trois sommets A, B et C reliés les uns aux autres, la transitivité de seuil 1 de A montre la liaison entre B et C alors que l'utilisation de la fonctionnalité « voisins directs » permet de masquer ce lien, comme le montre la Figure 78.

***Figure** 78. Différence entre le calcul de la fermeture transitive de seuil 1 et l'affichage des voisins directs.*

L'étude de la structure du graphe se base en partie sur la transitivité. Cependant, le nœud n'étant qu'une représentation de la donnée textuelle, il est indispensable de pouvoir revenir à tout moment au document de base, à partir duquel la donnée a été extraite. Pour ce faire, l'outil VisuGraph est enrichi par le retour aux documents.

3.3.6. Retour aux documents

VisuGraph est doté d'une fonctionnalité permettant de favoriser l'exploration locale d'un sommet. Suite à la sélection d'un nœud spécifique, un pop up[26] apparait indiquant l'intitulé du sommet, la valeur de sa métrique, ainsi que le libellé des nœuds auxquels il est lié, ainsi que la valeur des liens. Cet affichage est complété par le retour aux documents. Cette fonctionnalité permet, à partir d'un sommet sélectionné, d'afficher dans un éditeur de texte, les documents contenant l'item choisi. Ainsi, si l'utilisateur clique sur un sommet, tous les documents du corpus initial concernant ce nœud s'afficheront dans un éditeur de texte (Loubier, 2007), (Loubier et Carbonnel, 2007).

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, pour un sommet sélectionné, tous les synonymes ou variation orthographique de ce dernier seront retournées.

Un autre aspect important de la synonymie, dans un contexte de « retour aux documents », est la proximité des termes. En effet, un même terme peut être écrit sous des formes différentes, telles que les erreurs de saisie, abréviations, …. Par exemple, le nom de notre institut peut être écrit sous la forme « IRIT » ou « Inst. de Rech. en Inf. de Toulouse » ou « Institut de Recherche en Informatique de Toulouse » ou encore « UMR 5505 ».

[26] Nouvelle fenêtre s'ouvrant automatiquement au dessus de la fenêtre de navigation actuelle.

Or, dans le cas d'un croisement entre auteurs et documents, lors de la recherche des documents dans lesquels un auteur particulier apparaît, le système doit retourner ceux comportant au moins une parmi toutes les formes possibles d'écriture de son nom, comme le montre la Figure 78.

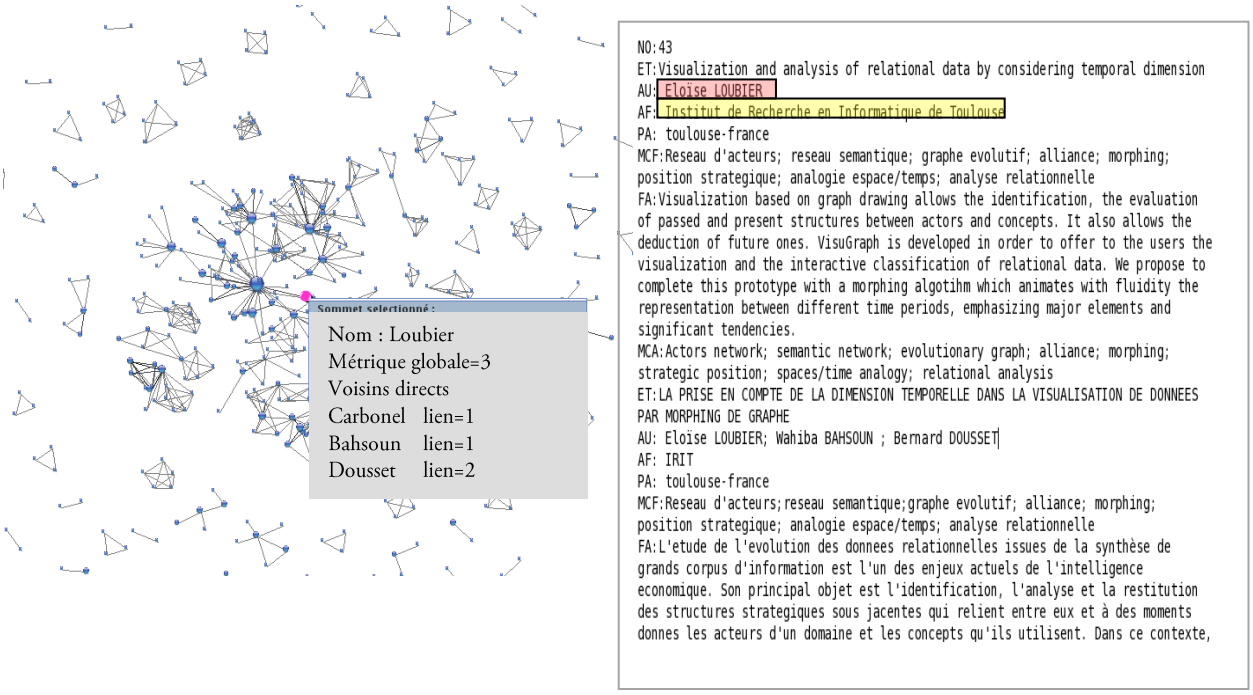


Figure 79. *Exploration d'un sommet spécifique et retour aux notices.*

Le retour au document ne doit pas se limiter à un seul nœud spécifique. Dans le cas de classes connexes, ou encore dans le cas de concurrents, il est intéressant de visualiser les notices de plusieurs acteurs afin de pouvoir les comparer mais aussi d'élargir la recherche. Cette fonctionnalité se nomme la focalisation.

3.3.7. Focalisation

La focalisation 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 focalisation sont proposés sous deux formes complémentaires : le retour aux notices et la représentation sous forme de graphe.

✓ ***Elaboration***

Afin d'adapter au mieux l'analyse de données en fonction des besoins de l'utilisateur, nous proposons une méthode (Loubier et Carbonnel, 2007) qui permet de rechercher les informations issues d'un corpus, répondant à un ou plusieurs critères d'une requête, de type union / intersection.

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

- sélection graphique de sommets ;
- sélection des noms des sommets via des listes dans un menu.

La sélection graphique permet d'obtenir une liste des sommets sélectionnés organisée par champ. Cette liste est affichée et l'utilisateur peut interagir avec sa sélection afin de supprimer ou d'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) [21]

(A1 ou A2 ou A3 ou A4) et (J1 ou J2) [22]

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. Un article ne peut être publié que dans un seul journal. Dans la seconde reequête, 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 d'intersection ou d'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 tels que la date, les auteurs, les journaux, les termes, les titres, …. Si l'utilisateur privilégie une dimension dans son analyse, il est libre de déplacer en conséquence les sommets la représentant, afin d'obtenir un graphe organisé. La fonctionnalité de focalisation permet donc de restreindre le domaine de recherche en fonction des besoins de l'utilisateur.

✓ ***Restitution des résultats***

Le retour aux notices restitue les documents répondant à la requête par le même procédé que celui détaillé dans la section précédente. Les documents répondant aux critères de la focalisation 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 (Loubier et Carbonnel, 2007).

La représentation, 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 les sommets avec la souris.

Ainsi l'utilisateur maîtrise pleinement son analyse dans chacune des étapes de focalisation, que ce soit dans la sélection d'information ou encore dans la manipulation des résultats, restitués sous forme textuelle et graphique. La Figure 80 illustre ce principe de focalisation par la recherche d'un article co-écrit par les auteurs « *Loubier, Dousset et Bahsoun* ». L'ordre d'apparition des auteurs dans la formulation de la requête n'a pas de conséquence sur leur ordre d'apparition dans les résultats, restitués sous forme graphique et textuelle.

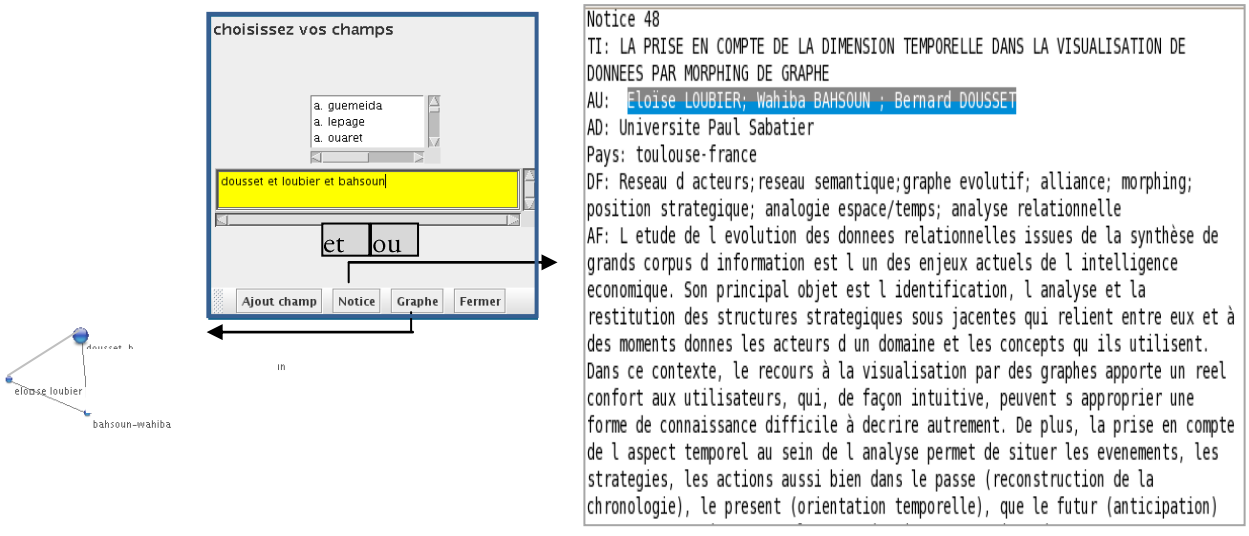


Figure 80. *Restitution des résultats graphiques et textuels, suite à la sélection de plusieurs auteurs.*

Cette réalisation se décompose en deux fonctions, une d'élaboration f_e permettant de cibler les éléments recherchés dans le graphe ou dans le corpus et une de restitution f_r qui fournit une visualisation des résultats sous forme graphique ou textuelle selon le choix effectué.

$$f_e\ (G,\{<x_i,op,x_j>\}\ )=R$$

$$f_r\ (R,c)=Res$$

Soit $G=(X,A)$ le graphe initial et x_i, x_j les sommets appartenant à l'ensemble X.
op est un opérateur $\in O=\{et, ou\}$
R est la formulation des éléments recherchés.
c correspond à l'option émise pour la nature de la restitution des résultats, sous forme de Graphe ou de Notice.

$Res \in R=\{G^e\ (v^e,e^e),\ C^e\ (C,\ R)\}$,
où $G^e=(x^e\ a^e)$, le graphe composé des sommets compris dans R *et* a^e, l'ensemble des liens liant l'ensemble des v^e.
C^e est le corpus C composé d'attributs, répondant aux contraintes imposées par R.

Dans l'exemple que nous avons illustré dans la Figure 80, l'élaboration s'effectue comme suit :

$F_e\ (G,\{<dousset,\ et,\ loubier>,<loubier,\ et,\ bahsoun>,\ <dousset,\ et,\ bahsoun>\})=dousset\ et\ loubier\ et\ bahsoun$

La restitution s'effectue sous forme de graphe :

$$fr\ (dousset\ et\ loubier\ et\ bahsoun,Graphe)\ =\ G^e(\{dousset,\ loubier,\ bahsoun\},$$
$$\{<dousset,loubier>,<loubier,bahsoun>,<bahsoun,\ dousset>\})$$

ou sous forme textuelle :

$$fr(dousset\ et\ loubier\ et\ bahsoun,Notice)=C^e(C,\{dousset\ et\ loubier\ et\ bahsoun\})$$

3.3.8. Partitionnement de graphes

✓ ***Principe général***

Dans un contexte d'analyse de graphe, la lisibilité et l'interprétation deviennent de plus en plus complexe, lorsque le volume de données augmente. Il est indispensable d'avoir à disposition des techniques permettant de réduire le nombre de données représentées, lorsque ce dernier devient trop important. Le partitionnement de graphe, intégré à VisuGraph, propose une solution.

Créer une partition consiste à répartir un ensemble d'objets en plusieurs sous-ensembles. Il est utile de pouvoir comparer ces sous-ensembles entre eux. Ainsi, chaque objet va être associé à d'autres objets, et les associations résultants, ou liens, doivent pouvoir être quantifiées. Le but des méthodes de classification est de construire une partition d'un ensemble d'objets. La classification a pour hyponyme le partitionnement de données, qui se traduit en anglais par data clustering.

Une *k-partition* d'un ensemble X est définie par une famille de sous ensembles $\{x_1, \ldots, x_k\}$ vérifiant

$$\bigcup_{i=1}^{k} V_i = V\ et\ V_i \cap V_j = \emptyset, \forall i \neq j \qquad [23]$$

Un graphe clusterisé est un graphe $G = (X, A)$ pour lequel on dispose d'une partition $\{x_1,\ldots,x_k\}$ de l'ensemble des sommets où les x_i sont des clusters.

✓ ***Markov Clustering intégré à VisuGraph***

Afin de faciliter l'analyse, les données les plus fortement liées doivent être regroupées en classes homogènes. Parmi les travaux effectués sur le partitionnement de graphe, les travaux de (Alpert et Kahng, 1995),

(Kuntz et Henaux, 2000), (Jouve et al., 2001) se basent sur des approches spectrales alors que les algorithmes de la famille METIS (Karypis et Kumar, 1998) se basent sur le partitionnement multi niveaux. Le Markov Clustering (MCL) consiste en l'alternance d'un mouvement en deux étapes -*expansion* et *inflation*- afin d'atteindre la convergence d'une matrice stochastique par laquelle un réseau entier est subdivisé en « clusters durs » sans aucun chevauchement. Le sous-réseau de chaque cluster de Markov est de type « étoile » dont le centre est le nœud de plus haut degré et les autres nœuds ne sont reliés qu'à celui-ci.

Ce type d'approche est basé sur la notion des déplacements aléatoires dans un graphe, selon un processus stochastique à temps discret par lequel on se déplace d'un sommet à un autre, choisi aléatoirement.

Soit un graphe G de n sommets, M sa matrice d'adjacence. La matrice de Markov associée à G, notée T_g est formellement définie par les colonnes normalisées de M. Soit d la matrice diagonale correspondant aux poids des colonnes de M, donnée par $d_{kk} = \sum_i M_{ik}$ et $d_{ij} = 0 \; si \; i \neq j$. . Alors T_g est définie par :

$$T_g = Md^{-1} \qquad [24]$$

La matrice de Markov T_g correspond au graphe G', appelé le graphe de Markov, associé à G. La valeur $(T_g)_{ij}$ indique l'importance de l'attirance du sommet j par le sommet i.

La définition suivante généralise l'opération de normalisation.

Soit la matrice $M \in R^{k\times l}, avec \; M \geq 0 \; et \; r \in R_+. Soit \; l' opérateur \; \Gamma_r$ de $R^{k\times l} \rightarrow R^{k\times l}$ défini par:

$$(\Gamma_r M)_{pq} = \frac{(M_{pq})^r}{\sum_{i=1}^{k}(M_{iq})^r} \qquad [25]$$

Γ_r s'appelle l'opérateur d'inflation avec le coefficient puissance r.

La méthode de partitionnement utilisée dans VisuGraph est inspirée du Markov Clustering (Van Dongen, 2000) que nous avons aménagée pour pouvoir influencer le nombre de classes proposées (Karouach et Dousset, 2003).

Cette approche se base sur deux opérations matricielles simples, successivement itérées :

- La première calcule les probabilités de transition par des marches aléatoires de longueur fixée r et correspond à une élévation de la matrice à la puissance r, visant à élargir la capacité de l'arc entre deux nœuds.
- La seconde consiste à amplifier les différences en augmentant les transitions les plus probables et en diminuant les moins probables. Les transitions entre sommets d'une même communauté sont alors favorisées et les itérations successives des deux opérations conduisent à une situation limite dans laquelle seules les transitions entre sommets d'une même communauté sont possibles.

Algorithme de MCL (G, $[e_i]_{i\in N}$, $[r_i]_{i\in N}$)
$T_1 \leftarrow T_G$;
$k \leftarrow 0$;
Tant que T 2k+1 n'est pas (approx.) idempotente
$k \leftarrow k+1$;
$T_{2k} \leftarrow Exp_{ek}(T_{2k-1})$;
$T_{2k+1} \leftarrow \Gamma r_k (T_{2k})$;

Algorithme de MCL

Une fois cette algorithme appliqué, la manipulation du graphe réduit peut être effectué.

✓ ***Manipulation du graphe réduit***

Décomposer un graphe en structures élémentaires est un paradigme classique permettant d'appliquer des techniques telles que « diviser pour résoudre » (Dousset et Karouach, 2007).

L'évaluation de la méthode MCL présentée précédemment a montré la rapidité et la qualité de ses résultats (Enright et al., 2002). Le graphe final est alors un graphe de classe, pour lequel chaque sommet est en fait une des classes obtenues, permettant de travailler alors sur un graphe réduit. Les liens entre les sommets sont assimilés à des liaisons interclasses (Loubier, 2007b), (Loubier et Dousset, 2007).

Dans un second temps, l'attribution d'une couleur spécifique à chaque classe permet de visualiser le graphe de départ, en figeant un représentant par classe et en distribuant les autres sommets sur une couronne centrée sur ce dernier, permettant ainsi une première vue intra classe.

L'avantage d'un tel procédé est de pouvoir travailler alternativement sur un graphe dit réduit, facilement manipulable, beaucoup plus lisible sur le graphe initial avec un dessin initialisé par celui du graphe réduit. On peut ainsi passer d'une vue synthétique à des vues détaillées des classes redessinées autour de leurs centres et qui ne se recouvrent plus.

Afin d'agir sur le nombre de classes obtenues, l'utilisateur intervient au niveau du réglage de la densité du clustering. Pour ce faire, un curseur est ajouté à la fenêtre de paramétrage, permettant d'augmenter ou de diminuer cette densité, comme le montre la Figure 81.

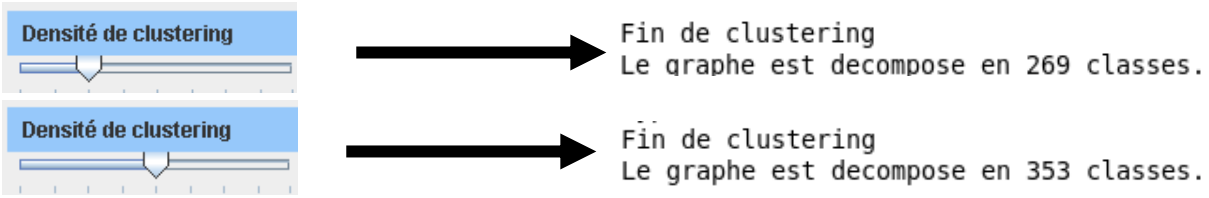


Figure 81. *Conséquence du changement de la valeur de la densité du clustering.*

Sur la fenêtre de visualisation, chaque classe obtenue par application de l'algorithme du MCL apparait sous forme d'un sommet de couleur. Le contenu de chaque classe peut être obtenu en détail en cliquant sur ce sommet. Une nouvelle fenêtre apparait, dans laquelle chacun des sommets composant la classe est focalisé.

De plus, VisuGraph offre la possibilité d'obtenir sous forme de liste textuelle tous les composants d'une classe. Cette fonctionnalité, applicable à partir du menu, permet d'obtenir le fichier texte résultat.

L'exemple, visible dans les figures 82, 83 et 84, illustre cette fonctionnalité. Le premier graphe permet de visualiser les données sous forme circulaire. Cette première représentation, bien que simple ne facilite pas l'analyse du graphe, l'étude de la centralité d'un sommet spécifique et de ses relations avec les autres données est une tâche fastidieuse.

Afin d'obtenir une meilleure visibilité, l'algorithme FDP semi paramétré est appliqué, permettant ainsi d'obtenir le second graphe, figure 83. Sous cette forme, la structure du graphe est clarifiée, cependant son analyse demeure difficile d'un point de vue global et local. Nous cherchons donc à partitionner ces données pour établir des ensembles de sommets dont la densité de connexions internes est plus forte que la densité de connexions vers l'extérieur, sans pour autant définir de seuil formel. Pour cela, nous appliquons l'algorithme du MCL et nous obtenons un méta graphe, Figure 83 qui est constitué de sommets représentatifs, de chaque regroupement de données. Le graphe est alors facilement analysé et manipulé.

L'algorithme de FDP semi structuré est appliqué pour obtenir une visualisation plus planaire. On distingue des *regroupements*, plus importants par leur taille, que d'autres. Il faut bien préciser que ce graphe est un méta-graphe et n'est composé que d'un représentant par groupe.

Pour obtenir le détail de chaque classe, deux solutions sont appliquées, selon le point de vue global ou local.

Dans le *cas global*, le retour à un graphe complet permet d'obtenir autour de chaque représentant, l'ensemble des sommets constituant la classe, comme visualisé dans le quatrième graphe, Figure 83, graphe situé en haut.

Dans le *cas local*, il est possible d'extraire la classe, en l'affichant dans une autre fenêtre de façon complète, c'est-à-dire en visualisant le représentant et les constituants de la classe, ainsi que les liens directs de la classe, Figure 83, graphe du bas.

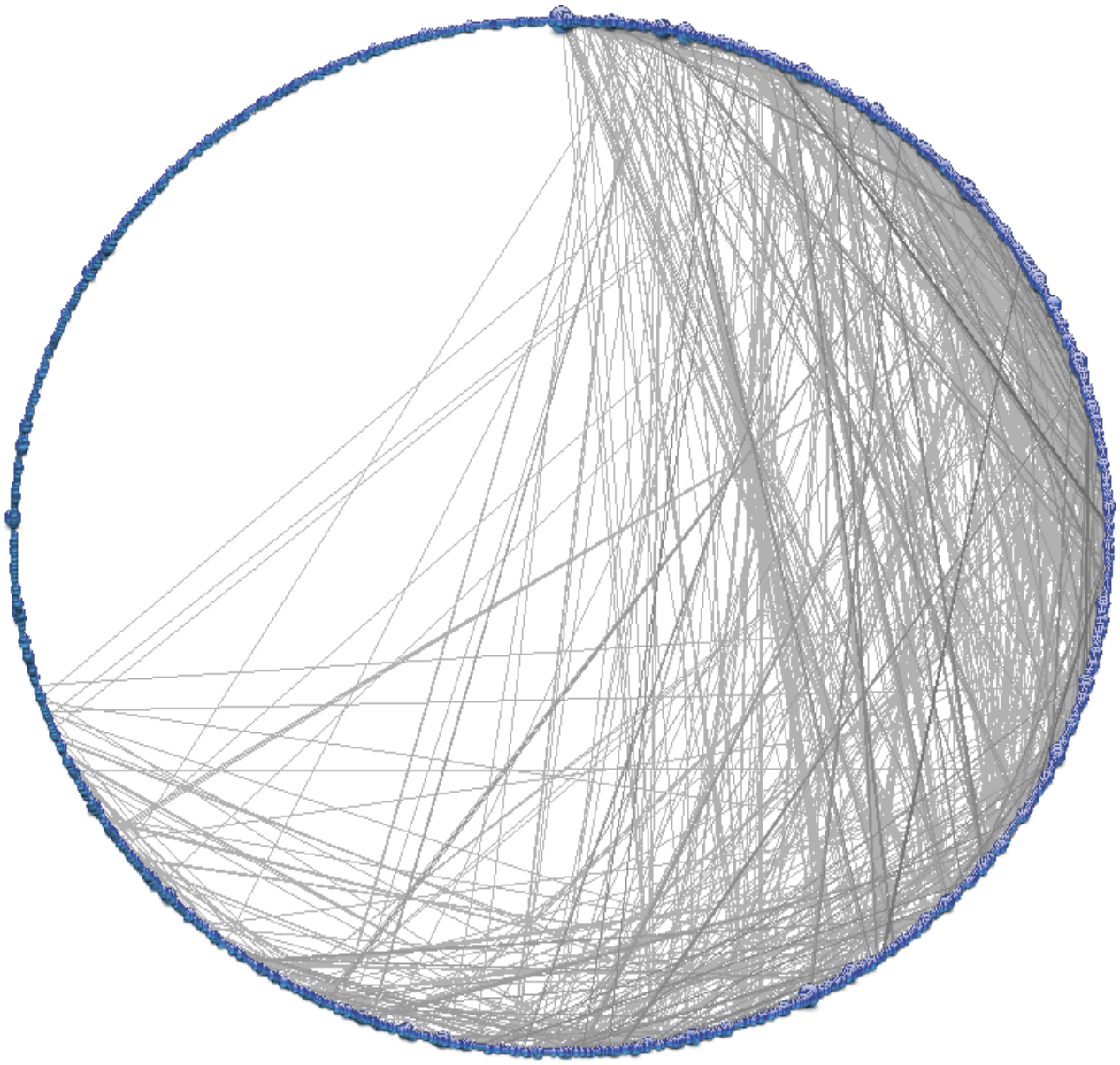


Figure 82. *Graphe initial circulaire, représentant la totalités des données relationnelles.*

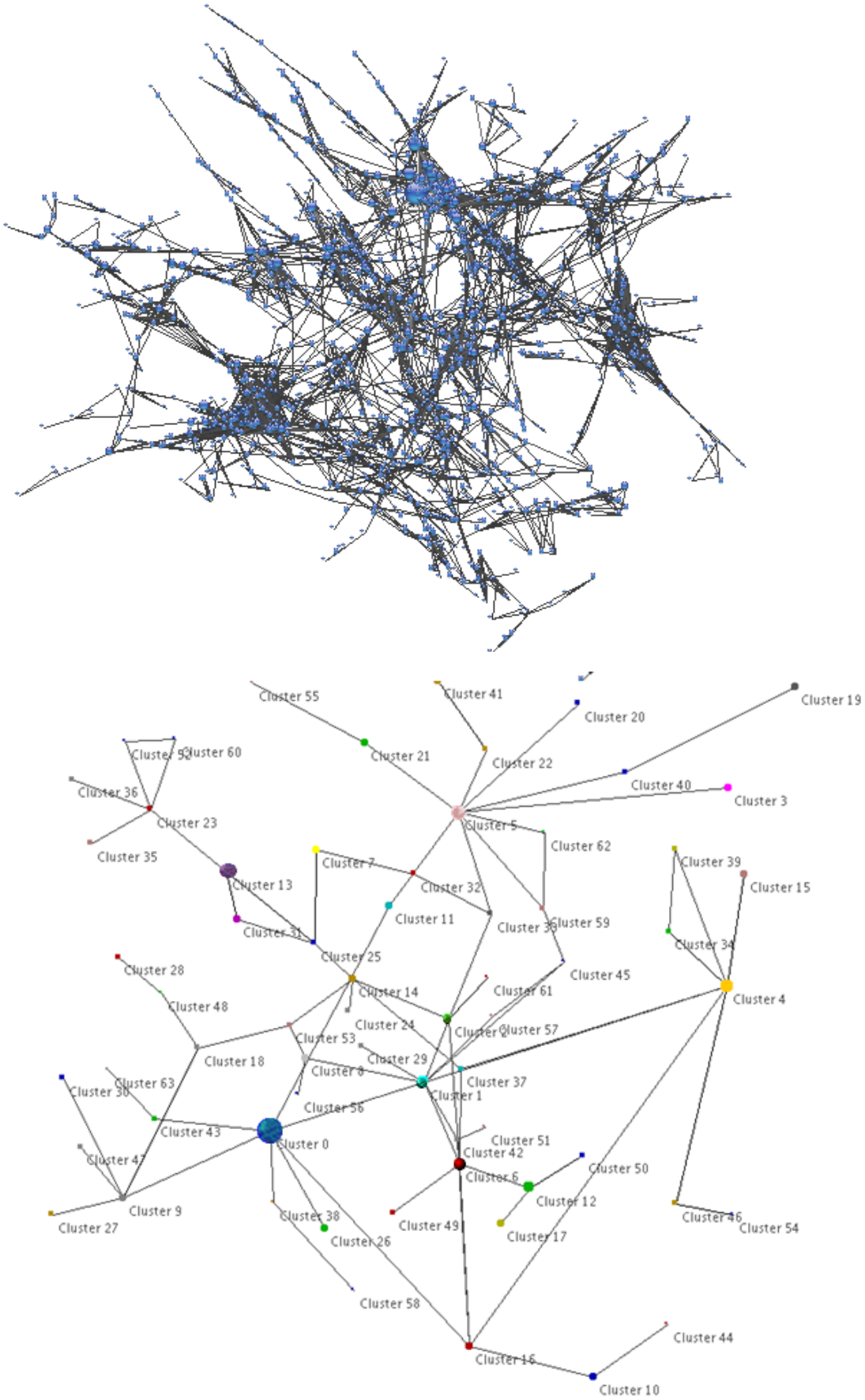
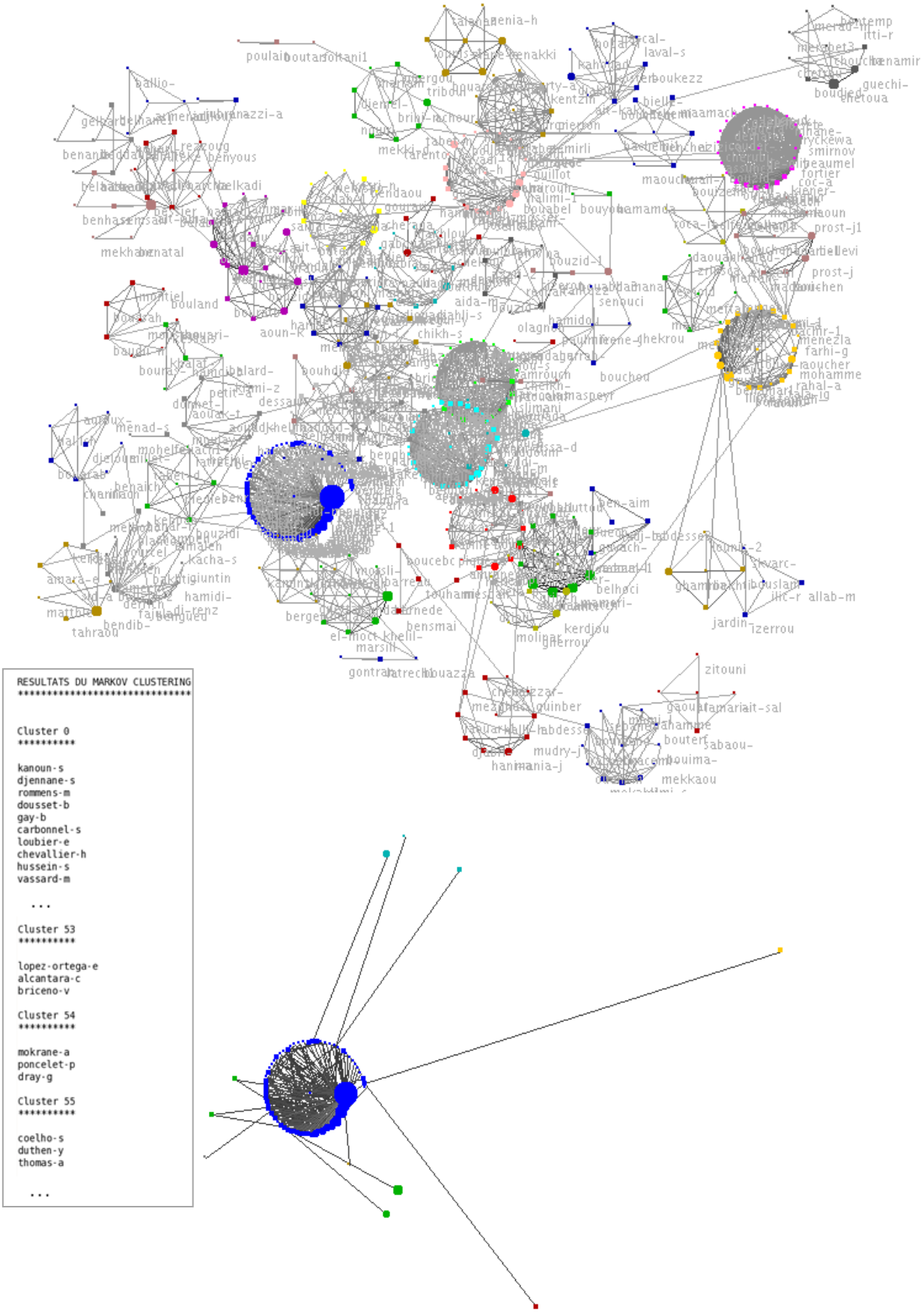


Figure 83. *Graphe de la figure précédente sur lequel l'algorithme FDP est appliqué (en haut). Puis application du MCL (en bas).*



```
RESULTATS DU MARKOV CLUSTERING
******************************


Cluster 0
**********

kanoun-s
djennane-s
rommens-m
dousset-b
gay-b
carbonnel-s
loubier-e
chevallier-h
hussein-s
vassard-m

  ...

Cluster 53
**********

lopez-ortega-e
alcantara-c
briceno-v

Cluster 54
**********

mokrane-a
poncelet-p
dray-g

Cluster 55
**********

coelho-s
duthen-y
thomas-a

  ...
```

Figure 84. *Affichage de la totalité des données partitionnées de la figure précédente (en haut), extrait de l'affichage sous forme d'un fichier texte de la composition des classes et extraction d'uns classe (en bas).*

Ce principe d'analyse de structure de graphe, par partitionnement est applicable à de nombreux domaines. Voici quelques exemples d'application à des domaines spécifiques :

- Réseaux sociaux afin d'identifier des sous-groupes homogènes d'individus (communautés) et la manière dont ils sont structurés entre eux ;
- World Wide Web, Recherche d'informations, pour grouper les sites Web par similarité pour faciliter l'identification de sites pertinents ;
- marketing, afin identifier des groupes d'individus ou des groupes de produits pour effectuer des conseils d'achats aux acheteurs;
- graphes de protéines, réseau métabolique, pour proposer des regroupements thématiques de protéines, d'enzymes ;
- réseaux sémantiques, pour donner du sens à des concepts sous-jacents ;
- réseaux de transport, pour identifier les solutions de déplacement les plus adéquats aux besoins des utilisateurs.

3.4. Conclusion

Dans ce chapitre, nous avons présenté notre contribution pour les graphes statiques, à travers la conception d'un outil de visualisation de données relationnelles, VisuGraph. Cet outil se base sur une représentation de sommets et de liens symbolisant leurs relations, par le biais de techniques de sémiologie préconisées dans le chapitre précédent, tels que la couleur, la forme, la taille… L'outil développé permet de représenter des graphes basés sur les matrices de cooccurrences présentées dans le chapitre 1. Ces derniers peuvent être des graphes simples, bipartis, orientés ou encore basés sur plusieurs matrices de nature différente. L'originalité de notre proposition repose sur toutes ces solutions possibles au sein d'un même outil et sur le panel de méthodes spécifiques permettant l'exploration et l'analyse des graphes. Ainsi, nous avons proposé trois algorithmes FDP, dont la caractéristique repose sur la recherche de valeurs adéquates pour les paramètres et sur des techniques permettant une représentation graphique plus spécifique. Nous avons proposé une variante, pour les graphes orientées, afin d'immobiliser une catégorie de sommets selon qu'il soit à l'origine ou à l'extrémité d'une liaison dirigée.

Suivant les besoins de l'utilisateur, la visualisation peut alors être orientée, via ces forces, vers l'étude de la similarité des sommets ou encore sur la structure globale des données. Cependant, ces représentations ont des limites que nous avons spécifiées et que nous reprendrons dans nos perspectives d'évolution.

Les fonctionnalités d'exploration intégrées à VisuGraph permettent l'analyse de la structure, par étude locale ou globale du voisinage des sommets caractéristiques.

L'interaction entre la représentation et l'utilisateur étant un souci majeur dans nos travaux, toutes les fonctionnalités présentées sont paramétrables via des sliders, afin que l'utilisateur reste maître de sa visualisation. Ce dernier choisit les méthodes à appliquer sur le graphe, il règle via la fenêtre de paramètre la fonctionnalité jusqu'à ce que le résultat visuel le satisfasse.

Les fonctionnalités d'exploration de graphe présentées dans ce chapitre sont la transitivité, permettant d'étudier le voisinage directe et indirect d'un sommet spécifique. Selon sa place dans la structure, le sommet peut alors être caractérisé comme étant un acteur majeur, dominant ou inversement sans influence. La *k-core* permet d'identifier les sous ensembles particuliers du graphe, il permet aussi d'obtenir un graphe dont les sommets ont un degré supérieur ou égal au degré fixé en paramètre et changeable à tout moment par l'utilisateur.

Ainsi, le *degré* d'un sommet correspond au nombre d'arcs issus de ce dernier, et constitue donc une mesure de la taille de son voisinage, composé de l'ensemble des sommets qui lui sont reliés. De ce fait, le degré d'un sommet peut donc être pris en général comme un indicateur de son intégration ou au contraire de son isolement dans l'ensemble du réseau, ou bien encore comme un indicateur de sa *centralité* : on mesurera ainsi la *centralité de degré* d'un sommet par le rapport entre son degré et le nombre de sommets auxquels il pourrait être relié.

Enfin, nous avons proposé une méthode de partitionnement par le Markov Clustering. Cette solution, initialement développée par (Karouach, 2003) a été améliorée par l'ajout de la visualisation textuelle de la composition des classes, sur l'exploration des résultats et plus particulièrement sur l'aspect évolutif, étudié dans le chapitre suivant.

Ainsi VisuGraph permet de visualiser des volumes de données importants et permet d'analyser les structures de graphe afin de répondre à des questions, telles que « s'agit-il d'un ensemble de données relationnelles formant des ensembles extrêmement soudés ?

Quel rôle a tel sommet ?

Quel est son niveau d'implication au sein de la structure ?

Chapitre 4.
Notre contribution pour les graphes dynamiques

« Quand les vérités sont trop lasses
Pour douter du temps qui passe
D'autres ont déjà pris leur place
Alors elles s'effacent. »
Mobilis in Mobile (L'Affaire Louis Trio, 1993)

4.1. Introduction 144
4.2. Spécifications 145
4.3. Espace de représentation des données temporelles 146
4.4. Métriques 146
4.4.1. Métriques associées aux sommets 146
4.4.2. Métriques associées aux arêtes 147
4.4.3. Codage des métriques 147
4.5. Positionnement temporel 148
4.6. Algorithme de représentation de graphe 150
4.7. Analyse de structure temporelle 153
4.8. Morphing de graphe : du graphe global au graphe de période 155
4.8.1. Définition 155
4.8.2. Principe 156
✓ Représentation globale 156
✓ Représentation par période 157
4.8.3. Morphing sans transition 162
4.8.4. Morphing avec transition 164
✓ Le point de vue du repère temporel de l'instance considérée 164
✓ Le point de vue de la structure temporelle 164
4.8.5. Discussion sur le morphing de graphe 167
4.9. Les graphes orientés 168
4.10. Fonctionnalités 170
4.10.1. Points de vues 170
✓ Le point de vue global 170
✓ Le point de vue évolutif 170
4.10.2. Filtrage 171
4.10.3. K-Core 172
4.10.4. Transitivité 173
4.10.5. Partitionnement de graphe 173
4.11. Application de notre contribution à des analyses non temporelles 175
4.12. Synthèse 177

4.1. Introduction

Au cours de la dernière décennie, notre capacité à recueillir et à stocker des données a dépassé notre capacité à les traiter, les analyser et à les exploiter. Anticiper les évolutions de son environnement est vital pour maintenir ou développer sa compétitivité pour se positionner dans une perspective de performance et de création. L'information est au cœur d'une telle démarche de veille stratégique, technique, scientifique et concurrentielle.

Les travaux de (Jakobiac, 2004), concluent sur les besoins de :

- Réaliser la surveillance systématique des secteurs technologiques, du marché de la concurrence, puis l'exploitation rationnelle des données captées ;
- Savoir collecter et exploiter l'information informelle ;
- Privilégier le côté offensif : saisir les opportunités de développement sans négliger la nécessité de détecter les dangers et de s'en protéger.

Notre contribution pour les graphes statiques, présentée dans le chapitre 3, nous permet de répondre à ce besoin d'analyse de l'information et du domaine à une période donnée unique. Cependant, dès que l'activité d'analyse cible plusieurs périodes et surtout vise à étudier l'évolution des données par la distinction de signaux faibles, l'aspect statique se limite rapidement.

L'objectif de nos travaux est de proposer une représentation des informations, permettant l'identification et l'analyse des structures stratégiques sous jacentes qui relient entre eux et à des moments donnés les acteurs d'un domaine et les concepts qu'ils utilisent. La prise en compte de l'aspect temporel au sein de l'analyse permet de situer les événements, les stratégies, les actions aussi bien dans le passé par reconstruction de la chronologie, le présent par orientation temporelle, que le futur par anticipation, pour tout ce qui concerne les organisations successives d'un réseau, telles que les collaborations, alliances, fusions, acquisitions, co-citations, co-signatures, co-occurrences de tous ordres. Le temps peut apporter plusieurs sortes de changements. D'abord, la structure de données peut être modifiée par l'ajout ou le retrait d'un ou plusieurs nœuds. Il est important de préserver l'image mentale de l'utilisateur, c'est-à-dire le graphe de base avant toute modification de la représentation. Par ailleurs, l'écoulement du temps peut en lui-même être un indicateur intéressant. Afin de mieux analyser les données évolutives des différents réseaux d'acteurs ou sémantiques, VisuGraph se base sur une fonctionnalité portant sur le morphing de graphe. L'objectif est de faire ressortir les tendances significatives en se basant sur la représentation d'un graphe global toutes périodes confondues et en réalisant une animation entre les visualisations successives des graphes attachés à chaque période.

Notre contribution repose sur les quatre étapes de découverte interactive de connaissances proposées par (Newell et Simon, 1972) :

- La caractérisation de l'environnement. Il s'agit d'extraire les propriétés et opérations pertinentes et accessibles à l'utilisateur pour la découverte de connaissances ;
- Le choix d'une représentation formelle pour définir l'espace de navigation;
- L'implémentation informatique de cette représentation formelle : son codage et sa représentation visuelle;
- L'implémentation de la procédure de découverte de connaissances : l'interactivité.

Dans ce chapitre, nous présentons notre contribution pour les graphes dynamiques, permettant l'analyse de données relationnelles évolutives, via l'outil VisuGraph.

Tout d'abord, nous exposons notre approche dans la section 4.2, en spécifiant les axes étudiés pour répondre à ce besoin de visualisation de données relationnelles évolutives.

Dans la section 4.3, nous présentons notre perception de l'espace de représentation des données temporelles.

Nous reprenons les principes de notre contribution statique, vues dans le chapitre 3, et nous étudions leur correspondance pour un cas dynamique. Puis, nous proposons une stratégie de positionnement des sommets selon leurs caractéristiques temporelles.

Par ce biais, il est plus facile de caractériser les entités d'un point de vue évolutif. De plus, cette solution permet de regrouper les sommets dont les caractéristiques temporelles sont semblables.

Dans un second temps, nous présentons la partie de notre contribution pour les données relationnelles et temporelles, à savoir le morphing de graphe. Après avoir défini cette notion, nous expliquons son principe. Puis, nous étudions deux aspects distincts du morphing à savoir la visualisation de l'évolution des données par transition ou non.

Enfin, nous concluons sur ce chapitre de contribution, en effectuant un bilan de ces apports et en les discutant.

4.2. Spécifications

Dans le premier chapitre, lors de la création de cubes de données, nous avons vu la possibilité de diviser le temps en périodes, afin de pouvoir le considérer comme troisième dimension. Nous considérons effectivement la dimension temporelle, comme une segmentation des données selon leur appartenance à un instant *t*. Dans tout ce chapitre, nous nommons « instance » ou encore « période » ou « moment », toute unité de temps considérée. Ainsi, si nous étudions plusieurs années individuellement, elles seront qualifiées d'instance. De même, si l'étude porte sur un regroupement d'années, alors nous nommerons par le même qualificatif cet ensemble d'années.

Notre approche, illustrée dans la Figure 85, repose sur le contrôle total de la représentation graphique par l'utilisateur. Les principaux axes concernent le prétraitement des données, dont le résultat est une matrice temporelle, la navigabilité, via des méthodes, des approches d'analyse spécifiques, l'interactivité, par le paramétrage des fonctions.

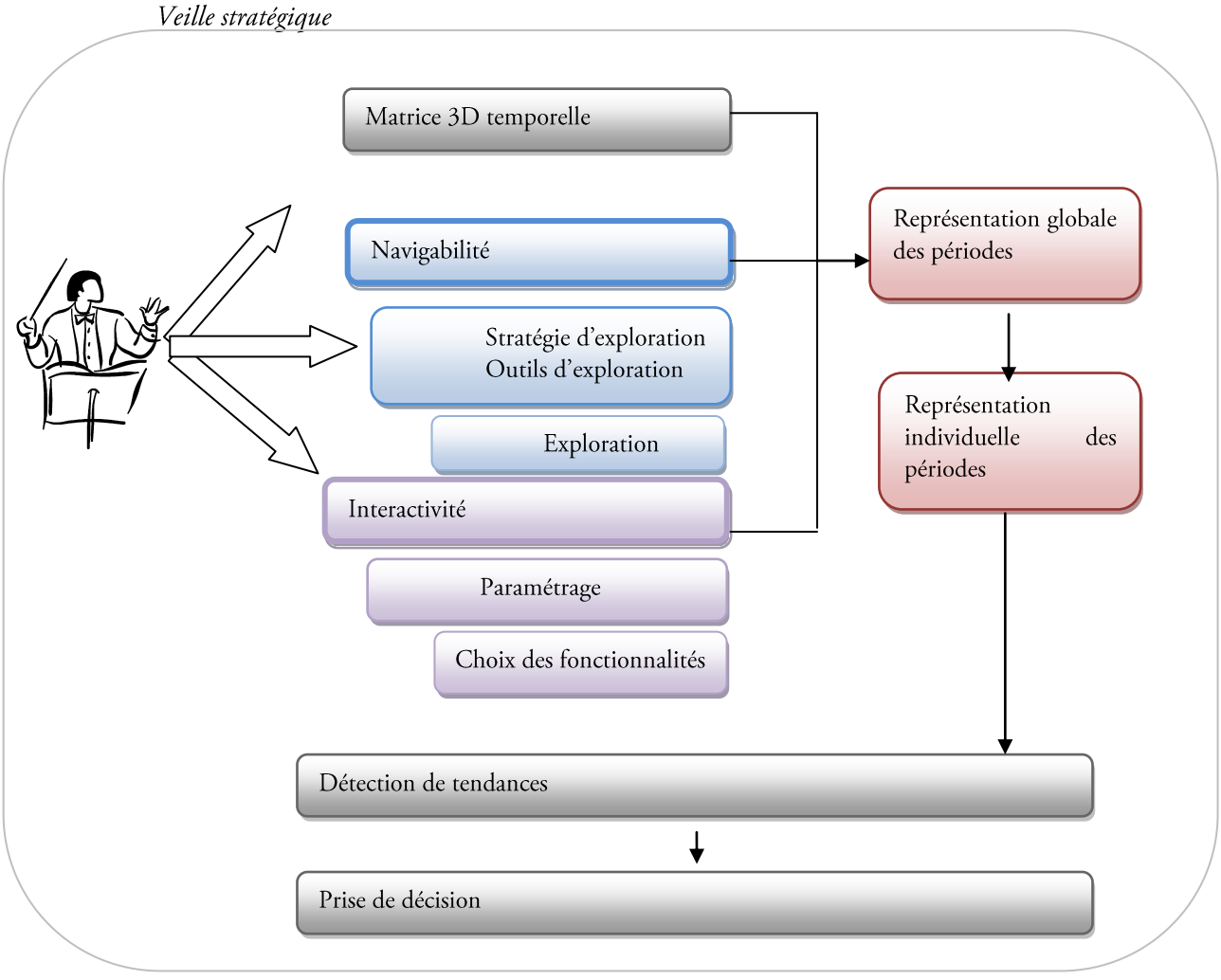


Figure 85. *Notre approche pour les graphes dynamiques.*

4.3. Espace de représentation des données temporelles

Selon la définition de l'espace de représentation étudiée dans le chapitre 1, nous définissons la structure de représentation des données temporelles. Pour les matrices symétriques

$$\mathcal{S}_g = \{<x_j, a_{<xj,xk,ph>}, x_k, \{P_h\}>\}$$

x_j l'ensemble des sommets liés à x_k par l'arête $a_{<Vi,Vj>}$. Pour $x_j, x_k \in X$

P_h représente les périodes durant lesquelles respectivement x_j et x_k sont valués et liés, pour $h \in \{1,\ldots,r\}$. r est le nombre d'instances considérées. P_h peut être sous forme d'années, de mois, de semaines, mais aussi le regroupement de ces périodes.

Au niveau des données en entrée, nous utilisons les matrices temporelles obtenues selon la méthode étudiée dans le chapitre 1. Nous illustrons cette structure de l'espace de représentation par un exemple simple. Considérons cinq auteurs, que nous nommerons *A1,A2,A3,A4,A5* et trois journaux, *J1, J2, J3*. Ces données sont étudiées durant deux années : *2007* et *2008*.

Les matrices obtenues suite à l'extraction des informations sont les suivantes :

2007

	A1	A2	A3	A4	A5
J1	2	3	0	2	0
J2	0	0	3	1	0
J3	1	0	0	2	0

2008

	A1	A2	A3	A4	A5
J1	2	0	0	0	2
J2	0	0	0	0	1
J3	1	1	0	3	5

Figure 86. *Exemple de matrices temporelles sur lesquelles repose la structure de l'espace de représentation des données.*

x_j *est* l'ensemble des sommets liés à x_k par l'arête $e_{<xi,xj>}$. Pour $j,k \in \{1\ldots 5\}$.

$$P_h \in \{2007, 2008\}\ h \in \{1,2\}.$$

Le graphe peut alors être décrit comme l'ensemble $\mathcal{S}$, de la manière suivante : *{<A1,2,J1,2007>,<A2,3,J1,2007>,<A4,2,J1,2007>,<A3,3,J2,2007>,<A4,1,J2,2007>,<A1,1,J3,2007>,<A4,2,J3, 2007>,<A1,2,J1,2008>,<A5,2,J1,2008>,<A5,1,J2,2008>,<A1,1,J3,2008>,<A2,1,J3, 2008>, <A4, 3, J3, 2008>, <A5, 5, J3, 2008>}*

4.4. Métriques

4.4.1. Métriques associées aux sommets

La notion de « métrique » étudiée dans le chapitre précédent est adaptée aux données temporelles.
Pour cela, nous la définissons V_j un sommet caractérisé par un ensemble de valeurs que prend cette donnée pour toutes les périodes considérées. Tout comme pour le cas statique, nous nous basons sur une matrice A de cooccurrences.

$$\mathcal{S}_7 = \{<x_j, m_g\{< x_j, m_t, t>\}>\}$$

pour $t \in \{1,\ldots, r\}$. r est le nombre d'instances considérées.

Avec $x_j \in \{x_1,\ldots,x_n\}$. n est le nombre de sommets du graphe.

m_g est la métrique globale, définie par :

$$m_g = \sum_{t=0}^{r} m_t$$

Et m_t chaque métrique attribuée au sommet x_j pour chaque période considérée.

Si nous reprenons l'exemple illustré dans la Figure 86, le graphe peut être décrit ainsi :

<A1,6,<J1,2,2007>, <J3,1,2007>,<J1,2,2008>,<J3,1,2008>>,
<A2,4,<J1,3,2007>,<J3,1,2008>>,
<A3,3,<J2,3,2007>>,
<A4,8,<J1,2,2007>,<J2,1,2007>,<J3,2,2007>,<J3,3,2008>>,
<A5,8,<J1,2,2008>,<J2,1,2008>,<J3,5,2008>>

4.4.2. Métriques associées aux arêtes

Le principe de métriques associées aux arêtes, vu dans le chapitre précédent, est légèrement modifié pour être applicable au cas temporel. Chaque paire de sommets adjacents est liée par une arête à un instant *t*, pour *t* appartenant à l'ensemble des périodes considérées, la pondération de cette dernière se traduit par la valeur a_{jk}, métrique associée à l'arête (Loubier et Dousset, 2008d) à l'instant *t*. Ainsi, si deux sommets x_j et x_k sont liés à ce moment considéré, la valeur de la métrique de l'arête les joignant est égale à la valeur du croisement entre x_j et x_k, dans la matrice de cooccurrences associée.

Dans l'exemple de la figure 86, l'auteur *A3* est lié au journal *J2* par une arête dont la métrique a la valeur 3, en *2007*. Par contre, en *2008*, cette arête disparait puisque le lien entre *A3* et *J2* est nul.

La structure de métrique associée aux arêtes, pour un graphe composé de *n* sommets de la première variable et *p* sommets de la seconde *est* caractérisée par:

$$\mathcal{S}_A = \{<a_{<xj,xk>}, m_g, <m_{t,<xj,xk>},t>>\}$$

pour $j,k \in \{1,\ldots,n\}$

m_g est la métrique globale définie par $m_g = \sum_{t=0}^{r} m_{t,<X_j,X_k>}$

Et $m_{t,<xi,xj>}$ la métrique attribuée à l'arête $a_{<xi,xj>}$ à l'instant *t*

pour $t \in \{1,\ldots, r\}$ *r* est le nombre d'instances considérées.

Appliqué à notre exemple de la Figure 85, les arêtes du graphe peuvent être décrites de la manière suivante :

$<a_{<A1,J1>},4,<2,2007>,<2,2008>>,<a_{<A2,J1>},3,<3,2007>>,<a_{<A4,J1>},2,<2,2007>>$
$<a_{<A5,J1>},2,<2,2008>>,<a_{<A3,J2>},3,<3,2007>>,<a_{<A4,J2>},1,<1,2007>>,$
$<a_{<A5,J2>},1,<1,2008>>,<a_{<A1,J3>},2,<1,2007>,<1,2008>>,<a_{<A2,J3>},1,<1,2008>>,$
$<a_{<A4,J3>},5,<2,2007>,<3,2008>>,<a_{<A5,J3>},5,<5,2008>>$

4.4.3. Codage des métriques

Afin de coder ces métriques, nous les proposons, de la même manière que celle décrite dans le chapitre précédent. Les principes de sémiologie, préconisés par (Bertin, 1970) sont repris :

La variation de taille permet de traduire parfaitement les variations quantitatives d'une période à une autre (Heer et al., 2009). Afin de pouvoir aisément comparer les valeurs de métrique de la donnée, nous utilisons la taille de l'objet représenté, pour montrer son évolution au cours du temps. La taille est relative à la valeur de la métrique du sommet. Ainsi les sommets les plus importants se distinguent par leur taille prédominante.

Nous n'utilisons pas de *grain* spécifique mais plutôt des surfaces lisses afin de ne pas surcharger la lisibilité du graphe, visible dans la Figure 87.

L'orientation choisie pour chacun des sommets est la position verticale. Chaque instance de la donnée se base sur le même segment, afin de faciliter la comparaison des tailles des valeurs de métriques aux différentes périodes, comme le montre la Figure 87.

Les formes choisies sont les barres verticales. Elles permettent de constituer, pour chaque donnée, un histogramme, illustré dans la Figure 87.

Cette forme valorise la détection des tendances car la comparaison des barres successives est plus facile que toute autre forme de représentation. Ainsi les avantages de la représentation linéaire sont conservés pour la visualisation de chaque sommet.

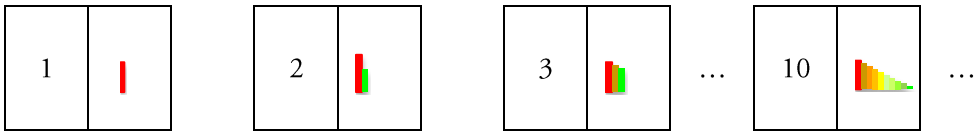


Figure 87. *Représentation d'une donnée temporelle, selon le nombre de périodes considérées.*

Les couleurs traduisent l'appartenance à une ou plusieurs périodes et permettent d'ordonner les données par le biais de dégradés. Dans le cas temporel, la couleur est utilisée pour faciliter la comparaison des données à un instant donné mais aussi la variation au cours du temps. En effet, pour chacune des périodes considérées, une couleur spécifique est attribuée. Ainsi, toutes les valeurs des données, pour la période *t* étudiée, sont représentées de la même couleur, facilitant ainsi leur visualisation et leur comparaison. Par convention, le choix de la couleur rouge est attribuée à la première période et le vert à la dernière, quelque soit le nombre de périodes. Les autres instances sont colorées selon un dégradé allant du rouge au vert, dans l'ordre temporel du plus ancien au plus récent, comme le montre l'exemple de la figure 87, selon le nombre de périodes considérées.

4.5. Positionnement temporel

Contrairement aux graphes statiques, pour lesquels la position initiale n'est pas significative, mais juste en adéquation avec des critères d'esthétiques, le placement de chacune des données, dans un contexte évolutif, est justifié.

Suite aux observations du chapitre 2, sur les différentes formes de représentation du temps, la visualisation la plus naturelle apparait comme étant la forme cyclique. Basé sur ce principe, dans VisuGraph, le positionnement des données s'effectue de façon circulaire afin de faciliter l'analyse temporelle, selon le principe de l'horloge. Dans un premier temps, chaque période considérée est assimilée à un sommet nommé « *repère temporel* ». Chacun de ces repères est placé de façon circulaire près des bords de la fenêtre, tous comme le sont les heures sur un cadran d'horloge.

La fonction permettant ce placement circulaire se base sur les paramètres suivants:

- *Instance,* le nombre de périodes considérées ;
- *teta*, variable divisant *Pi* par *instance*;
- x_i et y_i, les coordonnées du repère *i* pour $i \in \{1,\ldots,r\}$;
- *largeur* et *longueur*, respectivement la longueur et la largeur de la fenêtre de représentation ;
- *sin*, la fonction Sinus ;
- *cos*, la fonction Cosinus.

Ces paramètres sont utilisés pour calculer les coordonnées des repères, positionnés sur un cercle dont le périmètre est divisé par le nombre de périodes.

```
double teta= π/instance);
Pour i allant de 1 à instance
{
      int xi=(largeur)/2)*(1+(0.95* sin(teta));
      int yi=(longueur)/2)*(1-0.95* cos(teta));
      Sommet s = new Sommet(xi,yi,nom_reperei);
      teta=n*( π /instance);
      n+=2;
}
```

Algorithme de placement des repères temporels.

Une fois ces repères positionnés, chaque donnée est placée à leur proximité selon leur appartenance à chaque période. Ainsi, si une donnée à une valeur de métrique nulle pour une période, son positionnement n'est pas proche du repère de cet instance. Inversement, plus la valeur de sa métrique est importante pour une ou plusieurs périodes, plus la donnée est proche des repères correspondant.

L'intérêt est de placer les sommets représentant nos données, à une distance, des repères temporels, relative à leur appartenance à chaque période.

Ainsi les coordonnées temporelles, de chaque sommet du graphe sont calculées en fonction de celles des repères, obtenues précédemment.

Dans la formule [26], qui indique ces coordonnées temporelles pour chaque sommet du graphe :

128n est le nombre de périodes considérées ;
x_i, y_i sont les coordonnées du repère pour $i \in \{1,\ldots,r\}$;
m_{ik} est la valeur de la métrique du sommet x_k pour la période i, avec $x_k \in X$.

$$(x_k, y_k) = \left(\frac{\sum_n^{i=1} x_i \,.\, m_{i,k}}{\sum_n^{i=1} m_{i,k}} \,;\, \frac{\sum_n^{i=1} y_i \,.\, m_{i,k}}{\sum_n^{i=1} m_{i,k}} \right) \quad [26]$$

Une fonction est mise en place afin d'éviter la superposition des sommets. En effet, il est possible que deux sommets aient des appartenances identiques aux différentes périodes considérées. Selon les formules proposées, ces deux sommets ont les mêmes coordonnées, ce qui signifie qu'ils se superposent, empêchant ainsi leur distinction.

Il convient donc, lors de chaque calcul de coordonnées, de vérifier si les abscisses et ordonnées résultantes ne sont pas déjà attribuées à un autre sommet. Si tel est le cas, nous incrémentons, au pixel près, la coordonnée afin d'éviter toute superposition.

Le placement stratégique des sommets permet, non seulement, de les situer dans le temps mais aussi d'en évaluer la persistance et d'en déduire la tendance.

Une fois ces positionnements effectués, il est important de garder en mémoire cette attirance relative des repères. Pour ce faire, nous créons des liaisons entre chaque donnée et son repère, via des arêtes. Ainsi, pour toute période durant laquelle la donnée *d* est valuée, une arête la lie à ses repères. L'intérêt de ces arêtes temporelles est explicité dans la section 4.6.

Afin de ne pas surcharger le graphe déjà complexe, suite à la représentation de toutes les instances pour chaque donnée, les arêtes temporelles sont automatiquement masquées, les repères le sont optionnellement, mais pas supprimés du graphe. Leur présence est donc prise en compte malgré leur invisibilité.

La représentation des sommets, ainsi que le positionnement temporel facilitent le caractère évolutif d'une donnée. Cette information est très importante, mais dans un contexte de veille stratégique, elle est insuffisante, puisqu'elle ne donne pas de renseignement sur l'état de cette donnée vis-à-vis des autres. Le focus de visualisation est agrandi et la donnée peut alors être étudiée dans son contexte environnemental et évolutif. Par sa position dans le graphe global, cette donnée se distingue par son émergence, comparée à une majorité qui se caractérise par une certaine présence permanente et se situe vers le centre du graphe, par attirance simultanée de plusieurs repères.

L'exemple simple, illustré dans la Figure 88, montre la conception d'un graphe temporel sous VisuGraph. Les contours de la fenêtre de visualisation, ainsi que du cercle tracé pour placer les repères sont affichés de couleur très claire. Les repères sont disposés à équidistance les uns des autres et dans un ordre croissant. Puis les coordonnées d'un sommet temporel *dt* sont changées, engendrant son déplacement.

Les valeurs de métriques, pour chacune des périodes, sont indiquées dans le tableau accompagnant cet exemple. La valeur de métrique la plus importante concerne la période 4, puis par ordre décroissant, les périodes 3 et 1.
La formule [25] est appliquée et le positionnement graphique est illustré dans la Figure 88, à des distances des repères proportionnelles aux valeurs des métriques pour chaque période. Les arêtes temporelles sont affichées et graduées d'intensité faible pour illustrer la distance relative entre le sommet et chaque repère.

Période	*Valeur de métrique*	*Distance entre le repère et le sommet*
1	m_1	*2,5 .n*
2	*0*	
3	*2 . m_1*	*1,5 .n*
4	*1,5 X m_3= 2,5 . m_1*	*n*

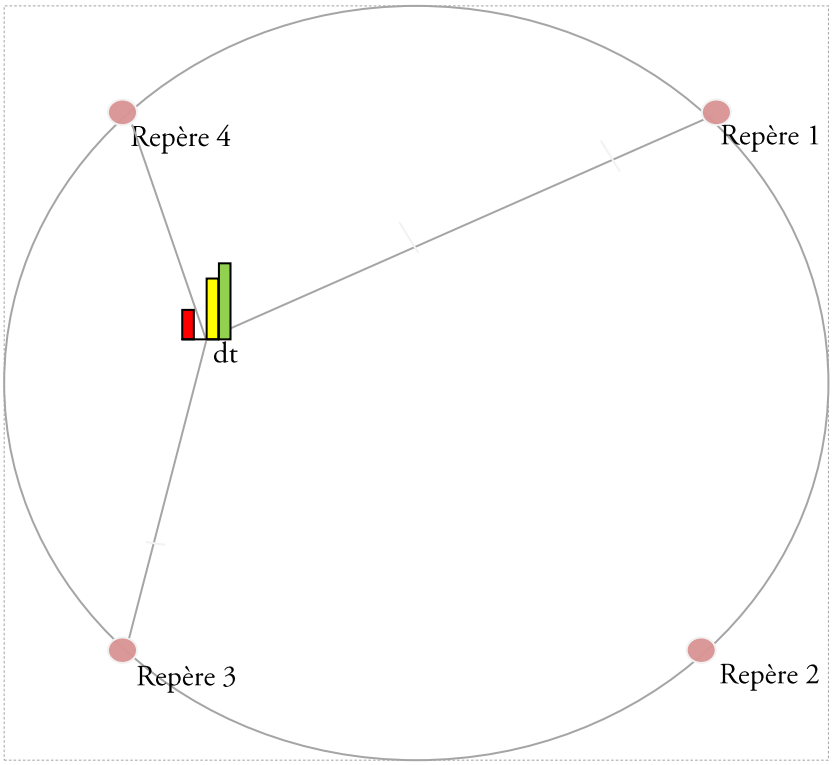


Figure 88. *Conception d'un graphe temporel sous VisuGraph.*

4.6. Algorithme de représentation de graphe

Tout comme pour les graphes statiques, afin d'obtenir davantage d'interactivité entre le système et l'utilisateur et surtout pour permettre à ce dernier de contrôler pleinement sa représentation graphique, l'utilisateur peut accentuer l'attraction ou la répulsion entre les données. Pour cela, des règles graduées sont mises à disposition, dans le menu de l'outil, permettant l'augmentation ou la diminution de ces deux types de forces. Le système dispose ainsi d'une règle graduée spécifique aux forces d'attractions, correspondant au paramètre $\alpha_a$et un pour les forces de répulsion, ajustant la valeur de α_r. Un troisième, appelé *Force repères temporels*, est ajouté pour contrôler l'attirance des sommets vers les repères temporels. Cet algorithme est donc interactif et l'utilisateur reste maître de sa représentation graphique.

Dans un premier temps, la force de répulsion entre tous les sommets est calculée. Dans un second temps, toutes les attractions sont prises en compte, pour toute paire de sommets liés. Ces deux forces, d'attraction et de répulsion, engendre le déplacement des sommets, c'est-à-dire un changement des coordonnées.

Si la valeur de la règle graduée temporelle est valuée et si le sommet appartient à la période considérée, alors une attraction entre ce dernier et le repère est appliquée, engendrant son déplacement vers le repère.

```
Pour tout sommet u
{
    Si u est visible
    Alors{
        Calcul de la distance d(u,v) ;
        pour tout sommet v{
            f_r(u,v, d(u,v));
            S'il y a une arête entre u et v{
                f_a(u, v);
                Si (u ou v est un repère temporel)
                Slider_force_reperes_temporels X f_a(u, v, d(u,v));
            }
        }
    }
}
Pour tout sommet u {
    Si (u n'est pas un repère)
    Déplacement du sommet ;
}

/* *Vérification de la non superposition des sommets par comparaison des coordonnnées **/
Pour tout sommet u{
    Pour tout sommet v{
        Si (x_u,y_u) == (x_v, x_y)
        alors changer les coordonnées de v.
    }
}
```

Algorithme de placements dirigés par des forces en prenant en compte la dimension temporelle.

Dans cet algorithme de FDP, les paramètres ont été étudiés pour obtenir des résultats pertinents.

Ainsi, pour le calcul de la force d'attraction :

β est une constante, initialisée à 2 ;

d_{uv}^{ae} est la distance entre *u* et *v*, où α_a correspond à la valeur du slider permettant d'interagir sur l'attraction.

Pour le calcul de la force de répulsion :

α_r correspond à la valeur du slider, permettant d'interagir sur la répulsion ;

c est une constante, initialisée à 1,5.

Dans l'exemple illustré par la Figure 88, nous étudions des auteurs spécifiques au domaine du data-mining sur quatre périodes : 2005, 2006, 2007, 2008-2009. Nous ne nous attardons pas sur l'origine des données utilisées, cet exemple a pour objectif de montrer une application de l'algorithme et non pas les résultats d'une analyse spécifique.

Pour chacune des périodes, un repère est attribué en couleur rouge, représenté sur la Figure 88, et chaque sommet ayant une métrique valuée pour une période est alors lié au repère correspondant par un arc invisible.

Sur le graphe de gauche de la Figure 88, aucune force d'attraction et de répulsion n'a été appliquée. Sur le graphe de droite, elles sont appliquées et la valeur du slider permettant d'augmenter l'attirance vers les périodes est élevée. On constate que chaque partie du graphe de droite est spécifique à une caractéristique temporelle.

Plus un sommet est proche d'un repère, plus il est fortement caractérisé par cette période. Les sommets situés à équidistance de deux repères révèlent une appartenance aux deux périodes. Ainsi, sur la figure suivante, il est facile de distinguer les sommets spécifiques à 2005 puisqu'il s'agit de l'ensemble des sommets situé autour du repère, proches des bords de la fenêtre de visualisation et caractérisé par une seule barre d'histogramme en première position.

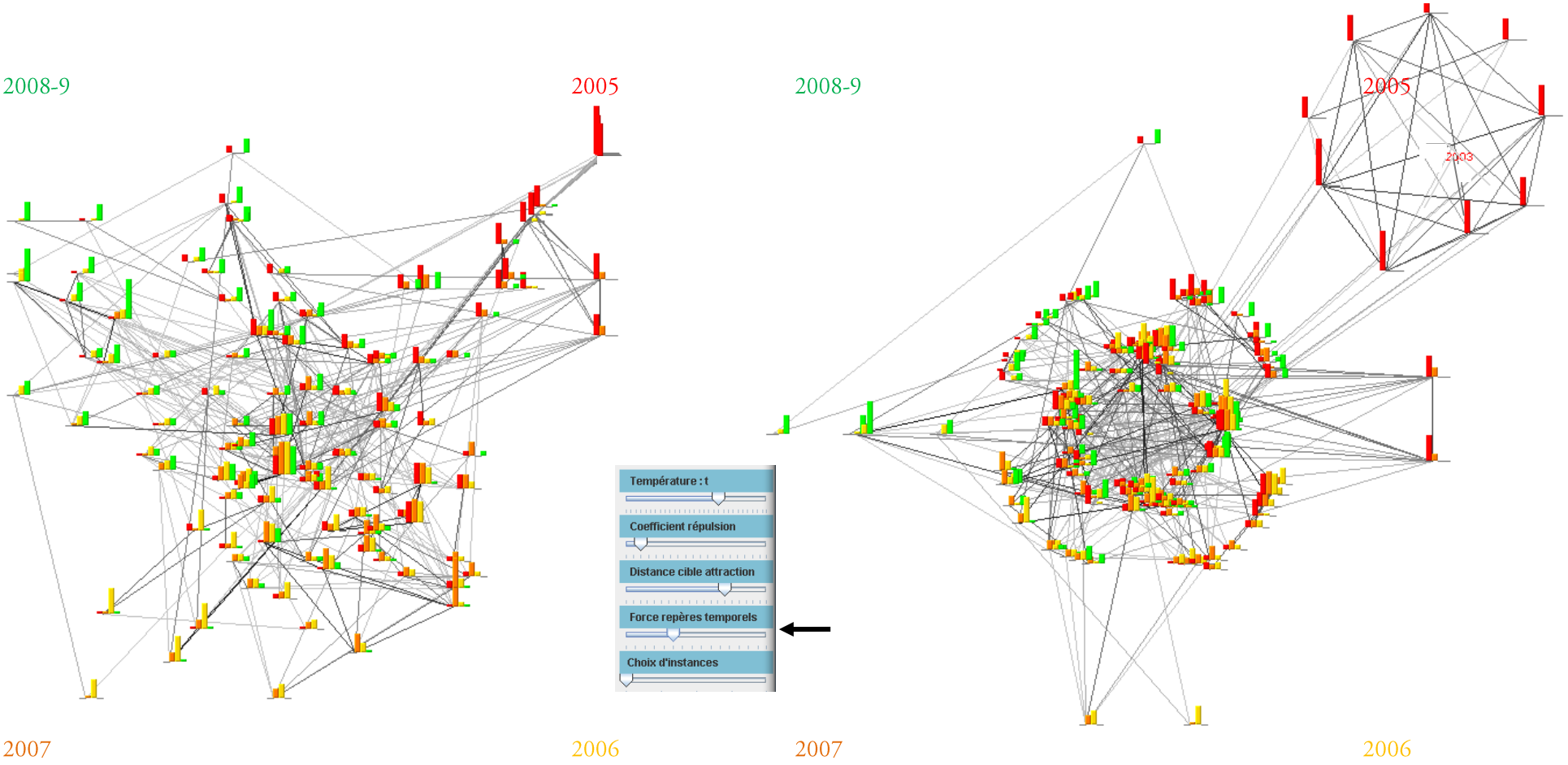


Figure 89. *Sur un graphe évolutif (à gauche), application de l'algorithme présenté, paramétré à l'aide du slider « Force repères temporels » (graphe de droite).*

De même pour les autres périodes, plus un sommet est situé proche du centre de la Figure 89, plus le nombre de périodes auxquelles il appartient est important. Ainsi, les auteurs représentés au centre de la Figure 89 sont les plus persistants. Ceux représentés proches du repère 2005 sont les plus anciens et ceux qui sont les plus proches de 2008-2009 sont les auteurs les plus émergents.

Les préconisations présentées dans le chapitre 3 sur l'utilisation des forces dirigées, en trois étapes, par application d'une forte valeur d'attraction des repères temporels et de la température, puis la réduction de l'attraction et l'augmentation de la répulsion sont valables de façon similaire dans le contexte temporel.

4.7. Analyse de structure temporelle

L'analyse de structure temporelle repose sur plusieurs points :

- Le signalement explicite.

Certaines informations se suffisent à elles mêmes, Après leur validation et leur mise en forme, elles sont communiquées au bon moment et à la bonne personne. Elles sont présentées à titre signalétique.

- Le recoupement.

L'information est étudiée par rapprochement avec d'autres données de même nature.

- Le phénomène de rupture.

La disparition brutale d'un sous domaine, d'une thématique, d'une discipline, d'une équipe, d'un acteur majeur peut être une information stratégique. Il peut alors s'agir d'effectuer une réorientation thématique, un changement d'alliance ou tout simplement l'arrêt d'une collaboration.

La visualisation de données temporelle doit être en mesure d'apporter la bonne information, au bon moment, à la bonne personne et sous la meilleure forme possible pour qu'elle soit associée au processus de décision. Un des objectifs visés par VisuGraph est la surveillance dynamique d'un système, à savoir l'évolution de ses performances, la détection des signaux faibles, les changements de collaborations, d'alliances ou encore d'associations.

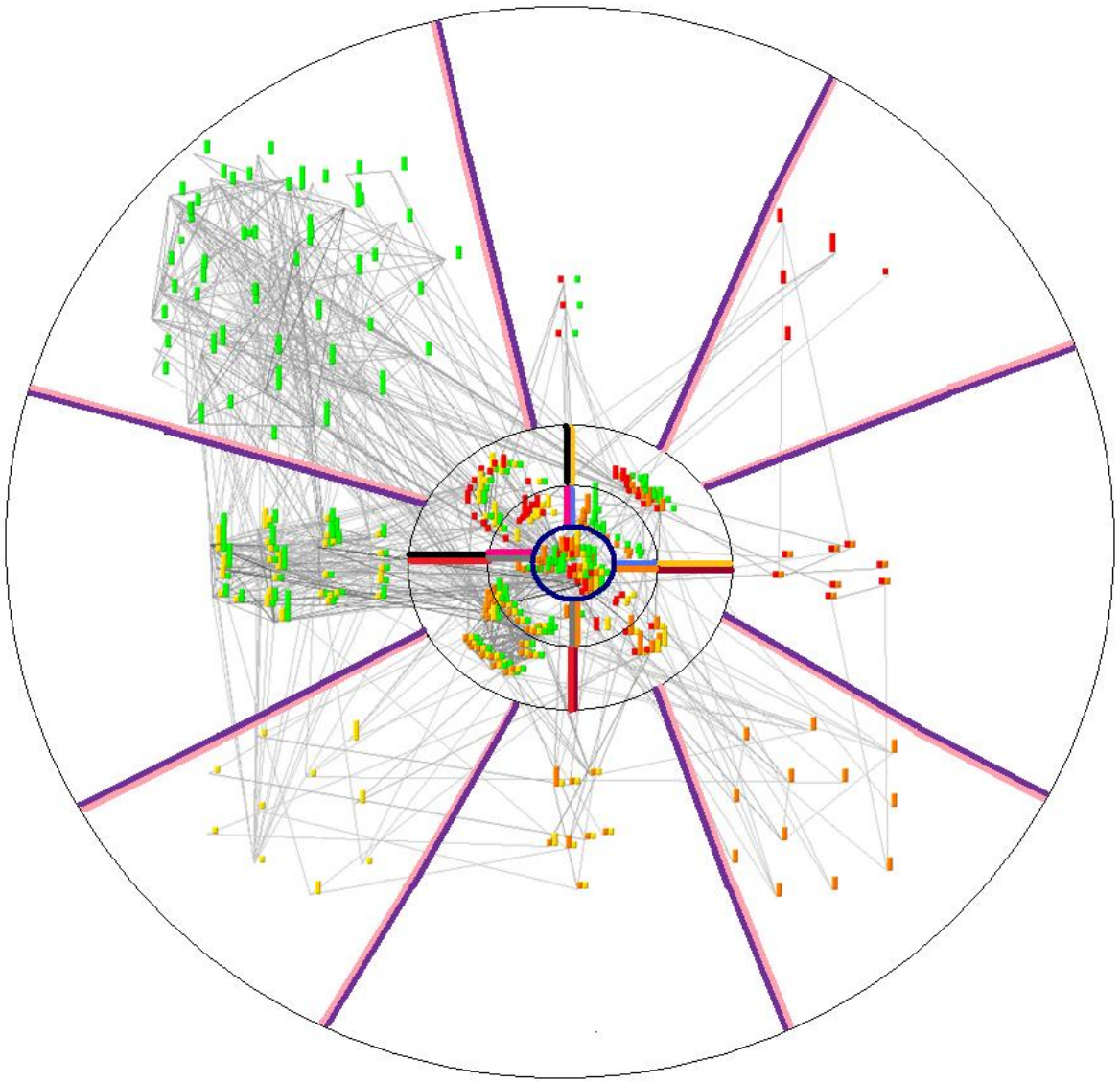
La façon la plus immédiate de représenter le temps dans les graphes de données est d'utiliser une représentation spatiale du temps. L'information temporelle que contiennent les données est alors transformée en une information spatiale, c'est-à-dire une position caractérisant temporellement l'information.

La prise en compte du temps est réalisée en permettant l'accès à l'état précédent du graphe, ainsi qu'à la période successive.

Notre contribution vise à pouvoir permettre à l'utilisateur d'effectuer une étude en trois périodes, avec une analyse directe, abordable par simple visualisation du graphe globale, puis un second niveau, par compréhension de l'évolution des données. Cette étape demande un effort cognitif plus important, puis qu'elle atteint l'analyse stratégique de la structure. Enfin la dernière partie de l'analyse concerne l'extrapolation des résultats, à savoir l'aspect « anticipatif » des données, spécifique au contexte de veille stratégique.

Dans l'analyse de la structure dans un graphe évolutif proposé par VisuGraph, chaque partie de la fenêtre du graphe traduit une caractéristique temporelle. Nous pouvons effectuer une généralisation, quelque soit le nombre de périodes étudiées, de ces caractéristiques selon leur positionnement. L'aide graphique que nous proposons peut alors être appliquée à tout graphe temporel pour faciliter son analyse. Pour décomposer les différentes catégories temporelles, nous procédons par élaboration de cibles applicables au graphe temporel.

Chacune des cibles peut être décomposée en autant de sous parties que d'instances étudiées. Dans l'exemple de la Figure 90, nous considérons n le nombre d'instances, égal à 4.



Légende

- Zone spécifique à l'appartenance à une seule période.
- Zone spécifique à l'appartenance à deux périodes.
- Zone spécifique à l'appartenance aux périodes *n*, 1 et 2.
- Zone spécifique à l'appartenance aux périodes 1, 2 et 3.
- Zone spécifique à l'appartenance aux périodes 2, 3 et 4.
- Zone spécifique à l'appartenance aux périodes 3, 4 et 1.
- Zone spécifique à l'appartenance aux périodes 2 et et une faible présence en période 1.
- Zone spécifique à l'appartenance aux périodes 1 et 3 et une faible présence en période 2.
- Zone spécifique à l'appartenance aux périodes 2 et n, avec une faible présence en période 3.
- Zone spécifique à l'appartenance aux périodes 1 et 3 une faible présence en 4.

Figure 90. *Décomposition du graphe en zones temporelles.*

Dans le cadre d'une analyse de la structure de graphe, les caractéristiques temporelles des entités dépendent de leur position.

Les données situées dans la plus grande cible sont divisées en deux sous catégories :

- Les données spécifiques à une seule période. Elles sont placées extrêmement proches des repères temporels. Dans le cas des données spécifiques à la dernière période, il est prévisible que ces dernières puissent apparaître de nouveaux dans les instances suivantes.
- Les données spécifiques à deux périodes. Elles se situent à équidistance de deux repères temporels. De même, si l'une des deux périodes auxquelles la donnée est rattachée, est la dernière instance, on peut prédire qu'elle aura des chances de persister à l'avenir.

Quelque soit le nombre de périodes étudiées, les données situées dans la cible intermédiaire sont généralement présentes durant les trois périodes, symbolisées par les trois repères temporels encadrant le sommet. Ces données peuvent être qualifiées comme des *pivots temporels*. Ils sont apparus durant la première période, ont confirmé leur position en seconde instance, puis ont disparu finalement, s'ils ne sont pas liés à la dernière instance. S'ils le sont, les données peuvent être évolutives et pourront être présentes dans les instances à venir. C'est à ce niveau là, que VisuGraph aide à la prédiction.

Les données situées dans le centre de la figure, à savoir la cible la lus petite. Les données temporelles sont spécifiques à une majorité de périodes étudiées. Leur positionnement symbolise leur proximité similaire aux différents repères. Ces sommets peuvent être qualifiés comme persistants, du fait qu'ils sont présents pour toutes les périodes. Ils peuvent laisser à penser qu'ils seront toujours présents dans les instances suivantes.

Ces caractéristiques sont spécifiques aux sommets pour une visualisation globale de toutes les périodes. Cependant, ces conclusions ne peuvent s'appliquer pour les liens, sur le graphe global. En effet, ce dernier représente un cumul des valeurs de métriques pour les sommets et pour les liens. Ainsi, deux données présentes pour toutes les périodes peuvent avoir un lien, mais il est impossible de détecter sur le graphe global, durant quelle instance ce lien a été crée et surtout s'il a persisté ou non. Pour effectuer une telle analyse, il faut passer par le morphing de graphe et étudier en détail chacune des représentations de période.

4.8. Morphing de graphe : du graphe global au graphe de période

4.8.1. Définition du morphing de graphe

La visualisation graphique des données relationnelles, qu'elles soient temporelles ou non remplit pleinement sa fonctionnalité de simplification d'analyse. La représentation graphique facilite l'exploration des données et des différentes tendances par analyse de la structure du graphe et particulièrement par voisinage des sommets. Cependant, dans le cas temporel, il est important de considérer dans un premier temps la visualisation globale des données, puis, dans un second temps, la représentation individuelle de chaque période, avec la possibilité de revenir à tout moment à n'importe quel type de graphe, général ou non.

Nous définissons le *morphing de graphe* comme la transformation géométrique T_g d'une représentation graphique, permettant le passage d'une visualisation de donnée au temps *t-1* à celle de t et inversement. Il s'agit d'une déformation de graphe continue. Le morphing (Loubier, et al., 2007) consiste à fabriquer une animation qui transforme de la façon la plus naturelle et la plus fluide possible un graphe initial vers un graphe final (Loubier et al., 2007).

L'objectif est de réaliser une lecture intuitive de l'évolution en répartissant séquentiellement les périodes de façon cyclique, permettant, à partir de la représentation du graphe global, de visualiser successivement chaque graphe de période, de façon animée et fluide. En se basant sur l'analogie espace/temps, il permet ainsi de détecter, comprendre et même prévoir les tendances significatives, au travers de la visualisation de l'évolution des données.

Soit $G_g = (X,A)$ le graphe global ;
Soit P_t un sous ensemble de X, noté $P_t = \{x_1, \ldots, xt\}$ *et* caractérisé par l'appartenance à une période spécifique.

On dit que P_t est un *graphe de période* issu de G_g.

Un sous-ensemble de X, élément de P_t peut être vide, dans le cas d'une période étudiée durant laquelle aucune des données n'a de valeur de métriques positive. Ce cas là ne présente que peu d'intérêt.

Les sous-ensembles de X, éléments de P_t ne sont pas obligatoirement disjoints deux à deux, dans le cas où des données sont valuées durant plusieurs périodes.

$$\forall (i,j) \in \{1,\dots,y\}, si\ i = j, V_i \cap V_j \neq \emptyset$$
$$ou\ si\ i \neq j, V_i \cap V_j = \emptyset$$

Notre contribution s'appuie sur les travaux de (Lamure, 1987) ou encore ceux qui portent sur le morphing sur images réalisés par (Sederberget et al. 1993), (Shapira et Rappoport,1995), (Goldstein & Gotsman,1995),(Floater et Gotsman1999), (Alexa et al., 2000), (Surazhsky et Gotsman, 2003). Leurs travaux aboutissent à des variations d'images comme le montre la Figure 91, notre contribution repose sur l'application de ce principe à la théorie des graphes, dans un contexte évolutif.

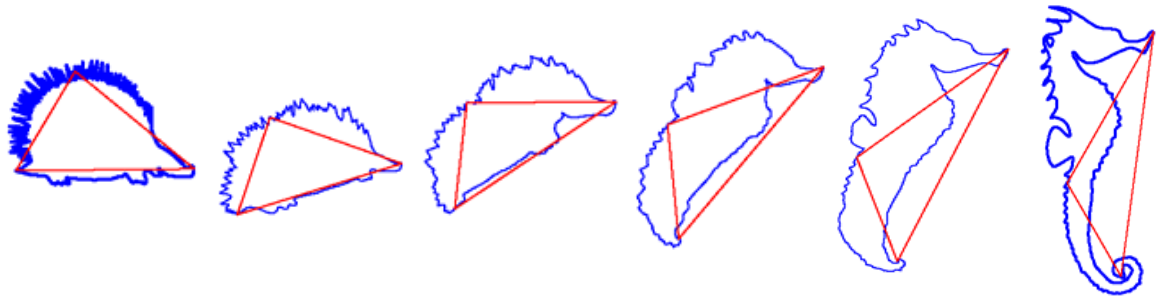


Figure 91. *Morphing d'images.*

4.8.2. Principe

Le morphing de graphe se base sur une représentation globale des données temporelles, en utilisant une structure permettant de détecter rapidement les caractéristiques temporelles des données, à savoir quelles sont les données persistantes et/ou celles apparaissant ?

L'animation des visualisations successives des différentes périodes, dans le sens chronologique similaire à l'analogie espace/temps d'une horloge, permet de créer une certaine dynamique, révélant l'évolution des données au cours du temps, trouvant ainsi un bon compromis entre la préservation de la carte mentale de l'utilisateur et la lisibilité du tracé. Le morphing de graphe repose sur deux types de visualisation : la représentation globale et par période.

✓ ***Représentation globale***

La représentation globale sert de carte mentale à l'utilisateur, elle est à l'origine de toute visualisation temporelle dans VisuGraph. Les données sont placées spécifiquement, comme indiqué dans la section 4.5, selon leurs spécificités temporelles.

« La stabilité est une notion complexe qui dépend des caractéristiques géométriques et combinatoires du tracé, mais aussi des facultés de perception et de mémorisation de l'utilisateur » (Pinaud et Kuntz, 2004).

Afin de maintenir une bonne interactivité avec l'utilisateur, il faut préserver au mieux la stabilité des tracés. Pour cela, l'utilisateur doit prendre appui sur le graphe global, c'est à dire sa carte mentale. Les perturbations apportées sur le nouveau tracé de graphe par période, par rapport aux précédents doivent être limitées. A travers le graphe global, l'utilisateur visualise toutes les données, pour toutes périodes confondues,

Il est donc indispensable que le graphe global de référence soit intelligible et clair, permettant, sans grand effort cognitif, de situer chaque graphe période dans le graphe global.

Il convient que ce dernier repose sur un placement fixe par défaut des sommets, selon des repères spatiaux précis, permettant une mémorisation simple de la représentation.

✓ ***Représentation par période***

Afin de pouvoir étudier chaque période individuellement, nous proposons, de réduire ce dernier à des graphes temporels appelés « graphes de périodes ». Pour une période spécifique, tous les sommets et toutes les arêtes n'appartenant pas à l'instance d'étude sont masqués. Il ne reste alors que les données propres au moment analysé. Le choix de la période à visualiser s'effectue par le biais d'un slider, actionné par l'utilisateur.

Le passage d'un graphe à l'autre s'effectue par disparition progressive des sommets appartenant au graphe d'origine mais pas à la représentation finale, par apparition progressive des éléments nouveaux et par évolution des persistants (Gay et Loubier, 2008).

La

illustre ce principe, extrait à partir d'un graphe global sur lequel les sommets temporels ont été positionnés selon leurs appartenances aux différentes périodes.

Dans la structure suivante, nous nommons

G_g le graphe global, représentant toutes les périodes confondues ;

G_t le graphe de période t avec $t \in \{1, \ldots, r\}$ et r le nombre d'instances considérées.

$$G_g = \{ G_t \}$$

G_t est structuré de la façon suivante:

$$G_t = \{<X_t, A_t>\}$$

X_t est l'ensemble des sommets de la période t pour $X_t \subset X$ et $A_t \subset A$. X et A étant l'ensemble des sommets et des arêtes que nous considérons dans tous nos travaux.

f_m la fonction de morphing qui transforme un graphe de départ, nommé G_d en graphe transformé, appelé G_a, *avec* $d \subset t$ *et* $a \subset t$.

$$\text{Ainsi } f_m(G_d)=G_a$$

$$f_m(G_a)^{-1} = G_d$$

Appliqué à un exemple simple, si deux périodes sont considérées, basées sur les deux matrices suivantes, croisant des auteurs A1, A2, A3 *:*

Période 1			
	A1	A2	A3
A1	1	1	0
A2	1	2	0
A3	0	0	0

Période 2			
	A1	A2	A3
A1	1	0	1
A2	1	0	0
A3	1	0	1

$n=2$, c'est-à-dire le nombre de périodes

$t \in \{1, 2\}$

$d \in \{1, 2\}$

$a \in \{1, 2\}$

$G_1 = \{<A1,A2>, <E_{<A1,A2>}>\}$,

$G_2 = \{<A1,A3>, <E_{<A1,A3>}>\}$

$G_{glob} = \{G_1, G_2\}$

$f_m(G_1) = G_2$

$f_m(G_2)^{-1} = G_1$

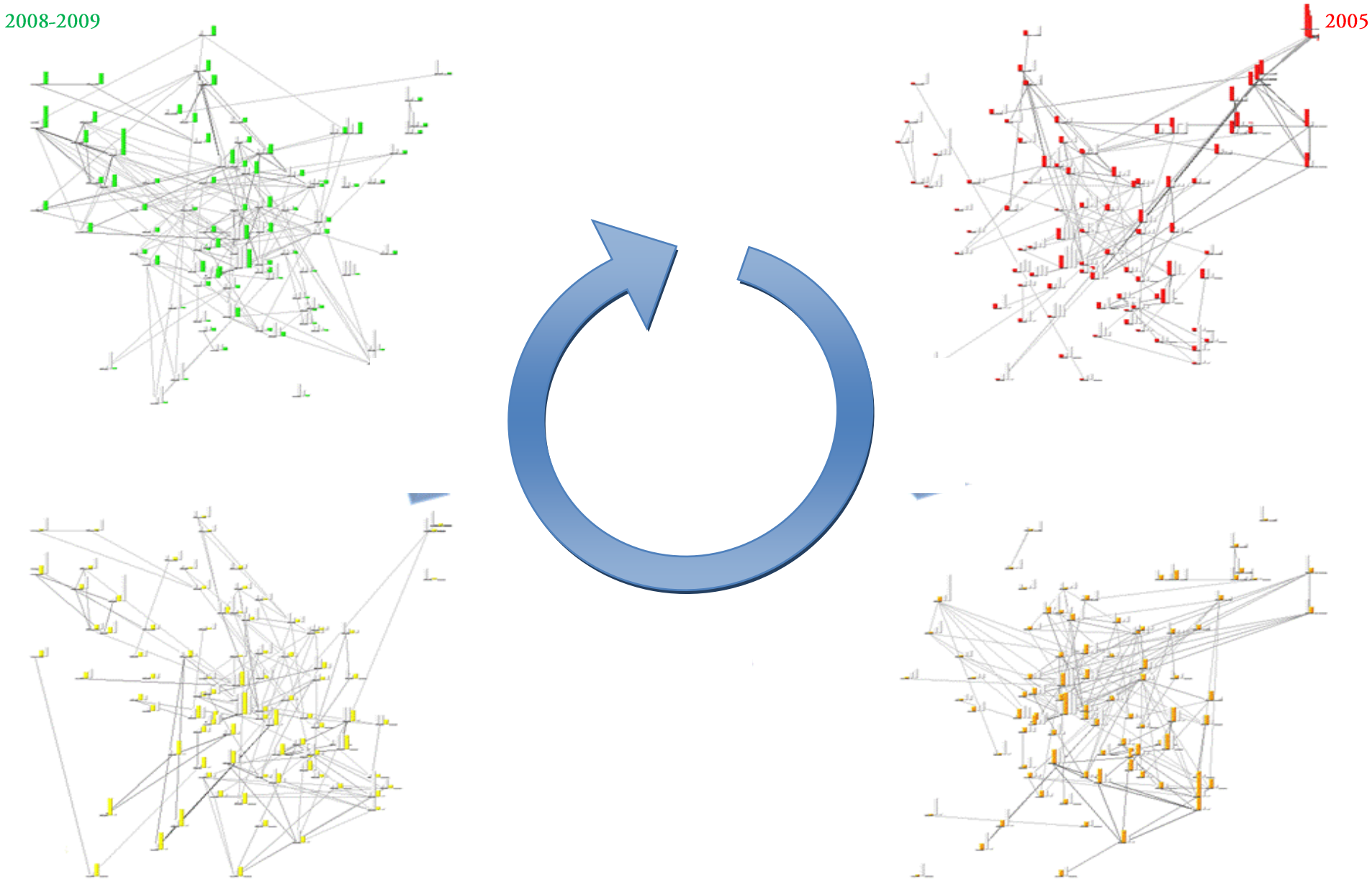


Figure 92. *Graphes de période.*

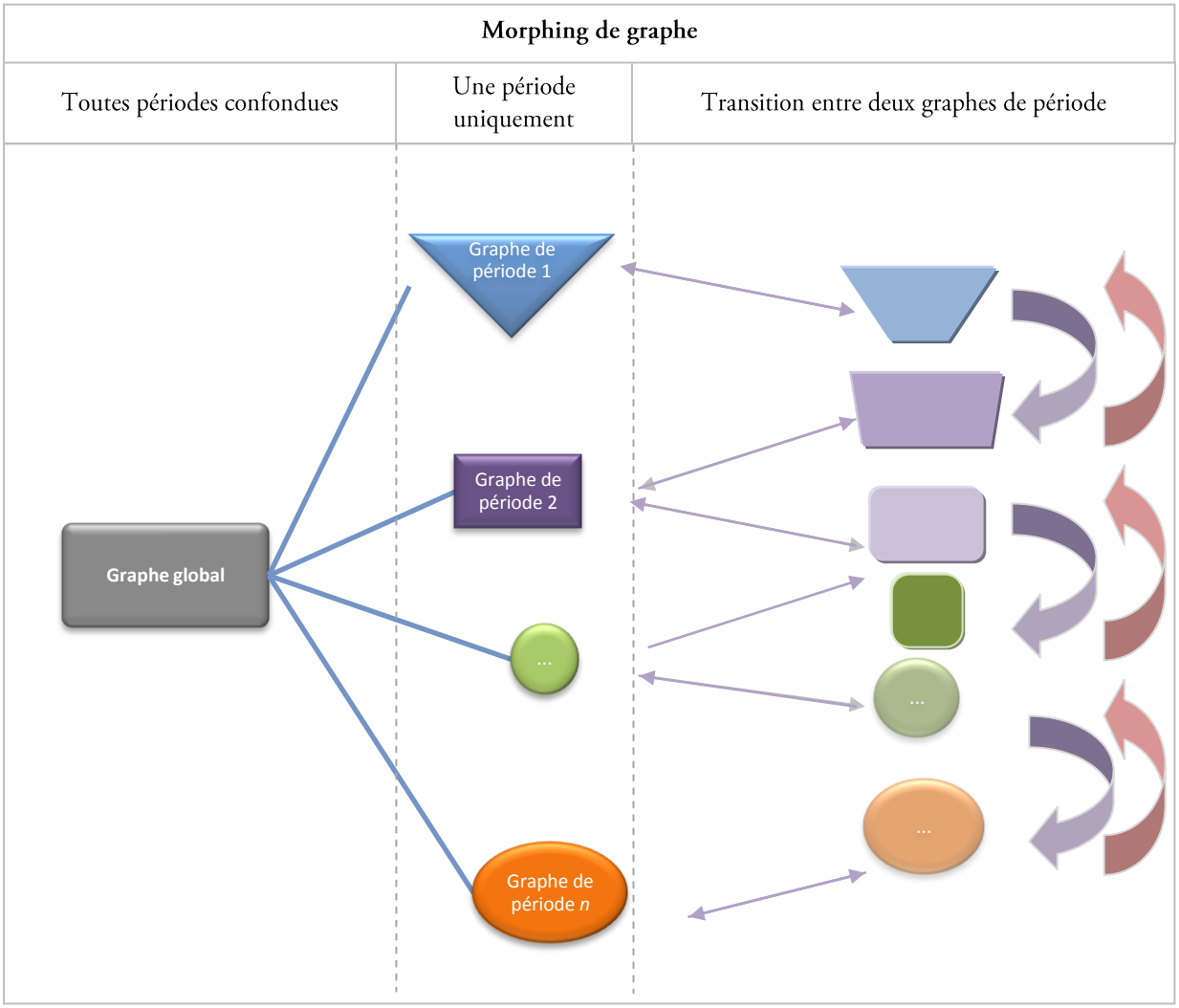


Figure 93. *Principe du morphing de graphe.*

Dans le cadre du morphing, le passage entre les graphes temporels, s'effectue par décomposition en dix étapes de la transformation du premier graphe en second. Ce nombre a été choisi expérimentalement, permettant la décomposition du changement de graphe de façon progressive.

Pour cette animation, on considère trois groupes de sommets/arêtes (Loubier et al., 2008).

- Ceux persistants, présents au cours de la période de départ et de celle d'arrivée;
- Ceux disparaissant, présents au cours de la première période et disparaissant dans la seconde;
- Ceux apparaissant, absent durant la première période mais présents dans celle d'arrivée.

Afin d'obtenir une déformation de graphe fluide, les plus gros sommets, n'apparaissant pas dans la période d'arrivée, disparaissent en premier, par diminution progressive de leur taille. Les plus petits sommets sont traités en dernier. Pour les sommets apparaissant, il en est de même. Les plus gros sommets naissent dès le début des dix périodes et les plus petits apparaissent au dernier moment. Pour les sommets persistants, la progression/diminution des sommets se fait progressivement au cours des dix étapes.

L'utilisateur peut régler la vitesse de l'animation, par le biais d'une règle graduée. Plus elle est élevée, plus l'apparition/disparition/changement des sommets et des arêtes se fait rapidement et inversement.

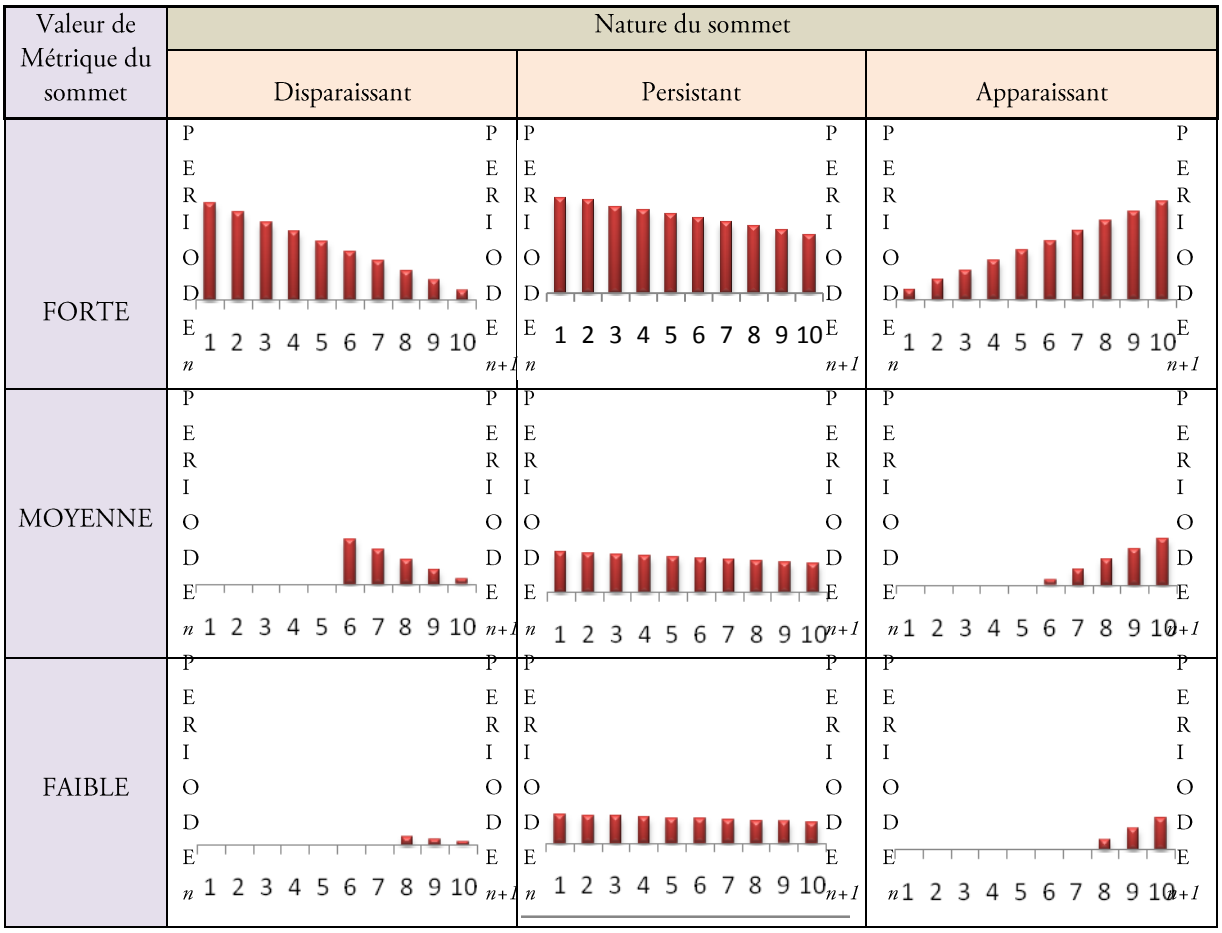


Figure 94. *Schématisation de l'apparition/ disparition / persistance d'un sommet selon sa valeur de métrique, en dix étapes, par morphing de graphe.*

Le changement de couleur d'un sommet lors du passage d'une période à une autre doit être progressif. La solution la plus simple semble être l''ajout/soustraction d'un incrément fixe aux valeurs RVB[27] du sommet. Par exemple, la couleur rouge et le vert sont codés en RVB de la manière suivante

Rouge→(255 ;0 ;0)

Vert →(0 ;255 ;0)

Prenons l'exemple de *2* périodes considérées. Pour attribuer une couleur dégradée du rouge au vert à chacune des instances, la couleur attribuée pourrait être élaborée selon l'algorithme suivant. Pour chaque étape, la valeur correspondant au rouge est décrémentée et le seuil du vert est incrémenté, permettant une disparition du rouge pour laisser place au vert. « *temp* » est une variable temporaire dont la valeur finale indiquera le dosage de rouge et de vert finaux.

```
Int temp = 255;
Int r = 255;
Int v = 0;
Pour i allant de 1 à 2
{
        couleur_i = (r,v,0);
        r-=temp ;
        v+=temp ;
}
```

[27] Le système RVB (initiales de Rouge-Vert-Bleu) permet d'obtenir par mélange toutes les couleurs. Le système RVB est une des façons de décrire une couleur en informatique. Ainsi le triplet {255, 255, 255} donnera du blanc, {255, 0, 0} un rouge pur, {100, 100, 100} un gris, etc. Le premier nombre donne la composante rouge, le deuxième la composante verte et le dernier la composante bleue.

Cependant, l'application de cet algorithme est contrainte à des difficultés de distinction des couleurs. Nous avons codé cet algorithme en langage java, afin d'obtenir la Figure 95, contenant des dégradés de couleur allant du rouge au vert, qui s'effectue en dix étapes.

Dans ce cas là, temp = $\frac{255}{9}$ = 28.

Cet exemple met en avant l'effet de la subjectivité sur l'appréciation des couleurs. Chez l'homme, la rétine est sensible aux rayonnements électromagnétiques. Par ailleurs, la réponse de l'œil à la lumière n'est pas linéaire. Nous sommes plus sensibles à des changements qui se produisent dans les basses lumières que dans les hautes lumières. En pratique cela se manifeste par une vision beaucoup plus précise et détaillée dans les zones d'ombres que dans les zones brillantes. Ce phénomène a un impact direct sur la restitution des couleurs sur un écran.

Deux teintes successives sont difficilement distinguables et le pas entre deux teintes est trop régulier, rendant moins nette la séparation entre les deux périodes. Lorsqu'une donnée passe d'une étape *n* à *n+1,* il est important que la couleur traduise l'effet de présence dans la première instance, sa diminution, sa disparition puis son apparition dans la nouvelle période jusqu'à sa stabilisation. Dans un contexte d'optimisation de la qualité de perception de notre outil de visualisation, cette solution ne peut être retenue, puisqu'elle complique la tâche de l'utilisateur dans sa perception des tendances et plus particulièrement dans l'étude de l'évolution individuelle des données.

étapes	Rouge	Vert	Bleu	Résultat
1	255	0	0	
2	227	28	0	
3	198	57	0	
4	170	85	0	
5	142	113	0	
6	113	142	0	
7	85	170	0	
8	57	198	0	
9	28	227	0	
10	0	255	0	

Figure 95. *Problème d'ergonomie dans cette palette de couleurs allant du rouge au vert de façon incrémentale.*

Pour éviter ces contraintes d'impression de fusion visuelle de certaines couleurs proches, il est important de créer une répartition non linéaire.

Pour cela, considérons les notations suivantes.

k est une constante dont la valeur permet de réduire la linéarité de la courbe résultant de la fonction établie .

n est le nombre d'étapes *;*

t est une constante $\in \{\frac{1}{n}, \frac{2}{n}, \frac{3}{n}, \dots, \frac{n}{n}\}$;

x est la saturation de la couleur.

$$x = \frac{t + k(\cos\ (\Pi \times t) + 1}{2k + 1} \quad [27]$$

Appliqué à l'exemple des dix étapes entre deux périodes vues précédement, les résultas obtenus sont les suivants :

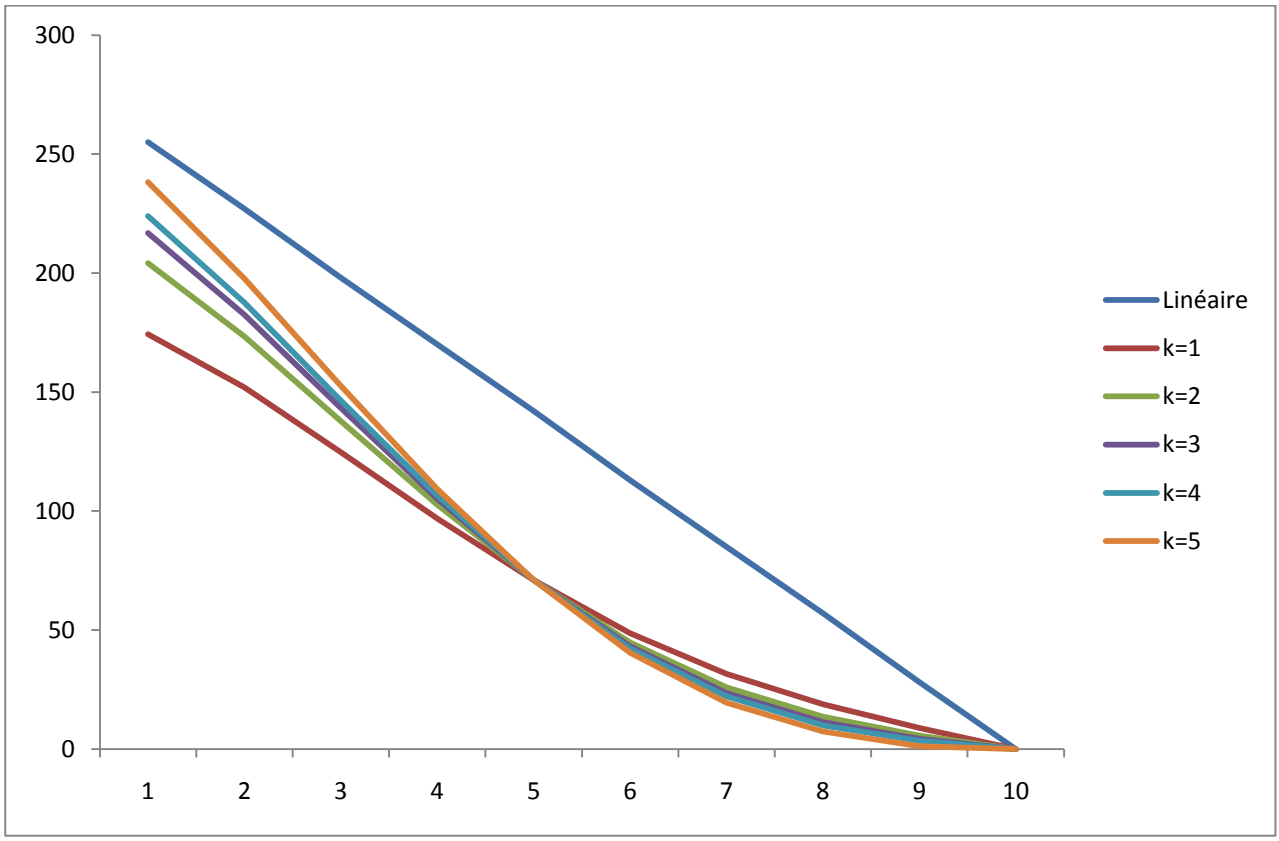


Figure 96. *Répartition non linéaire de la teinte rouge.*

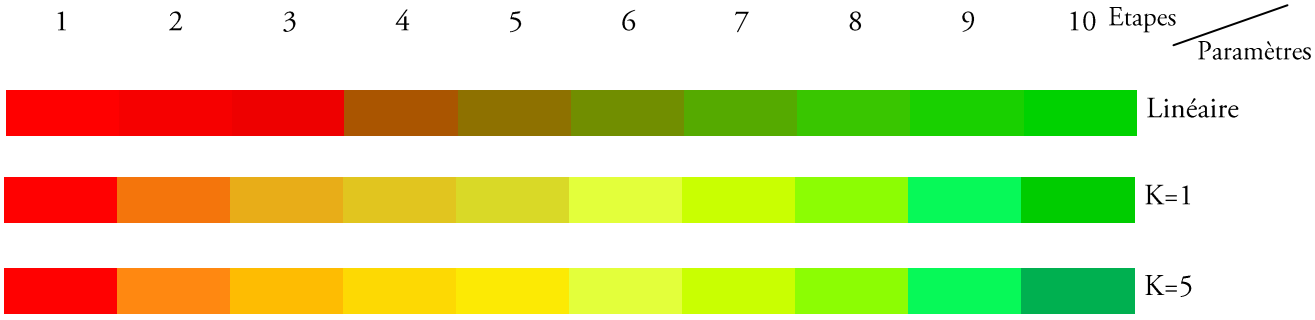


Figure 97. *Comparaison entre le passage du rouge au vert linéaire et non liénaire*

Ainsi, la formule proposée permet de distinguer plus instinctivement la phase de disparition de la période vers la suivante. Dans VisuGraph, la couleur de transition de chacune des données durant le morphing est calculée avec un paramètre *k=5*. Cette valeur met en avant les couleurs des extrémités, étapes 1 et 10, et facilite la distinction des différentes étapes.

4.8.3. Morphing sans transition

Le passage entre deux périodes peut s'effectuer par un morphing simple sans transition spécifique. Au départ se trouve un graphe G_d et à l'arrivée se trouve G_a.

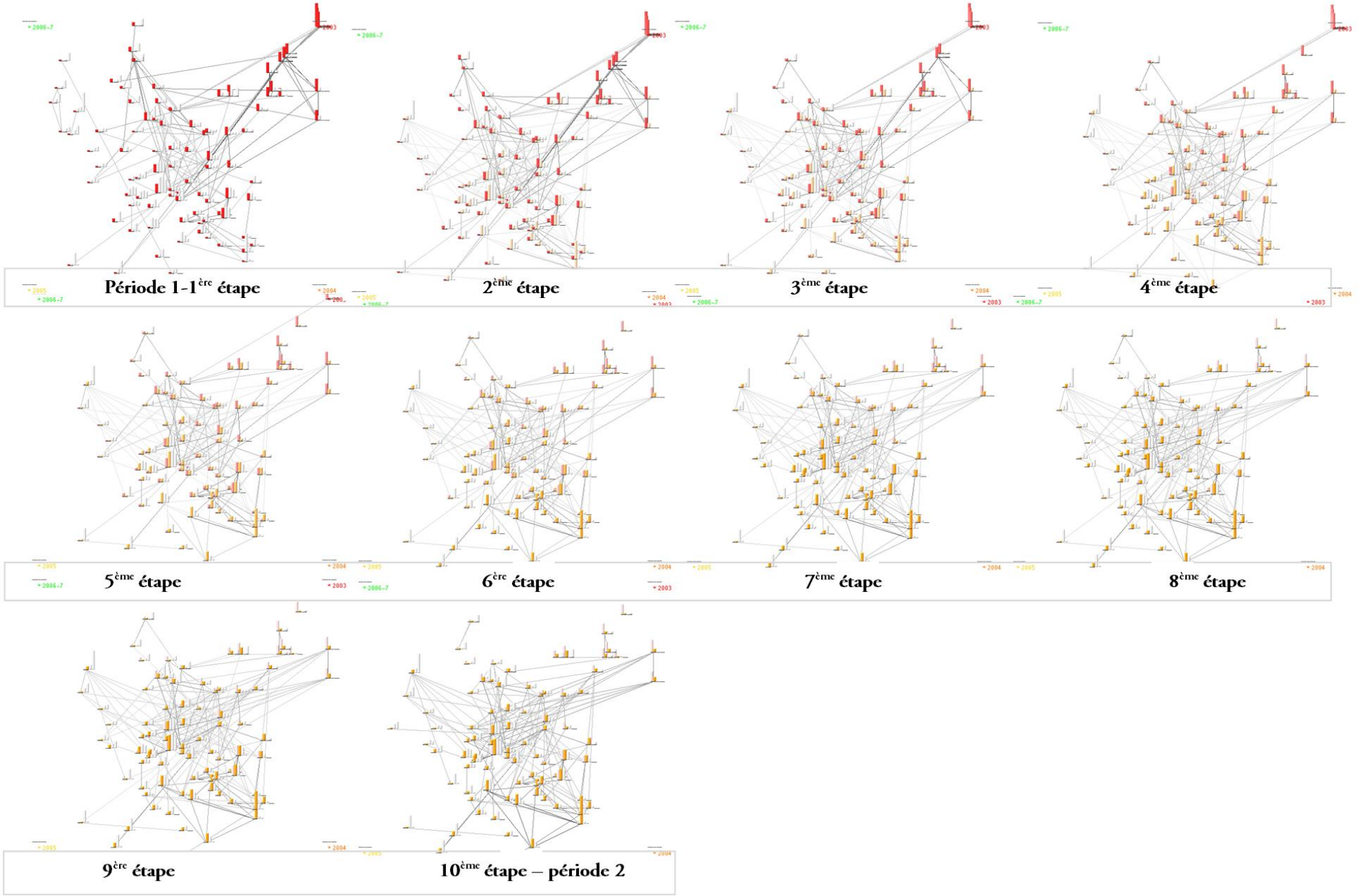


Figure 98. *Passage du graphe de première période à celui de la seconde, par morphing de graphe en 10 étapes.*

Le passage de l'un à l'autre s'effectue par disparition successive des sommets présents dans G_d et absent dans G_a, par affichage progressif des données apparaissant en G_a et par changement progressif des éléments persistants mais dont la valeur peut varier.

Ainsi, l'aperçu de morphing conserve la carte mentale puisque les sommets de G_d et G_a conservent la même position que dans la représentation globale.

L'utilisateur conserve ses repères visuels et peut aisément suivre l'évolution d'une donnée particulière, sans la confondre avec une autre. Une fois G_a obtenu, l'utilisateur peut appliquer toute une série de fonction, que nous détaillons en section 4.10. La déformation du graphe est possible afin d'améliorer la lisibilité du graphe, ou encore afin de l'étudier sous un autre point de vue. Si les positions temporelles des sommets ont été changées, l'application du morphing sur G_a passe dans un premier temps par un retour aux placements initiaux, puis par une transformation du graphe pour obtenir le graphe d'une autre période.

La Figure 98 illustre le passage d'une première période à la seconde par morphing de graphe en dix étapes. Les sommets du graphe de première instance sont positionnés selon leur appartenance aux différentes périodes, dans un contexte de graphe global. Les sommets n'appartenant qu'à la première période disparaissent en premier, de manière progressive, puis les sommets de la seconde période apparaissent, transformant progressivement le graphe de départ.

Le nombre d'entrecoupements d'arêtes est important et peut gêner l'utilisateur dans son analyse de la structure de graphe temporel. Un élément de réponse à ce problème est le morphing avec transition, illustré dans le paragraphe suivant.

4.8.4. Morphing avec transition

L'organisation des éléments du graphe dans un contexte de morphing sans transition peut sembler difficile à analyser pour l'utilisateur et il est important d'offrir des solutions rendant le graphe de période plus lisible. Pour ce faire, VisuGraph est doté de deux types de paramètres optionnels afin d'organiser le morphing selon différents points de vue.

- ✓ ***Le point de vue du repère temporel de l'instance considérée***

Dans ce cas là, le positionnement des sommets est relatif au repère temporel de la période représentée. Partant du graphe global, le morphing vers la représentation de la première instance s'effectue par déplacement et organisation des sommets autour du premier repère temporel, sans tenir compte de l'éventuelle appartenance aux autres instances. Le passage à la seconde période s'effectue par un retour des sommets en position initiale, afin de symboliser la fin de la première instance et de conserver la carte mentale, puis par déplacement et organisation des sommets vers le second repère temporel. Il en est de même pour les autres périodes étudiées. Cette solution, illustrée en Figure 99, a pour avantage de favoriser le contexte temporel en cours, mais d'un point de vue de la carte mentale, l'utilisateur peut avoir des difficultés pour repositionner un sommet sans afficher son nom. Pour cette raison, cette solution est préconisée principalement pour une étude globale des tendances, à savoir quelle quantité de données concerne telle période ? Les informations relatives à cette instance sont elles de fortes valeurs ? Sont-elles des signaux faibles ?

- ✓ ***Le point de vue de la structure temporelle***

Dans ce cas là, la structure évolutive est favorisée ainsi que la mise en évidence de *classes temporelles graphique*. L'objectif est de regrouper les données de même(s) nature(s) temporelle(s). Ainsi, les données spécifiques à une période particulière sont regroupées. Graphiquement, cette notion se traduit par la formation de clusters temporels et visuels, au sein de chaque graphe d'instance. Ce point de vue permet de décomposer la structure d'un graphe de période afin de l'analyser plus en détails. Pour obtenir de tels groupes, l'algorithme FDP temporel, présenté en section 4.6 est intégré avec une forte attirance vers les repères temporels, au morphing de graphe.

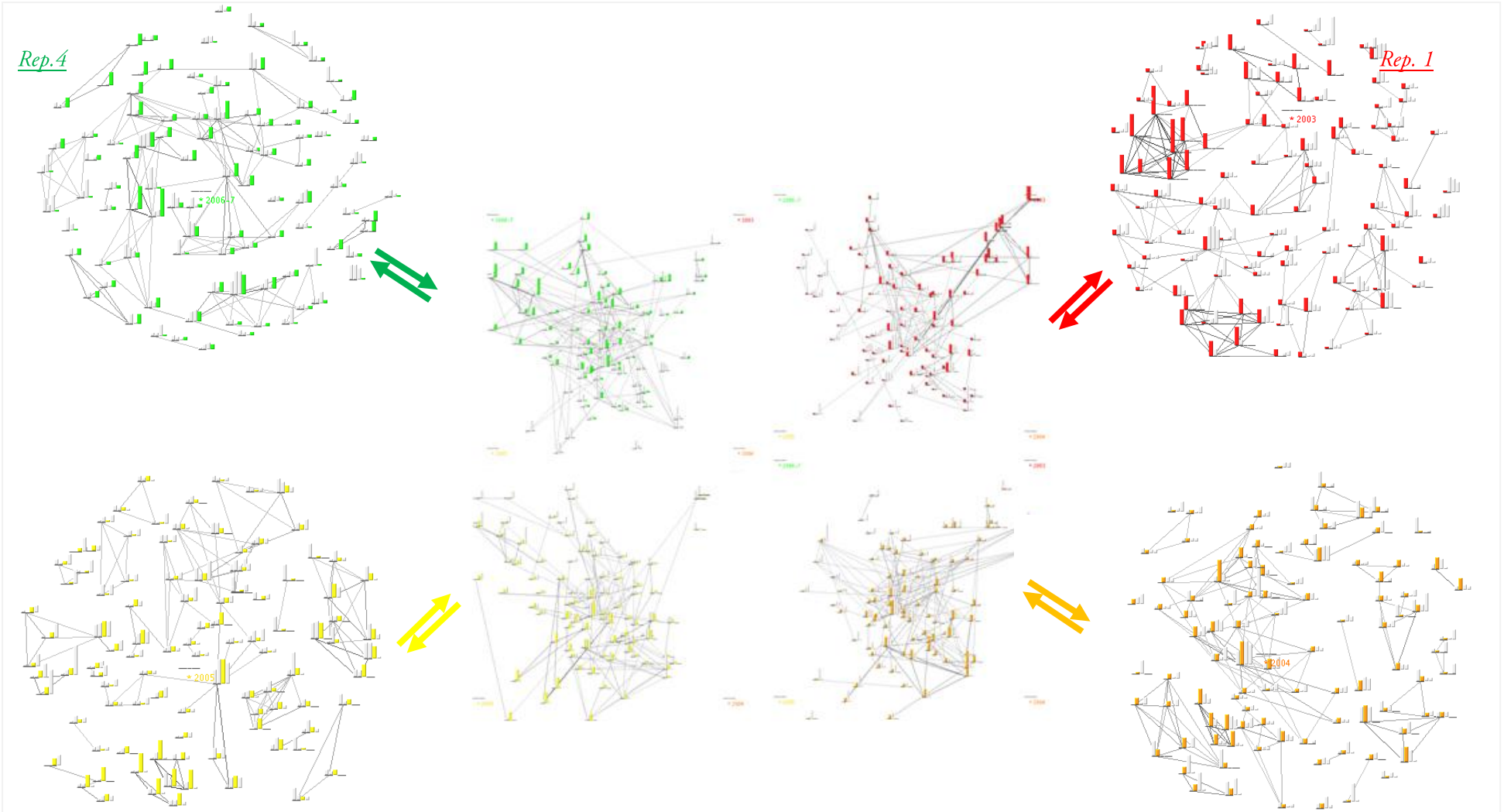


Figure 99. *Mophing avec transition, d'un point de vue « repères temporels ».*

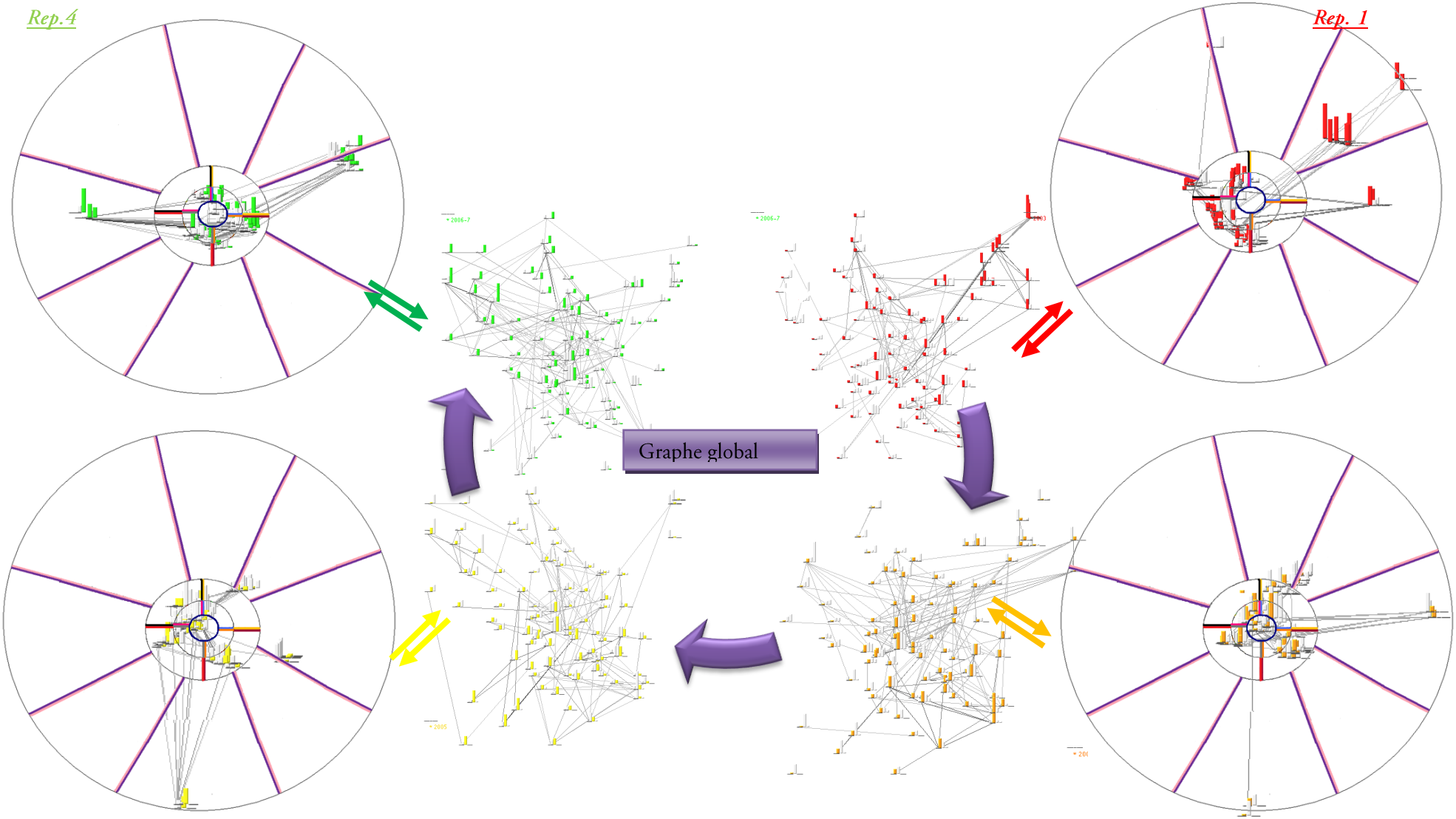


Figure 100. *Morphing avec transition, d'un point de vue de la structure temporelle. transition, d'un point de vue « repère temporel*

Cette solution déforme la structure du graphe global, en renforçant les liens temporels. Un retour à la structure du graphe global lors du changement de période permet de se remémorer le positionnement de chaque sommet et minimise l'effort cognitif.

La Figure 100 illustre ce principe par morphing de graphe de quatre périodes. La décomposition en zones temporelles, illustrée en Figure 100 est appliquée à chaque graphe de période, afin de faciliter l'interprétation évolutive de chaque groupes constitué. Chaque graphe d'instance est alors composé de groupes de sommets spécifiques à la période considérée, où alors à un ensemble ayant les mêmes caractéristiques temporelles.

4.8.5. Discussion sur le morphing de graphe

Le Tableau 19 synthétise les différences types de morphings de graphe proposés.

Type de morphing	Point de vue	Type d'analyse	Avantages	Inconvénients
Sans transition		Analyse de la structure évolutive de la période et de la suivante ou de la précédente	• Distinction claire des sommets spécifiques à la période. • Mise en avant de l'appartenance des sommets à plusieurs périodes. • Prédiction possibles des sommets qui seront présents dans la période suivante. • Etude des liens entre les données et de leur évolution. • Détection des signaux faibles. • Carte mentale préservée. • Retour aisé au positionnement d'un auteur dans le graphe global mais aussi dans les graphes de périodes. • Si perte de la carte mentale par application de méthodes, retour aux positionnements temporels initiaux.	• Graphe de départ et d'arrivée avec des entrecoupements d'arêtes. • Etude de la structure d'un point de vue du repère de la période restreint.
Avec transition	Repères temporels	Analyse de la période dans son contexte	• Détail de la structure de la période étudiée • Moins d'entrecoupements d'arêtes • Retour aux positionnements temporels avant le changement d'instance. • Mise en avant de l'importance de la période vis-à-vis de l'ensemble des données. • Grande lisibilité des liens • Mise en avant de la structure formée par les sommets.	• Regroupement des données ne favorisant pas l'analyse de structure temporelle. • Perte de la carte mentale.
	Structure temporelle	Analyse de la période dans un contexte global	• Distinction entre les zones temporelles. • Facilité de caractérisation évolutive des données. • Si perte de la carte mentale par application de méthodes, retour aux positionnements temporels initiaux. • Distinction claire des sommets spécifiques à la période. • Mise en avant de l'appartenance des sommets à plusieurs périodes. • Prédiction possibles des sommets qui seront présents dans la période suivante.	• Etude de l'évolution des liens non mise en avant. • Perte de lisibilité par superposition de sommets, si l'attraction des sommets est trop forte.

Tableau 19. *Synthèse des trois propositions de morphing*

Ces trois points de vue de morphing peuvent être obtenus sur le même graphe de façon successive. Ainsi, si l'analyse porte principalement sur la structure évolutive des différentes périodes, l'utilisateur peut dans un premier temps appliquer la première proposition. Puis, s'il s'intéresse plus précisément, à chacune des périodes individuelles, il peut passer à la seconde solution. Enfin, s'il désire analyser chaque période dans un contexte global, la troisième proposition est appliquée.

Il est possible à tout moment de revenir à un autre type de morphing, puisque cette fonction se base sur le positionnement initial temporel des sommets.

Le morphing de graphe permet de :

- *Comparer des éléments de l'espace de données.* Pour cela, l'utilisateur peut déplacer ou permuter des éléments visuels de l'espace de données afin de les rapprocher dans l'espace de représentation pour faciliter leur comparaison. Il peut superposer deux données pour comparer leurs histogrammes ou encore les disposer côte à côte afin de mieux les étudier.
- *Rechercher des effets de causalité entre données.* Cela consiste à étudier l'influence du comportement d'une donnée particulière sur une autre donnée, en mettant en évidence les relations de causalité suite à un événement, entre les changements de valeur, d'une donnée source et une autre donnée cible. Dans le cas où un effet de causalité est identifié, il convient alors d'étudier le temps selon le délai *e* de réaction de la donnée cible suite à un événement de la donnée source.
- *Anticiper pour agir.* Par détection des signaux faibles et par analyse temporelle des données, il est facile de prévoir quels seront les piliers du domaine du futur proche. Le morphing de graphe permet, par le positionnement de détecter, d'une part, les nouveaux individus apparaissant en dernière période et susceptibles d'être présent dans l'avenir, et d'autre part, les piliers, présents au cours de toutes les instances. Dans le cadre d'une étude de structure, il est alors possible de distinguer les orientations du groupe, de façon organisationnel et stratégique ou encore les émergences ou les digressions.
- *Mesurer la similarité entre données temporelles.* Cette comparaison s'effectue via les histogrammes représentant les sommets mais aussi, via le placement temporel des entités, caractérisant leur persistance au cours des différentes périodes étudiées. Une représentation appropriée de plusieurs données temporelles facilite l'identification visuelle des sous-séquences similaires.
- *Segmenter les données temporelles.* La segmentation d'une donnée temporelle consiste en sa partition en périodes individuelles.

4.9. Les graphes orientés

La prise en compte de la dimension temporel dans le graphe orienté permet d'étudier précisément l'évolution des relations entre éléments d'un ensemble. La relation entre un sommet x et un autre y peut être consolidée ou au contraire réduite au cours du temps. On parle alors de graphe orienté temporel.

Un graphe $G = (X, A, T)$ est déterminé par :

- un ensemble $X = \{x_1, x_2, \ldots, x_n\}$ dont les éléments sont des sommets ou nœuds ;
- un ensemble $A = \{a_1, a_2, \ldots, a_m\}$ dont les éléments sont des arcs.
- une période T, pouvant être discrète ou continue.

Les notions de successeurs et de prédécesseurs définit dans le cas non temporel sont repris, sachant qu'ils ne le sont que pour la période T considérée.

A l'origine d'un graphe orienté temporel, sous VisuGraph, se trouve une matrice asymétrique pour chaque période croisant les prédécesseurs et les successeurs. Le principe de morphing de graphe, décrit dans la section 4.8 est applicable sur ce type de graphe.

L'intérêt d'une telle représentation repose sur la décomposition d'une relation entre deux individus X_i et X_j à une période spécifique et de distinguer qui agit de qui subit.

Tout comme pour les graphes statiques, nous permettons dans un premier la prise en compte de la dimension temporelle pour des graphes orientés, en ajoutant un complément d'information pour les sommets. En effet, il est

possible de croiser deux types d'information pour un même sommet, comme par exemple son nom et son pays. Ainsi ces deux renseignements sont codés sous un même sommet à travers l'utilisation de barres pour étudier l'évolution de la donnée, illustré en 4.4, ou encore à travers l'utilisation de couleur spécifique pour coder la seconde information, par le soulignement de l'histogramme.

Dans l'exemple de la Figure 101, nous considérons des sociétés vendant des licences informatiques (LI), à l'origine de l'arc, et celles les achetant (LO), en fin d'arc représenté de couleur uniforme pour toutes les barres. La présence d'une flèche à double sens signifie que les deux sociétés s'achètent et se vendent mutuellement des licences. Pour chaque société, son pays est codé par un trait de couleur horizontal.

Le placement temporel, ainsi que l'application de l'algorithme FDP temporel permet d'obtenir le résultat suivant. Les licenciés et les licenciant spécifiques à une période sont représentés autours du repère temporel correspondant. Les sommets à mi chemin entre les deux repères appartiennent aux deux périodes. On distingue ainsi trois catégories temporelles :

- Les sommets n'appartenant qu'à 2006 ;
- ceux n'appartenant qu'à 2007 ;
- et ceux appartenant aux deux périodes.

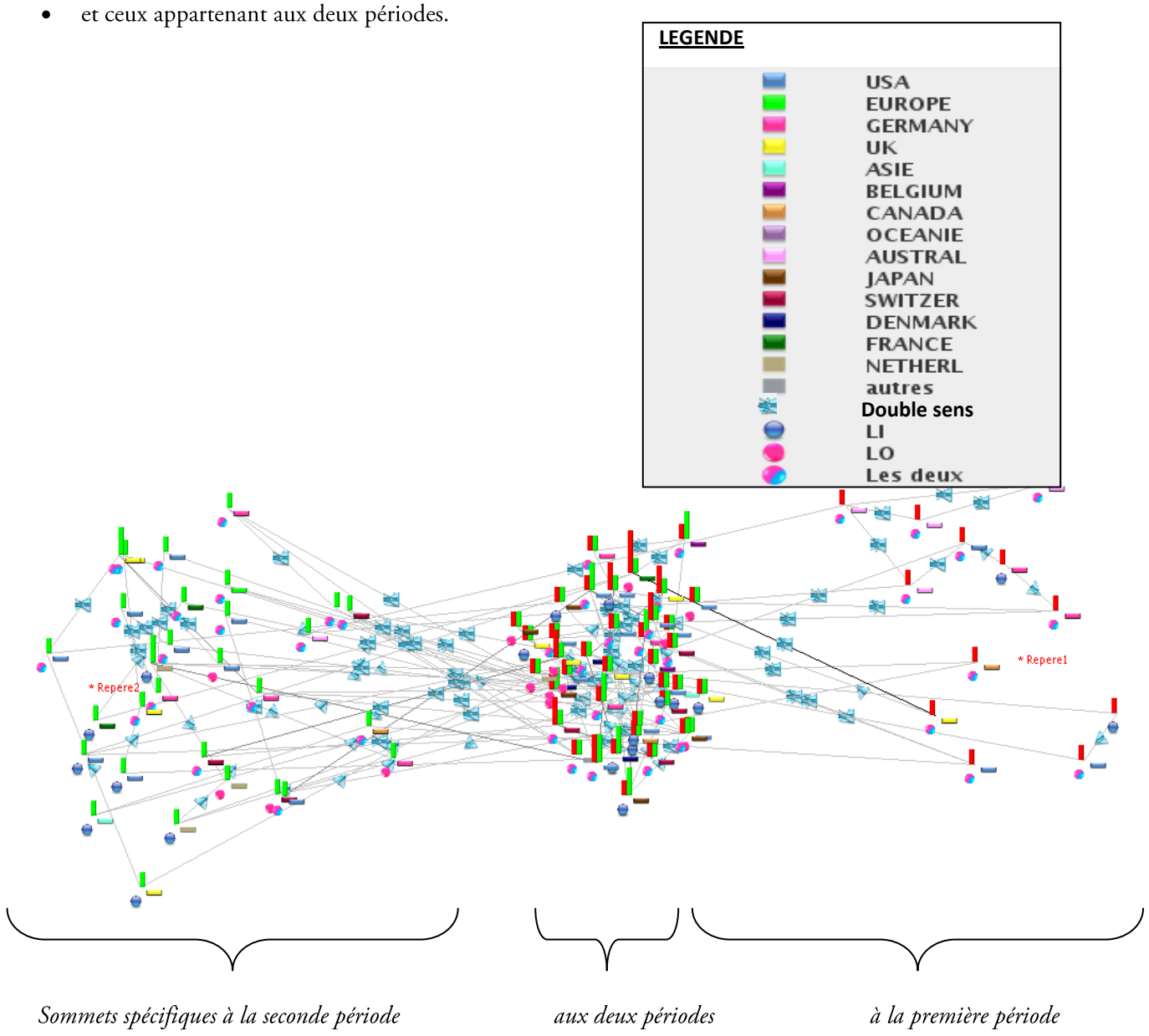


Figure 101. *Graphe orienté biparti temporel.*

Si l'utilisateur souhaite connaître de façon plus claire le domaine médical d'un sommet, il suffit d'afficher son nom ou la fenêtre indiquant les caractéristiques de la donnée, comme le montre la Figure 101.

L'application du morphing permet d'extraire de ce graphe global les périodes 2006 et 2007. Ainsi, il est facile d'étudier en quelle année la licence a été vendue et/ou achetée et si elle a été reconduite en 2007.

4.10. Fonctionnalités

Dans cette section, les fonctionnalités étudiées dans le chapitre 3 sont adaptées au cas temporel.

4.10.1. Points de vues

Dans le contexte des graphes évolutifs, deux sortes de points de vue sont considérés :

- ✓ ***Le point de vue global***

Ce premier point de vue concerne l'application des fonctions sur le graphe global, toutes périodes confondues. Dans ce contexte là, toutes les fonctions présentées dans le chapitre 3 sont applicables. Cependant il faut être conscient du risque d'erreur d'interprétation, provoqué par une analyse du graphe global. En effet, un graphe connexe qui représente toutes les données temporelles pour toutes les périodes cumulées peut ne jamais avoir été connexe s'il est étudié dans chaque cas d'instance individuelle. Ainsi, les résultats des fonctionnalités présentées dans le chapitre 3 sur le graphe global peuvent aussi être erronés si les données de la visualisation globale sont hétérogènes temporellement.

Ce point de vue discerne surtout une analyse de structure globale qui est ensuite étudiée en détail pour chaque période. Cette information peut être primordiale selon les besoins exprimés. De la structure globale étudiée peuvent être extraits les représentations de période, permettant d'analyser la visualisation générale dans un contexte individuel.

- ✓ ***Le point de vue évolutif***

Le point de vue évolutif cible l'application des fonctions sur chaque graphe de période de manière individuelle. Ainsi cette analyse se dissocie de l'étude du graphe global. Dans le cas du point de vue global, la structure générale est conservée et visualisée dans des contextes d'instances individuelles, ici nous ciblons chaque composition de manière indépendante. Afin d'illustrer ces deux points de vues, l'exemple illustré, Figure 102, cible le *k-core* appliqué dans un premier temps à un graphe global.

Dans la colonne de gauche sont représentés, de façon chronologique, les différents graphes de périodes d'un point de vue global. Le *k-core* n'est appliqué qu'au graphe global et le morphing est appliqué sur la structure obtenue. La visualisation générale se décompose en graphes d'instances de façon à ce que les sommets aient une valeur de métrique positive pour la période considérée et que la structure globale du *k-core* soit conservée. Dans le graphe global, on constate qu'une donnée concerne la période 2. Or, la deuxième ligne de la colonne de gauche montre un graphe vide. Cela s'explique par le fait que la structure obtenue à partir du seuil du *k-core,* fixé dans le graphe global ne persiste pas en deuxième instance.

Dans la colonne de droite se situent les graphes temporels sur lesquels est appliqué individuellement le *k-core* d'un point de vue évolutif, c'est-à-dire par application de cette fonctionnalité sur chacune de ces représentations. Ainsi la valeur de *k* varie d'un graphe d'instance à un autre, de manière à obtenir le coreness le plus élevé pour chacun des cas.

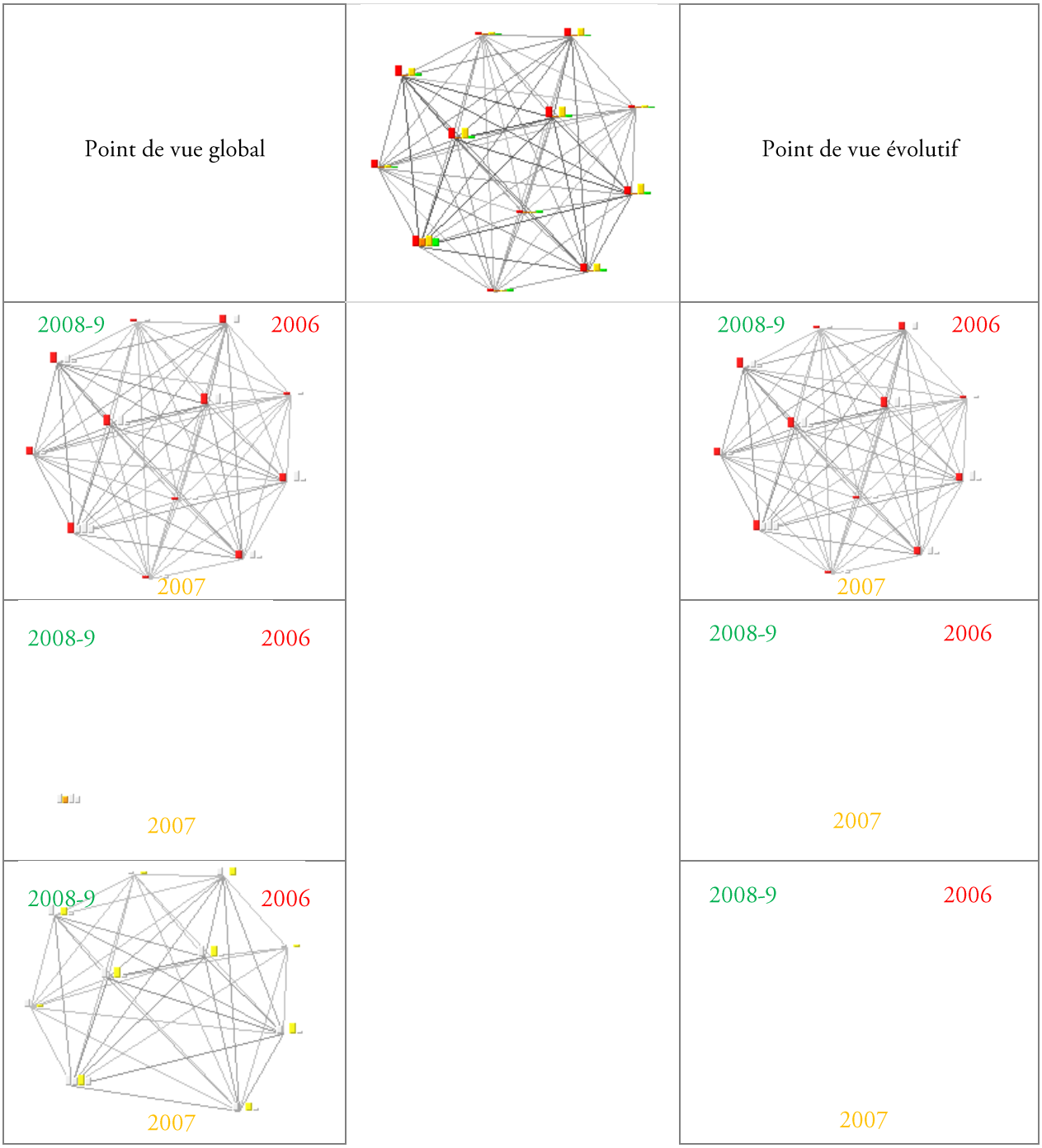


Figure 102. *Différences de point de vue sur l'application des fonctions.*

4.10.2.Filtrage

Quelque soit le point de vue favorisé, le filtrage s'applique aux deux cas, permettant deux focalisations différentes. Dans le cas « global », le filtrage est appliqué au graphe représentant toutes les périodes et le filtrage est appliqué, tel qu'il a été présenté dans le chapitre 3.

Une étude de chaque instance masque uniquement les données, ainsi que leurs liens, absents lors de cette période. Ainsi certains sommets peuvent apparaître seuls dans les visualisations de période, du fait qu'ils étaient liés dans le contexte global à d'autres données, mais qu'ils ne le sont plus dans un cas temporel individuel.

Ainsi, la fonction de filtrage, d'un point de vue général se formalise de la façon suivante :

$$f_F\,(G_{glob},\ Filtre{=}n){=}G_{gl}^{F}$$

ou G_g est le graphe global, $G_g = (X_g, A_g)$ avec X_g l'ensemble des sommets et A_g l'ensemble des liens, pour toutes périodes confondues.

Toute arête appartenant à A_g a une structure de la forme

$a_{<xi,xj,\ t>} = <x_i, x_j, q, t>$

x_i, x_j, sont les extrémités de l'arête

q est la valeur de la métrique de l'arête

t est l'instance à laquelle apparait l'arête.

G_g^F est le graphe résultant, c'est-à-dire filtré.

n est la valeur du seuil du filtre, c'est-à-dire la valeur minimale que devra prendre la valeur quantitative de $a_{<xi,xj,t>}$ dans le graphe filtré G_g^F.

Les graphes de période filtrés se caractérisent de la manière suivante :

$$G_t^F = G_g^F - \{X_0, A_0\}$$

Avec G_t^F le graphe filtré de la période t. $t \in \{1, .., r\}$ avec r le nombre d'instances composant le graphe global.

X_0 l'ensemble des sommets de X_g dont la valeur de métrique pour l'instance t est nulle.

A_0 l'ensemble des arêtes d'A_g dont la valeur de métrique pour l'instance t est nulle.

Dans le cas d'un point de vue évolutif, le filtrage est recalculé pour chacun des graphes de périodes, Indépendamment de la structure filtrée obtenue sur la représentation générale. La focalisation sur chaque instance met en avant l'évolution de la structure, c'est-à-dire les relations entre les données au cours du temps, plutôt que l'évolution individuelle des données. Dans ce cas là, chaque graphe filtré possède la même structure que celle définit dans le chapitre précédent.

4.10.3.K-Core

Le *k-core* appliqué au cas temporel a été illustré dans la Figure 102. L'élagage récursif des nœuds qui ont un degré plus petit que k s'effectue sur la visualisation générale, dans le point de vue global. Le graphe restant ne contient que des sommets de degré ≥ k.

Nommons f_{core} cette fonction :

$$f_c(G_g, seuil=n) = G_g^K$$

G_g, est le graphe global, selon la définition initialement posée dans laquelle $G_g = (X_g, A_g)$ avec X_g l'ensemble des sommets et A_g l'ensemble des liens, pour toutes périodes confondues

n est le seuil permettant de calculer le *k-core* et permettant de caractériser le *coreness*.

$n = \{0,\ldots,k\}$.

k est une valeur calculée dès le premier appel de la fonction. Elle correspond au *coreness* maximal pouvant être atteint.

G_g^K est le graphe obtenu pour le seuil fixé, composé de sommets ayant tous au moins n voisins.

Les graphes de période se caractérisent de la manière suivante :

$$G_t^K = G_g^K - \{X_0, A_0\} - \{<X_t, A_j, n>\}$$

Avec G_t^K le graphe résultant de l'application du *k-core* de la période t. $t \in \{1, .., r\}$ avec r le nombre d'instances composant le graphe global.

X_0 l'ensemble des sommets de X_g dont la valeur de métrique pour l'instance t est nulle.

A_0 l'ensemble des arêtes de A_g dont la valeur de métrique pour l'instance t est nulle.

$<X_t A_j, n>$ est l'ensemble des sommets de l'instance t dont le degré est supérieur ou égal à n et A_j l'ensemble des sommets correspondants.

D'un point de vue évolutif, cette fonction est appliquée à chaque graphe de période.

4.10.4. Transitivité

L'étude de la transitivité, d'un point de vue global, permet de retrouver sur un ensemble de périodes le voisinage d'un sommet. L'intérêt de ce point de vue, vis-à-vis de cette fonctionnalité est d'étudier la progression de ces différentes associations entre le sommet sélectionné et ses associés (Loubier et Gay, 2009).

Nommons f_t cette fonction de transitivité, appliquée à un graphe $G_g = (X_g, A_g)$.

$$f_t(G_g, x_g, seuil=n) = G_g^T$$

G_g est le graphe global sur lequel va être calculé la fermeture transitive.

x_g est le sommet du graphe global initialement choisi, appartenant à X.

n est le seuil fixé pour le calcul de la fermeture transitive. $n = \{1,\ldots,M\}$. M est le seuil de transitivité maximal fixé dès le premier appel de la fonction. Il s'agit du nombre de pas permettant d'obtenir le voisin le plus éloigné du sommet x_g.

G_g^T est le graphe résultant après calculé la transitivité de seuil n.

Dans le cas du point de vue évolutif, chaque fermeture transitive de chaque graphe de période est calculée indépendamment, permettant d'étudier le voisinage directement lié d'un même sommet, à des instances différentes. Cette approche met en avant l'évolution du nombre de collaborations du sommet sélectionné, à savoir les périodes caractérisées par de faibles collaborations et inversement les années fastes, ainsi que la persistance de ces liens.

4.10.5. Partitionnement de graphe

Le principe de partitionnement de graphe, par la méthode du Markov Clustering, présentée dans le chapitre 3 est adapté au cas temporel. Le graphe global sert de base lors de l'analyse. C'est sur ce dernier que s'effectue le partitionnement des données, sans prendre en compte les arcs liants les sommets aux repères temporels, afin de conserver les propriétés classiques de classification. Une fois le graphe réduit obtenu, les liens entre sommets et repères temporels sont instaurés et l'application de l'algorithme FDP permet de placer chaque représentant de classe selon ses caractéristiques temporelles.

La visualisation successive de chaque graphe de période permet de les comparer, au niveau de leur structure inter classes mais aussi au niveau des sommets composant les différentes classes (intra classe), (Loubier et al., 2007b). L'expérimentation effectuée se base sur un corpus portant sur les auteurs ayant publié un article lors des quatre dernières sessions du colloque VSST.

Il est important de noter que toutes les classes sont reliées pour l'ensemble des quatre périodes visualisées simultanément, laissant penser que le graphe est connexe. La représentation par période des données montre que les données ne sont pas forcément toutes reliées entre elles pour chaque tranche de temps, révélant ainsi la non connexité du graphe initial.

Par masquage des données du graphe global n'appartenant pas à la première tranche de temps, le graphe de la première période est obtenu. Dans la Figure 103, les classes situées dans la partie Nord-Est de la fenêtre sont des classes présentes principalement dans les premières périodes, mais pas dans les dernières tranches de temps. A l'inverse, les classes situées dans le cadran Nord-Ouest laissent à penser qu'il s'agit de classes émergentes, qui sont en pleine extension, puisqu'elles n'étaient pas aussi développées dans les premières périodes. Les classes des zones Sud-Est et Sud-Ouest correspondent aux périodes de transition où les classes vont évoluer que ce soit au niveau des liens inter classes, mais aussi au niveau interne (évolution de la première période à la dernière). Les classes situées vers le centre de la figure sont les classes persistantes, présentes pour toutes les périodes (Loubier et Dousset, 2008c).

Alors que leurs liens inter classes varient au cours des quatre périodes, leur évolution interne au sein de chacune varie faiblement, comparé aux classes situées en périphérie.

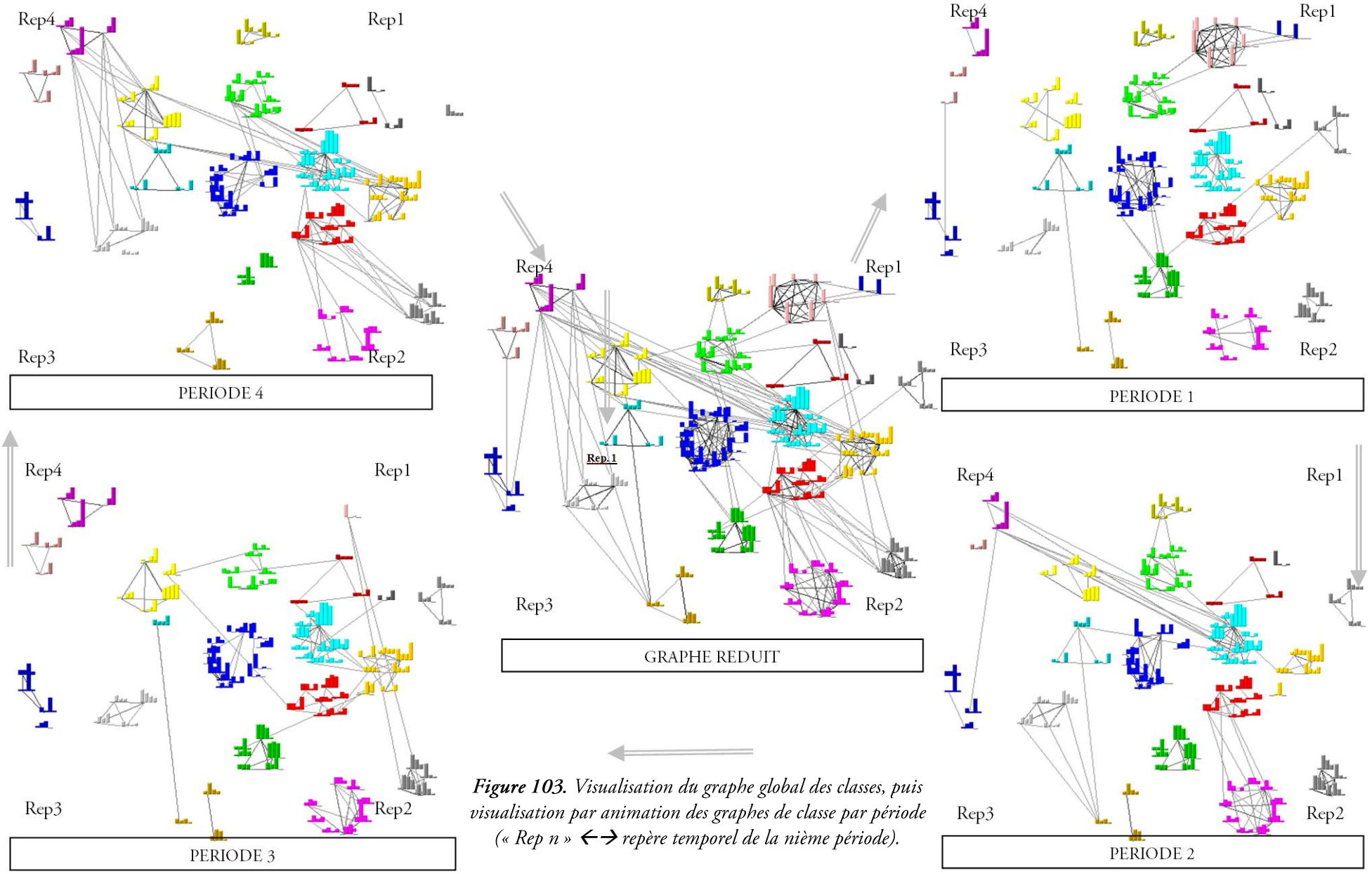


Figure 103. *Visualisation du graphe global des classes, puis visualisation par animation des graphes de classe par période (« Rep n » ←→ repère temporel de la nième période).*

4.11. Application de notre contribution à des analyses non temporelles

Notre contribution, axée sur le morphing de graphes temporels peut être adapté à des études ne portant pas forcément sur l'évolution des données. Le principe de visualisation globale, puis successive par transformation progressive de la représentation peut être appliqué à des notions autres que temporelles. Par exemple, le temps peut être remplacé par des zones géographiques, telles que des continents, ou encore par des multi-termes spécifiques à un domaine, ... Suivant le sujet étudié il est important que l'utilisateur effectue auparavant un ordonnancement des repères. Par exemple, si l'étude porte sur la répartition d'auteurs ayant publié sur un sujet précis et qu'on s'intéresse à l'origine géographique de ces individus, il est important de placer les pays, sous forme de repère, selon un ordre stratégique. Afin de minimiser l'effort cognitif de l'utilisateur, il est important que le positionnement des données et principalement des repères soit judicieux et stratégique. Dans l'exemple choisi, un calcul des fréquences d'apparition dans le corpus des données placées en repères peut être utile. Ainsi, la chronologie croissante utilisée pour le morphing de graphes temporels laisse place, dans un contexte géographique, à un ordonnancement croissant ou décroissant de la fréquence d'apparition des états.

L'exemple de la Figure 104 illustre graphiquement ce principe de morphing de graphes géographiques. Dans ce contexte, chaque matrice de croisement est spécifique à un lieu géographique spécifique, comme le montre la structure générique du Tableau 20.

$Repere_i$

	d_1	d_2	d_3	…	d_v
d_1	$<d_1,d_1>$	$<d_1,d_2>$	$<d_1,d_3>$	…	$<d_1,d_v>$
d_2	$<d_2,d_1>$	$<d_2,d_2>$	$<d_2,d_3>$	…	$<d_2,d_v>$
d_3	$<d_3,d_1>$	$<d_3,d_2>$	$<d_3,d_3>$	…	$<d_3,d_v>$
…	…	…	…	…	…
d_v	$<d_v,d_1>$	$<d_v,d_2>$	$<d_v,d_3>$	…	$<d_v,d_v>$

Tableau 20. *Matrice à la base du morphing de graphe.*

$i \in \{1,\ldots,r\}$ avec r le nombre de repères sur lesquels se base le morphing
$Repere_i \in \{dg_1, dg_2,\ldots,dg_n\}$ avec dg_i l'intitulé du repère i.
d_j représente l'entité croisée, pour $j \in \{1,\ldots,v\}$ avec v le nombre de données.
$<d_j,d_k>$ est la valeur de la cooccurrence, issue du croisement entre d_j et d_k.

Dans l'exemple illustré en Figure 103,

$i \in \{1,\ldots,8\}$
$Repere_i \in$ {USA, Canada, Allemagne, Japon, Suisse, France, Espagne, Italie}

La figure suivante est issue du croisement des compagnies pharmaceutiques, première dimension, avec elles mêmes, seconde dimension et leur pays d'appartenance, troisième dimension. Le morphing de graphe permet, dans ce cas là, de distinguer clairement les états dominant dans ce secteur pharmaceutique.

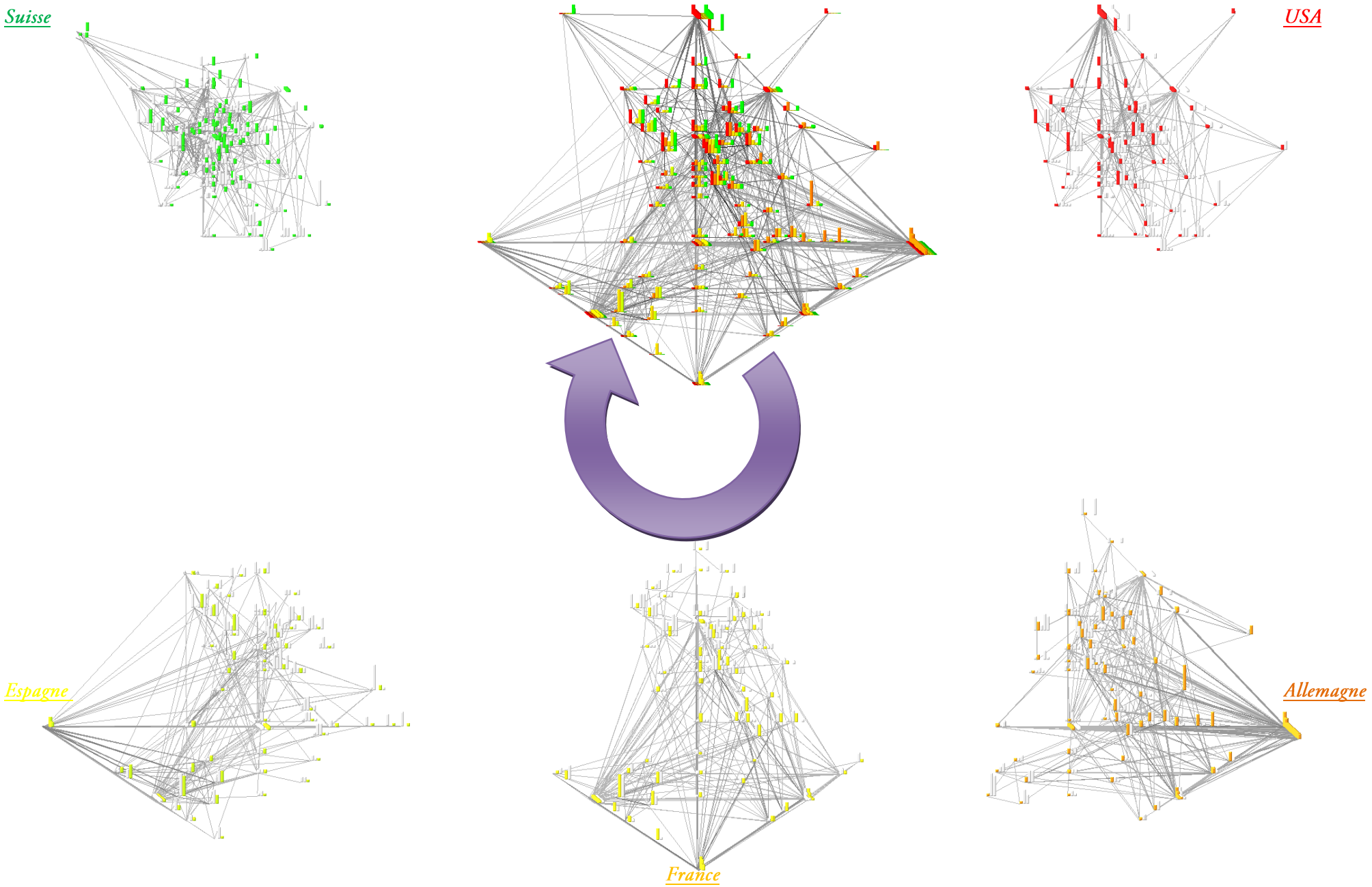


Figure 104. *Application du morphing de graphe dans un contexte géographique.*

4.12. Synthèse

Dans ce chapitre, nous avons présenté notre contribution pour les graphes dynamiques. La sémiologie utilisée repose sur le choix d'histogramme pour représenter l'évolution individuelle des données. Le recours à des couleurs spécifiques pour chacune des barres de l'histogramme, représentant les périodes considérées, permet de faciliter la comparaison des différentes instances pour l'utilisateur, qui reste au centre de nos préoccupations.

Afin d'organiser stratégiquement la représentation graphique de données temporelles, pour chaque instance considérée un repère est attribué, sous forme de sommet. Ces indicateurs sont situés par ordre chronologique sur les contours de la fenêtre de visualisation. Chaque donnée temporelle est ensuite placée à une distance de chaque repère, relative à son appartenance aux différentes périodes correspondantes.

Notre contribution repose en partie sur le morphing de graphe, reposant sur ce principe de positionnement temporel des sommets mais aussi sur une succession de représentations complémentaire :

- le graphe global, toutes périodes confondues
- les graphes de période dont la transition pour le passage d'une instance à la suivante s'effectue par transformation progressive, de la même façon que pour les morphings d'images.

Un nouvel algorithme de placement dirigé des sommets est proposé, paramétrable par l'utilisateur. Il favorise l'attirance des sommets vers les repères temporels. Paramétré avec une forte valeur d'attirance des repères, le graphe obtient une typologie qui met en avant les caractéristiques temporelles des données sous forme de regroupement de ces dernières. Ainsi, chaque portion de la fenêtre de visualisation est spécifique à des caractéristiques temporelles. Les données proches d'un repère sont caractérisées par l'appartenance à la période correspondante, uniquement. Les sommets situés à équidistance de plusieurs repères apparaissent aussi intensément dans chacune des périodes concordantes.

Les fonctionnalités proposées dans la première partie de notre contribution, dans le chapitre 3, sont adaptées au contexte temporel selon deux points de vue spécifiques :

- global, cas où les méthodes sont appliquées sur le graphe de toutes les périodes cumulées, et le morphing n'est appliqué que pour suivre l'évolution de la structure globale, dans un contexte d'instance individuelle.
- Evolutif, cas où les fonctionnalités sont appliquées indépendamment sur chaque graphe de période.

Chapitre 5. Expérimentations et évaluation de VisuGraph

> *En essayant continuellement, on finit par réussir. Donc plus ça rate, plus on a de chances que ça marche.*
>
> Shadok : devise numéro 1

5.1. Présentation des expérimentations 180
5.2. Architecture et conception du prototype 180
5.3. Evaluation de la représentation graphique 181
5.3.1. Evaluation sur la représentation des données temporelles 181
5.3.2. Evaluation sur l'espace des données 183
5.4. Expérimentation de VisuGraph 184
5.4.1. Méthodologie 184
5.4.2. Technique utilisée 184
5.4.3. Population étudiée 186
5.4.4. Accès à l'expérimentation 187
5.4.5. Résultats 188
✓ Durée d'expérimentation 188
✓ Niveau de difficulté par type de données 188
✓ Analyse des autres résultats 201
✓ Conséquences du sondage 202
5.5. Synthèse de l'expérimentation et de la validation de VisuGraph 202

5.1. Présentation des expérimentations

Ce chapitre présente la démarche de validation scientifique mise en place pour appliquer la contribution des travaux présentée dans les chapitres 3 et 4, centrés sur les tâches des utilisateurs pour la visualisation de données temporelles.

Précisons que les techniques de représentation proposées se basent sur l'extraction des connaissances présentées dans le chapitre 1, des processus de visualisation exposés dans le chapitre 2 et des tâches utilisateur, des fonctionnalités et des méthodes présentées dans les chapitres 3 et 4, synthétisés dans la Figure 105.

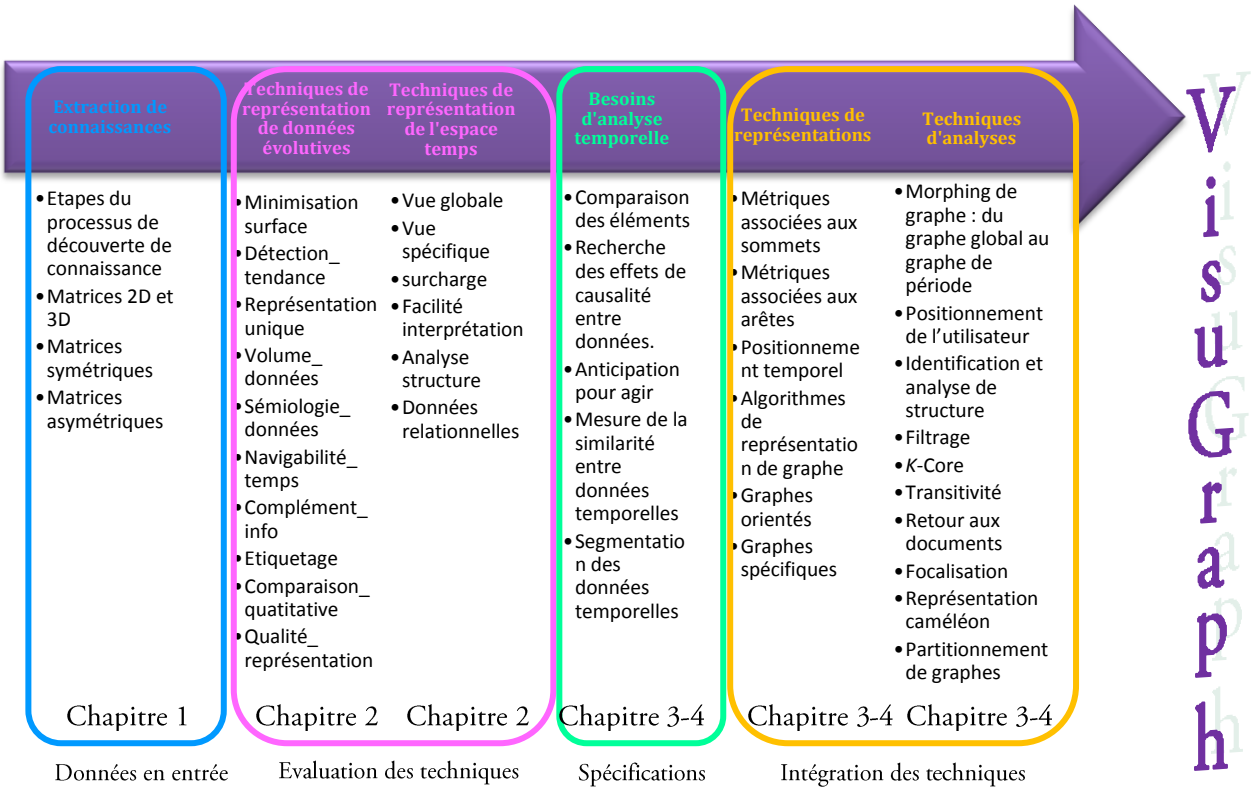


Figure 105. *Etapes du processus de conception et de réalisation de l'outil VisuGraph.*

Dans ce chapitre, nous présentons, en 5.2, l'architecture de l'outil VisuGraph, à savoir le processus de développement, sa décomposition en termes de modules et son codage.

Dans un second temps, en section 5.3, notre contribution est évaluée selon les critères retenus pour qualifier les outils de visualisation étudiés dans le chapitre 2.

Dans la section 5.4, nous présentons l'expérimentation effectuée pour valider la conception de l'outil, c'est-à-dire l'enquête réalisée auprès d'une population ciblée. Les résultats de ces évaluations nous ont servi dans un premier temps à améliorer l'utilisabilité des différentes techniques de visualisation, puis d'envisager de nouvelles perspectives pour nos travaux à venir.

Enfin, une synthèse des expérimentations et évaluations de VisuGraph est proposée.

5.2. Architecture et conception du prototype

Nous avons modélisé et conçu une architecture dans le but de faciliter le travail d'analyse de corpus, en particulier dans un contexte de veille stratégique. Les méthodes et approches présentées dans les chapitres 3 et 4 ont démontré la faisabilité technique de cette architecture, implantée dans l'outil VisuGraph.

Concevoir et réaliser un outil de visualisation de données temporelles ne se résume pas à élaborer une représentation de la donnée accompagnée d'une méthode de navigation.

Il faut, avant tout, comprendre et analyser les objectifs de la visualisation et quelles sont les méthodes de représentation et de navigation les plus adaptées.

Pour spécifier la conception de l'outil VisuGraph, nous avons minutieusement précisé les besoins de l'utilisateur, afin de concevoir une architecture y répondant au mieux. Le choix des méthodes de représentation et de navigation adaptées à la tâche de l'utilisateur mènent à la mise en place d'un scénario de visualisation.

Ainsi, cette analyse des besoins en termes de conception de représentation et de navigation doit être le résultat d'une collaboration entre utilisateur et concepteur. Pour ce faire, l'écoute de besoin d'experts de la veille stratégique, appliquée à des secteurs divers tels que les biotechnologies, le domaine pharmaceutique, aéronautique, commercial nous a permis de cibler la réalisation de VisuGraph.

Le développement de VisuGraph repose sur la programmation objet et plus particulièrement sur le langage de programmation Java (Eckel, 1998). La programmation orientée objet consiste à modéliser un ensemble d'éléments du monde réel. Un programme java est constitué de packages qui sont des ensembles de classes. Une classe définit la structure d'un objet, c'est-à-dire la déclaration de l'ensemble des entités qui le composent. Un objet est donc « issu » d'une classe. En réalité on dit qu'un objet est une instanciation d'une classe, caractérisé par plusieurs notions :

- Les *attributs*: Il s'agit des données caractérisant l'objet. Ce sont des variables stockant des informations d'état de l'objet ;
- les *méthodes* : Les méthodes d'un objet caractérisent son comportement, c'est-à-dire l'ensemble des actions (appelées opérations) que l'objet est à même de réaliser. Ces opérations permettent de faire réagir l'objet aux sollicitations extérieures (ou d'agir sur les autres objets). De plus, les opérations sont étroitement liées aux attributs, car leurs actions peuvent dépendre des valeurs des attributs, ou bien les modifier

Le développement de l'outil repose ainsi sur 14 classes java, permettant la création et la manipulation de chacun des composants de la visualisation, que ce soit au niveau de la fenêtre de représentation ou encore du graphe lui-même.

5.3. Evaluation de la représentation graphique

Dans cette section, nous reprenons les critères retenus dans le chapitre 2 qui nous ont permis d'évaluer les outils graphiques existant. Nous les utilisons pour évaluer VisuGraph et confirmer l'apport de nos contributions.

5.3.1. Evaluation sur la représentation des données temporelles

Avant d'apporter une appréciation sur la qualité de visualisation, il est important d'évaluer la représentation des données, dans un contexte temporel. Dans VisuGraph, il s'agit de porter une appréciation sur l'assimilation des entités temporelles à des histogrammes. Dans le Tableau 21, les critères auxquels répond VisuGraph sont signalés par un onglet.

Codage du critère	VisuGraph
Minimisation_surface	✓
Détection_tendance	✓
Représentation_unique	
Volume_données	
Sémiologie_donnée	✓
Navigabilité_temps	✓
Complément_info	✓
Etiquetage	✓
Comparaison_quantitative	✓
Qualité_représentation	✓

Tableau 21. *Evaluation de VisuGraph*

La *minimisation de la surface de représentation* est possible, selon la représentation des données choisies, que ce soit de forme de cercles, de barres, …, pour les cas non temporels, ou d'histogramme, pour les cas évolutifs. D'autre part, le recours au partitionnement de graphe, exposés dans les chapitres 3 et 4, permet de réduire la surface de représentation.

La *détection des tendances* s'effectue à travers le placement temporel des données, mais aussi à travers la comparaison des barres des histogrammes représentant les différentes valeurs de métriques par instance. Il est alors plus facile de caractériser les entités, à savoir si elles sont en émergence, en digression et/ou en stagnation. A travers l'étude de la structure de graphe, VisuGraph facilite la détection des acteurs importants du domaine étudié.

La *représentation des données* n'est pas unique, comme nous l'avons vu dans les chapitres 3 et 4. La possibilité d'avoir recours à d'autres formes pour la visualisation de l'entité permet, selon le volume de données à faciliter l'approche analytique.

Au niveau du *volume de données*, VisuGraph rencontre quelques faiblesses. Au niveau de la lisibilité, plus le nombre d'entités est important, plus le graphe est chargé, provoquant des chevauchements entre les libellés des sommets.

La *sémiologie des données* se traduit par le recours à des couleurs spécifiques, surtout dans le cas temporel. L'histogramme nous permet de connaître en seul regard la tendance de l'entité par comparaison des différentes barres et par attribution d'une couleur spécifique et significative à chacune.

La *navigabilité dans le temps* se traduit par le morphing, permettant de passer d'un graphe global à un graphe de période. Ainsi, la granularité du temps peut être réduite. Par exemple, le graphe global peut porter sur un cumul de dix années, détaillé en dix graphes de période. Cependant, il n'est pas possible de travailler sur plus de deux granularités à la fois. Par exemple, si le graphe global correspond à une période de dix années, divisées en dix représentations de période dont l'unité est l'année, il ne sera pas possible de changer la visualisation en mois ou en semaines sans réinitialiser les graphes sur lesquelles repose la représentation.

L'*accessibilité à l'information complète* s'effectue par la fonction de focalisation, par retour aux notices contenant l'entité sélectionnée ou par réduction du graphe selon le(s) sommet(s) choisi(s).

L'*étiquetage clair et complet des données* est obtenu par affichage du libellé court ou complet du nom du sommet, ou encore par visualisation de la fenêtre pop-up indiquant le libellé du sommet, sa valeur de métrique et la nature de ses relations avec ses voisins au sein de la structure de graphe.

La *comparaison des données temporelles* s'effectue en étudiant les barres de chaque histogramme, qualifiant ainsi l'importance de l'entité.

La *qualité de la représentation d'un point de vue général* est évaluée vis-à-vis des structures de graphes obtenues. La disposition des sommets au sein de VisuGraph permet la mise en avant des regroupements d'entités, révélant le rôle de chacun au sein de la structure, qu'elle soit temporelle ou non.

Alors que ces critères caractérisent la représentation des données à titre individuel ou comparatif, il est important d'analyser l'apport des visualisations proposées dans l'outil VisuGraph de manière globale, à travers l'évaluation de l'espace des données.

5.3.2. Evaluation sur l'espace des données

Les données sont étudiées dans un contexte de structure et non pas sur des entités individuelles. L'évaluation porte sur l'outil complet.

Rappelons les critères retenus, dans le chapitre 2:

- *Une vue globale,* c'est-à-dire une perception efficace de l'ensemble des périodes étudiées. La première représentation proposée dans VisuGraph, dans un contexte évolutif, permet d'obtenir une première idée des structures formées par les données au cours du temps et des caractéristiques des entités. Cette vue globale constitue la carte mentale de l'utilisateur, sur laquelle repose toute étude temporelle.
- *Une vue spécifique*, permettant d'analyser unitairement chacune des données et de la quantifier à tout moment, via les graphes de période, mais aussi grâce à l'extraction de graphes réduits, notamment lors du partitionnement de graphe, où il est possible d'étudier une classe dans une autre fenêtre de visualisation ;
- *Une surcharge du graphe,* le rendant difficile à analyser. Le partitionnement de graphe, permet de simplifier la visualisation, par affichage d'un représentant unique après classification des entités relationnelles. Cette solution permet d'éviter de surcharger le graphe et de le rendre moins lisible. Cependant, si ce partitionnement n'est pas effectué, la surcharge du graphe croit de façon relative au volume de données. Plus le nombre d'entités, mais aussi de liens, représentés sont importants, moins les structures sont lisibles.
- *Une facilité d'interprétation du graphe.* Le recours à des méthodes mathématiques spécifiques permet d'explorer les structures de graphes plus facilement. L'étude du voisinage d'un sommet ou encore de l'évolution d'un groupement d'entités est aisément réalisable.
- *Une comparaison des changements de structures au cours du temps.* Le fait de conserver une carte mentale comme repère, c'est-à-dire le graphe global, permet d'obtenir un environnement structural spécifique. Le morphing de graphe permet d'extraire de façon progressive la composition de la structure étudiée à tout instant t. On peut alors facilement détecter les nouveaux membres, ceux disparaissant et ceux persistants. Des erreurs d'interprétation peuvent ainsi être évitées. Admettons que sur un graphe global de cinq ans, un individu est relié à 20 personnes. Une première appréciation serait de qualifier cette personne comme « noyau actif, permanant d'une équipe». Or, la décomposition en année peut révéler que la première année, cet individu a collaboré avec 15 personnes, la deuxième avec 5 autres, et avec aucune le reste du temps. Cet individu n'est donc pas permanent, et son activité est importante que pour une année, ce qui contredit notre hypothèse de base. Alors que sur le graphe global cet individu semblait important, seules les deux premières années peuvent le qualifier ainsi.
- Une application permettant la *comparaison claire de « données relationnelles »*. Dans la section précédente, nous avons vu que les données pouvaient être comparées via les barres d'histogramme. Ici, nous nous intéressons davantage à l'association des entités, c'est à dire à la nature de leurs liens. Le recours à des artifices visuels tels que le changement d'intensité du lien en fonction de la valeur de sa métrique ou encore l'attirance des sommets les plus fortement liés, via les forces FDP, permet de caractériser des ensembles d'entités relationnelles.

5.4. Expérimentation de VisuGraph

5.4.1. Méthodologie

Afin d'expérimenter VisuGraph, un questionnaire a été réalisé afin de tester cet outil dans sa globalité. Les objectifs de cette expérimentation sont:

- Cibler la population, à savoir si l'outil est destiné à un public plus ou moins restreint ;
- tester toutes les fonctionnalités proposées, dans tous les cas possibles, c'est dire sur tous les types de matrices de croisement étudiées jusqu'ici ;
- recenser les lacunes de l'outil ;
- être à l'écoute des futurs utilisateurs, quelque soient leur origine scientifique, afin de prendre en compte leurs besoins ;
- valider notre contribution et particulièrement l'apport de nos travaux pour la veille stratégique ;
- prévoir des perspectives d'évolution et cibler davantage nos orientations de recherche à venir.

5.4.2. Technique utilisée

Pour cette expérimentation, un sondage est réalisé. Il s'agit d'une enquête ponctuelle qui consiste à construire un échantillon à partir d'une population ciblée. Les personnes faisant partie de l'échantillon sont interrogées à l'aide d'un questionnaire et les réponses obtenues sont ensuite extrapolées à la population de base.

La méthodologie élaborée consiste dans un premier temps à sélectionner des données de qualité, à savoir des matrices symétriques et asymétriques, 2D et 3D bien construites, pour que l'expérimentation repose entièrement sur la visualisation et non pas sur l'extraction et la mise en forme de connaissances.

Nous avons effectué une analyse précise de chacune des matrices et un questionnaire a été formulé permettant de tester si la population interrogée trouvait les mêmes résultats que nous. Cette enquête se décompose sur plusieurs axes. Comme le montre la Figure 106, chaque expérimentation est chronométrée afin d'évaluer la difficulté d'utilisation, selon le profil utilisateur. Un « testeur » doit remplir une fiche de renseignement dans laquelle il indique son niveau d'enseignement supérieur, son secteur d'activité, son niveau de connaissance dans le domaine de la visualisation.

L'outil VisuGraph étant destiné à un public assez ciblé, qui a un minimum d'acquis dans le domaine de la fouille de données, en particulier dans la représentation graphique de données.

Dans un second temps, les données sont chargées. Les critères d'évaluation sont la facilité de lancement de l'outil et son temps d'exécution, c'est-à-dire la le temps de réponse de l'outil avant d'obtenir la première visualisation.

Les différentes fonctionnalités sont testées et l'utilisateur doit évaluer le niveau de difficulté pour les appliquer, en cochant pour chacune d'entre elles la case correspondante, à savoir :

- très facile
- facile
- sans difficulté
- difficile
- très difficile

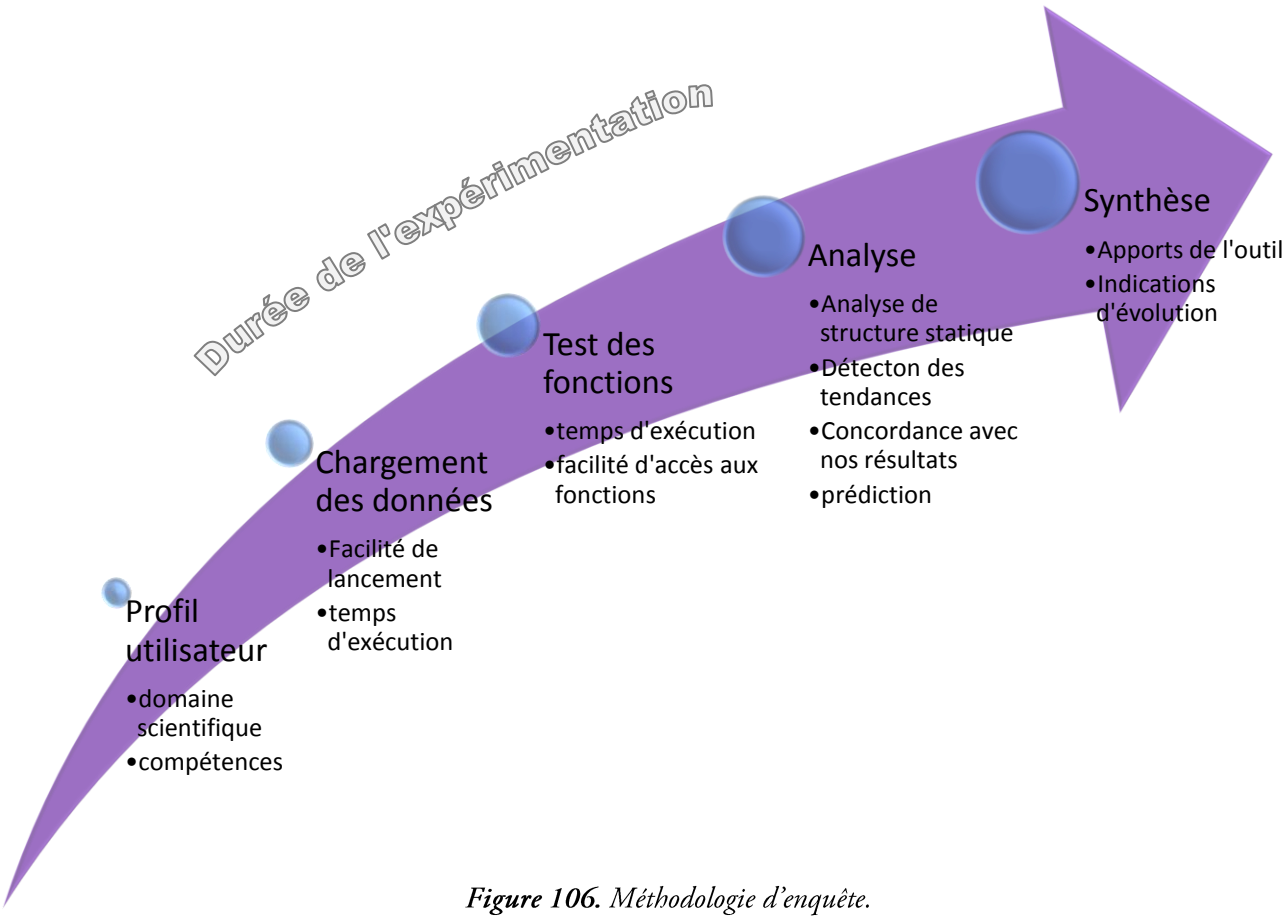


Figure 106. *Méthodologie d'enquête.*

Les questions posées mènent la personne interrogée à effectuer une analyse, portant dans un premier temps sur l'analyse de structures statiques puis sur l'étude évolutive d'entités, à savoir la prédiction et la détection des tendances. Ces résultats doivent être comparés aux nôtres afin de s'assurer de la bonne utilisation de l'outil et de l'absence d'ambiguïté dans l'interprétation des graphes.

Enfin, l'utilisateur interrogé évalue librement notre contribution en qualifiant l'intérêt de VisuGraph, s'il souhaiterait l'utiliser à l'avenir, ses points forts, ses points faibles et les recommandations pour faire évoluer cet outil.

Bases du questionnaire

Après avoir élaboré la méthodologie, en listant les axes sur lesquels porte l'enquête et avoir définit pour chaque thème quelles sont les informations primaires recherchées, les questions sont rédigées selon le type d'information recherchée.

L'enquête est réalisée en suivant les règles suivantes :

- choisir des mots adaptés à la population interrogée.
- Utiliser un langage spécifique à une population ayant des bases scientifiques.
- Ne pas favoriser un type de réponse.
- Les questions sont indépendantes et le formulaire est divisé en plusieurs parties autonomes.
- Limiter l'appel à la mémoire.
- Ordonner les questions par thèmes. Pour chaque partie du questionnaire, les questions sont ordonnées selon un ordre chronologique d'utilisation de l'outil. Dans un premier temps la représentation des

données, puis les questions portent sur l'amélioration du dessin de graphes et enfin sur l'utilisation des méthodes.

- Essayer le questionnaire auprès d'un échantillon réel et réduit. Trois personnes, expertes du domaine ont testé dans un premier temps le questionnaire afin de pouvoir apprécier le niveau de difficulté de l'enquête, le temps d'application, la cohérence des points étudiés.
- S'assurer que toutes les questions sont intelligibles.
- Corriger le questionnaire et refaire le test si nécessaire.

Plusieurs corrections ont été effectuées suite aux essais du questionnaire de l'échantillon réduit, d'un point de vue de l'adéquation entre les questions et les résultats graphiques, mais aussi au niveau de la précision de certaines formulations.

Le questionnaire est visible en annexe ou sur le site « http://www.irit.fr/~Eloise.Loubier ».

Les caractéristiques de l'enquête sont indiquées dans le

Tableau 22.

axes étudiés / Types de questions et réponses	Profil utilisateur	Questions générales d'utilisation	Matrice symétrique 2D	Matrice asymétrique 2D	Matrice symétrique 3D	Matrice asymétrique 3D	Synthèse	Total
Réponse à choix	4	9	11	11	11	9	10	60
Réponse libre	2	0	6	5	10	8	1	31
Nombre questions	6	9	11	11	14	9	11	60

Tableau 22. *Nombre de questions et réponses possibles.*

Certaine questions ont des réponses « à choix », c'est-à-dire que plusieurs types de réponses sont données et la personne interrogée doit sélectionner celle qui correspond le plus à son opinion.

Pour les 9 questions générales d'utilisation, les 11 portant sur matrices symétriques et 11 sur les asymétriques, 2D et 11 et 9 sur les 3D, les réponses à choix sont composées de 5 items, allant de « très facile » à « très difficile ». 51 questions ont des réponses de ce type. Nous considérons un second type de question, nécessitant une réponse rédigée par l'utilisateur, que nous nommons « réponse libre ».

5.4.3. Population étudiée

On considère une population bien déterminée et une variable formalisant l'information qui nous intéresse appelée variable d'intérêt, définie sur chaque individu de cette population. L'individu est l'unité de base à laquelle on s'intéresse et la population est l'ensemble des individus.

La taille de la population est fixe et connue, elle est notée *N*. on prend pour acquis qu'il existe pour chaque individu de cette population, une information d'une nature quelconque permettant de la repérer précisément et sans aucune ambiguïté.

Les critères de qualité des sondages sont relativement clairs sur le plan théorique. La base de sondage, c'est-à-dire la liste composant l'échantillon, doit idéalement comprendre tous les membres de la population à représenter.

La choix de la population interrogée consiste à fixer des quotas basés sur les orientations scientifiques. La sélection des personnes interrogées repose sur les individus qui répondent aux critères permettant de remplir les quotas fixés. Ainsi, nous fixons plusieurs catégories de personnes interrogées selon leur domaine de compétence et de niveau supérieur ou égal à bac +2.

Elles constituent ainsi des Sous-ensemble d'une population donnée, considéré comme un véritable modèle réduit de la population étudiée.

- Les informaticiens;
- Les Statisticiens et informaticiens décisionnels;
- Les individus de l'Intelligence Economique ;
- Les autres, de profils scientifiques diverses telles que la sociologie, l'aéronautique, la médecine,…

Le

Tableau **23** indique, selon les disciplines, le nombre de personnes interrogées, la proportion ayant des connaissances dans le domaine de la visualisation, celles ayant déjà manipulé Tétralogie.

disciplines	Nombre De personnes interrogées	% Notions de visualisation sur le total par catégorie	% Notions de visualisation sur le total global	% ayant manipulé Tétralogie sur le total par catégorie	% ayant manipulé Tétralogie sur le total global
Informaticiens	25	60	23	20	7
Statisticiens	25	80	30	100	38
IE	6	0	0	100	9
Autre	10	20	3	0	0
Total	66		56		54

Tableau 23. *Caractéristiques de la population étudiée par catégorie.*

Le

Tableau **23** indique que 56% des personnes interrogées possèdent des connaissances dans le domaine de la visualisation.

5.4.4. Accès à l'expérimentation

Le module VisuGraph, intégré à Tétralogie, a été installé sur le serveur nommé « Tétralogie », accessible à distance à partir d'un terminal X ou à partir de PC ou encore d'une station UNIX.

Le formulaire est accessible sur la page « http ://www.irit.fr/~Eloise.Loubier », les données sont sur le serveur Tétralogie, dans le répertoire « donnees/Enquete ».

La connexion à tétralogie s'effectue par la commande

```
ssh -X tetralogie.irit.fr -l [nom_utilisateur].
```

`[nom_utilisateur]`est le login utilisé pour la connexion. Un mot de passe est nécessaire pour accéder à la plateforme, ainsi protégée.

Nous avons demandé aux personnes interrogées de répondre au questionnaire sur le document téléchargé sur le site et de nous le retourner à l'adresse loubier@irit.fr.

5.4.5. Résultats

✓ ***Durée d'expérimentation***

Le premier point étudié est la durée nécessaire à chaque profil de personnes interrogées pour effectuer l'étude, afin de pouvoir situer le niveau de difficulté, selon l'utilisateur. Les résultats sont restitués dans le Tableau **24**, dont la dernière colonne correspond aux résultats des trois personnes, expertes du domaine, qui ont testé dans un premier temps notre sondage afin de l'améliorer et qui servent de comparatif. Il est normal que ces trois personnes aient mis nettement moins de temps que les autres sondés puisqu'il s'agit de personnes expertes du domaine.

Les personnes travaillant dans l'intelligence économique ont tous mis deux heures, puisqu'elles ont effectué l'expérimentation en même temps, sur un créneau de 2h. Les résultats étant proches, pour cet aspect temporel, la difficulté d'utilisation reste raisonnable, par comparaison entre le temps d'application pour des personnes habituées à manipuler l'outil et celles qui ne le sont pas.

Discipline / Durées	Informaticiens	Statisticiens	Intelligence Economique	Autres	Echantillon réduit
Durée minimale	2h00	2h00	2h00	1h00	0h50
Durée maximale	2h30	2h30	2h00	2h45	1h00
Moyenne	2h05	2h01	2h00	2h15	0h55
Ecart-type	0,11	0,06	0	0,63	5

Tableau 24. *Résultats pour la durée de l'expérimentation.*

✓ ***Niveau de difficulté par type de données***

Afin de restituer les résultats de notre enquête, le morphing de graphe est utilisé, dans un contexte non temporel. Les repères utilisés correspondent au niveau de difficulté des tâches, à savoir :

- très facile
- facile
- moyen, sans difficulté
- difficile
- très difficile

L'association des 51 questions aux différentes réponses est visualisée par le biais d'un lien. Ces questions sont classées en cinq groupes dans les Tableau 25, Tableau 26, Tableau 27, Tableau 28 et Tableau 29.

Dans un second temps, les questions sont croisées avec la discipline personnes interrogées, à savoir:

- Les informaticiens;
- Les Statisticiens;
- Les individus de l'Intelligence Economique ;
- Les autres, c'est-à-dire les individus n'appartenant pas aux catégories précédemment citées.

Ainsi, sur un même graphe, il est plus facile de distinguer les réponses selon la discipline.

Les critères retenus, toutes catégories confondues, sont les suivants :

Questions	Codage
Combien de sommets composent ce graphe ?	nb_sommets
Affichez les noms des nœuds.	nom_noeud
Affichez le poids des arêtes	poids_lien
Agrandissez la taille de la police de caractère.	taille_police
Eclaircissez la police de caractère.	couleur_police
Changez la couleur du fond d'écran.	couleur_fond
« Masquez » les nœuds non liés.	masquage
Déplacez un nœud dans la fenêtre de représentation.	deplacement_noeud
Citez deux auteurs importants de ce graphe.	auteurs_importants

Tableau 25. *Questions d'ordre général.*

Questions	Codage
représentation graphique de la matrice *symetrique*	Vue_mat
« Filtrez » le graphe en ne conservant que les liens les plus forts.	filtrage
Quels sont les auteurs ayant le plus collaboré ?	forte_collaboration
Revenir au mode sans filtrage (filtrage nul).	retour_filtrage
K-core : indiquez combien d'équipes se distinguent par leur taille imposante.	kcore
Quel est le seuil du KCore atteint ?	seul_kcore
Via le menu, retrouvez l'auteur « Lopez m».	retour_auteur
Quelle valeur est associée à cet auteur ?	metrique_auteur
Combien de voisins directs a-t-il (passer par la transitivité) ?	nb_voisins
Retrouvez ses collaborations par transitivité.	transitivité
Effectuez un clustering.	MCL
Combien de classes trouvez-vous ?	nb_Classes_MCL
Effectuez une représentation circulaire des données.	representation_circulaire

Tableau 26. *Questions sur les représentations de matrices symétriques.*

Questions	Codage
Lancez la représentation graphique de la matrice *AC-AC-DP* du répertoire *Enquete*.	matrice_sym_temp
Combien de périodes sont représentées dans ce graphe ?	nb_periodes
De quelle couleur sont représentés Les sommets appartenant à la nième période.	couleur_perioden
Citez un sommet spécifique de la nième période	sommet_periode1
Quelle spécificité ont les sommets situés au centre ?	specificite_centre
Quelle spécificité ont les sommets situés en haut à gauche de l'écran ?	sommets_gauche
Quels sont les auteurs émergents ?	sommets_emergents
Affichez uniquement les auteurs présents durant la nième période.	auteur_specifique_perioden
Trouvez un auteur de faible valeur mais présent durant toutes les périodes.	auteir_present_faible
Appliquez les forces paramétrées d'attraction:	FDP
semi-paramétrées	FDP_SP
paramétrées	FDP parametree
Obtenez un graphe révélant les différentes équipes.	structure

Tableau 27. *Questions sur les représentations de matrices symétriques temporelles.*

Questions	Codage
Lancez la représentation graphique de la matrice *CO-PA* du répertoire *Enquete*	matrice_asym
De quelle couleur sont représentés les pays ?	couleur_pays
Combien de pays sont affichées ? Les citer	nb_pays
De quelle couleur sont représentées les compagnies?	couleur_compagnie
affichez les libellés longs des nœuds.	libelles_longs
Citez une compagnie importante.	compagnie_importante
Citez une compagnie moins importante.	cie_pas_importante
Représentation des données sous forme de nuances.	nuances
Retrouvez le sommet représentant la France.	retour_sommet

Tableau 28. *Questions sur les représentations de matrices asymétriques.*

Questions	Codage
Lancez la représentation graphique de la matrice *AL-PA-DP* du répertoire *Enquete*	mat_asym_temp
De quelle couleur sont représentés les pays ?	couleur_pays
Appliquer la force paramètre attraction. Combien de pays sont affichées ? Les citer.	FDP_morphing
Quel est le pays le plus présent sur l'ensemble de toutes les périodes ?	sommet_majeur
Appliquer le morphing. Quels sont les caractéristiques des auteurs émergents ?	effet_morphing
Quels sont les pays émergents ? (citer les principaux ?)	pays_emergents
Que peut-on dire du domaine étudié sur l'ensemble des périodes .	etude_domaine

Tableau 29. *Questions sur les représentations de matrices asymétriques temporelles.*

La décomposition matricielle des réponses s'effectue par critère d'évaluation. Chaque matrice correspond à une appréciation spécifique de difficulté et chacune des valeurs des tableaux recense le nombre de réponses pour un type de difficulté, de la catégorie en colonne, pour le critère en ligne.

Critères	Informaticiens	statisticiens	IE	Autres
Vue_mat	19	19	3	5
nb_sommets	22	22	4	7
nom_noeud	12	14	1	2
poids_lien	13	14	1	2
taille_police	22	22	3	5
couleur_police	19	22	6	10
couleur_fond	19	20	4	7
…	…	…	…	…
etude_domaine	18	21	2	3

Tableau 30. *Matrices des réponses « Très Facile », croisant les questions et les catégories d'individus.*

Critères	Informaticiens	statisticiens	IE	Autres
Vue_mat	13	16	3	5
nb_sommets	0	10	2	3
nom_noeud	14	19	3	5
poids_lien	19	19	2	3
taille_police	10	13	3	5
couleur_police	13	13	0	0
couleur_fond	15	17	2	3
…	…	…	…	…
etude_domaine	0	6	1	2

Tableau 31. *Matrice des réponses « Facile », croisant les questions et les catégories d'individus.*

Critères	Informaticiens	statisticiens	IE	Autres
Vue_mat	2	0	0	0
nb_sommets	1	3	0	0
nom_noeud	0	2	2	3
poids_lien	1	2	3	5
taille_police	0	0	0	0
couleur_police	0	0	0	0
couleur_fond	1	0	0	0
…	…	…	…	…
etude_domaine	0	0	1	2

Tableau 32. *Matrice des réponses « Sans difficulté », croisant les questions et les catégories d'individus.*

Critères	Informaticiens	statisticiens	IE	Autres
Vue_mat	1	0	0	0
nb_sommets	1	0	0	0
nom_noeud	1	0	0	0
poids_lien	1	0	0	0
taille_police	1	0	0	0
couleur_police	1	0	0	0
couleur_fond	0	0	0	0
…	…	…	…	…
etude_domaine	0	0	0	0

Tableau 33. *Matrice des réponses « Difficile », croisant les questions et les catégories d'individus.*

Critères	Informaticiens	statisticiens	IE	Autres
Vue_mat	0	0	0	0
nb_sommets	0	0	0	0
nom_noeud	0	0	0	0
poids_lien	0	0	0	0
taille_police	0	0	0	0
couleur_police	0	0	0	0
couleur_fond	0	0	0	0
…	…	…	…	…
etude_domaine	0	0	0	0

Tableau 34. *Matrice des réponses « Très Difficile », croisant les questions et les catégories d'individus.*

Graphiquement, les repères d'appréciations sont placés du plus favorable au plus difficile, suivant une logique similaire à celle d'un contexte temporel. En effet, dans le cas de données évolutives, les repères sont placés chronologiquement et la dernière période mène à la détection des éléments émergents qui doivent être étudiés précisément puisqu'il s'agit des entités majeures du futur proche. Dans la Figure 107, toutes les catégories de personnes interrogées sont représentées, pour tout type de réponse données, de « Très Facile » à « Très Difficile ». On distingue clairement que la majorité des manipulations ont été effectuées facilement pour l'ensemble des individus ayant participé à l'enquête.

Dans le cadre de ce sondage, les manipulations sont classées selon l'intérêt que nous leur portons. En effet, les questions dont la réponse a été « très facile » à trouver nous intéresse le moins. Il en est progressivement de même pour les réponses « faciles » ou « sans difficultés » puisque cela signifie que les utilisateurs sont satisfaits des résultats et que nous devons nous concentrer sur d'autres points plus sensibles, spécifiques aux graphes des réponses « difficile » et « très difficile ». Ces derniers révèlent les points à améliorer sur l'outil, afin de pouvoir dans un avenir proche effectuer une nouvelle enquête et vérifier l'amélioration de ces fonctionnalités. Afin de détailler les résultats du dépouillement du sondage, un morphing est appliqué sur l'ensemble des catégories des personnes interrogées en Figure 108.

Puis le morphing de graphe est appliqué pour chaque groupe afin de pouvoir les comparer mais aussi pour étudier la relation entre le domaine des personnes interrogées et la difficulté de manipulation de certaines fonctionnalités.

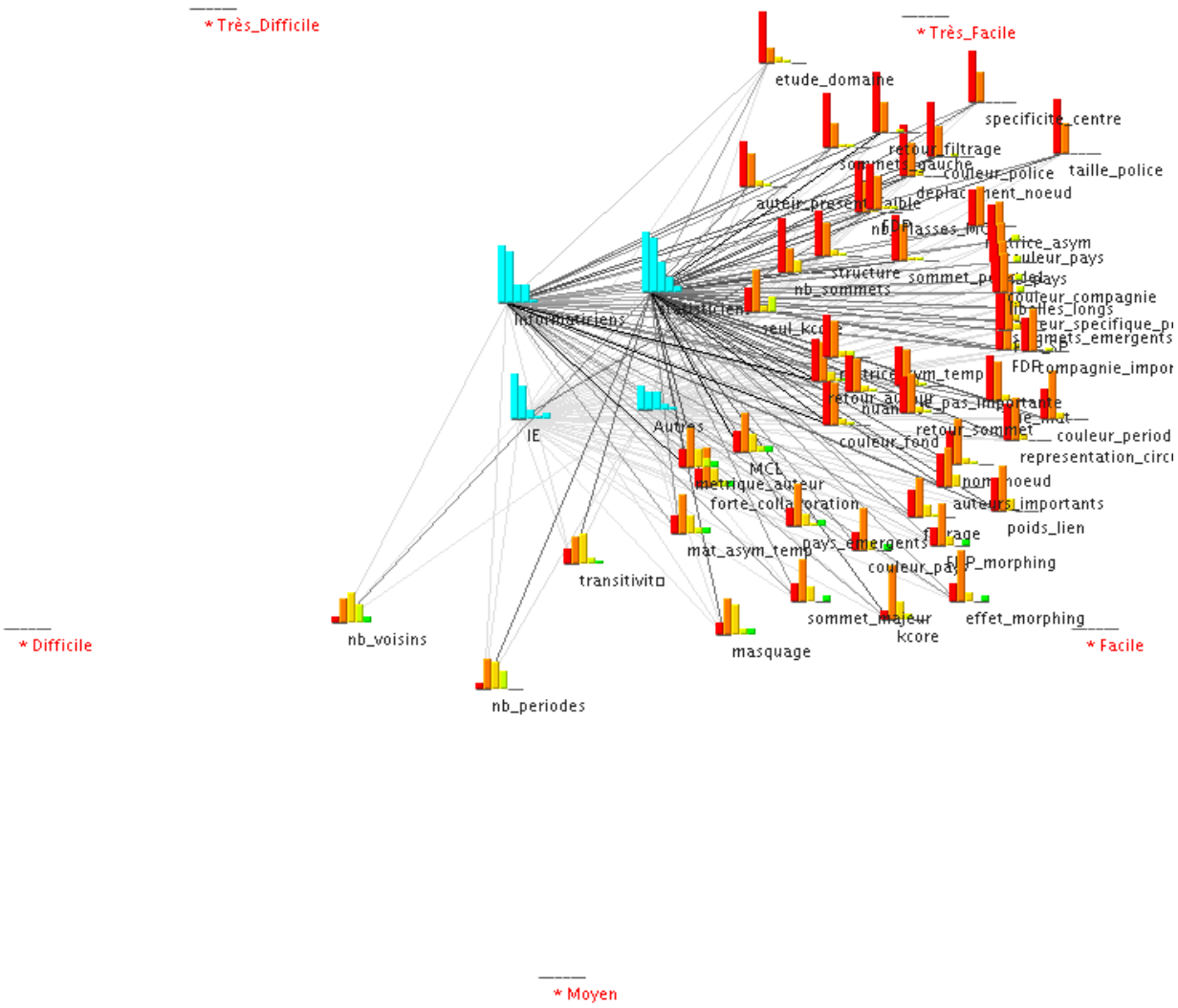


Figure 107. *Graphe globale de toutes les réponses données de notre sondage.*

Ainsi, en Figure 109, le morphing de graphes est appliqué aux réponses données par les informaticiens, la Figure 110, aux statisticiens, la Figure 111 aux individus travaillant dans le domaine de l'Intelligence Economique et enfin, la Figure 112, pour toutes les autres personnes. La présence d'un lien entre une catégorie de personnes interrogées et un critère, pour un graphe d'instance, indique que une ou plusieurs personnes ont répondu que la difficulté pour répondre à cette question est de niveau de l'instance.

Le morphing de graphe de chacune des catégories révèle la similarité des réponses de tous les groupes. En effet, tous les groupes trouvent très facile la manipulation de l'outil sur un nombre important de points. Le morphing de graphe sur toutes les catégories permet ainsi de détecter la tendance générale de la population interrogée, c'est-à-dire un niveau de difficulté de l'utilisation de l'outil assez faible. De ce point de vue, la validation de l'outil est positive puisque les résultats des utilisateurs d'origines diverses concordent.

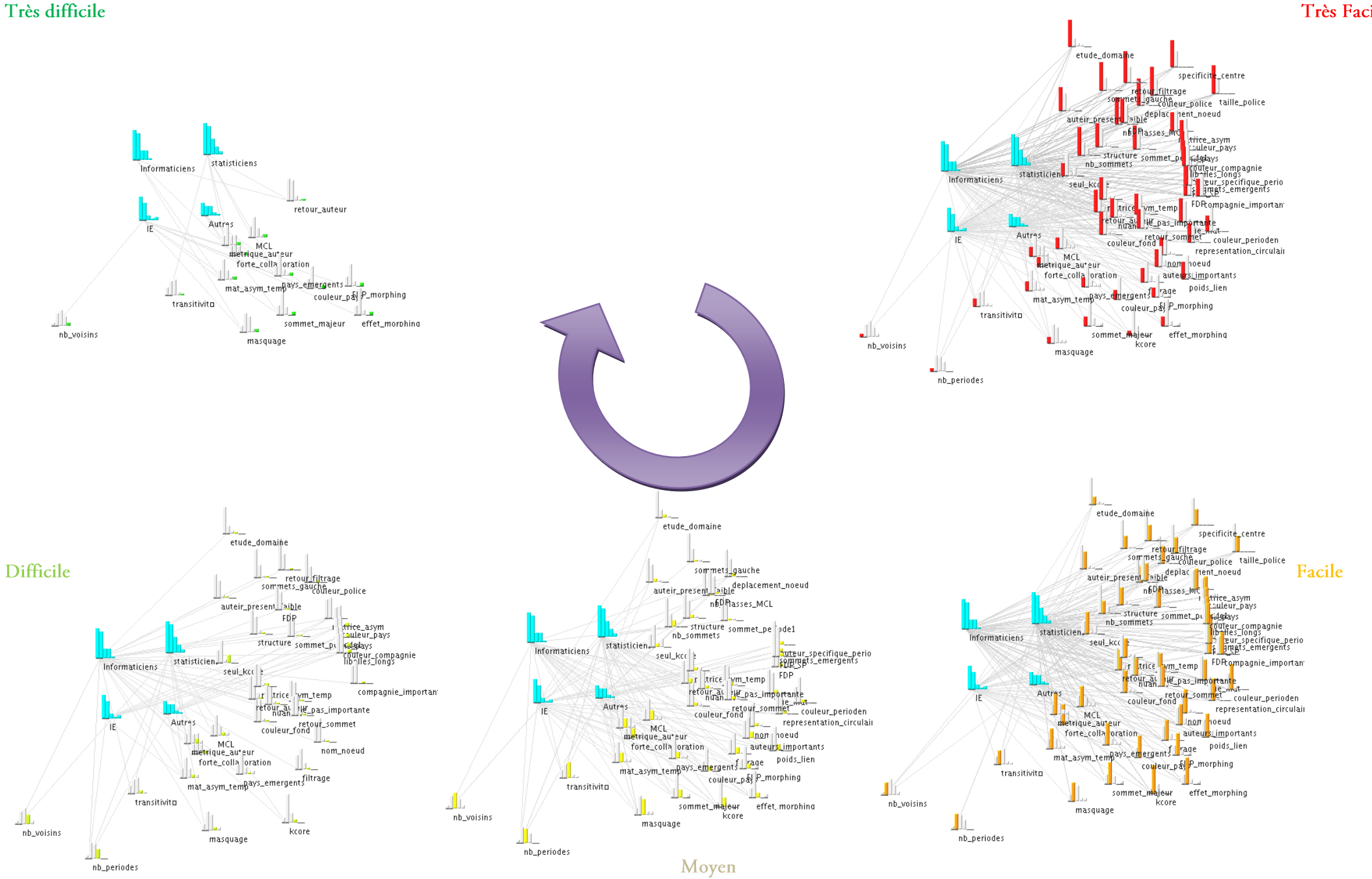


Figure 108. *Morphing de toutes les réponses.*

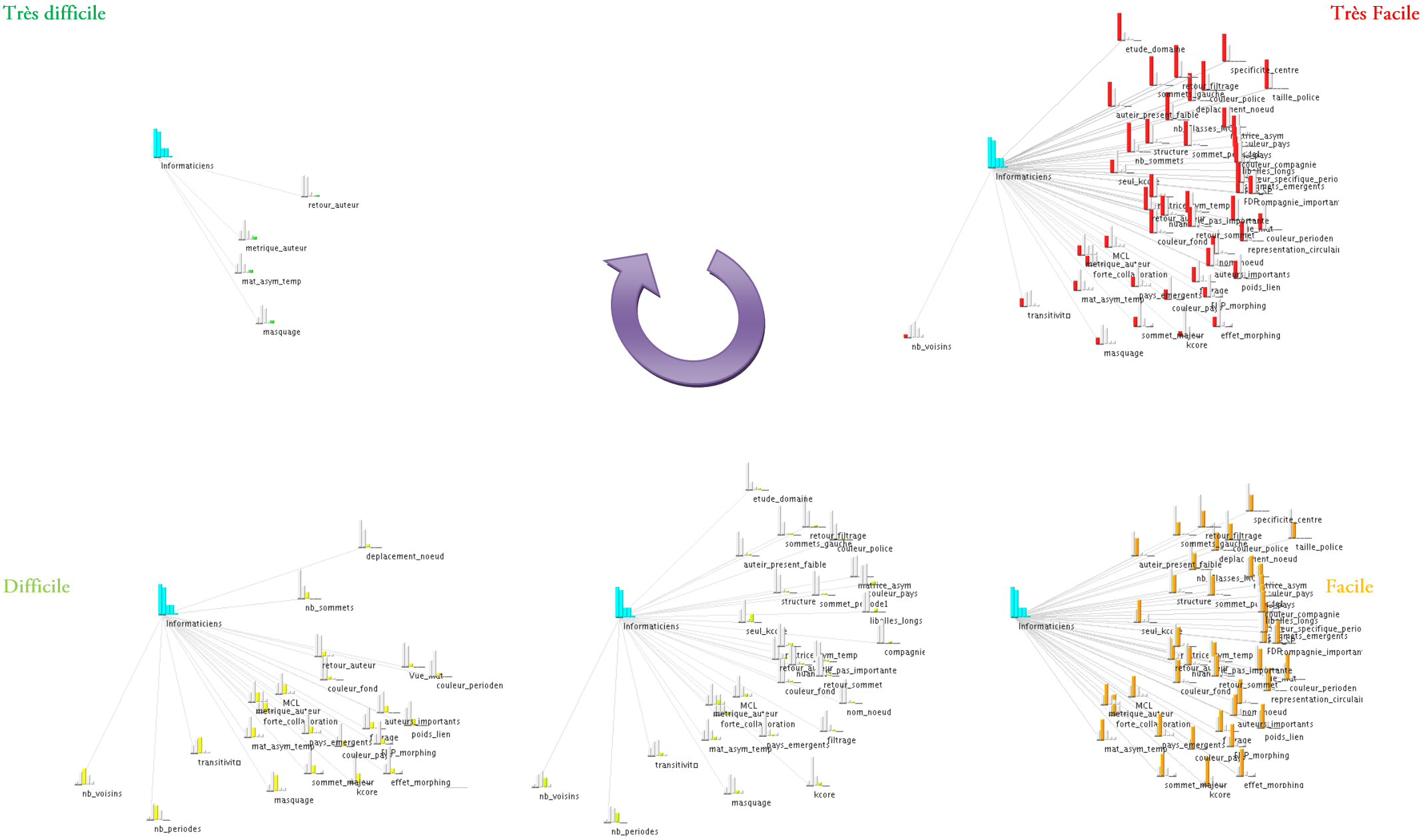


Figure 109. *Morphing des réponses des informaticiens.*

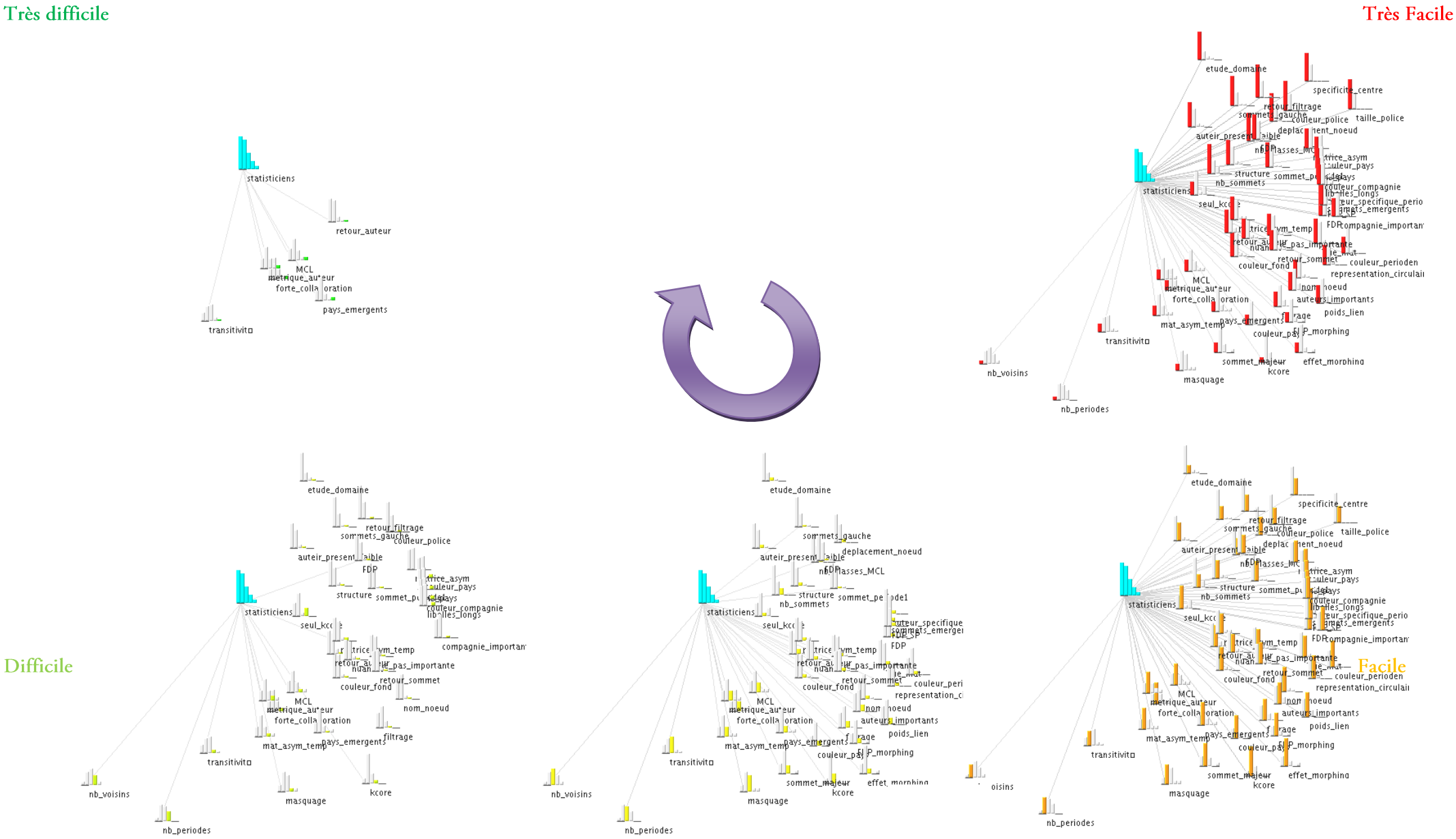


Figure 110. *Morphing des réponses des statisticiens.*

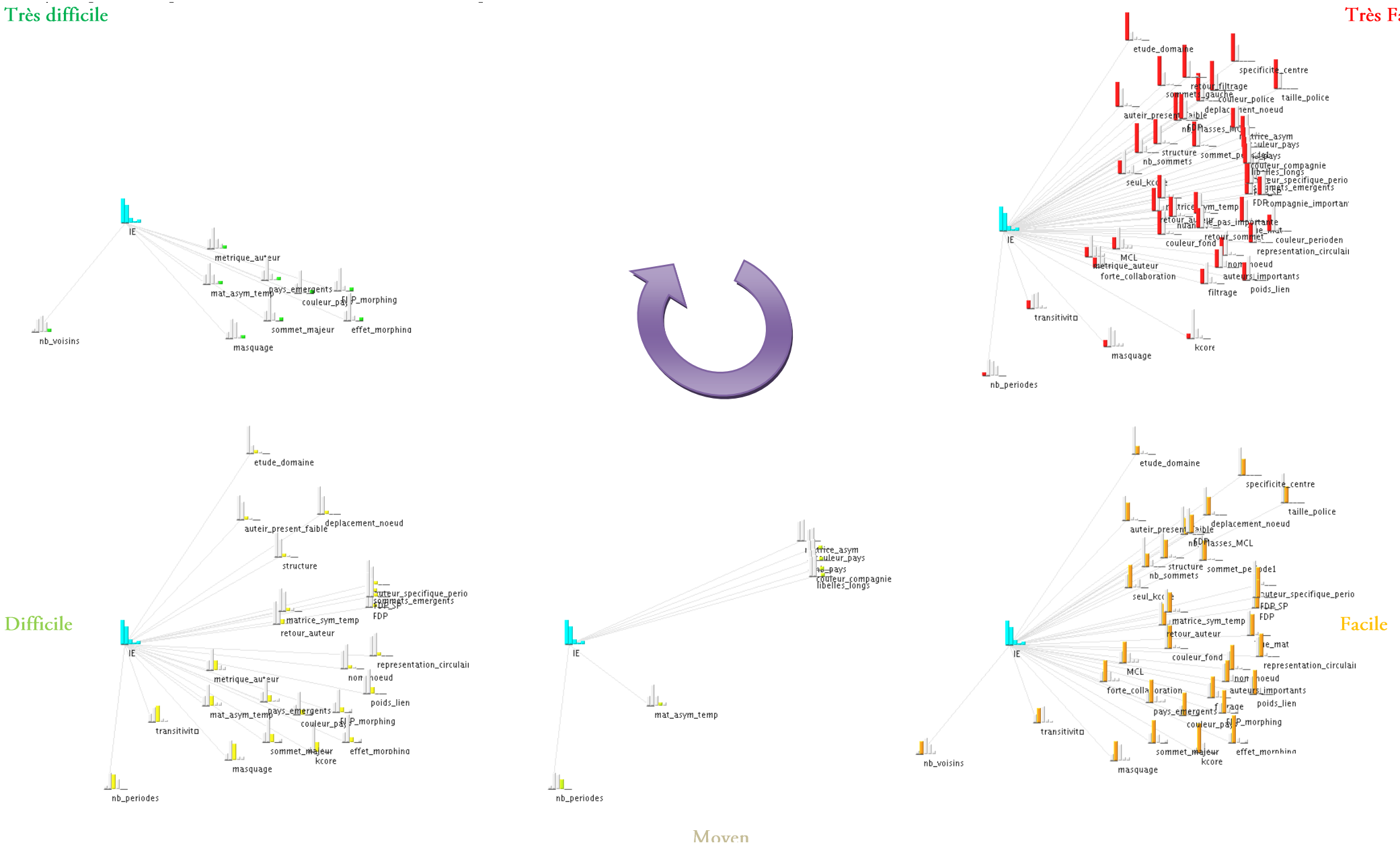


Figure 111. *Morphing des réponses des personnes travaillant dans le domaine de l'IE.*

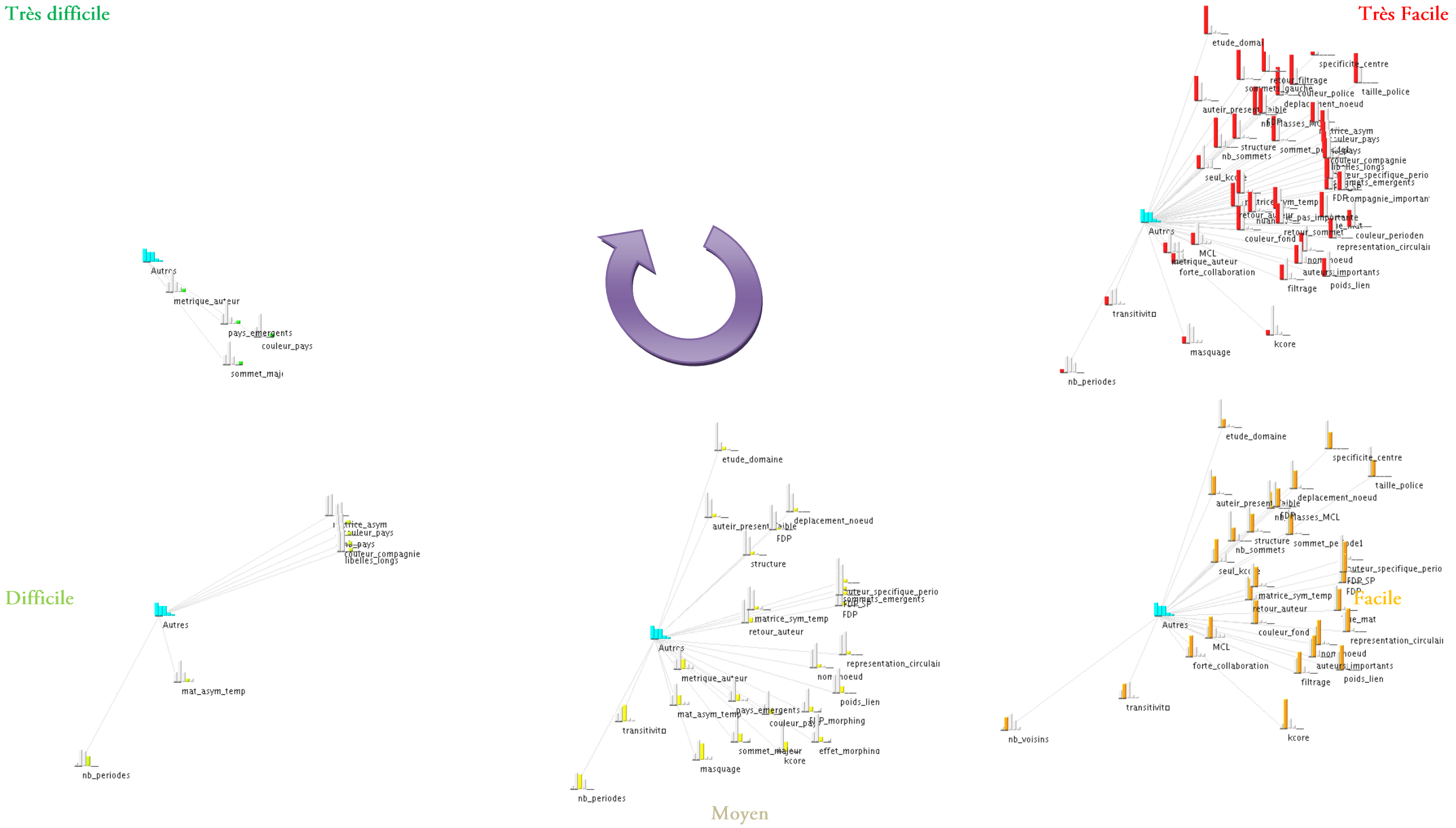


Figure 112. *Morphing des réponses des autres personnes interrogées.*

Si nous nous intéressons aux points à améliorer, qui sont considérés par les utilisateurs, des questions sont à se poser sur certaines difficultés:

- la non connaissance de la fonctionnalité avant ce sondage ? Par exemple, les personnes qui ne connaissent pas obligatoirement le principe des *k-cores* ou encore de la *transitivité*.
- Est-ce un problème ergonomique ? La fonctionnalité est-elle trop difficile à appliquer ? les résultats ne sont-ils pas clairs ? Quelles seraient les solutions, dans ce cas là pour améliorer l'échange entre le système et l'utilisateur ?
- La découverte de l'outil ? On considère dans ce cas là que l'utilisateur connait le principe de la fonctionnalité, pour ne pas retomber dans le cas de la première question. Ainsi, après avoir expliqué la manipulation à l'utilisateur, il faut vérifier que ce dernier reproduit seul cette manipulation sans difficulté.

Pour répondre à ces questions, en tenant compte des réponses de l'enquête, les tableaux 35, 36, 37, 38 et 39 synthétisent les résultats « difficile » et « très difficiles ». *Total D* correspond au nombre de personnes qui ont trouvé l'application de la fonctionnalité difficile sur les 66 interrogées. Elle indique le total de ceux qui ont trouvé la tâche très difficile.

Nous ne nous intéressons qu'aux critères ayant plus de 6 personnes ayant signalé une difficulté pour réaliser la tâche. En effet, 6 personnes ne représentent que 9%, alors que 7 personnes représentent environ 10,6% de la population totale. Nous nous fixons un seuil pour lequel si au moins 10% de la population trouve le critère très difficile d'application, alors nous le prenons en compte. Les lignes contenant ces valeurs sont surlignées dans les tableaux

Ainsi, les résultats synthétisés dans les *Tableaux 35, 36, 37, 38* et *39* montrent bien que les testeurs ayant répondu à notre enquête rencontrent des difficultés pour appliquer les fonctionnalités suivantes.

La représentation circulaire n'a pas satisfait 12 personnes soit 18% de la population globale étudiée. Afin de trouver l'origine de cette difficulté, la majorité de ces personnes nous ont informé que cette représentation fortement utile doit être plus facilement accessible et davantage mise en avant.
Le nombre de périodes n'a pas été trouvé par 8 personnes, soit 12% de la population générale, sur le graphe global. Les individus interrogés indiquent que ce procédé étant innovant, sans pré-requis la distinction entre les repères et les sommets n'est pas spontanée. Après leur avoir expliqué le principe, 100% de ces personnes ont été capable de répondre aux questions sur ce sujet.
La couleur d'une période spécifique, soit 12 personnes, ont affirmé rencontrer une difficulté, allant dans le même sens que le point précédent sur le nombre de périodes étudiées.
Pour la découverte du sommet d'une période spécifique, 7 personnes, soit 10,6% n'ont pas su répondre correctement. Pour cette question, une ambigüité est apparue dans le questionnaire. En effet, la réponse attendue doit être un sommet ayant une valeur de métrique positive pour la période considérée et nulle pour toutes les autres. Parmi ces 7 personnes, 5 ont fourni un sommet ayant une très forte valeur pour l'instance considérée et une faible valeur pour les autres. Leur solution ne relève pas d'un problème d'utilisation de VisuGraph mais plus d'une mauvaise compréhension du sujet.
La détection d'auteurs faiblement présents durant toutes les périodes étudiées, a été difficile pour 12 personnes. La réponse attendue ciblait un auteur en position centrale, attiré par tous les repères. La discussion avec ces individus a abouti à un souci de distinction entre un individu central de faible valeur et un individu central de forte valeur. Le fait que ces deux sommets ont des positions similaires a perturbé les testeurs. La lecture de la métrique d'un auteur dans un contexte temporel a été difficile pour 7 personnes, tout comme le point précédent sur la détection d'auteurs présents pour toutes les périodes. Dans un contexte asymétrique, ces personnes omettent que la valeur de la métrique est égale à la somme des liens du sommet avec son entourage.
Une fois l'affichage des informations sur le nœud apparu, les personnes pensent que la valeur de métrique est clairement indiquée pour chaque instance alors que ce sont les valeurs des liens pour chaque période.

	Question	Info.	Stat.	IE	Autres	Total D	Info.	Stat.	IE	Autres	Total TD
Fonction générale	Vue_mat	1	0	0	0	1	0	0	0	0	0
	nb_sommets	1	0	0	0	1	0	0	0	0	0
	nom_noeud	1	0	0	0	1	0	0	0	0	0
	poids_lien	1	0	0	0	1	0	0	0	0	0
	couleur_police	1	0	0	0	1	0	0	0	0	0
	couleur_fond	1	0	0	0	1	0	0	0	0	0
	deplacement_noeud	2	1	0	0	3	0	0	0	0	0
	auteurs_importants	1	0	0	0	1	0	0	0	0	0

Tableau 35. *Synthèse des résultats des morphing par catégorie de personnes interrogées, pour les fonctionnalités générales de VisuGraph.*

	Question	Info.	Stat.	IE	Autres	Total D	Info.	Stat.	IE	Autres	Total TD
Matrice symétrique	filtrage	1	0	0	0	1	0	0	0	0	0
	retour_filtrage	1	0	0	0	1	0	0	0	0	0
	kcore	2	2	0	0	4	0	0	0	0	0
	seul_kcore	1	0	0	0	1	0	0	0	0	0
	nb_Classes_MCL	2	2	0	0	4	0	0	0	0	0
	representation_circulaire	6	6	0	0	12	0	0	0	0	0

Tableau 36. *Synthèse des résultats des morphing par catégorie de personnes interrogées, pour les représentations des matrices symétriques de VisuGraph.*

	Question	Info.	Stat.	IE	Autres	Total D	Info.	Stat.	IE	Autres	Total TD
Matrice symétrique temporelle	nb_periodes	4	4	0	0	8	0	0	0	0	0
	couleur_periode n	6	6	0	0	12	0	0	0	0	0
	sommet_periode 1	4	3	0	0	7	0	0	0	0	0
	specificite_centre	3	2	0	0	5	0	0	0	0	0
	sommets_emergents	1	0	0	0	1	0	0	0	0	0
	auteur_specifique_perioden	1	1	0	0	2	0	0	0	0	0
	auteur_present_faible	7	6	1	2	16	0	0	0	0	0
	FDP	1	0	0	0	1	0	0	0	0	0
	FDP parametree	1	0	0	0	1	0	0	0	0	0

Tableau 37. *Synthèse des résultats des morphing par catégorie de personnes interrogées, pour les représentations des matrices symétriques temporelles de VisuGraph.*

	Question	Info.	Stat.	IE	Autres	Total D	Info.	Stat.	IE	Autres	Total TD
Matrice asymétrique	couleur_compagnie	1	0	0	0	1	0	0	0	0	0
	libelles_longs	1	0	0	0	1	0	0	0	0	0
	cie_pas_importante	1	0	0	0	1	0	0	0	0	0
	nuances	0	0	1	2	3	0	0	0	0	0
	retour_sommet	1	0	1	2	4	0	0	0	0	0
	etude_domaine	0	0	1	2	3	0	0	0	0	0
	transitivité	1	0	1	2	4	0	1	0	0	1
	retour_auteur	1	0	1	2	4	1	1	0	0	2
	MCL	1	0	0	0	1	0	2	0	0	2
	forte_collaboration	1	0	0	0	1	0	3	0	0	3
	nb_voisins	0	0	0	0	0	0	0	2	0	2
	FDP_morphing	1	0	0	0	1	0	0	2	0	2
	effet_morphing	1	1	1	2	5	0	0	2	0	2
	masquage	1	0	0	0	1	1	0	2	0	3

Tableau 38. *Synthèse des résultats des morphing par catégorie de personnes interrogées, pour les représentations des matrices asymétriques de VisuGraph.*

	Question	Info.	Stat.	IE	Autres	Total D	Info.	Stat.	IE	Autres	Total TD
Mat. asymétrique temporelle	mat_asym_temp	1	0	0	0	1	1	0	2	0	3
	couleur_pays	1	0	0	0	1	0	0	2	1	3
	sommet_majeur	1	0	0	0	1	0	0	2	1	3
	pays_emergents	1	0	0	0	1	0	1	2	1	4
	metrique_auteur	0	0	0	0	0	2	2	2	1	7

Tableau 39. *Synthèse des résultats des morphing par catégorie de personnes interrogées, pour les représentations des matrices asymétriques temporelles de VisuGraph.*

✓ ***Analyse des autres résultats***

La dernière partie de l'enquête porte sur le ressenti des testeurs vis-à-vis de VisuGraph. Les résultats obtenus sont les suivants.

Questions	Utile / Oui(%)	Assez utile (%)	Inutile /Non (%)	sans réponse (%)
Qualification de l'outil	33,33	66,66	0,00	0,00
représentation des données	0,00	92,60%	0,00	11,11
Facilite l'exploration des données	81,78	0,00	0,00	22,22
détection des tendances	88,88	0,00	3,70	11,11
interactif	85,18	0,00	14,81	3,70
ergonomique	40,74	0,00	33,33	25,92
accès fonctionnalités facile	66,67	0,00	14,81	11,11
déjà effectué des analyse via tétralogie	55,5	0,00	33 ,33	7,41
Ajout à Tétralogie	62,50	25,00	0,00	12,50
réutilisabilité professionnelle	62,50	25,00	0,00	12,50

Tableau 40. *Synthèse sur l'évaluation de VisuGraph.*

Les résultats du Tableau 40 montrent une importante satisfaction de l'expérimentation de VisuGraph. Ces réponses nous permettent d'établir des plans de prévisions sur les axes à améliorer, tels que l'ergonomie, l'accès aux fonctionnalités et davantage d'interactivité. Pour la dernière question, l'interrogation portait sur la volonté des testeurs d'utiliser ultérieurement VisuGaph dans leur domaine professionnel. 62,50% des personnes interrogées ont répondu qu'ils réutiliseront VisuGraph dans le cadre de leur emploi et l'intégration de ce module à tétralogie leur semble utile, i.e. « Ajout à Tétralogie ».

✓ ***Conséquences du sondage***

Suite au dépouillement des données et à l'étude des résultats, des mesures sont prises pour améliorer VisuGraph.

Deux types de mesures sont effectués :

- Les modifications directes. Des changements sont applicables facilement, sans provoquer une restructuration trop importante de l'outil. Elles concernent les points suivants :
  La représentation circulaire. Suivant les suggestions des personnes ayant répondu à l'enquête, cette visualisation est appliquée par défaut aux graphes non temporels, qu'ils soient bipartis ou non. Ainsi, les utilisateurs trouvent que cette représentation est beaucoup plus claire qu'une représentation moins ordonnée.
  L'ajout d'une fonctionnalité permettant de modifier le libellé des repères facilite la maîtrise de l'outil pour une personne peu initiée. Ainsi, si un utilisateur préfère appeler un repère « Repère n » plutôt que par l'année qu'il représente, cas par défaut dans un cas temporel, il peut facilement le faire.
- Les modifications à long terme. Elles entrainent soit un changement de la structure du programme, soit une nouvelle formulation de l'approche, ou encore la recherche et le test de nouvelles solutions. Elles concernent principalement l'ergonomie de l'outil, l'accès aux fonctionnalités, et l'extension du morphing de graphe. Nous détaillons ces mesures dans nos perspectives, dans la dernière partie de ce document.

5.5. Synthèse de l'expérimentation et de la validation de VisuGraph

Dans ce chapitre, nous avons précisé la conception de l'outil de représentation d'entités évolutives ou non, dans la perspective de l'analyse visuelle de données temporelles. Nous avons dans un premier temps présenté l'architecture élaborée.

Dans un second temps, les critères retenus dans les chapitres précédents pour évaluer les outils de visualisation de données temporelles, disponibles dans ce domaine. Nous avons alors évalué VisuGraph par rapport à ces critères, afin de pouvoir le positionner par rapport aux autres outils existants.

Puis, une enquête a été élaborée, basée sur un questionnaire précis, portant sur la manipulation de l'outil pour visualiser les connaissances organisées sous forme de matrice, comme vu dans le premier chapitre.

La population interrogée à été présentée afin de spécifier les catégories d'individus la composant. Nous avons ciblé des personnes du domaine, et plus spécifiquement des informaticiens, des statisticiens, des personnes de l'intelligence économique mais aussi des personnes extérieures, dont l'utilisation de l'outil pour être utile professionnellement.

Le protocole de réalisation de l'enquête a été expliqué, c'est-à-dire la situation contextuelle de l'expérimentation, les notions et le vocabulaire de base, les objectifs et la finalité de cette étude.

Les résultats ont ensuite été dépouillés et analysés. Tout d'abord, le temps mis par chaque catégorie de la population étudiée a été analysé afin d'étudier les éventuelles difficultés, selon le domaine professionnel. Nous nous sommes rendu compte que les résultats ne variaient pas beaucoup, ce qui nous indique que l'outil est accessible par un bon nombre d'individu d'origines scientifiques distinctes.

Puis, le morphing de graphe a été appliqué sur les résultats de l'enquête. La notion de temps a été remplacée par celle de notation, allant de « très facile » à « très difficile ». Ainsi, chaque graphe d'instance correspond à la représentation des manipulations jugées selon la qualification spécifique à la visualisation.

Par exemple, le graphe d'instance « très facile » correspond à la visualisation des manipulations jugées très aisées à effectuer. Par ces représentations, le constat d'une satisfaction plutôt générale, pour toutes les catégories s'est traduite par une facilité à utiliser l'outil.

Dans la dernière partie du sondage, nous avons demandé aux testeurs de qualifier l'outil, à savoir si son intégration au sein de Tétralogie est utile, si la manipulation est aisée. Les résultats sont encourageants et les points à améliorer ont été catégorisé selon la possibilité d'effectuer des correctifs immédiats, sans que cela implique un remaniement de la structure de base de VisuGraph et les changements plus conséquents. Les premiers ont été corrigés et ces modifications ont été présentées. Les second sont présentés dans le chapitre suivant qui conclut ce manuscrit et présente nos perspectives.

Ces évaluations nous ont permis d'évaluer l'utilisabilité de chaque technique de visualisation temporelle au regard de chaque tâche utilisateur. Les résultats de ces évaluations sont exploités par l'algorithme de sélection des techniques de visualisation par rapport aux tâches utilisateur.

Enfin, nous notons que VisuGraph est de l'ordre de dix mille lignes de code Java.

Conclusion et perspectives

> *«Ce n'est pas la fin. Ce n'est même pas le commencement de la fin. Mais, c'est peut-être la fin du commencement»* (Winston Churchill, 1953)

6.1. Résumé des contributions ____ 206
6.2. Limites et approches envisagées ____ 208
6.2.1. Volume de données ____ 208
6.2.2. Ergonomie ____ 208
6.3. Perspectives de recherche ____ 208
6.3.1. Données en entrée ____ 208
6.3.2. Ajout de fonctionnalités ____ 209
✓ Classification ... 209
✓ 3 dimensions (3D) ... 210
✓ Amélioration du morphing de graphe ... 210
✓ Granularité temporelles ... 210
✓ Image de synthèse ... 210
✓ Réalisation d'un rapport automatique ... 210

Tout comme dans la nature, dans la société ou encore chez l'homme, le risque est présent à chaque instant de la vie de l'entreprise et des organisations, sous toutes les formes, souvent imprévues. La récente crise en est la confirmation.

L'évolution du monde, sa globalisation, sa complexité font apparaître de nouveaux risques ou de nouvelles formes de risques. Elles sont autant d'opportunités, mais les réponses ne sont plus les mêmes. Le décideur doit multiplier les sources d'information, les hypothèses, innover pour trouver les démarches adaptées, saisir sa chance.

L'approche scientifique préconisée pour ce type de démarche se fonde sur la collecte qualifiée des informations nécessaires, leur synthèse efficace et hiérarchisée, l'analyse objective des causes et l'évaluation réaliste des conséquences. Pour cela, le recours à la visualisation de données relationnelles et temporelles permet d'analyser de façon claire de grands volumes de données, difficilement exploitables sous forme textuelle.

De nombreux outils de visualisation de données relationnelles sont accessibles pour répondre à un besoin de représentation globale des données afin de faciliter leur analyse. Une grande majorité de ces outils se basent sur des visualisations statiques munie de fonctionnalités d'analyses performantes mais qui peuvent mener à des erreurs d'interprétation si l'utilisateur ne maîtrise pas complètement la notion de temporalité.

En effet, si la visualisation repose sur plusieurs périodes cumulées les différentes associations décelées ne sont pas forcément simultanées. Pour éviter ces soucis d'interprétation, des outils de représentation de données relationnelles et temporelles sont nombreux et variés. La multiplicité des techniques d'interaction dédiées aux données temporelles témoigne du dynamisme de cet axe d'étude. Nous proposons ici, en conclusion, un résumé de nos contributions en soulignant leur originalité.

Une analyse critique des résultats permet d'envisager pour nos travaux de multiples perspectives, que nous organisons en deux parties : les extensions et les prolongements à plus long terme.

6.1. Résumé des contributions

Afin de pouvoir disposer des avantages de la représentation statique mais aussi dynamique, notre contribution repose sur l'accès à ces deux types de visualisation, complétés par des méthodes d'analyse permettant l'exploration des structures de graphes, qu'elles soient temporelles ou non. L'intérêt de notre approche repose sur l'étendu des étapes d'extraction et de visualisation des données. Comme nous l'avons vu dans le premier chapitre, suite à une collecte de textes semi-structurés, sur un sujet défini, les informations sont extraites, prétraitées via la plateforme de veille stratégique Tétralogie, développée par notre équipe. Notre démarche d'analyse s'appuie en premier lieu sur les caractéristiques des données manipulées, en l'occurrence des données temporelles. Or ces dernières sont d'une part manipulées par des utilisateurs et d'autre part gérées par un système informatique. L'intérêt de notre analyse tient à sa capacité à prendre en compte à la fois l'utilisateur et le système.

Les entités sont croisées en matrices de cooccurrences, dont la valeur indique le nombre de fois où la présence simultanée de deux informations est observée. Suivant la granularité temporelle choisie, la matrice peut être décomposée en plusieurs périodes homogènes. Dans ce cas là, nous avons autant de matrices de cooccurrences que de périodes considérées. Une fois les entités organisées, notre contribution repose sur l'étude des techniques d'interaction, à savoir les sémiologies utilisées, ainsi que les fonctionnalités et approches structurales, dans le chapitre 2.

L'intérêt de notre contribution repose sur trois axes principaux :

- La représentation graphique de données, enrichie en information afin de permettre une exploration plus pertinente. La conception d'un espace de représentation dans lequel sont placées spécifiquement les données relationnelles et la visualisation de ces dernières par des procédés de codage sémiologique des valeurs de métrique facilitent le travail de l'utilisateur.
- La mise en place de techniques d'analyse de structure de graphes communes à de nombreux outils graphiques, complétées et compatibles avec des méthodes performantes. L'utilisateur étant au centre de nos travaux, chacune de ces fonctionnalités est paramétrable, via des règles graduées. Notre contribution repose aussi sur la proposition d'algorithmes et de fonctionnalités qui améliorent le tracé du graphe et facilite son interprétation, mais aussi sur la représentation de graphes qui offrent davantage d'informations que les outils de visualisation classique. Nous validons la faisabilité de notre contribution de conception ergonomique avec la création de VisuGraph, incluant les techniques d'interaction conçues.
- Le développement d'une animation temporelle de graphe. . L'élément principale de notre contribution repose sur le morphing de graphe, c'est à dire la transformation géométrique d'une représentation graphique, permettant le passage d'une visualisation de donnée au temps t-1 à celle de t et inversement, de façon fluide et naturelle. Pour ce faire, chaque donnée temporelle est représentée selon une sémiologie facilitant la détection des tendances de cette dernière à savoir son degré de persistance. Le morphing repose sur une première représentation globale, toutes périodes confondues, puis individuelles. L'animation des visualisations successives des différentes périodes, dans le sens chronologique similaire à l'analogie espace/temps d'une horloge, permet de créer une certaine dynamique, révélant l'évolution des données au cours du temps, trouvant ainsi un bon compromis entre la préservation de la carte mentale de l'utilisateur et la lisibilité du tracé.

La Figure 113 synthétise notre contribution en précisant la spécificité de chacune et illustre deux exemples d'interactions entre les composants de notre contribution permettant une analyse pertinente.

Notre contribution est expérimentée, en se basant sur les critères qui nous ont permis d'évaluer les outils dans l'état de l'art, puis à travers un sondage réalisé auprès d'une population ciblée. Cette enquête repose sur des analyses effectuées par notre équipe. Les sujets interrogés ont effectué ces mêmes analyses, en évaluant le niveau de difficulté d'utilisation de l'outil sous différents aspects. Leurs résultats ont été comparés aux nôtres afin de s'assurer de la validité des interprétations les plus intuitives des représentations fournies. Les résultats prouvent l'intérêt de VisuGraph dans un contexte de veille stratégique.

Contributions	Spécificités	Références		Exemples d'interactions intéressantes
FONCTIONNALITES				
k-core	Affichage des plus grosses équipes	Chapitres 2 et 3		Cette complémentarité de fonctionnalités permet, à partir d'un sommet sélectionné, d'étudier son voisinage direct, puis indirect, dans un contexte temporel. L'application des FDP permet d'obtenir un graphe lisible, dont les arcs permettent de distinguer l'influence du sommet sur les autres et inversement, confirmé par la représentation caméléon. Le morphing de graphe permet de détecter les caractéristiques temporelles du sommet à savoir s'il est persistant, émergent ou inversement en régression. Le nombre, l'évolution et la nature des relations permettent de distinguer les propriétés du sommet initialement sélectionné.
transitivité	Sélection d'un sommet et affichage de son voisinage par pas.			
FDP statique/dynamique	placement (temporel ou non) sommets de façon à réduire les croisements d'arêtes. Réglage via des slider			
Filtrage	Masquage des sommets et arêtes de valeur inférieure au seuil fixé via le slider.			
Focalisation	Affichages de sommets, relations spécifiques			
Retour documents	Affichages des notices dans lesquels apparait la donnée sélectionnée			
Partitionnement de graphe	MCL, méta-graphe, extraction d'une classe spécifique, exportation des classes.			
TYPE DE GRAPHES				
Orientés	Orientation des arcs et caractérisation de chaque nœud par croisement avec une matrice complémentaire.	Chapitre 2 et 3		Le Kcore, pour un k maximal permet de détecter le centre de la structure du graphe, à savoir l'équipe la plus importante. L'application des FDP permet de rendre la lisibilité de cette structure plus planaire. La visualisation spécifique permettant de qualifier quantitativement et qualitativement les liens, il est alors directement possible de dire si les relations entre les individus de la structures sont strictement les mêmes ou si elles différent. De plus, la qualification des sommets permet de distinguer une information complémentaire sur la donnée, comme par exemple sa nationalité ou encore son laboratoire de rattachement.
Non orientés	Graphe simple ou biparti avec possibilité d'ajouter des informations complémentaire : possibilité de croiser 4 dimensions.			
caméléon	Distinction entre les prédécesseurs et les successeurs dans les cas orientés avec positionnement relatif à l'influence d'un sommet sur l'autre.	Chapitre 2		
spécifique	Liens caractérisés qualitativement et quantitativement avec possibilité de croiser plusieurs dimensions.			
ANIMATION				
Morphing de graphe	Animation chronologique des différents graphes de période. Détection des tendances, analyse de l'évolution.	Chapitre 3		

Figure 113. *Synthèse de notre contribution et apports de la combinaison de fonctionnalités.*

6.2. Limites et approches envisagées

Le développement d'un outil de visualisation de données relationnelles intégrant plusieurs techniques d'interaction différentes est une tâche ambitieuse dans le temps imparti d'une thèse. Aussi VisuGraph connait certaines faiblesses. Nos perspectives à court terme visent donc la complétude et l'amélioration de VisuGraph d'un point de vue ergonomique mais aussi au niveau de la réalisation logicielle, entre autres, par l'ajout de fonctionnalités.

6.2.1. Volume de données

Dans un certain nombre de cas, le dessin de graphes et le calcul d'attributs numériques et/ou d'indices visuels s'avèrent impuissants face au volume des données à visualiser et à explorer. Plus les matrices utilisées en entrée de VisuGraph sont importantes, plus le système est ralenti et le graphe chargé. Des solutions pour analyser des graphes de grande taille ont été présentées dans ce manuscrit, telles que la conception de méta graphe via le Markov Clustering, l'application de forces d'attraction et de répulsion pour former des structures spécifiques,... Notre contribution n'est pas de concevoir des graphes de taille très importante mais plutôt de favoriser l'aspect interactif et fonctionnel des besoins en visualisation. Une perspective d'évolution de VisuGraph est l'optimisation lors de la lecture des matrices en entrée et des différentes fonctionnalités.

Un autre aspect contraignant du volume important de données concerne le ralentissement lors du déplacement des sommets suite à l'application des algorithmes de forces de placements dirigés. Plus le nombre de sommets est conséquent, plus le déplacement est saccadé. Chaque changement de position d'un sommet entraine le rafraichissement de l'ensemble du graphe, ralentissant considérablement le système. Des tests sur l'affichage d'un déplacement sur deux pour un sommet se sont révélés efficace au niveau des performances du systèmes mais pas au niveau de la fluidité du mouvement. Cette solution engendre la réduction du nombre de calcul, solution clé de cette limite. Une nouvelle perspective est la conception d'un algorithme se basant sur la valeur des liens et sur la distance entre les sommets. Si deux entités sont fortement éloignées et peu, ou pas du tout, liés alors leur répulsion ne sera pas calculée, la distance les séparant étant significative. Cela implique la délimitation de seuils au niveau de la distance et de la valeur des liaisons qui doivent être expérimentalement justifiées.

6.2.2. Ergonomie

La recherche du bon équilibre entre une fonction, un matériel et son utilisateur est à la base du travail de l'ergonomie. Lors de l'enquête, des problèmes d'ergonomie ont été découverts au niveau du menu. Il convient donc d'améliorer ce dernier.

6.3. Perspectives de recherche

6.3.1. Données en entrée

Dans le chapitre 1, les matrices de croisement ont été présentées. Elles sont conçues via la plateforme Tétralogie et ne permettent pas, à ce jour, de concevoir directement les matrices qualificatives étudiées dans le chapitre 3, en 3.2.7, pour les graphes spécifiques. Cependant, une autre approche, via la plateforme, permet d'obtenir un résultat similaire.

En effet, dans le cas de VisuGraph, la matrice de cooccurrences est discrétisée en périodes homogènes. Le même procédé peut être appliqué pour les qualificatifs des liens. Ainsi, la discrétisation se fera par type qualitatif de lien.

Chaque graphe qualitatif peut alors être représenté via la fonctionnalité de morphing et le graphe global permet alors d'obtenir une visualisation, pour toutes données confondues, mais dont les liens sont typés.

Cette approche est testée mais n'a pas encore été complètement validée par notre équipe. Elle nécessite donc d'être davantage testée afin de s'assurer de sa conformité.

6.3.2. Ajout de fonctionnalités

L'ajout de nouvelles fonctionnalités est une perspective pour l'évolution de tout outil logiciel. Suite aux réponses obtenues lors du sondage d'évaluation de nos travaux, des améliorations doivent être apportées au niveau des fonctionnalités d'exploration d'un nœud, de ses caractéristiques dans un contexte temporel, telles que son voisinage, ses valeurs de métrique pour les différentes périodes considérées.

- ✓ ***Classification***

Actuellement, l'outil VisuGraph est doté d'une fonctionnalité de partitionnement, à travers le Markov Clustering. Il est intéressant de compléter cet outil par d'autres approches de classification.

L'algorithme *k-means* (Diday, 1971) permet le partitionnement des données d'une image en k clusters. Le *k-means* est un algorithme itératif qui minimise la somme des distances entre chaque objet et le centroïde de son cluster. Contrairement à d'autres méthodes dites hiérarchiques, qui créent une structure en « arbre de clusters » pour décrire les groupements, le *k-means* ne crée qu'un seul niveau de clusters. L'algorithme renvoie une partition des données, dans laquelle les objets à l'intérieur de chaque cluster sont aussi proches que possible les uns des autres et aussi loin que possible des objets des autres clusters. Cette classification se base sur la notion de distance entre les données.

Une autre approche concerne la Classification Ascendante Hiérarchique (CAH) qui considère initialement toutes les observations comme étant des clusters ne contenant qu'une seule observation, et leur distance est alors le plus souvent définie comme étant leur distance euclidienne. La première étape consiste donc à réunir dans un cluster à deux observations les plus proches. Puis la CAH continue, fusionnant à chaque étape les deux clusters les plus proches au sens de la distance choisie. Le processus s'arrête quand les deux clusters restant fusionnent dans l'unique cluster contenant toutes les observations. La CAH étant déjà fonctionnelle sous la plateforme Tétralogie, les résultats de cette dernière seraient réutilisés dans le contexte de VisuGraph et les classes seraient visualisées sous forme de graphe de clusters.

Les variables étant toutes numériques, on recourt instinctivement à la distance euclidienne. Mais d'autres distances sont possibles (Diday, 2008) et habituellement disponibles, sans qu'il existe d'argument fort en faveur de l'une ou de l'autre. Dans le cas de la représentation de matrices de cooccurrences, une approche intéressante serait d'appliquer dans un premier temps l'algorithme FDP afin de disposer les sommets selon la nature de leur relation. Deux sommets fortement liés sont alors placés proches l'un de l'autre et inversement. Cette disposition peut servir de notion de distance et à partir des coordonnées obtenues de chaque sommet, le k-means, ou encore la CAH pourrait être appliquées. Cette perspective étant dans un contexte simple, sans prise en compte de la dimension temporelle, elle pourrait être appliquée par la suite au cas dynamique. Comme nous l'avons vu dans ce manuscrit, les données temporelles sont positionnées selon leur appartenance aux différentes périodes.

Appliqué à partir des coordonnées temporelles, le k-means, ou encore la CAH, permettrait d'obtenir un ensemble de classes temporelles.

Si nous comparons ces deux classifications, l'algorithme du k-means et ses variantes sont très rapides, bien adaptés à de grands modèles, mais non déterministes et la classification hiérarchique est justement une opposition à ce principe.

C'est pourquoi une approche intéressante serait d'associer ces deux approches en commençant par l'algorithme du k-means pour obtenir quelques dizaines ou centaines classes ; puis une classification hiérarchique de ces classes et non pas des données initiales, pour trouver le nombre de classes. Enfin, il serait possible d'affiner le résultat avec l'algorithme des k-means sur les classes nouvellement obtenues.

✓ ***3 dimensions (3D)***

Le passage à la 3D permet de fluidifier et d'améliorer les résultats des algorithmes FDP. En effet, lorsque deux arêtes sont croisées en 2D, le point d'intersection entre les deux représente un cap difficile à franchir. Dans le cas de la 3D, ce croisement est visible mais n'est pas réel, puisque la troisième dimension permet d'éviter tout point commun de croisement entre deux arêtes.

Des tests ont été effectués pour ajouter la 3D à VisuGraph. Ils ont été concluants, mais l'implication d'une troisième coordonnée rend le système plus complexe et les résultats obtenus nécessiteraient d'être approfondis et testés.

✓ ***Amélioration du morphing de graphe***

En ce qui concerne le morphing de graphe, contrairement au tracé statique où les contraintes liées au tracé sont connues à l'avance et ne changent pas dans le temps, le tracé dynamique doit prendre en compte le fait que l'utilisateur connaît le tracé présenté à l'instant t-1. Ce tracé doit donc être pris en compte lors de la conception du nouveau tracé. Donc, à chaque instant t, le nouveau tracé doit non seulement être intelligible selon les critères énoncés à la section précédente mais il doit, de plus, permettre une transition aisée avec le tracé présenté à la période précédente. Ceci permet à l'utilisateur d'éviter de perdre son temps et une énergie cognitive importante pour la découverte du nouveau tracé.

L'inconvénient de cette méthode est qu'elle est très coûteuse en temps de calcul, et doit être suffisamment lente pour permettre à l'utilisateur de voir et d'enregistrer les différents changements opérés sur le graphe. Le réglage de la vitesse de transition permet de remédier à ce problème. Cependant, le fondu entre les deux graphes manque de fluidité et nos perspectives d'amélioration porteront sur une animation plus naturelle au niveau des transitions, tout en prenant en compte l'optimisation du nombre de calculs pour ne pas ralentir le système.

✓ ***Granularité temporelle***

Enfin, une autre perspective repose sur la granularité temporelle. En effet, pour le moment, une unité de temps peut être décomposée en sous périodes, via les graphes d'instances. Cependant, une seule unité peut être choisie. Ainsi, si chaque graphe de période correspond à une année, ce dernier ne peut pas être décomposé en mois ou en trimestres. Une évolution serait de permettre cette focalisation temporelle.

✓ ***Image de synthèse***

L'amélioration du morphing de graphe passe par la fluidité de l'animation. La réalisation d'images de synthèse de l'évolution d'une donnée et de son environnement faciliterait grandement son analyse. Une perspective d'évolution serait la réalisation d'un scénario temporel, permettant de visualiser la structure d'une donnée et sa transformation naturelle au cours du temps.

✓ ***Réalisation d'un rapport automatique***

Afin de faciliter le travail de l'analyste, la réalisation automatique d'une synthèse sur les données visualisées est une solution.

Ce document texte comprendrait la classification croissante des sommets selon plusieurs critères choisis par l'utilisateur, tels que

- par valeur de métrique globale ;
- dans un contexte temporel, un classement par période ;
- dans un cas non orienté, un classement par nombre de liens. L'individu ayant le plus grand nombre de voisins serait positionné en tête de liste ;

- dans le cas orienté, un double classement par classement du nombre de liens des prédécesseurs puis des successeurs ;
- la liste des composants des structures les plus importantes. Ce résultat serait obtenu après l'application d'une forte valeur pour le *k-core* ;
- …

Liste des figures

Figure 1. *Notre approche de travail.* 17
Figure 2. *Principe de veille et découverte de connaissance.* 21
Figure 3. *Le concept d'intelligence économique* 22
Figure 4. *De l'IE aux différents types de veille.* 22
Figure 5. *Processus de veille type adapté de l'Afnor X50-053* 23
Figure 6. *ECD et RI, deux logiques différentes* 25
Figure 7. *Les étapes du processus d'Extraction de Connaissances à partir de Bases de Données (ECBD) ((Fayyad et al., 1996)* 26
Figure 8. *Qualification des multi-termes à conserver dans le dictionnaire (Roux, 1998).* 31
Figure 9. *Approche de la fouille de données.* 35
Figure 10. *AFC basée sur des croisements d'entreprises.* 45
Figure 11. *Zoom de deux données de l'AFC sur les croisements d'entreprise.* 46
Figure 12. *Extrait de la matrice de cooccurrences entre des entreprises.* 46
Figure 13. *Typologie de l'information* 48
Figure 14. *Principe de l'algorithme de ressorts.* 59
Figure 15. *Classification de (Keim, 2002).* 60
Figure 16. *Structure et caractéristiques de l'espace temps.* 62
Figure 17. *TimeLine sur des articles de recherche (Morris et al., 2003).* 63
Figure 18. *Starfiel, système dédié à la visualisation d'une base de données de films.* 64
Figure 19. *ThemeRiver (Havre et al., 1999), (Havre et al., 2000).* 64
Figure 20. *System TimeMapViewer (Johnson et Wilson, 2002).* 65
Figure 21. *Système OITL utilisant le principe du timeline (Bui et al., 2001).* 65
Figure 22. *LifeLines (Plaisant et al. 1998).* 66
Figure 23. *Timelines dynamiques pour la visualisation d'historique de photographies (Kullberg, 1995).* 66
Figure 24. *Deux axes pour représenter le temps (Wijk et Selow, 1999).* 67
Figure 25. *Timeline conçue à partir de l'outil AllOfMe.* 67
Figure 26. *Timeline obtenue à partir de Bee Doc's TimeLine.* 68
Figure 27. *Dandelife : un réseau social basé sur la biographie.* 68
Figure 28. *Timeline obtenue avec l'outil Mnemograph.* 69
Figure 29. *Vues non uniformes : la fonction de transformation permet de déformer les structures visuelles.* 70
Figure 30. *Vues non uniformes : la fonction et le facteur de magnification (Leung et Apperley, 1994).* 71
Figure 31. *Mur en perspective (Mackinlay et al., 1991).* 72
Figure 32. *Mur fuyant (Robertson et Mackinlay, 1993).* 72
Figure 33. *Recours à une spirale pour visualiser un calendrier (Mackinlay, 1994).* 73
Figure 34. *Visualisation 3D représentant huit periodes différentes, basées sur un axe temps central (Tominski et al., 2003).* 73
Figure 35. *Visualisation de l'intensité solaire dans le temps (Weber, 2001).* 74
Figure 36. *Représentation d'une spirale (Renieris et Reiss, 1999).* 75
Figure 37. *Technique SpiraClock (Dragicevic et al., 2002).* 75
Figure 38. *Technique des cercles concentriques (CCT) (Daassi et al., 2000).* 76
Figure 39. *Spiral indentée, utilisant l'axe temps comme support pour représenter les données temporelles.* 76
Figure 40. *Représentation cyclique des données avec l'outil Spiral Graph. (weber et al., 2001).* 77
Figure 41. *Graphe des critères et des outils de visualisation de données temporelles, effectué sous VisuGraph.* 80
Figure 42. *Widgets des dix critères de comparaison, obtenus sous VisuGraph (Loubier et Dousset, 2008).* 81
Figure 43. *Widgets des dix critères de comparaison, obtenus sous VisuGraph (Loubier et Dousset, 2008).* 82
Figure 44. *AFC des critères et des outils de visualisation de données temporelles.Les critères sont en gras et soulignés.* 83
Figure 45. *Timesearcher (Hochheiser, 2002a),(Hochheiser, 2002b).* 85
Figure 46. *Graphe des co-citations d'auteurs scientifiques (Chen, 2004).* 86
Figure 47. *Forme cyclique du temps, représentant plusieurs dimensions structurelles par rapport à un seul espace temps.* 86
Figure 48. *Technique Lexil Pencils (Brian, 2003).* 87
Figure 49. *Data Tube permettant de visualiser un grand nombre de données. (Ankerst, 2000), (Ankerst, 2001).* 87
Figure 50. *Plusieurs données visualisées sur un axe de temps unique (Browser, 2003).* 88
Figure 51. *Représentation 3D d'un réseau évolutif : (Brandes et Corman, 2003).* 88
Figure 52. *Représentation 2D de l'évolution de collaborations par TGRIP (Erten et al., 2004).* 89
Figure 53. *Représentation séparée des réseaux évolutifs (Chen et Carr, 1999).* 89
Figure 54. *Notre approche pour les graphes statiques.* 97

Figure 55. *Valeur de métrique pour chacun des sommets, dans un cas symétrique.* 99
Figure 56. *Valeur de métrique dans un cas asymétrique.* 100
Figure 57. *Prise en compte de la non linéarité de la fonction de codage.* 102
Figure 58. *Graphe basé sur une matrice symétrique (à gauche) et sur une matrice asymétrique (à droite).* 103
Figure 59. *Graphe sur lequel l'algorithme FDP totalement paramétré est appliqué.* 107
Figure 60. *Graphe sur lequel est appliqué l'algorithme FDP semi paramétré.* 109
Figure 61. *Graphe sur lequel l'algorithme FDP est appliqué, basé sur le paramétrage de l'attraction.* 111
Figure 62. *Préconisations sur les différentes étapes de réglage des FDP.* 112
Figure 63. *Prédécesseur et successeur.* 114
Figure 64. *Matrice des prédécesseurs X successeurs.* 115
Figure 65. *Matrice des {predecesseurs U successeurs} X Pays.* 115
Figure 66. *Graphe orienté biparti.* 116
Figure 67. *Matrice symétrique croisant les individus de X.* 117
Figure 68. *Matrice asymétrique croisant les individus de E et de P.* 117
Figure 69. *Matrice symétrique dont les croisement sont de nature qualitative.* 118
Figure 70. *Graphe sur des matrices symétrique et asymétrique dont les liens sont valués qualitativement et quantitativement.* 119
Figure 71. *Application de l'algorithme Caméléon sur le graphe orienté.* 121
Figure 72. *L'utilisateur, contrôlant la visualisation des données relationnelles.* 123
Figure 73. *Filtrage des liens du graphe de co-signatures d'articles scientifiques.* 126
Figure 74. *Décomposition en k-core d'un graphe. Figure réalisée sous VisuGraph* 127
(en bas à gauche 1-core, en haut à droite 2-core, au milieu 3-core, en haut à droite 4-core, en bas à droite 5-core). 127
Figure 75. *Décomposition de la structure d'un graphe.* 129
Figure 76. *Calcul des seuils de transitivité d'un sommet en « fin de structure », représenté en rose.* 130
Figure 77. *Calcul des seuils de transitivité d'une rotule, représentée en rose pour la distinguer des autres sommets.* 131
Figure 78. *Différence entre le calcul de la fermeture transitive de seuil 1 et l'affichage des voisins directs.* 131
Figure 79. *Exploration d'un sommet spécifique et retour aux notices.* 132
Figure 80. *Restitution des résultats graphiques et textuels, suite à la sélection de plusieurs auteurs.* 133
Figure 81. *Conséquence du changement de la valeur de la densité du clustering.* 136
Figure 82. *Graphe initial circulaire, représentant la totalités des données relationnelles.* 137
Figure 83. *Graphe de la figure précédente sur lequel l'algorithme FDP est appliqué. Puis application du MCL.* 138
Figure 84. *Affichage de la totalité des données partitionnées de la figure précédente* 139
Figure 85. *Notre approche pour les graphes dynamiques.* 145
Figure 86. *Exemple de matrices temporelles sur lesquelles repose la structure de l'espace de représentation des données.* 146
Figure 87. *Représentation d'une donnée temporelle, selon le nombre de périodes considérées.* 148
Figure 88. *Conception d'un graphe temporel sous VisuGraph.* 150
Figure 89. *Sur un graphe évolutif (à gauche), application de l'algorithme présenté, paramétré à l'aide du slider « Force repères temporels » (graphe de droite).* 152
Figure 90. *Décomposition du graphe en zones temporelles.* 154
Figure 91. *Morphing d'images.* 156
Figure 92. *Graphes de période.* 158
Figure 93. *Principe du morphing de graphe.* 159
Figure 94. *Schématisation de l'apparition/ disparition / persistance d'un sommet en dix étapes, par morphing de graphe.* 160
Figure 95. *Problème d'ergonomie dans cette palette de couleurs allant du rouge au vert de façon incrémentale.* 161
Figure 96. *Répartition non linéaire de la teinte rouge.* 162
Figure 97. *Comparaison entre le passage du rouge au vert linéaire et non liénaire* 162
Figure 98. *Passage du graphe de première période à celui de la seconde, par morphing de graphe en 10 étapes.* 163
Figure 99. *Mophing avec transition, d'un point de vue « repères temporels ».* 165
Figure 100. *Morphing avec transition, d'un point de vue structure temporelle. transition, d'un point de vue « repère temporel* 166
Figure 101. *Graphe orienté biparti temporel.* 169
Figure 102. *Différences de point de vue sur l'application des fonctions.* 171
Figure 103. *Visualisation du graphe global des classes, puis visualisation par animation des graphes de classe par période* 174
Figure 104. *Application du morphing de graphe dans un contexte géographique.* 176
Figure 105. *Etapes du processus de conception et de réalisation de l'outil VisuGraph.* 180
Figure 106. *Méthodologie d'enquête.* 185
Figure 107. *Graphe globale de toutes les réponses données de notre sondage.* 193
Figure 108. *Morphing de toutes les réponses.* 194
Figure 109. *Morphing des réponses des informaticiens.* 195

Figure 110. *Morphing des réponses des statisticiens.* _____ *196*
Figure 111. *Morphing des réponses des personnes travaillant dans le domaine de l'IE.* _____ *197*
Figure 112. *Morphing des réponses des autres personnes interrogées.* _____ *198*
Figure 113. *Synthèse de notre contribution et apports de la combinaison de fonctionnalités.* _____ *207*
Figure 115. *Graphe symétrique croisant des compagnies pharmaceutiques.* _____ *251*
Figure 116. *Graphe asymétrique orienté.* _____ *252*
Figure 117. *Graphe asymétrique des licenciés/ licenseurs.* _____ *253*
Figure 118. *Graphe asymétrique, orienté et issu du croisement de trois variables.* _____ *254*
Figure 119. *Graphe temporel orienté croisant des compagnies pharmaceutiques ainsi que leur pays d'appartenance* _____ *255*
Figure 120. *Figure précédente dont l'orientation des arcs à été masquée, ainsi que les icônes traduisant la temporalité.* _____ *255*
Figure 121. *Graphe symétrique des dépôts de brevet dans le domaine de l'aéronautique.* _____ *256*
Figure 122. *Graphe symétrique des inventeurs de brevet dans le domaine des véhicules hybrides.* _____ *257*

Liste des tableaux

Tableau 1. *Relations d'ordre d'informations géographiques.* 33
Tableau 2. *Génération de synonymie.* 33
Tableau 3. *Les différents types de variables.* 36
Tableau 4. *Matrice de contingence.* 38
Tableau 5. *Matrice de présence/ absence.* 39
Tableau 6. *Matrice de cooccurrence simple.* 39
Tableau 7. *Matrice symétrique de cooccurrence simple pondérée.* 40
Tableau 8. *Matrice asymétrique de cooccurrence multiple pondérée.* 41
Tableau 9. *Décomposition d'une matrice Auteurs X Auteurs en trois périodes homogènes.* 42
Tableau 10. *Type de matrice(s) possible(s) et domaine d'application (en italique) selon le type des variables croisées.* 47
Tableau 11. *Problèmes et complexités temporelles des graphes (Tamassia, 1997).* 56
Tableau 12. *Sémiologie graphique selon (Bertin, 1970).* 56
Tableau 13. *Critères choisis pour comparer les outils de visualisation de données temporelles, basés sur les trois formes du temps étudiées.* 77
Tableau 14. *Matrice de présence/ absence des critères de qualité des techniques de visualisation de données temporelles.* 78
Tableau 15. *Correspondance entre widgets et critères.* 79
Tableau 16. *Comparaison des travaux sur la représentation de l'espace temps et des données temporelles.* 90
Tableau 17. *Description des types de données identifiés dans la taxonomie de B.Schneiderman* 92
Tableau 18. *Comparatif des algorithmes FDP proposés.* 113
Tableau 19. *Synthèse des trois propositions de morphing* 167
Tableau 20. *Matrice à la base du morphing de graphe.* 175
Tableau 21. *Evaluation de VisuGraph* 182
Tableau 22. *Nombre de questions et réponses possibles.* 186
Tableau 23. *Caractéristiques de la population étudiée par catégorie.* 187
Tableau 24. *Résultats pour la durée de l'expérimentation.* 188
Tableau 25. *Questions d'ordre général.* 189
Tableau 26. *Questions sur les représentations de matrices symétriques.* 189
Tableau 27. *Questions sur les représentations de matrices symétriques temporelles.* 190
Tableau 28. *Questions sur les représentations de matrices asymétriques.* 190
Tableau 29. *Questions sur les représentations de matrices asymétriques temporelles.* 190
Tableau 30. *Matrices des réponses « Très Facile », croisant les questions et les catégories d'individus.* 191
Tableau 31. *Matrice des réponses « Facile », croisant les questions et les catégories d'individus.* 191
Tableau 32. *Matrice des réponses « Sans difficulté », croisant les questions et les catégories d'individus.* 191
Tableau 33. *Matrice des réponses « Difficile », croisant les questions et les catégories d'individus.* 192
Tableau 34. *Matrice des réponses « Très Difficile », croisant les questions et les catégories d'individus.* 192
Tableau 35. *Synthèse des résultats des morphing par catégorie de personnes interrogées, pour les fonctionnalités générales de VisuGraph.* 200
Tableau 36. *Synthèse des résultats des morphing par catégorie de personnes interrogées, pour les représentations des matrices symétriques de VisuGraph.* 200
Tableau 37. *Synthèse des résultats des morphing par catégorie de personnes interrogées,* 200
pour les représentations des matrices symétriques temporelles de VisuGraph. 200
Tableau 38. *Synthèse des résultats des morphing par catégorie de personnes interrogées,* 201
pour les représentations des matrices asymétriques de VisuGraph. 201
Tableau 39. *Synthèse des résultats des morphing par catégorie de personnes interrogées,* 201
pour les représentations des matrices asymétriques temporelles de VisuGraph. 201
Tableau 40. *Synthèse sur l'évaluation de VisuGraph.* 201

Cette thèse a donné lieu à la publication de 5 rapports et 20 articles dont :

1 article de revues internationales

GHALAMALLAH I., LOUBIER E., DOUSSET B., *Business intelligence_a proposal for a tool dedicated to the analysis relational. SciWatch Journal*, **hexalog**, Barcelona - Spain, Vol. 3, (en ligne), 2008.

6 articles de conférences et workshops internationaux

GAY B., LOUBIER E., *Dynamics and Evolution Patterns of Business Networks.* International Conference on Advances in Social Networks Analysis and Mining, Athens, IEEE Computer Society, juillet 2009 (à paraître).

GAY B., LOUBIER E., Dynamic Analysis of Complex Network Structures of Business Connections. ICC International Conference on Network Modelling and Economic Systems, Lisbonne, support électronique, 2008.

GHALAMALLAH I., LOUBIER E., DOUSSET B., Competitive Intelligence: Approaches and proposal of a tool specific to relational analysis. Colloque européen d'intelligence économique, Lisbonne, 27/03/2008-28/03/2008, support électronique, 2008.

GUENEC N., LOUBIER E., GHALAMALLAH I., DOUSSET B., Management and analysis of chinese database extracted knowledge. BCS IRSG Symposium: Future Directions in Information Access, Londres, 22/01/2008, British Computer Society, (support électronique), 2008.

LOUBIER E., BAHSOUN W., DOUSSET B., Visualization and analysis of large graphs. ACM International Workshop for Ph.D. Students in Information and Knowledge Management (ACM PIKM 2007), Lisbonne - Portugal, 06/11/2007-09/11/2007, ACM, (support électronique), 2007.

LOUBIER E., DOUSSET B., Visualisation and analysis of relationnal data by considering temporal dimension. International Conference on Enterprise Information Systems (ICEIS 2007), Funchal, Madeira - Portugal, 12/06/2007-16/06/2007, Vol. ISAS, INSTICC Press, p. 550-553, 2007.

12 articles de conférences et workshops nationaux

LOUBIER E., Proposition d'un algorithme de placements temporels des sommets d'un graphe évolutif. Colloque Veille Stratégique Scientifique et Technologique (VSST 2009), Nancy, 30/03/2009-31/03/2009, IRIT, support électronique, mars 2009.

LOUBIER E., VisuGraph : un outil pour l'analyse du relationnel. Colloque Veille Stratégique Scientifique et Technologique (VSST 2009), Nancy, 30/03/2009-31/03/2009, IRIT, support électronique, 2009.

LOUBIER E., DOUSSET B., BAHSOUN W., Interactive methods for graph exploration. Conférence internationale Systèmes d'Information d'Intelligence Economique (SIIE 2009), Hammamet, 12/02/2009-14/02/2009, 2009.

LOUBIER E., BAHSOUN W., DOUSSET B., VisuGraph : un outil pour la visualisation de données temporelles. MAnifestation des Jeunes Chercheurs STIC (MajecStic 2008), Marseille, 29/10/2008-31/10/2008, Aline Cauvin, Abbas Chamseddine, Nicolas Faessel, Sébastien Fournier (Eds.), Laboratoire des Sciences de l'Information et des Systèmes (LSIS), support électronique, 2008.

LOUBIER E., DOUSSET B., La prise en compte de la dimension temporelle dans la classification de données. Journées Francophones Extraction et Gestion de Connaissances (EGC 2008), Sophia Antipolis, 29/01/2008-01/02/2008, Vol. II, Cépaduès Editions, p. 559-564, 2008.

LOUBIER E., DOUSSET B., Temporal and relational data representation by graph morphing. Conférence internationale Systèmes d'Information d'Intelligence Economique (SIIE 2008), Hammamet, 14/02/2008-16/02/2008, 2008.

LOUBIER E., Analyse et visualisation de données relationnelles évolutives. Rencontres Inter-Associations (RIA'S 2007), Toulouse, 12/03/2007-13/03/2007, IRIT, (en ligne), 2007.

LOUBIER E., BAHSOUN W., DOUSSET B.,. La prise en compte de la dimension temporelle dans la visualisation de données par morphing de graphe. Colloque Veille Stratégique Scientifique et Technologique (VSST 2007), marrakech, 21/10/2007-25/10/2007, IRIT, (support électronique), 2007.

LOUBIER E., CARBONNEL S., VisuGraph : Un outil d'exploration de données relationnelles évolutives. Congrès Informatique des Organisations et Systèmes d'Information et de Décision (INFORSID 2007), Perros-Guirec, 22/05/2007-25/05/2007, Vol. 1, Hermès, p. 53-68, 2007.

LOUBIER E., CARBONNEL S., DOUSSET B., Influence du prétraitement textuel sur la représentation graphique dans un contexte d'analyse de données relationnelles. Colloque Veille Stratégique Scientifique et Technologique (VSST 2007), Marrakech, 21/10/2007-25/10/2007, IRIT, (support électronique), 2007.

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, Vol. 1, Association Francophone de Recherche d'Information et Applications (ARIA), p. 149-164, 2007.

LOUBIER E., BAHSOUN W., DOUSSET B., Visualisation de l'évolution des informations relationnelles par morphing de graphe. Journées Francophones Extraction et Gestion de Connaissances (EGC 2007), Namur, Belgique, 23/01/2007-26/01/2007, Cépaduès Editions, p. 43-54, 2007.

1 article de conférence sans actes publiés

LOUBIER E., DOUSSET B., Mieux comprendre les enjeux stratégiques liés à l'analyse relationnelle: le morphing de graphe. Journées IST, Université de Marne la Vallée, 23/06/2006-24/06/2006.

5 Rapports de contrat

LOUBIER E., DOUSSET B., Analyse et visualisation d'information relationnelle par morphing de graphe – Prise en compte de la dimension temporelle. Rapport de recherche, final, IRIT, 2008.

LOUBIER E., DOUSSET B., Analyse et visualisation d'information relationnelle par morphing de graphe – Prise en compte de la dimension temporelle. Rapport de contrat, 4, IRIT, 2008.

LOUBIER E., DOUSSET B., Analyse et visualisation d'information relationnelle par morphing de graphe – Prise en compte de la dimension temporelle. Rapport de contrat, 3, IRIT, 2007.

LOUBIER E., DOUSSET B., Analyse et visualisation d'information relationnelle par morphing de graphe – Prise en compte de la dimension temporelle.. Rapport de contrat, 2, IRIT, 2007.

LOUBIER E., DOUSSET B., Analyse et visualisation d'information relationnelle par morphing de graphe – Prise en compte de la dimension temporelle.. Rapport de contrat, 1, IRIT, 2006.

BIBLIOGRAPHIE

(Abello et al., 2001) **ABELLO J., KORN J.**, *Mgv: A system for visualizing massive multi-digraphs.* Transactions on Visualization and Computer Graphics, 2001.

(Agrawal, 1997) **AGRAWAL R., GUPTA A., SARAWAGI A.**, *Modeling Multidimensional Databases*, ICDE'97.

(Aiello et al., 2000) **AIELLO W., CHUNG F., LU L.**, *A random graph model for massive graphs.* Proceedings of the thirty-second annual ACM symposium on Theory of computing, pages 171–180. ACM Press. Cité page(s) 18, 20, 21, 30, 2000.

(Alexa et al., 2000) **ALEXA M, COHEN-OR D, LEVIN D.**, *As-rigid-as-possible polygon morphing.* Proceedings of SIGGRAPH 2000, New Orleans, LA, pages 157–164, 2000.

(Alpert et Kahng, 1995) **ALPERT C.J., KAHNG A.B.**, *Recent developments in netlist partitioning : A survey.* The VLSI journal, vol. 19, pages 1-18, 1995.

(Alvarez et al., 2005) **ALVAREZ-HAMELIN J.I., DALL'ASTA L., BARRAT A., VESPIGNANI A.**, *k-core decomposition: a tool for the visualization of large scale networks.* Cité page(s) 41, 52, 53, 54, 55, 2005.

(Amaral et al., 2000) **AMARAL L., SCALA A., BARTHELEMY A., STANLEY H.**, *Classes of small-world networks.* Proceedings of National Academy of Science, 97(21):11149–11152. Cité page(s) 17, 19, 84, 2000.

(Andrews, 2002) **ANDREWS K.**, *Information Visualisation.* http://www2.iicm.edu/ivis/ivis.pdf, 2002.

(Ankerst, 2000) **ANKERST M.**, *Visual Data Mining.* Thèse de doctorat en Informatique de l'Université Ludwig-Maximilians, München, 2000.

(Ankerst, 2001) **ANKERST M.**, *Visual Data Mining with pixel-oriented Visualization Techniques.* ACM SIGKDD Workshop on Visual Data Mining, San Francisco, CA, 2001.

(Ankerst et al, 2008) **ANKERST M, KAO A., TJOELKER R., WANG C.**, *DataJewel: Integrating Visualization with Temporal Data Mining.* Visual Data Mining, pages 312-330, 2008.

(Ansoff, 1975) **ANSOFF, H.I.**, Managing strategic Surprise by Response to Weak Signals. Management Review, V.XVIII, n°2, page(s) 21-33, California, 1975.

(Auber, 2001) **AUBER D.**, *Tulip.* Mutzel P., Jünger M., Leipert S., editors, 9th Symp. Graph Drawing, volume 2265 of Lecture Notes in Computer Science. Springer-Verlag, pages 335-337, 2001.

(Auray et al., 2001) **AURAY J., DURU G., LAMURE M., NICOLOYANNIS N.**, Extension du concept de métrique et structures topologiques associées. VIIIème Congrès de la Société Francophone de Classification (SFC'01), Pointe-à-Pitre, Guadeloupe, France, pages 14-18, 2001.

(Bachimont, 2000) **BACHIMONT B.**, Engagement sémantique et engagement ontologique: conception et réalisation d'ontologies en ingénierie des connaissances. Ingénierie des connaissances. Jean Charlet, INRIA, 2000.

(Bahsoun et BeyLagoun, 2006) **BAHSOUN W., BEYLAGOUN M.**, *Experimentation d'une classification contextuelle de documents.* Séminaire Veille Stratégique Scientifique et Technologique (Séminaire VSST 2006), ENIC Telecom Lille 1, 16/01/2006-17/01/2006, IRIT, support électronique, 2006.

(Balmisse, 2005) **BALMISSE G.**, *REFLEXIONS - Visualisation de l'information : quelques repères.* http://www.gillesbalmisse.com/blog/index.php?2005/02/24/35-visualisation-de-linformation-quelques-reperes, 2005.

(Barabasi et Albert, 1999) **BARABASI AL, ALBERT R.**, *Emergence of scaling in random networks.* Science, 286:509–512. Cité page(s) 8, 17, 20, 31, 37, 1999.

(Barabasi et al., 2002) BARABASI A.L., JEONG H., NÉDA Z., RAVASZ E., SCHUBERT A., VICSEK T., *Evolution of the social network of scientific collaborations.* Physica A 311, pages 590–614. Cité page(s) 18, 20, 21, 22, 2002.

(Batagelj et Mrvar, 1998) BATAGELJ V., MRVAR A., *PAJEK: Program for Broad Network Analysis.* Connections, pages 47-57, 1998.

(Batagelj et Zaversnik, 2002) BATAGELJ V., ZAVERSNIK M., *Generalized cores.* 2002.

(Bavelas, 1950) BAVELAS A., *Communication patterns in task-oriented groups.* D. Cartwright et A.Zander (Eds), Groupe Dynamics, Nex York: Row-Peterson, pages 493-506, 1950.

(Bellot, 2004) BELLOT P., *Classification de documents et enrichissement de requêtes.* Méthodes avancées pour les systèmes de recherche d'informations. Editions Hermes, Volume 2, 2004.

(BenAmmar et Dousset, 1999) 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, 1999.

(Bertin, 1967) BERTIN J., *La sémiologie graphique.* Paris, Gauthier-Villars, 1967.

(Bertin, 1970) BERTIN J., *Sémiologie graphique.* La Haye, Mouton, 1970.

(Bertin, 1977) BERTIN J., La Graphique et le Traitement Graphique de l'information. Flammarion, 1977.

(Besancon et al., 1999) BESANCON R., RAJMAN M., CHAPPELIER J.C., *Textual similarities based on a distributional approach.* International Workshop on Similarity Search (IWOSS99), Italy, 1999.

(Bonnet, 1989) BONNET C., *Traité de Psychologie Cognitive 1.* Chapitre 1- La perception visuelle des formes. Dunod, 1989.

(Borgatti, 2002) BORGATTI S.P., NetDraw Network Visualization Software. 2002.

(Boughanem et al., 2000) BOUGHANEM M., CHRISMENT C., MOTHE J., SOULÉ-DUPUY C., TAMINE L., *Connectionist and genetic approaches to perform IR.* F. Crestani and G. Pasi editors. Soft computing in information retrieval: techniques and applications. Physica Verlag, pages 173-198, Heidelberg, 2000.

(Boughanem et Dousset, 2001) BOUGHANEM M., DOUSSET B., *Relation entre le push adaptatif et l'optimisation des abonnements dans les centres de documentation.* Veille stratégique, scientifique et technologique : VSST'01, pp 239-252, Tome 1, (Barcelone, Espagne), 2001.

(Bouroche, 1989) BOUROCHE J.M., SAPORTA G., *L'analyse des données*, 1989.

(Brandes et Corman, 2003) BRANDES U., CORMAN S., *Visual unrolling of network evolution and the analysis of dynamic discourse.* InfoVis'02 Vol. 2, N°1, pages 40-50, 2003.

(Bridgeman et Tamassia, 2000) BRIDGEMAN S., TAMASSIA R., *Interactive Difference Metrics for Orthogonal Graph Drawing Algorithms.* Newspaper off Graph Algorithms and Applications.

(Bui et al., 2001) BUI A., ABERLE D, MCNITT-GRAY M., CARDENAS A., GOLDIN J., *Problem-oriented Prefetching for an Integrated Clinical Imaging Workstation.* Journal of the American Medical Informatics Association. Pages 242-253, 2001.

(Bulinge, 2002) BULINGE F., Pour une culture de l'information dans les petites et moyennes organisations : un modèle incrémental d'intelligence économique. Thèse de Doctorat, Université du Sud, Toulon, 2002.

(Burt, 1992) BURT R., *Structural Holes: The Social Structure of Competition.* Boston, MA: Harvard University Press, 1992. http://www.cs.brown.edu/publications/jgaa/, vol. 4, No 3, pages 47-74, 2000.

(Cailleteau et Plaisant, 1999) CAILLETEAU L., PLAISANT C., *Interfaces for Visualizing Multi-Valued Attributes: Design and Implementation Using Starfield Displays.* Rapport pour l'obtention du Diplôme d'Etudes Approfondies, 1999.

(Card et al., 1999) CARD S.K., MACKINLAY J.D., SHNEIDERMAN B., *Readings in Information Visualization.* Morgan Kaufmann, San Francisco, 1999.

(Carlis et Konston, 1998) CARLIS J.V., KONSTON J.A., *Interactive Visualization of Serial Periodic Data.* UIST'98, ACM, San Francisco, Ca, 1998.

(Castellani et al, 2008) CASTELLANI U., GAY-BELLILE V., BARTOLI A., *Robust deformation capture from temporal range data for surface rendering.* Journal of Visualization and Computer Animation (JVCA), volume 19, pages 591-603, 2008.

(Chen, 2004) CHEN C., *Searching for intellectual turning points : Progressive Knowledge Domain Vizualisation.* Proceedings of the National Academy of Sciences of the United States of America, 101(suppl. 1), pages 5303-5316, 2004.

(Chen et Carr, 1999) CHEN C., CARR L., *Visualizing the evolution of a subject domain: A case study.* Proceedings of IEEE Visualization '99, (San Francisco, CA, 1999), IEEE Computer Society, pages 449-452, 1999.

(Chen et al., 2009) CHEN M., EBERT D., HAGEN H., LARAMEE R., VAN LIERE R., MA K., RIBARSKY W., SCHEUERMANN G., SILVER D., *Data, Information, and Knowledge in Visualization.* IEEE Computer Graphics and Applications (CGA) volume 29, pages 12-19, 2009.

(Chevallet et Bruandet, 1997) CHEVALLET J. P., BRUANDET M. F., *Impact de l'utilisation de multi termes sur la qualité des réponses d'un système de recherche d'information à indexation automatique.* Organisation des connaissances en vue de leur intégration dans les systèmes de représentation et de recherche d'information Collection UL3 Lilles, ISBN 2-84467-002-4, pages 223–238, Lille, 1997.

(Chittaro, 2006) CHITTARO L., Visualization of patient data at different temporal granularities on mobile devices. AVI 2006, pages 484-487, 2006.

(Chrisment, 1997) CHRISMENT C., DKAKI T., DOUSSET B., MOTHE J., *Extraction et Synthèse de Connaissances à partir de Données Hétérogènes.* Ingénierie des Systèmes d'Information, Hermès Science Publications, Vol. 5, N. 3, pages 367-400, 1997.

(Chrisment et al., 2004) CHRISMENT C., DOUSSET B., KAROUACH S., MOTHE J., *Information mining : extracting, exploring and visualising geo-referenced information.* Workshop on Geographic Information Retrieval, Sheffield , NA, juillet, 2004.

(Chrisment et al., 2006) CHRISMENT C., HERNANDEZ N., GENOVA F., MOTHE J., *D'un thesaurus vers une ontologie de domaine pour l exploration d un corpus.* AMETIST, INIST, Vol. 0, pages 59-92, septembre 2006.

(Chrisment, 2007) CHRISMENT C. *Recherche en systèmes d'information.* Axes prioritaires. Ingénierie des Systèmes d'Information (ISI) 12(4):11-20, 2007.

(Chrisment et al., 2008) CHRISMENT C., HAEMMERLE O., HERNANDEZ N., MOTHE J., *Méthodologie de transformation d'un thesaurus en une ontologie de domaine.* Revue d'Intelligence Artificielle (RIA) 22(1):7-37, 2008.

(Church et Hanks, 1990) CHURCH W. K., HANKS P., *Word association norms, mutual information, and lexicography.* Computational Linguistics, volume 16, pages 22–29, 1990.

(Coleman et al, 1966) COLEMAN J.S., KATZ E., MENZEL H., *Medical innovation : a diffusion study.* New York, Bobbs-Merril, 1966.

(Coleman, 1973) COLEMAN, J. S., *Loss of Power.* American Sociological Review 38, pages 1-17, 1973.

(Compieta et al., 2007) COMPIETA P., DI MARTINO S., BERTOLOTTO M., FERRUCCI F., TAHAR KECHADI T., *Exploratory spatio-temporal data mining and visualization.* Journal of Visual Languages and Computing, Volume 18, 2007, pages 255-279, 2007.

(Cosmin Porumbel et al, 2007) COSMIN PORUMBEL D., HAO J., KUNTZ P., *A Study of Evaluation Functions for the Graph K-Coloring Problem.* Artificial Evolution 2007: 124-135, 2007.

(Cosmin Porumbel et al, 2009) COSMIN PORUMBEL D., HAO J., KUNTZ P., *Diversity Control and Multi-Parent Recombination for Evolutionary Graph Coloring Algorithms.* EvoCOP 2009: 121-132, 2009

(Cross et al., 20009) CROSS II J., HENDRIX D., BAROWSKI L., *Integrating Multiple Approaches for Interacting with Dynamic Data Structure Visualizations.* Electronic Notes in Theoretical Computer Science, Volume 224, pages 141-149, 2009.

(Daassi et al., 2000) DAASSI C., DUMAS M., FAUVET M.-C., NIGAYZ L., SCHOLL P.C., *Visual Exploration of Temporal Object Databases.* Proceeding of BDA 2000, Blois, France, pages 159-178, 2000.

(Daassi, 2003) DAASSI C., *Techniques d'interaction avec un espace de données temporelles.* Thèse de doctorat en Informatique, université Joseph Fourier, Grenoble I, 2003.

(Daniels, 1986) DANIELS J.P., *Cognitive Models in Information Retrieval – An Evaluation Review.* Journal of Documentation, vol 42, n° 4, pages 272-304, 1986.

(Davidson et Harel, 1996) DAVIDSON R., HAREL D., *Drawing Graphs Nicely Using Simulated Annealing.* ACM Transactions On Graphics, 15, pages 301–331, 1996.

(Degenne et Forsé, 2004) DEGENNE A., FORSE M., *Les réseaux sociaux.* Paris, Armand Colin, 2004.

(Di Battista et al., 1999) DI BATTISTA G., EADES P., TAMASSIA R., TOLLIS I.G., *Graph drawing - Algorithms for the visualization of graphs.* Prentice Hall, 1999.

(Diday, 1971) DIDAY E., *La méthode des nuées dynamiques.* Thèse de doctorat, Université de Paris VI, paris, 1971.

(Diday, 2008) DIDAY E., *Spatial classification.* Discrete Applied Mathematics (DAM) 156(8):1271-1294, 2008.

(Diday, 2008b) DIDAY E., *Principes d'Analyse des données symboliques et application à la détection d'anomalies sur des ouvrages publics.* EGC 2008, pages 211-212, 2008

(Diday, 2005) DIDAY E., *De la statistique des données à la statistique des connaissances : avancées récentes en Analyse des Données Symboliques.* EGC 2005, page 703, 2005

(Dkaki et al., 1991) DKAKI T., DOUSSET B., KOUSSOUBE S., *Les apports de la représentation de la quatrième dimension en analyse de données multidimensionnelles.* Journées d'études sur les systèmes d'informations élaborées: Bibliométrie - Informatique stratégique - Veille technologique. pp 98-105, 1991.

(Dkaki, 1993) DKAKI T., *Outils informatiques et méthodes automatiques pour la veille technologique.* Thèse de l'Université Paul Sabatier, Toulouse, 1993.

(Dkaki et Dousset, 1995) DKAKI T., DOUSSET B., *Tetralogie : A new method for Competitive Intelligence.* International Conference on Industrial Engineering and Management (IEPM'95) Marrakech, 1995.

(Dkaki et al., 1997) DKAKI T., DOUSSET B., MOTHE J., *Recherche de l'information stratégique dans les bases de données : veille scientifique et technique.* 15ème congrès INFORSID, INFORSID'97, pp 673-690, 1997.

(Dkaki et al., 2000) DKAKI T., DOUSSET B., EGRET D., MOTHE J., *Information discovery from semi-structured sources - Application to astronomical literature.* Computer Physics Communication, 2000.

(Dobrowolski, 1964) DOBROWOLSKI Z., *Etude sur la construction des systèmes de classification.* Préface d'Eric de Grolier, Paris : Gauthier-Villars, 1964.

(Dousset et Benjamaa, 1988) DOUSSET B., BENJAMAA T., *Trilogie logiciel d'analyse de données.* Journées d'études sur les systèmes d'informations élaborées: Bibliométrie - Informatique stratégique Veille technologique. Ile Rousse, 1988.

(Dousset et al., 1995) DOUSSET B., ROMMENS M., SIBUE D., *Application du logiciel de veille technologique Tétralogie aux huiles de poissons.* Symposium International, Omega-3, Lipoprotéines et athérosclérose, 1995.

(Dousset et kanoun, 1998) DOUSSET B., KANOUN S., Optimisation du choix de la terminologie pour la reformulation de requêtes: cas des multi-termes. VSST'98, pp 107-119, 1998.

(Dousset et Karouach, 2002) DOUSSET B., KAROUACH S., *Collaboration interactive entre classifications et cartes thématiques ou géographiques.* 9ièmes rencontres de la société francophone de classification, 2002.

(Dousset, 2003) DOUSSET B., *Intégration de méthodes interactives de découverte de connaissances pour la veille stratégique.* Habilitation à diriger des recherches, Université Paul Sabatier, Toulouse, 2003.

(Dousset et Karouach, 2005) DOUSSET B., KAROUACH S., *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, 2005.

(Dousset, 2006) DOUSSET B., *TETRALOGIE: a platform for scientific and technological survey. International Workshop on Webometrics, Infometrics and Scientometrics & Seventh COLLNET Meeting, Nancy, 10/05/2006-12/05/2006* (conférencier invité), LORIA, 2006.

(Dousset et Karouach, 2007) DOUSSET B., KAROUACH S., *Apports de la classification dans l'analyse des graphes de grande taille.* VSST 2007, CD-ROM, 2007.

(Dousset, 2009) DOUSSET B., *Extraction de l'information implicite par analyse textuelle de sites Internet en UNICODE,* VSST 2009, CD-ROM, 2009.

(Dragicevic et al., 2002) DRAGICEVIC P., HUOT S., *Spiraclock: A continuous and non-instrusive display for upcoming events.* In CHI 2002, pages 604, Minneapolis, Minnesota USA, 2002.

(Dreyfus et al., 2004) DREYFUS G., MARTINEZ J.M., SAMUELIDES M., GORDON M., BADRAN F., THIRIA S., HERAULT L., *Réseaux de neurones, méthodologie et applications.* Eyrolles, 2e édition, 2004.

(Dunning, 1993) DUNNING T., *Accurate Methods for the Statistics of surprise and Coincidence.* Computational Linguistics, vol. 19, pages 61-74, 1993.

(Durand, 1979) DURAND D., *La systémique.* Que sais-je ? Presses Universitaires de France, 1979.

(Eades, 1984) EADES P., *A Heuristic for Graph Drawing.* Congressus Numerantium, vol. 42, pages 149-160, 1984.

(Eades et al., 1991) EADES P., LAY W., MISUE K., SUGIYAMA K., *Mental Preserving the map of a diagram.* Proceedings of Compugraphics 91, pages 24-33, 1991.

(Eades et Lin, 2000) EADES P., LIN X., *Spring algorithms and symmetry.* Theoretical Computer Science. v.240 n.2, pages 379-405, 2000.

(Enright et al., 2002) ENRIGHT A.J., VAN DONGEN S., OUZOUNIS C.A., *An efficient algorithm for large-scale detection of protein families.* Nucleic Acids Research, vol. 30, pages 1575-1584, 2002.

(Erten et al., 2004) ERTEN C., HARDING P., KOBOUROV S., WAMPLER K., YEE G., *Exploring the computing literature using temporal graph visualization.* Conference on Visualization and Data Analysis, 2004.

(Eve, 2002) EVE M., *Deux traditions d'analyse des réseaux sociaux.* Réseaux, 20,115, pages185-212, 2002.

(Fagan, 1987) FAGAN J. L., Experiments in Automatic Phrase Indexing for Document Retrieval : a Comparison of Syntactic and Non-Syntactic Methods. Thèse de doctorat, Université de Cornell, New York, 1987.

(Farago et al., 2004) FARAGO F., GUILLAUD F., LEMOINE M., MORANA C., *Philosophie Tles L,ES,S.* Editeur Bréal. ISBN 2-7495-0276-4, page 124, 2004.

(Fayyad et al., 1996) FAYYAD U., PIATETSKY-SHAPIRO G., SMYTH P., *From Datamining to Knowledge Discovery.* Chapitre 1, 1996.

(Fayyad, 2002) FAYYAD U., GRINSTEIN G.G., WIERSE A., *Information visualization in data mining and knowledge discovery.* Morgan Kaufmann, 2002.

(Fekete et Lecolinet, 2006) FEKETE J.D., LECOLINET E., *Visualisation pour les bibliothèques numériques.* Volume 9-2006/2, PAGES 7-11, 2006.

(Floater et Gotsman, 1999) FLOATER MS, GOTSMAN C., *How to morph tilings injectively.* Comput Appl Math 101:117–129, 1999.

(Fong et al., 2006) QIYUE FONG, FOO MENG NG, ZHIYONG HUANG., *Spatio-temporal Visualization of Battlefield Entities and Events.* Computer Graphics International 2006, pages 622-629, 2006.

(Ford et Fulkerson, 1956) FORD L. R. JR., FULKERSON D. R., *Maximal Flow Through a Network*. Canadian Journal of Mathematics, 8, pages 399-404, 1956. http://wisl.ece.cornell.edu/ECE794/Jan29/ford_fulkerson/ff1956.pdf.

(Frank et al., 2001) FRANK A., RAPPER J., CHEYLAN J.-P., *Life and Motion of Socio-economic Units*. New York, Londres: Taylor and Francis, GISDATA Series n° 8, ISBN: 0-7484-0845-2, 2001.

(Freeman, 1979) FREEMAN L.C., *Centrality in social networks: conceptual clarification*. Social networks, vol.1, pages 215-239, 1979.

(Freeman et al., 1991) FREEMAN, L. C., BORGATTI, S.P., WHITE, D. R., *Centrality in Valued Graphs: A Measure of Betweennes Based on Network Flow*. Social Networks 13, 141– 154, 1991.

(Freeman, 1997) FREEMAN E., *The Lifestreams Software Architecture*. http://www.cs.yale.edu/homes/freeman/dissertation/etf.pdf, 1997.

(Frick et al., 1991) FRICK A., SANDER G., WANG K., *Simulating Graphs ace physical systems*. Dr. Dobb' S Jornal, pages 58-64, 1999.

(Friedkin, 1991) FRIEDKIN N. E., *Theoretical foundations for centrality measures*. American Journal of Sociology, 1991.

(Fruchterman et Reingold, 1991) FRUCHTERMAN TMJ., REINGOLD EM., *Graph drawing by force_directed placement*. Software – Practice and experience, 21, pages 1129-1164, 1991.

(Fung et McKeown, 1997) FUNG P., MCKEOWN K., *A technical word and term translation aid using noisy parallel corpora across language groups*. Machine Translation, volume 12, pages 53-87, 1997.

(Gao et revesz, 2006) GAO J., REVESZ P., *Visualization of Temporal-Oriented Datasets*. GMAI 2006, pages 57-62, 2006.

(Gay et Dousset, 2006) GAY B., DOUSSET B., *Cartographie de réseaux d'alliances et analyse stratégique*. Revue des sciences et technologies de l'information, série ingénierie des systèmes d'information (ISI), systèmes d'information stratégique, Hermes-Lavoisier, vol. 11, n° 2/2006, pages 37-51, 2006.

(Gay et Loubier, 2008) GAY B., LOUBIER E., *Dynamic Analysis of Complex Network Structures of Business Connections*. ICC International Conference on Network Modelling and Economic Systems, Lisbonne, en ligne, 2008.

(Gayte et al., 1997) GAYTE 0., LIBOUREL T., CHEYLAN JP., LARDON S., *Conception des systèmes d'information sur environnement*. Paris, Edition Hermès, Collection géomatique, 1997.

(Ghalamallah et al., 2007) GHALAMALLAH I., GRIMEH A., DOUSSET B., *Processing data stream by relational analysis. Data mining, statistique et analyse de* données, INRIA, Vol. 36, 2007.

(Ghalamallah et al., 2008) GHALAMALLAH I., LOUBIER E., DOUSSET B., *Business intelligence_a proposal for a tool dedicated to the relational analysis. SciWatch Journal*, **hexalog**, Barcelona - Spain, Vol. 3, (en ligne), 2008.

(Ghalamallah et al., 2008b) GHALAMALLAH I., LOUBIER E., DOUSSET B.,*Competitive Intelligence: Approaches and proposal of a tool specific to relational analysis*. Colloque européen d'intelligence économique, Lisbonne, 27/03/2008-28/03/2008, support électronique, 2008

(Goldstein et Gotsman, 1995) GOLDSTEIN E, GOTSMAN C., *Polygon morphing using a multi resolution representation*. Proceedings of Graphics Interface '95, 247-254, 1995.

(Goria, 2006) GORIA S., L'expression du problème dans la recherche d'informations : application à un contexte d'intermédiation territoriale. Thèse de Doctorat, Université de Nancy 2, Nancy, 2006.

(Graham et al., 2008) GRAHAM D., WILLIAMS J., CHRISTEN P., *ReDSOM: Relative Density Visualization of Temporal Changes in Cluster Structures Using Self-Organizing Maps*. ICDM 2008, pages 173-182, 2008.

(Granovetter, 1973) GRANOVETTER M. S., *The Strength of Weak Ties*. American Journal of Sociology, 78, pages 1360-1380, 1973.

(Grefenstette, 1992) **GREFENSTETTE G.**, *Use of suntactic context to produce term association lists for text retrieval.* Annuel International ACM SIGIR conference on research and development in information retrieval (SIGIR'92), ACM press, pages 89–97, Denmark, 1992.

(Gribaudi, 1998) **GRIBAUDI M.**, Espaces temporalités stratifications - Exercices sur les réseaux sociaux. Ed. EHESS, Paris, 1998.

(Grinstein, 1996) **GRINSTEIN G.**, *Harnessing the human in knowledge discovery.* Proceedings Of the 2nd International Conference on Knowledge Discovery and Data Mining, pages 384–385, AAAI Press, 1996.

(Grinstein et Ward, 2002) **GRINSTEIN G., WARD M.**, *Introduction to Data Visualization.* Fayyad, U., Grinstein, G., Wierse, A. (eds.). Information Visualization in Data Mining and Knowledge Discovery. Morgan Kaufmann Publishers, San Francisco, USA, 2002.

(Groves et al., 1990) **GROVES L., MICHALEWICZ Z., ELIA P., JANIKOW C.**, *Genetic algorithms for drawing directed graphs.* Proceedings of the International Fifth Symposium one Methodologies off Intelligent Systems, Knoxville, 25-27 October, pages 268-276, 1990.

(Gruber, 1993) **GRUBER T.**, *A translation Approach to portable ontology specifications.* Knowledge Acquisition, Vol. 5, pages 199-220, 1993.

(Gruber, 1993b) **GRUBER T.**, *Towards Principles for the Design of Ontologies Used for Knowledge Sharing.* International Workshop on formal Ontology in Conceptual Analysis and Knowledge Representation. Padova, LADSEB-CNR, 1993.

(Guarino, 1999) **GUARINO N., MASOLO C., VETERE G.**, *OntoSeek: Content-Based Access to the Web.* IEEE Intelligent Systems, pages 70-80, 1999.

(Guenec et al., 2008) **GUENEC N., LOUBIER E., GHALAMALLAH I., DOUSSET B.**, *Management and analysis of chinese database extracted knowledge.* BCS IRSG Symposium: Future Directions in Information Access, Londres, 22/01/2008, British Computer Society, support électronique, septembre 2008.

(Habert, 2000) **HABERT B.**, *Des corpus représentatifs : de quoi, pour quoi, comment ?* Presses Universitaires de Perpignan : Études et réflexions. Linguistique sur corpus. Edition M. Bilger, pages 11-58, Perpignan, 2000.

(Hachul et Jünger, 2005) **HACHUL S., JÜNGER MR.**, *Experimental An Comparison off Fast Algorithms for General Drawing Broad Graphs.* Proceedings of the 13th Symposium on Graph Drawing (GD' 05), pages 235-250, 2005.

(Hansen, 2001) **HANSEN H.**, *A Time-Series of Urban Growth in Copenhagen Area.* ScanGIS'2001, The 8th Scandinavian Research Conference on Geographical Information Science, Norway, pages 225-235, 2001.

(Harel et Koren, 2002) **HAREL D., KOREN Y.**, *A Fast Multi-Scale Algorithm for Drawing Broad Graphs.* Graph Algorithms and Applications, vol. 6, No 3, pages 179-202, 2002.

(Havre et al., 1999) **HAVRE S., HETZLER B., NOWELL L.**, *ThemeRiver : In Search of Trends, Patterns, and Relationships.* IEEE Symposium on Information Visualization, InfoVis '99, San Francisco CA, 1999.

(Havre et al., 2000) **HAVRE S., HETZLER B., NOWELL L.**, *ThemeRiver: Visualizing Theme changes over Time.* Proceedings of the IEEE Symposium on Information Visualization, 2000.

(Hearst, 1992) **HEARST M. A.**, *Automatic acquisition of hyponyms from large text corpora.* Proceedings of the Fourteenth International Conference on Computational Linguistics, France, Nantes, 1992.

(Heer et al., 2009) **HEER J., KONG N., AGRAWALA M.**, *Sizing the horizon: the effects of chart size and layering on the graphical perception of time series visualizations.* CHI 2009, pages 1303-1312, 2009.

(Henry, 1992) **HENRY T. R.**, *Interactive Graph Layout: The Exploration of Large Graphs.* Departme ntof Computer Science. Tucson, University of Arizona, 1992.

(Herman et al., 2000) **HERMAN I., MARSHALL M. S., MELANÇON G.**, *Graph Visualisation and Navigation in Information Visualisation: A Survey.* IEEE Transactions on Visualization and Computer Graphics 6(1), pages 24-43, 2000.

(Hewagamage et al., 1998) HEWAGAMAGE, K., HIRAKAWA M., ICHIKAWA T., *Interactive Visualization of Spatiotemporal Patterns Using Spirals on a Geographical Map*. Proceedings of the IEEE Symposium on Visual Languages, 1998.

(Himsolt, 1995) HIMSOLT MR., *Comparing and Evaluating Layout Algorithms within GraphEd*. Newspaper of Visual Languages and Computing, 1995.

(Hinum et al., 2005) HINUM K., MIKSCH S., AIGNER W., OHMANN S., POPOW C., POHL M., RESTER M., *Gravi : Interactive Information Visualization to Explore Highly Structured Temporal Data*. J. UCS (JUCS), volume 11, pages 1792-1805, 2005.

(Hochheiser, 2002a) HOCHHEISER H., *Interactive Querying of Time Series Data*. Proceedings of CHI 2002, Minneapolis, Minnesota, USA, 2002.

(Hochheiser, 2002b) HOCHHEISER H., SHNEIDERMAN B., *A dynamic Query Interface for Finding Patterns in Time Series Data*. Proceedings CHI 2002, Minneapolis, Minnesota, USA, 2002.

(Hopfielf, 1982) HOPFIELD J.J., *Neural networks and physical systems with emergent collective computational abilities*. Proceedings of the National Academy of Sciences, pages 2554-2558, 1982.

(Huang et al., 2005) HUANG X., EADES P., LAI W., *A Framework of Filtering, Clustering and Dynamic Layout Graphs for Visualization*. ACSC 2005, pages 87-96, 2005.

(Hubert et al., 2001) HUBERT G., MOTHE J., BENAMAR A., DOUSSET B. , DKAKI T., KAROUACH S., *Textual document Mining using graphical interface. International Human Computer Interaction*. HCI International 2001 , New Orleans (USA). Lawence Erlbaum Associates - Publishers , Mahwah - New Jersey , pages 918-922 (volume 1), 2001.

(Hudon, 1994) HUDON M., *Le thésaurus : Conception, élaboration, gestion*. Montréal, ASTED, 1994.

(Hussein et al., 2004) HUSSEIN S., SALLES S., DOUSSET B., *Les besoins des PME en Intelligence Economique : définition de profils types*. 4ièmes journées VSST, (Toulouse, France), 2004.

(Huyn et al., 2007) HUYN X., GUILLET F., BLANCHARD J., KUNTZ P., GRAS R., BRIAND H. *A graph-based clustering approach to evaluate interestingness measures: a tool and a comparative study*. Quality measures in Data Mining, Chapitre 2, Springer Verlag, 2007.

(Jakobiak, 1995) JAKOBIAK F., *L'information scientifique et technique*. Que sais-je, n°3015, 1995.

(Jakobiac, 2004) JAKOBIAC F., *L'intelligence économique, la comprendre, l'implanter, l'utiliser*. Editions d'Organisation, Paris, page 335, 2004.

(Jeong et al., 2000) JEONG H., TOMBOR B., ALBERT R., OLTVAI Z.N., BARABASI A.L., *The large-scale organization of metabolic networks*. Nature, pages 631:651, Cité page(s) 18, 2000.

(Jog et Schneiderman, 1995) JOG N., SHNEIDERMAN B., *Starfield Information Visualization with Interactive Smooth Zooming*. Proceedings of Visual Databases Systems, pages 1-10, Lausanne, Suisse, 1995.

(Johnson et Wilson, 2002) JOHNSON I., WILSON A., *The TimeMap Project: Developing Time-Based GIS Display for Cultural Data*. Journal of GIS in Archaeology Vol 1. ESRI Inc., Redlands. http://www.archaeology.usyd.edu.au/, 2002.

(Johnson et al., 2006) JOHNSON C., MOORHEAD R., MUNZNER T., PFISTER H., RHEINGANS P., YOO T. S. NIH-NSF, *Visualization Research Challenges Report*. IEEE Computer Society, 2006.

(Jolibois et al., 2000) 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.

(Josselin et Fabrikant, 2003) JOSSELIN D., FABRIKANT S., *Cartographie animée et interactive*. Revue internationale de géomatique, vol. 13, n° 1, 2003.

(Jouve et al., 2001) JOUVE B., KUNTZ P., VELIN F., *Extraction de structures macroscopiques dans des grands graphes par une approche spectrale.* ECA, Hermès Science publication édition, vol. 1, pages 173-184, 2001.

(Kamada et Kawai, 1989) KAMADA T., KAWAI S., *An Algorithm for General Drawing Undirected Graphs.* Information Processing Letters, 31, pages 7-15, 1989.

(Kanoun, 1998) KANOUN S., Reformulation de requêtes : choix de la terminologie pertinente et recherche de corrélations entre termes. Rapport de Stage de DEA, Université Paul Sabatier, Toulouse, 1998.

(Kapusova, 2004) KAPUSOVA D., *Visualisation de l'information dans le portail STAF18.* http://tecfaseed.unige.ch/staf18/modules/ePBLjolan/uploads/proj15/paper%20(et%20dispositif)6.xml, 2004.

(Karouach, 2003) KAROUACH S., *Visualisations interactives pour la découverte de connaissances : concepts, méthodes et outils.* Thèse de Doctorat en informatique, Université Paul Sabatier, France, 2003.

(Karouach et Dousset, 2003) KAROUACH S., DOUSSET B., *Les graphes comme représentation synthétique et naturelle de l'information relationnelle de grandes tailles.* Workshop sur la recherche d'information, associé à INFORSID'2003, Nancy, 03/06/2003-06/06/2003, INFORSID, pages 35-48, juin, 2003.

(Karypis et Kumar, 1998) KARYPIS G., KUMAR V., *Multilevel k-way partitioning scheme for irregular graphs.* Journal of Parrallel and distributed Computing, vol. 48, pages 96-129, 1998.

(Katz, 1953) KATZ L., *A new status index derived from sociometric analysis*. Psychometrika, *18*, pages 39-43, 1953.

(Kaur Padda et al., 2009) KAUR PADDA H., SEFFAH A., MUDUR S., *Investigating the Comprehension Support for Effective Visualization Tools - A Case Study.* ACHI 2009, pages 283-288, 2009.

(Keim et al., 1994) KEIM D.A., KRIEGEL H.P., *Using visualization to support data mining of large existing databases.* Lecture Notes in Computer Science, pages 210 -229, 1994.

(Keim et Kriegel, 1994) KEIM D. A., KRIEGEL H.P., *Using visualization to support data mining of large existing databases.* Lecture Notes in Computer Science, 871, pages 210 -229, 1994.

(Keim, 2002) KEIM DA., *Information Visualization and visual datamining.* IEEE Transactions on Visualization and Computer Graphics.
http://fusion.cs.uni-magdeburg.de/pubs/TVCG02.pdf, 2002.

(Keim et al., 2005) KEIM D, MANSMANN F., SCHRECK T., *Analyzing Electronic Mail Using Temporal, Spatial, and Content-based Visualization Techniques.* GI Jahrestagung 2005, pages 434-438, 2005.

(Keogh, 2005) KEOGH E., *Visualization and Mining of Temporal Data.* IEEE Visualization 2005, page 126, 2005.

(Kleinberg, 1999) KLEINBERG J. M., *Authoritative sources in a hyperlinked environment.* Journal of the ACM, pages 604–632, 1999.

(Knocke, 1990) KNOKE D., *Political networks : the structural perspective*. Cambridge, Cambridge University Press, 1990.

(Koenig , 1996) KOENIG G., *Management stratégique: paradoxes, interactions et apprentissage*. Paris, Ed; Nathan, page 544, 1996.

(Koren et al., 2003) KOREN Y., CARMEL L., AND HAREL D., *Drawing Huge Graphs by Algebraic Multigrid Optimization.* Multiscale Modeling and Simulation, vol. 1, No 4, pages 645-673, 2003.

(Koussoube et al., 1992) KOUSSOUBE S., DKAKI T., DOUSSET B., *Outils et méthodologie d'étude de la cohérence et de la complétude dans les bases de connaissances.* Journées Francophones de la Validation et de la Vérification des Systèmes à Base de connaissances. (Dourdan) pp 17-24, 1992.

(Kraaij et R. Pohlmann, 1998) KRAAIJ W., POHLMANN R., *Comparing the effect of syntactic vs. statistical phrase indexing strategies for dutch.* Proceedings of Second European Conference on Research and Advanced Technology for Digital Libraries ECDL'98. Editeurs: Christos Nicolaou and Constantine Stephanidis, pages 605–614, 1998.

(Krackhardt et al., 1994) KRACKHARDT D., BLYTHE J., MCGRATH C., *KrackPlot 3.0: Year Improved Network Drawing Program.* Connections 17, pages 53-55, 1994.

(Kramer et Jozsa, 1998) KRAMER T., JOZSA J., *Visualization and analysis of timedependent hydrometric data in windows environment.* Proceedings of the 3rd International Conference on Hydroinformatics, Copenhagen, Danemarque, A.A. Balkema, 1998.

(Kullberg, 1995) KULLBERG R.L., *Dynamic Timelines: Visualizing Historical Information in Three Dimensions.* Master of science in media arts and sciences at the Massachusetts institute of technology, Massachusetts, 1995.

(Kuntz et Henaux, 2000) KUNTZ P., HENAUX F., *Numerical comparaison of two spectral decomposition for vertex clustering.* Data Analysis, Classification and Related Methods, Proceeding Of IFCS'2000, Springer Verlag, pages 581-586, 2000.

(Kuntz, 2003) KUNTZ P., Découverte de règles d'association et de structures dans des réseaux de relations par des approches non supervisées automatiques et interactives. Habilitation à diriger des recherches, Université de Nantes, 2003.

(Kuntz et al., 2006) KUNTZ P., PINAUD B., LEHN R., *Minimizing crossings in hierarchical digraphs with a hybridized genetic algorithm.* Journal of Heuristics, vol. 2, n°1-2, pages 23-36, 2006.

(L'Affaire Louis Trio , 1993) L'AFFAIRE LOUIS TRIO., *Mobilis In Mobile.* Paroles et musique: C. Boris, Label: universal, ASIN : B000007WZ2, 1993

(Lafon, 1984) LAFON P., *Dépouillements et Statistiques en Lexicométrie.* Genève-Paris : Slatkine-Champion, 1984.

(Lamure, 1987) LAMURE M. Espaces abstraits et reconnaissance des formes : application au traitement des images digitales. Thèse en informatique théorique, Université de Lyon I, 1987.

(Langran, 1993) LANGRAN G., *Time in Geographic Information Systems.* Londres: Taylor & Francis, ISBN: 0-7484-0059-1, 1993.

(Lazéga, 1998) LAZEGA E., *Réseaux sociaux et structures relationnelles.* Paris, PUF, 1998.

(Le Coadic, 2004) LE COADIC Y.F., *La science de l'information.Paris : PUF.* Que sais je ?, n°2873, 2004.

(Lesca, 1986) LESCA H., Système d'information pour le management stratégique de l'entreprise. Ed. Mac Graw Hill, page 146, 1986.

(Lesca, 1997) LESCA H., *Veille stratégique, concepts et démarche de mise en place dans l'entreprise.* Ministère de l'Education Nationale, de la Recherche et de la Technologie, ADBS, page 27, http://membres.lycos.fr/jeanlucmoya/Veille%20Lesca.doc, 1997.

(Leung et Apperley, 1994) LEUNG, Y. K. APPERLEY, M. D., *A review and taxonomy of distortion-oriented presentation techniques.* ACM Transactions on Computer.Human Interactions. 1, pages 126-160. DOI= http://doi.acm.org/10.1145/180171.180173, 1994.

(Lévy-Strauss, 1964) LEVY-STRAUSS C., *Le cru et le cuit.* Editions Plon, Paris, 1964.

(Leavitt, 1951) LEAVITT H.J., *Some effects of certain communication pattern on group performance.* Journal of abnormal and social psychology, 46, pages 38-50, 1951.

(LeMoigne, 1984) LE MOIGNE J.L., *La théorie du système général.* Paris, Presses Universitaires de France, 1984.

(Lopez et al., 2001) LOPEZ N., KREUSELER M., SCHUMANN H., *A scalable framework for information visualization.* Transactions on Visualization and Computer Graphics, 2001.

(Loubier et Dousset, 2006) LOUBIER E., DOUSSET B., *Mieux comprendre les enjeux stratégiques liés à l'analyse relationnelle: le morphing de graphe.* Journées IST, Université de Marne la Vallée, 2006.

(Loubier et Dousset, 2006b) LOUBIER E., DOUSSET B., *Analyse et visualisation d'information relationnelle par morphing de graphe – Prise en compte de la dimension temporelle.* Rapport de contrat 1, IRIT, 2006.

(Loubier, 2007) **LOUBIER E.**, *Analyse et visualisation de données relationnelles évolutives.* Rencontres Inter-Associations (RIA'S 2007), Toulouse, 12/03/2007-13/03/2007, IRIT, (en ligne), mars 2007. URL : http://www.irit.fr/RIA07/intervenants.html, 2007.

(Loubier, 2007b) **LOUBIER E.**, *Visualization and analysis of large graphs.* Conference on Information and Knowledge Management (CIKM 2007), Lisbonne - Portugal, 06/11/2007-09/11/2007, ACM, (support électronique), 2007.

(Loubier et al., 2007) **LOUBIER E., BAHSOUN W., DOUSSET B.**, *La prise en compte de la dimension temporelle dans la visualisation de données par morphing de graphe.* Colloque Veille Stratégique Scientifique et Technologique (VSST 2007), Marrakech, IRIT, (support électronique), 2007.

(Loubier et al., 2007b) **LOUBIER E., BAHSOUN W., DOUSSET B.**, *Visualisation de l'évolution des informations relationnelles par morphing de graphe.* Journées Francophones Extraction et Gestion de Connaissances (EGC 2007), Cépaduès Editions, pages 43-54, 2007.

(Loubier et al., 2007c) **LOUBIER E., BAHSOUN W., DOUSSET B.**, *Visualization and analysis of large graphs.* ACM International Workshop for Ph.D. Students in Information and Knowledge Management (ACM PIKM 2007), ACM, support électronique, 2007.

(Loubier et Bahsoun, 2007) **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, Vol. 1, Association Francophone de Recherche d'Information et Applications (ARIA), pages 149-164, 2007.

(Loubier et Carbonnel, 2007) **LOUBIER E., CARBONNEL S.**, *VisuGraph : Un outil d'exploration de données relationnelles évolutives.* Congrès Informatique des Organisations et Systèmes d'Information et de Décision (INFORSID 2007), Perros-Guirec, 22/05/2007-25/05/2007, Vol. 1, Hermès, pages 53-68, 2007.

(Loubier et Carbonnel, 2007b) **LOUBIER E., CARBONNEL S.**, *Influence du prétraitement textuel sur la représentation graphique dans un contexte d'analyse de données relationnelles.* Colloque Veille Stratégique Scientifique et Technologique (VSST 2007), Marrakech, IRIT, (support électronique), 2007.

(Loubier et Dousset, 2007) **LOUBIER E., DOUSSET B.**, *Analyse et visualisation d'information relationnelle par morphing de graphe – Prise en compte de la dimension temporelle.* Rapport de contrat 3, IRIT, 2007.

(Loubier et Dousset, 2007b) **LOUBIER E., DOUSSET B.**, *Visualisation and analysis of relational data by considering temporal dimension.* International Conference on Enterprise Information Systems (ICEIS 2007), Funchal, Madeira - Portugal, 12/06/2007-16/06/2007, Vol. ISAS, INSTICC Press, pages 550-553, 2007.

(Loubier et Dousset, 2007c) **LOUBIER E., DOUSSET B.**, *Analyse et visualisation d'information relationnelle par morphing de graphe – Prise en compte de la dimension temporelle.* Rapport de contrat 2, IRIT, 2007.

(Loubier et al., 2008) **LOUBIER E., DOUSSET B., BAHSOUN W.**, *VisuGraph : un outil pour la visualisation de données temporelles.* MAnifestation des Jeunes Chercheurs STIC (MajecStic 2008), Aline Cauvin, Abbas Chamseddine, Nicolas Faessel, Sébastien Fournier (Eds.), Laboratoire des Sciences de l'Information et des Systèmes (LSIS), support électronique, 2008.

(Loubier et Dousset, 2008) **LOUBIER E., DOUSSET B.**, *Analyse et visualisation d'information relationnelle par morphing de graphe – Prise en compte de la dimension temporelle.* Rapport de recherche, final, IRIT, 2008.

(Loubier et Dousset, 2008b) **LOUBIER E., DOUSSET B.**, *Analyse et visualisation d'information relationnelle par morphing de graphe – Prise en compte de la dimension temporelle.* Rapport de contrat 4, IRIT, 2008.

(Loubier et Dousset, 2008c) **LOUBIER E., DOUSSET B.**, *La prise en compte de la dimension temporelle dans la classification de données.* Journées Francophones Extraction et Gestion de Connaissances (EGC 2008), Sophia Antipolis, 2008.

(Loubier et Dousset, 2008d) LOUBIER E., DOUSSET B., *Temporal and relational data representation by graph morphing.* Safety and Reliability for managing Risk (ESREL 2008), Hammamet, 2008.

(Loubier, 2009) LOUBIER E., *Proposition d'un algorithme de placements temporels des sommets d'un graphe évolutif.* Colloque Veille Stratégique Scientifique et Technologique (VSST 2009), IRIT, support électronique, 2009.

(Loubier, 2009b) Eloïse Loubier., *VisuGraph : un outil pour l'analyse du relationnel.* Colloque Veille Stratégique Scientifique et Technologique (VSST 2009), IRIT, support électronique, 2009.

(Loubier et al., 2009) LOUBIER E., DOUSSET B., BAHSOUN W., *Interactive methods for graph exploration.* Conférence internationale Systèmes d'Information d'Intelligence Economique (SIIE 2009), Hammamet,CD-ROM, 2009.

(Loubier et Gay, 2009) LOUBIER E., GAY B., *Dynamics and Evolution Patterns of Business Networks.* International Conference on Advances in Social Networks Analysis and Mining. Athens, IEEE Computer Society, 2009 (à paraître).

(Lovins, 1968) LOVINS J. B., *Development of a Stemming Algorithm.* Mechanical Translation and computational Linguistics, pages 22.31, 1968.

(Lyman et Varian, 2003) LYMAN P., VARIAN H. R., *How Much Information?* http://www.sims.berkeley.edu/how-much-info-2003, 2003.

(MacEachren et al., 1998) MACEACHREN A.M., BOSCO F.P, HAUG D., PICKLE L.W., *Geographic Visualization: Designing Manipulable Maps for Exploring Temporally Varying Georeferenced Statistics.* Proceedings of the IEEE Symposium on Information Visualization InfoVis, Research Triangle Park, USA, IEEE Computer Society, pages 87-94, 1998.

(Mackinlay et al., 1991) MACKINLAY J.D., ROBERTSON G., CARD K., *The Perspective Wall: Detail and Context Smoothly Integrated.* Proceedings of the CHI `91, ACM Press, pages 173- 179, 1991.

(Mackinlay et al., 1994) MACKINLAY J.D., ROBERTSON G.G., *DeLine R.. Developing Calendar Visualizers for the Information Visualizer.* Proceedings of UIST '94, 1994.

(Malhotra, 2000) MALHOTRA Y., *Knowledge Management & New Organization Forms: A Framework for Business Model Innovation.* Information Resources Management Journal, volume 13, pages 5-14, 2000.

(Maniez et Grolier, 1991) MANIEZ J., DE GROLIER E., *A decade of research in classification.* 1991.

(Manning & Schütze, 1999) MANNING C.D., SCHÜTZE H., *Foundations of Statistical Natural Language Processing.* The MIT Press, Massachusetts, 1999.

(Maron et Kuhns, 1960) MARON M., KUHNS J., *On relevance, probabilistic indexing and information retrieval.* Journal of the Association for Computing Machinery, pages 216–244, 1960.

(Martinet, 1984) MARTINET A. C., *Management stratégique : organisation et politique.* McGraw-Hill, Paris, 1984.

(Martinet et Ribault, 1989) MARTINET B., RIBAULT J.M., *La veille technologique, concurrentielle et commerciale.* Paris : les Editions d'organisation, 1989.

(Martinez et guillaume, 1998) MARTINEZ J., GUILLAUME S. Colour image retrieval fitted to classical querying. Networking and Information Systems Journal (NISJ), 1(2-3):251–278, 1998.

(Martinez et Loisant, 2002) MARTINEZ J., LOISANT E. Browsing image databases with Galois' lattices. In Proceedings of the ACM International Symposium on Applied Computing (SAC'02), pages 971–975, Madrid, Spain, March 11-14. ACM Computer Press, 2002.

(Martinez et Marchand, 1998) MARTINEZ J., MARCHAND S. Towards intelligent retrieval in image databases. In B. BERRA,réd., Proceedings of the 5th IEEE International Workshop onMulti-Media Data BaseManagement Systems (MMDBMS'98), pages 38–45, Dayton, Ohio, August 5-7. IEEE Computer Press, 1998

(Martinez et mouaddib, 1999) MARTINEZ J., MOUADDIB N. Multimedia and databases: A survey. Networking and Information Systems Journal (NISJ), 2(1):89–123, Hermès Science, 1999.

(Martre, 1994) MARTRE H., *Intelligence Economique et Stratégie des entreprises.* Travaux du groupe dirigé par Henri Martre pour le Commissariat Général du Plan. La documentaion Française, 1994.

(Marwah et al., 2009) MARWAH M., SHARMA R., SHIH R., PATEL C., BHATIA V., MEKANAPURATH M., VELUMANI R., VELAYUDHAN S., *Data analysis, visualization and knowledge discovery in sustainable data centers.* Bangalore Compute Conference 2009, Page 2, 2009.

(McCulloch et pitts, 1943) MCCULLOCH, W.S., ET PITTS, W., *A logical calculus of the ideas imminent in nervous activity.* Bulletin of Mathematical Biophysics, pages 115–133, 1943.

(Melancon et al., 1999) MELANÇON G., HERMAN I., DELAST M., *Indices visuels et métriques combinatoires pour la visualisation de données hiérarchique.* Proceedings of the IHM'99 Workshop, Montpellier, pages 166-173, 1999.

(Meliker et al., 2005) MELIKER J., SLOTNICK M., AVRUSKIN G., KAUFMANN A., JACQUEZ J., NRIAGU J., *Improving exposure assessment in environmental epidemiology: Application of spatio-temporal visualization tools.* Journal of Geographical Systems (JGS), volume 7, pages 49-66, 2005).

(Mendelzon et Vaisman, 2000) MENDELZON A.O., VAISMAN A.A., *Temporal Queries in OLAP.* 26th international conference on Very Large Data Bases. VLDB 20000, Egypt, 2000.

(Mercklé, 2004) MERCKLE P., *Sociologie des réseaux sociaux.* La Découverte, Paris, 2004.

(Milgram, 1967) MILGRAM S., *The small world problem.* Psychology Today, pages 60–67. Cité page(s) 18, 21, 1967.

(Minsky et Papert, 1988) MINSKY M., PAPERT S., *Perceptrons : an introduction to computational geometry.* MIT Press, expanded edition, 1988.

(Mitra, 1997) MITRA M., BUCKLEY C., SINGHAL A., CARDI C., *An analysis of statistical and syntactic phrases.* Proceedings of RIAO'97 computer-Assisted Information Searching on Internet, McGill University, pages 200–214, Montreal, 1997.

(Morin, 1999) MORIN E., *Extraction de lien sémantique entre termes à partir de corpus de textes techniques.* Thèse de doctorat, Institut de recherche en informatique de Nantes, 1999.

(Morris et al., 2003) MORRIS S.A., ASNAKE B., YEN G., *Optimal dendrogram seriation using simulated annealing.* Information Visualization, vol. 2, pages 95-104, 2003.

(Mothe, 1994) MOTHE J., Modèle connexionniste pour la recherche d'informations – Expansion dirigée de requêtes et apprentissage. Thèse de doctorat, Université Paul Sabatier, 1994.

(Mothe, 2000) MOTHE J., *Recherche et exploration d'informations -Découverte de connaissances pour l'accès à l'information.* Habilitation à diriger des recherches, Université Paul Sabatier, Toulouse, 2000.

(Mothe, 2006) MOTHE J., CHRISMENT C., DKAKI T., DOUSSET B., KAROUACH S., *Combining mining and visualization tools to discover the geographic structure of a domain.* Computers, Environment and Urban Systems, Elsevier, Numéro spécial : Geographic Information Retrieval, Vol. Hors-série N. 4, pages 460-484, 2006.

(Munzer, 2003) MUNZNER T., *Penser par la vision.* Horizon0 : art et culture numériques au Canada, vol.6, 2003.

(Myers, 2000) MYERS D.G., *Psychology(6th edition).* Worth Publishing, 2000.

(Neches, 1991) NECHES R., FIKES R., FININ T., GRUBER T., PATIL R., SENATOR T., SWARTOUT W. R., *Enabling Technology for Knowledge Sharing.* AI Magazine. Pages 36-56, 1991.

(Newell et Simon, 1972) NEWELL A., SIMON H.A., *Human Problem Solving.* Prentice Hall, 1972.

(Newman et al., 1998) NEWMAN M.E.J, STROGATZ. S. H., WATTS D., *Random graphs with arbitrary degree distribution and their applications.* In Physical Review, volume 4, page 131. Cité page(s) 17, 18, 20, 21, 22, 1998.

(Newman, 2001) NEWMAN M.E.J., *The structure of scientific collaboration networks.* In National Academy of Science of the United States of America, pages 404-409, 2001.

(Nielsen, 1990) NIELSEN J., *Hypertext and hypermedia*. Academic Press, 1990.

(Nieuwbeerta et Flap, 2000) NIEUWBEERTA P., FLAP H., Crosscutting social circles and political choice : Effects of personal network composition on voting behavior in The Netherlands. Social Networks 22, pages 313–335, 2000.

(Nowell et al., 2001) NOWELL L., HAVRE S., HETZLER B., WHITNEY P., *Themeriver: Visualizing thematic changes in large document collections*. Transactions on Visualization and Computer Graphics, 2001.

(Paice, 1984) PAICE C., *Soft evaluation of boolean search queries in information retrieval systems*. Information Technology : Research and Development, pages 33–42, 1984.

(Paice, 1996) PAICE C., *Method for evaluation of stemming algorithms based on error counting*. Journal of the American Society for Information Science, pages 632.349, 1996.

(Paque, 2007) PAQUE D., *Gestion de l'historicité et méthodes de mise à jour dans les SIG*. Cybergeo, Cartographie, Imagerie, SIG, article 278, 2007.

(Paternostre et al., 2002) PATERNOSTRE M., FRANCQ P., SAERENS M., LAMORAL J., WARTEL D., *Carry, un algorithme de désuffixation pour le français*. URL : http://www.galilei.ulb.ac.be, 2002.

(Péry-Woodley, 1995) PERY-WOODLEY M.P., *Quels corpus pour quels traitements automatiques ?* Traitement Automatique des Langues, pages 213-232, 1995.

(Piaget, 1970) PIAGET J., *Epistémologie des sciences de l'homme*. Ed. Galimard, Coll. Idée, 1970.

(Pinaud et Kuntz, 2004) PINAUD B., KUNTZ P., *Un guide sur la Toile pour sélectionner un logiciel de tracé de graphes*. Congrès VSST'2004 : veille stratégique scientifique & technologique : Systèmes d'information élaborée, bibliométrie, linguistique, intelligence économique , pages 546 ; 540 , 2004.

(Pitrat, 1990) PITRAT J., *Métaconnaissance*. Futur de l'intelligence artificielle, Editions Hermès, Paris, 1990.

(Plaisant et al., 1996) PLAISANT C., MILASH B., ROSE A., WIDOFF S., SHNEIDERMAN B., *LifeLines: Visualizing Personal Histories*. Proceedings of CHI'96, pages, Vancouver, Canada, ACM Press, pages 221-227, 1996.

(Plaisant et al., 1998) PLAISANT C., MUSHLIN R., SNYDER A, LI J., HELLER D., SCHNEIDERMAN B., *LifeLines: Using Visualization to Enhance Navigation and Analysis of Patient Records Revised version in 1998*. American Medical Informatic Association Annual Fall Symposium, Orlando. http://hcil.cs.umd.edu/trs/98-08/98-08.html, 1998.

(Polanco, 2002) POLANCO X., *La notion de visualisation de l'information et le modèle de référence*. Actes du colloque Cartographie de l'information, 2002.

(Porter, 1980) PORTER M. F., *An algorithm for suffix stripping*. Program, pages 130–137, 1980.

(Porter, 1982) PORTER M., Choix stratégiques et concurrence, Économica, 1982.

(Poulet et Kuntz, 2006) POULET F., KUNTZ P.,*Visualisation en extraction de connaissances*. Revue des Nouvelles Technologies de l'Information, Numéro spécial, Cepaduès Edition, 2006.

(Purchase, 2000) PURCHASE H., Effective information visualization: a study of graph drawing aesthetics and algorithms. Interacting with computers, pages 127-145, 2000.

(Quillian, 1968) QUILLIAN R., *Semantic memory*. Semantic information processing. Pages 227-270, 1968.

(Rainsford et Roddick, 2000) RAINSFORD C.P., RODDICK J.F., *Visualisation of Temporal Association Rules*. Proceedings of 2nd International Conference on Intelligent Data Engineering and Automated Learning IDEAL, Springer, Hong Kong, Chine, pages 91-96, 2000.

(Redner, 1998) REDNER S., *How popular is your paper? An empirical study of citation distribution*. In European Physical Journal, volume 4, pages 131–134, Cité page(s) 18, 20, 21, 1998.

(Renieris et Reiss, 1999) RENIERIS M., REISS S.P., *ALMOST: Exploring Program Traces*. Workshop on New Paradigms in Information Visualization and Manipulation NPIVM, Kansas City, USA, ACM Press, pages 70-77, 1999.

(Robertson, 1977) ROBERTSON S.E., *The probability ranking principle in IR.* Journal of Documentation. pages 294-304, 1977.

(Robertson et Mackinlay, 1993) ROBERTSON G.G., MACKINLAY J.D., *The Document Lens.* Proceedings of. ACM Symposium on User Interface Software and Technology (UIST'93), pages 101–108, 1993.

(Rocchio, 1971) ROCCHIO J., *Relevance feedback in information retrieval.* Prentice Hall Inc., Englewood Cliffs, NJ, 1971.

(Rousseau-Hans, 1998) ROUSSEAU-HANS F., *L'analyse de corpus d'information comme support de la veille stratégique.* Document numérique. Volume 2, page 189, 1998.

(Roux, 1998) ROUX C., DOUSSET B., Une méthode de détection des signaux faibles: application à l'émergence des Dendrimères. VSST'98, pp 349-357, 1998.

(Roy, 1959) ROY B., *Transitivité et connexité.* CRAS 249, pages 216-218, 1959.

(Salton, 1971) SALTON G., *A comparison between manual and automatic indexing methods.* Journal of American Documentation. Pages 61–71, 1971.

(Salton et McGill, 1984) SALTON G., MCGILL M., *Introduction to modern information retrieval.* McGraw-Hill Int. Book Co, 1984.

(Sarkar & Brown, 1992) SARKAR M., BROWN M.H., *Graphical fisheye views of graphs.* Proceedings of CHI'92 ACM, New York, pages 83-91, 1992.

(Sauvagnat, 2005) SAUVAGNAT K., Modèle *flexible pour la recherche d'information dans des corpus de documents semi-structurés.* Thèse de doctorat, Université Paul Sabatier, Toulouse, 2005.

(Schindler et Scheuerell, 2002) SCHINDLER D.E., SCHEUERELL M.D., *Habitat coupling in lake ecosystems.* OIKOS pages 177-189, 2002.

(Sederberg et al, 1993) SEDERBERG T., GAO P., WANG G., MU H., *2-D shape blending: an intrinsic solution to the vertex path problem.* Proceedings of the 20th annual conference on Computer graphics and interactive techniques, pages 15-18, 1993.

(Shahar et Cheng, 1999) SHAHAR Y., CHENG C., *Intelligent Visualization and Exploration of Time-Oriented Clinical Data.* Topics in Health Information Management. Aspen Publishers, Vol. 20, No. 2, pages 15-31, 1999.

(Shahar et cheng, 2000) SHAHAR Y., CHENG C., *Model-Based Visualization of Temporal Abstractions.* Computational Intelligence. Blackwell Publishing, Vol. 16, No. 2, pages 279-306, 2000.

(Shannon et Weaver, 1975) SHANNON C., WEAVER W., *La théorie mathématique de la communication.* Paris, Retz-CEPL, 1975.

(Shapira et Rappoport, 1995) SHAPIRA M., RAPPOPORT A., *Shape Blending Using the Star-Skeleton Representation.* IEEE Computer Graphics and Applications, v.15 n.2, pages44-50, 1995.

(Shneiderman, 1996) SHNEIDERMAN B., *The eyes have it: A task by data type taxonomy for information visualizations.* Proceedings of the IEEE Symposium on Visual Languages,pages. IEEE Computer Society Press, pages 336–343, 1996.

(Simoni, 2000) SIMONI J. L., Accès à l'information à l'aide d'un graphe de termes construit automatiquement (Intégration de l'interrogation et de la navigation). Thèse de doctorat, Université Paris 7, 2000.

(Singh et al., 2006) SINGH M., BASU A., MANDAL M., *Temporal Alignment of Time Varying MRI Datasets for High Resolution Medical Visualization.* ISVC 2006, pages 222-231, 2006.

(Sowa, 1984) SOWA J.F., *Conceptual Structures: Information.* Processing in Mind and Machine. Addison-Wesley Publishing Company, USA, 1984.

(Spence, 2000) SPENCE R., *Information Visualization.* ACM Press. 2000.

(Surazhsky et Gotsman, 2003) SURAZHSKY V., GOTSMAN C., *Intrinsic morphing of compatible triangulations.* International Journal of Shape Modeling, 9(2):191--201, 2003.

(Thalmann et Thalmann, 1994) THALMANN N., THALMANN D., *Computer Animation : a Key Issue for Time Visualization.* Scientific Visualization, Academic Press, 1994, pages 201-222, 1994.

(Tamassia et al., 1988) TAMASSIA R., DIBATTISTA G., BATINI C., *Automatic graph drawing and readability of diagrams.* IEEE Transactions on Systems, Man and Cybernetics, pages 61-79, 1988.

(Tamassia, 1997) TAMASSIA R., *Graph Drawing.* CRC Handbook of Discrete and Computational Geometry, 1997.

(Tamine et al., 2007) TAMINE L., ZEMIRLI N., BAHSOUN W., *Approche statistique pour la définition du profil d'un utilisateur de système de recherche d'information.* Information - Interaction - Intelligence, Cépaduès Editions, Vol. 7, N. 1, (en ligne), 2007.

(Tang et al., 2001) TANG D., STOLTE C., HANRAHAN P., *Polaris: A system for query, analysis and visualization of multi-dimensional relational Databases.* Transactions on Visualization and Computer Graphics, 2001.

(Tominski et al., 2003) TOMINSKI C., ABELLO J., SCHUMANN H., *Interactive Poster: Axes-Based Visualizations for Time Series Data.* Poster Compendium of InfoVis'03, IEEE, (2003), 68-69.

(Toussaint, 2004) TOUSSAINT Y., *Extraction de connaissances à partir de textes structurés.* Document numérique, vol. 8, no 3, pages 11–34, 2004.

(Tufte, 1983) TUFTE E., *The visual display of quantitative information.* Graphic Press. Cheshire, page 198, Connecticut, 1983.

(Tufte, 1990) TUFTE E., *Envisioning Information.* Graphics Press, 1990.

(Tufte, 1997) TUFTE E., *Visual Explanations.* Graphics Press, 1997.

(Tutte, 1963) TUTTE W.J., *How to Draw has graph.* Proceeding in London Maths, Plowshare, Series 3,13, pages 743-768, 1963.

(Valente, 1995) VALENTE T. W., *Networks models of the diffusion of innovations.* Cresskill, Hampton Press, 1995.

(Van Dongen, 2000) VAN DONGEN S., *Graph Clustering by Flow Simulation.* Thèse de doctorat, Université d'Utrecht, Allemagne, 2000.

(Véronis, 2003) VERONIS J., *Cartographie lexicale pour la recherche d'information.* Actes de TALN 2003, pages 265-274, 2003.

(Véronis, 2004) VERONIS J., *Hyperlex : lexical cartography for information retrieval.* Computer, Speech and Language. Volume 18/3, pages 223-252, 2004.

(Vincke, 1992) VINCKE P., *Multicriteria Decision-aid.* J. Wiley and Sons, 1992.

(Walshaw, 2003) WALSHAW C., *A Multilevel Algorithm for Force-Directed Graph-Drawing.* J. Graph Algorithms Appl. 7, pages 253-285, 2003.

(Ware, 2000) WARE C., Information Visualization, perception for design. MorganKaufmann, 2000.

(Warshall, 1962) WARSHALL S., *A theorem on Boolean matrices.* Journal of the ACM, January, 9, pages 11–12, 1962.

(Wasserman et Faust, 1994) WASSERMAN S., FAUST K., *Social Network Analysis, Methods and Applications.* Cambridge, Mass., Cambridge University Press, 1994.

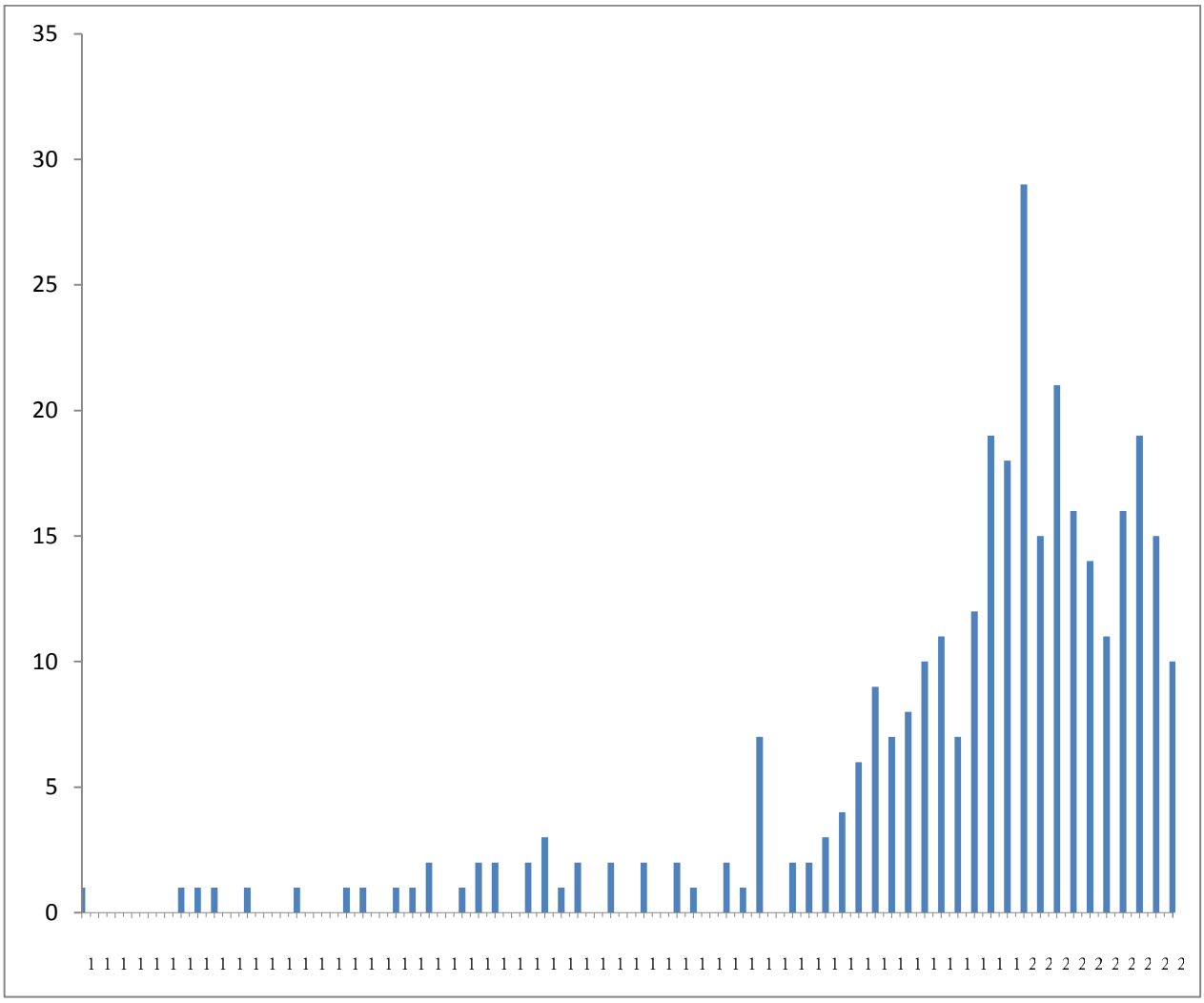
(Watts et Strogatz, 1998) WATTS D. J., STROGATZ S. H., *Collective dynamics of small-world networks.* Nature, pages 440–442. Cité page(s) 8, 14, 19, 21, 22, 28, 34, 35, 81, 84, 131, 1998.

(Weber, 2001) WEBER M., ALEXA M., MÜLLER W., *Visualizing time-series on spirals.* IEEE Symposium on Information Visualization (InfoVis 2001), IEEE Computer Society Pres, pages 7– 14, 2001.

(Wijk et Selow, 1999) WIJK J., SELOW E., *Cluster and Calendar based Visualization of Time Series Data.* IEEE Symposium on Information Visualization (InfoVis'99), 1999.

(Wu et Zhou, 2003) WU H. & ZHOU M., *Synonymous collocation extraction using translation information.* Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics. Editions Hinrichs E. & Roth D., pages 120-127, 2003.

(Zipf, 1949) ZIPF G., *Human Behaviour and the Principle of Least-Effort.* Cambridge : Addison-Wesley, 1949.



1940	0	1950	1	1960	1	1970	2	1980	1	1990	6	2000	29
1941	0	1951	1	1961	0	1971	3	1981	0	1991	9	2001	15
1942	0	1952	0	1962	1	1972	1	1982	2	1992	7	2002	21
1943	1	1953	1	1963	1	1973	2	1983	1	1993	8	2003	16
1944	0	1954	0	1964	2	1974	0	1984	7	1994	10	2004	14
1945	0	1955	0	1965	0	1975	2	1985	0	1995	11	2005	11
1946	0	1956	1	1966	1	1976	0	1986	2	1996	7	2006	16
1947	0	1957	0	1967	2	1977	2	1987	2	1997	12	2007	19
1948	0	1958	0	1968	2	1978	0	1988	3	1998	19	2008	15
1949	1	1959	1	1969	0	1979	2	1989	4	1999	18	2009	10

Nombre de publications citées par années.

GLOSSAIRE

✓ ***Veille***

Acquisition de l'information (collecte) Phase du cycle de veille pendant laquelle les textes contenant les informations pertinentes sont recueillis et conservés.

Axe de surveillance *"Description ou caractérisation de thèmes d'information"*

Exemple : dans une veille concurrentielle, la présence de la société X dans les linéaires des grandes surfaces.

Benchmarking Etalonnage de l'entreprise par rapport à une autre entreprise (concurrente ou d'un autre secteur) considérée comme particulièrement performante sur une fonction (service clientèle, coûts industriels, ….). Le processus peut être continu et alors être intégré au système de veille.

Cartographie (ou Mapping) Représentation visuelle des résultats de recherche ou des rapports de veille. Ils se présentent généralement sous forme de cartes qui indiquent les relations directes et indirectes identifiées au cours du cycle de veille.

Champ de veille Ensemble homogène des données faisant l'objet d'une veille ; ex : dans une veille concurrentielle : les sociétés concurrentes directes ou les produits de substitution.

Information blanche *"Information aisément et licitement accessible"* (médias, manifestations commerciales, ...)

Information formelle Donnée qui a été rédigée ou diffusée sur un support (papier, multi média, son, image, …)

Information grise *"Information licitement accessible, mais caractérisée par des difficultés dans la connaissance de son existence ou de son accès"* (contenu de banques de données, interviews d'experts, …)

Information informelle Donnée recueillie auprès d'une source orale ou n'ayant pas été explicitement mise en forme pour publication.

Information noire *"Information à diffusion restreinte et dont l'accès ou l'usage est explicitement protégé*", nécessite une autorisation pour être obtenue légalement.

Intelligence économique *"Ensemble des actions coordonnées de recherche, traitement et de distribution en vue de son exploitation, de l'information utile aux acteurs économiques. Ces divers actions sont menées légalement avec toutes les garanties de protection nécessaires à la préservation du patrimoine de l'entreprise, dans les meilleurs conditions de qualité, de délais et de coût". "La notion d'intelligene économique implique le dépassement des actions désignées par les vocables de documentation, de veille, …"* selon (Martre, 1994).

Knowledge Management Management des connaissances Système mis en place dans l'entreprise pour identifier, collecter, traiter, stocker, diffuser les informations formelles détenues par les membres de l'entreprise ; également transformer de l'information informelle en information formelle.

Networking Mode d'organisation fondé sur le déploiement et l'utilisation de réseaux

Norme AFNOR XP X-50 053	Description de la méthodologie globale à conduire dans la mise en place d'un cycle de veille. L'Agence Française de NORmalisation des recommandations. Elle décrit les phases principales du cycle de veille.
Système de veille	*"Ensemble structuré réunissant les compétences répondant à des besoins de veille"* Il s'agit donc à la fois de la cellule de veille et des outils.
Signal	Elément d'information indiquant aux veilleurs une évolution de tendances dans le champ de veille. Le signal peut être qualifié de faible (germe annonciateur) ou fort (confirmation d'une tendance repérée - synthèse des positions et menaces d'un acteur)
Signal critique	*"Information critique qui génère le déclenchement d'une analyse stratégique"*
Source	Emetteur de l'information diffusée qui doit être mentionnée lorsqu'elle est sortie de son contexte.
Sourcing (ou sélection de sources)	Phase du cycle de veille de recherche et de sélection des sources. Elles doivent être surveillées et mises à jour régulièrement.
Veille	*"Activité continue et en grande partie itérative visant à une surveillance active de l'environnement* [de l'entreprise ou de l'organisation] *pour en anticiper les évolutions"* Il s'agit donc d'un processus récurrent de recherche et collecte de l'information dont les données sont traitées selon une finalité propre au destinataire, dans une démarche d'intelligence économique La veille peut concerner toutes le fonctions de l'entreprise . La démarche globale de veille est généralement appelée veille stratégique
Veille stratégique	Regroupement de l'ensemble des activités de veille, consistant en un processus anticipatif d'observation et d'analyse de l'environnement, suivi de la diffusion ciblée des informations utiles à la prise de décisions. La veille stratégique concerne les décisions qui engagent le devenir, l'évolution de l'entreprise ou de la collectivité en relation avec les changements de son environnement socio-économique.
Veille commerciale	Clients, fournisseurs, circuits de distribution
Veille concurrentielle	Entreprises concurrentes ou susceptibles de le devenir (nouveaux entrants)
Veille environnementale	Cadre sociopolitique dont doit tenir compte l'entreprise (dont veille réglementaire, veille sociétale, veille groupes de pression, …)
Veille technologique	Produits, procédés de fabrication, évolutions technologiques et scientifiques (brevets).
Veille sectorielle	Offre, demande, conjoncture etc. sur un secteur d'activité
Veille sociale	Monde du travail : syndicats, législation sociale, état des relations humaines dans l'entreprise.
Veille sociétale	Etat de la société, comportements sociaux (notions de catégories socio-professionnelles, de modes de vie, de sociologie….).

✓ ***Théorie des graphes***

Adjacent — Deux sommets sont adjacents s'ils sont connectés par une arête.

Arête — Une arête est un lien non orienté entre deux sommets.

Arc — Un arc est un lien orienté entre deux sommets.

Biparti — Un graphe est biparti si ses sommets peuvent être répartis en deux sous-ensembles disjoints U et V tels que chaque lien relie un sommet de U à un sommet de V. Un graphe biparti est complet si tous les sommets de U sont connectés à tous les sommets de V.

Boucle — Une boucle est une arête reliant un sommet à lui-même.

Chaîne — Une chaîne est une suite finie de sommets dont deux sommets successifs sont reliés par une arête. La chaîne est dite simple si elle n'utilise pas deux fois la même arête.

Chemin — Un chemin est une suite finie de sommets dont deux sommets successifs sont reliés par un arc.

Circuit — Dans un graphe orienté, on appelle circuit un chemin dont l'origine et l'extrémité sont identiques. Si le chemin est élémentaire, c'est-à-dire ne passe pas deux fois par un même sommet, on parle de circuit élémentaire.

Classe connexe — Dans un graphe non orienté, une classe simplement connexe est une classe d'équivalence pour la relation qui lie deux sommets par une chaîne.

Dans le cas d'un graphe orienté on parle de classe fortement connexe. La relation d'équivalence devient alors la relation qui lie deux sommets par un circuit.

Cycle — Un cycle est une chaîne simple dont les deux extrémités coïncident.

Degré — Le degré (ou la valence) d'un sommet est le nombre d'arêtes ayant une extrémité en ce sommet. Dans un graphe orienté, le degré est décomposé en demi-degré intérieur et demi-degré extérieur, (dont la somme est le degré du sommet du graphe non orienté correspondant).

Ensemble d'articulation — C'est un ensemble de sommets tels que leur suppression augmente le nombre de composantes simplement connexes du graphe initial. Les graphes connexes résultants sont appelés pièces relativement à l'ensemble d'articulation considéré.

Graphe simple — De manière intuitive, un graphe non orienté est un ensemble fini de points appelés sommets (ou nœuds) connectés par des liens appelés arêtes. De manière plus formelle, un graphe simple est constitué d'un ensemble de sommets V et d'un ensemble de couples non ordonnés d'éléments distincts de V appelés arêtes. Tous les graphes ne sont pas simples. Deux sommets peuvent être reliés par plusieurs arêtes créant ainsi un multigraphe. Un sommet peut également être relié à lui même par une arête appelée boucle, conduisant à un pseudographe. Enfin, les liens peuvent être orientés définissant ainsi un graphe orienté.

Graphe complet — Un graphe complet avec n sommets est un graphe pour lequel chaque sommet est relié à tous les autres. Il existe un lien (arête ou arc) entre chaque paire de sommets.

Graphe orienté — Un graphe orienté est un graphe composé de sommets dont les liens sont matérialisés par des arcs notés *(a, b)*. Le sommet *a* est le sommet origine de l'arc et *b* est le sommet extrémité.

Longueur — La longueur d'une chaîne (ou d'un chemin) est égale au nombre d'arêtes (ou d'arcs) la constituant.

Nœud — Un synonyme de sommet.

Graphe planaire — Un graphe est planaire s'il peut être dessiné sur un plan sans que les arêtes ou les arcs (cas orienté) ne se croisent.

Point d'articulation — Un point d'articulation est un sommet dont la suppression augmente le nombre de composantes connexes. C'est un ensemble d'articulation réduit à un seul sommet.

✓ ***Informatique***

Algorithmique — Étude de la résolution de problèmes par la mise en oeuvre de suites d'opérations élémentaires selon un processus défini aboutissant à une solution.

Base de données — Ensemble d'informations regroupées sous la forme d'enregistrements stockés sur un système de fichiers logique (fichiers) ou physique structurés et organisées de manière à pouvoir être facilement manipulées.

Base de connaissances — Une base de connaissance regroupe des connaissances spécifiques à un domaine spécialisé donné, sous une forme exploitable par un ordinateur. Elle peut contenir des règles (dans ce cas, on parle de base de règles), des faits ou d'autre représentations. Si elle contient des règles, un moteur d'inférence - simulant les raisonnements déductifs logiques - peut être utilisé pour déduire de nouveaux faits.

Donnée — Représentation d'une information sous une forme conventionnelle destinée à faciliter son traitement.

Génie logiciel — Ensemble des méthodes, des techniques et des outils concourant à la production d'un logiciel, au-delà de la seule activité de programmation.

Icône — Sur un écran, symbole graphique qui représente une fonction ou une application logicielle particulière que l'on peut sélectionner et activer à partir d'un dispositif tel qu'une souris.

Information — Élément de connaissance susceptible d'être représenté à l'aide de conventions pour être conservé, traité ou communiqué.

Interactivité — L'interactivité est une activité nécessitant la coopération de plusieurs êtres ou systèmes, naturels ou artificiels qui agissent en ajustant leur comportement. L'interactivité est souvent associée aux technologies permettant des échanges homme-machine (voir interface homme-machine). Toutefois elle est présente dans toutes les formes de communication et d'échange où la conduite et le déroulement de la situation sont liées à des processus de rétroaction, de collaboration, de coopération entre les acteurs qui produisent ainsi un contenu, réalisent un objectif, ou plus simplement modifient et adaptent leur comportement.

Interface Jonction entre deux matériels ou logiciels leur permettant d'échanger des informations par l'adoption de règles communes, physiques ou logiques.

Langage formel Langage qui utilise un ensemble de termes et de règles syntaxiques pour permettre de communiquer sans aucune ambiguïté (par opposition à langage naturel).

Morphing de graphe Transformation progressive d'un graphe en un autre par un traitement informatique.

Programmation par objets Mode de programmation dans lequel les données et les procédures qui les manipulent sont regroupées en entités appelées objets.

Requête Expression formalisée d'une demande.

Réseau local Ensemble connexe, à caractère privatif, de moyens de communication établi sur un site restreint pourvu de règles de gestion du trafic et permettant des échanges internes d'informations de toute nature, notamment sous forme de données, sons, images, etc. *Note :* Le *réseau local*, ainsi défini en informatique, ne doit pas être confondu avec la notion de réseau local de raccordement, utilisée dans les télécommunications.

Tableur Logiciel de création et de manipulations interactives de tableaux numériques.

Traitement automatique des données Ensemble des opérations réalisées par des moyens automatiques, relatif à la collecte, l'enregistrement, l'élaboration, la modification, la conservation, la destruction, l'édition de données et, d'une façon générale, leur exploitation.

✓ ***Statistique***

ACP Analyse en Composantes Principales. Méthode d'analyse s'appliquant aux variables numériques.

ACP réduite Lorsque les données sont de même nature la métrique euclidienne est utilisable, mais elle va privilégier les caractères à forte dispersion. Dans tous les autres cas, il convient de normaliser les données en les divisant par leur écart type. Cette technique offre de nombreux avantages, car elle permet de s'affranchir de la notion d'unité et elle conduit à un rééquilibrage entre les caractères analysés. Dans le cas des tableaux de contingence, cette normalisation conduit à traiter tous les caractères sur un même pied d'égalité, même si certains d'entre eux sont très peu présents.

ACM Analyse des Correspondances Multiples. Méthode d'analyse des correspondances s'appliquant aux variables qualitatives et à l'étude de tableaux disjonctifs complets. Tableaux binaires dont les lignes sont des individus ou observations et les colonnes la juxtaposition des modalités de réponse à des questions (les modalités de réponse à une question s'excluant mutuellement).

AF Analyse Factorielle. Famille de méthodes statistiques d'analyse multidimensionnelle constituée des méthodes d'analyse factorielle (ACP, AFC, ACM, …), s'appliquant à des tableaux de nombres, qui visent à extraire des "facteurs" résumant approximativement par

quelques séries de nombres l'ensemble des informations contenues dans le tableau de départ.

AFC Analyse Factorielle des Correspondances. Conçu principalement pour traiter des tableaux de fréquence. Peut être appliqué à d'autres types de tableaux.

AFCM Analyse Factorielle des Correspondances Multiples. On considère p variables qualitatives *(p >= 3)* notées *{Xj ; j=1, …, p}*, possédant respectivement cj modalités, avec c $=\sum_{j=1}^{p} c_j$. On suppose que ces variables sont observées sur les mêmes n individus, chacun affecté du poids *1/n*. Soit *X = [Xp] … [Xp]* le tableau disjonctif complet des observations (*X* est *n X c*) et *B = X'X* le tableau de Burt correspondant (*B* est carré d'ordre c, symétrique).On appelle Analyse Factorielle des Correspondances Multiples (AFCM) des variables (X^1,…, X^p) relativement à l'échantillon considéré, l'AFC réalisée soit sur la matrice *X* soit sur la matrice *B*.

Caractère Signe typographique utilisé pour l'encodage du texte sur un support lisible par l'ordinateur.

Caractères délimiteurs / non-délimiteurs Distinction opérée sur l'ensemble des caractères, qui entrent dans la composition du texte permettant aux procédures informatisées de segmenter le texte en occurrences* (suite de caractères non-délimiteurs bornée à ses extrémités par des caractères délimiteurs). On distingue parmi les caractères délimiteurs:

- les caractères délimiteurs d'occurrence (encore appelés "délimiteurs de forme") qui sont en général : le blanc, les signes de ponctuation usuels, les signes de préanalyse éventuellement contenus dans le texte.
- les caractères délimiteurs de séquence : sous-ensemble des délimiteurs d'occurrence correspondant, en général, aux ponctuations faibles et fortes contenues dans la police des caractères.
- les caractères séparateurs de phrase : (sous-ensemble des délimiteurs de séquence) qui correspondent, en général, aux seules ponctuations fortes.

Classification par analyse de connexité Dans ce cas, la matrice individus/variables ou la matrice de contingence croisant deux types de caractères sur une population donnée, est considérée comme celle d'un graphe non orienté dont on va rechercher les composantes simplement connexe. Comme la simple connexité est une relation d'équivalence, nous obtenons une partition de l'unité en différentes classes indépendantes les unes des autres. Si le graphe issu de la matrice est simplement connexe (une seule classe), il est alors possible de partiellement le déconnecter en enlevant les liens faibles (ici les arêtes de valeur 1 ou 2). Le tri fait alors apparaitre des classes faiblement liées que l'on peut considérer comme représentatives de la structure des données analysées.

Classification ascendante hiérarchique Les méthodes hiérarchiques produisent des partitions en classes imbriquées de plus en plus grandes, le nombre de classes n'est pas connu à priori, plusieurs partitions imbriquées peuvent être proposées.

La classification ascendante hiérarchique (C.A.H.) consiste à regrouper les individus en classes en fonction de deux critères:

- les individus d'une même classe sont le plus semblable possible,
- les classes sont les plus disjointes possibles.

Pour cela, nous avons besoin d'une mesure globale de la proximité des individus à l'intérieur de chaque classe et de la distance interclasse pour apprécier la qualité de la partition obtenue. Comme il n'est pas envisageable d'évaluer toutes les partitions pour ne garder que la meilleure au sens du critère choisi (problème trop fortement combinatoire), il est donc exclu de trouver cette meilleure partition

Cooccurrence — Présence simultanée, mais non forcément contiguë, dans un fragment de texte (séquence, phrase, paragraphe, voisinage d'une occurrence, partie du corpus, …) des occurrences de deux formes données.

Corpus — Ensemble limité des éléments (énoncés) sur lesquels se base l'étude d'un phénomène linguistique. En lexicométrie, il s'agit de l'ensemble de textes réunis à des fins de comparaison, servant de base à une étude quantitative.

Délimiteurs de séquence — Sous-ensemble des caractères délimiteurs de forme correspondant aux ponctuations faibles et fortes (en général – le point, le point d'interrogation, le point d'exclamation, la virgule, le point virgule, les deux points, les guillemets, les tirets et les parenthèses).

Dendogramme — Représentation graphique d'un arbre de classification hiérarchique, mettant en évidence l'inclusion progressive des classes.

Facteur — Variables artificielles construites par les techniques d'analyse factorielle permettant de résumer (de décrire brièvement) les variables actives initiales.

Forme ou "**forme graphique**" — L'archétype correspondant aux occurrences identiques dans un corpus de textes, c'est-à-dire aux occurrences composées strictement des mêmes caractères non-délimiteurs d'occurrence.

Fréquence (d'une unité textuelle) — Le nombre de ses occurrences dans le corpus.

Fréquence relative — La fréquence d'une unité textuelle dans le corpus ou dans l'une de ses parties, rapportée à la taille du corpus (resp. de cette partie).

Apax — Chose dite une seule fois. Forme dont la fréquence est égale à un dans le corpus (apax du corpus) ou dans une de ses parties (apax de la partie).

Identification — Reconnaissance d'un seul et même élément à travers ses multiples emplois dans des contextes et dans des situations différentes.

Index — Liste imprimée constituée à partir d'une réorganisation des formes et des occurrences d'un texte, ayant pour base la forme graphique et permettant de regrouper les références relatives à l'ensemble des occurrences d'une même forme.

Lemmatisation — Regroupement sous une forme canonique (en général à partir d'un dictionnaire) des occurrences du texte. En français, ce regroupement se pratique en général de la manière suivante :

- les formes verbales à l'infinitif,

- les substantifs au singulier,
- les adjectifs au masculin singulier,
- les formes élidées à la forme sans élision.

Lexical qui concerne le lexique ou le vocabulaire.

Occurrence Apparition d'un terme dans un corpus.

Profil (d'une ligne ou d'une colonne d'un tableau à double entrée) vecteur constitué par le rapport des effectifs contenus sur cette ligne (resp. colonne) à la somme des effectifs que contient la ligne (resp. la colonne).

Rotations procustéennes Le but est, ici, d'étudier l'évolution relative des points les uns par rapport aux autres, afin de connaître la typologie de leur dynamique. Pour cela, nous partons d'une matrice 3D croisant deux variables et le temps. Les instances du nuage sont modifiées, afin de les faire coïncider au mieux entre elles, et ce, en éliminant entre deux positions successives : la rotation moyenne, la translation moyenne, l'homothétie moyenne. Il est possible d'associer ces transformations avec les analyses factorielles. Pour les ACP et ACPr on peut indifféremment appliquer la transformation avant ou après l'analyse, par contre pour les AFC il faut impérativement analyser avant d'appliquer les rotations procustéennes car ces transformations génèrent des valeurs négatives incompatibles avec l'AFC. Au centre de la carte se trouvent les éléments qui évoluent de façon standard. A la périphérie certaines occurrences ressortent : elles traduisent des anomalies de l'évolution. Les trajectoires permettent éventuellement de détecter des stratégies.

Question fermée Question dont les seules réponses possibles sont proposées explicitement à la personne interrogée.

Question ouverte Question posée sans grille de réponse préétablie, dont la réponse peut être numérique ou textuelle.

Séparateurs de phrases Sous-ensemble des caractères délimiteurs de séquence correspondant aux seules ponctuations fortes (en général : le point, le point d'interrogation, le point d'exclamation).

Séquence Suite d'occurrences du texte non séparées par un délimiteur de séquence.

Syntagmatique qui concerne le regroupement des unités textuelles, selon leur ordre de succession dans la chaîne écrite.

Syntagme Groupe de mots en séquence formant une unité à l'intérieur de la phrase.

Tableau de contingence Tableau à double entrée dans lequel on dispose les modalités de *la* variable X en lignes et celles de Y en colonnes. Ce tableau est donc de dimension $r \times s$ et a pour élément générique le nombre $n_{\ell h}$ d'observations conjointes des modalités x_ℓ de X et y_h de Y ; les quantités $n_{\ell h}$ sont appelées les *effectifs conjoints*. La case à l'intersection de la ligne i et de la colonne j contient le nombre d'individus ayant la modalité i de la première variable et la modalité j de la seconde variable.

Annexes

Questionnaire d'évaluation de VisuGraph, outil graphique d'analyse de données relationnelles et évolutives. _____ 246
Graphe symétrique _____ 251
Graphe asymétrique orienté _____ 252
Graphe asymétrique, non orienté _____ 253
Graphe asymétrique, orienté et issu du croisement de trois variables _____ 254
Graphe temporel, orienté _____ 254
Graphe symétrique des dépôts de brevets _____ 256
Graphe symétrique dans le domaine des véhicules hybrides _____ 257

Questionnaire d'évaluation de VisuGraph, outil graphique d'analyse de données relationnelles et évolutives.

La visualisation de données relationnelles permet la transformation, le codage et la représentation graphique efficace de grands volumes de données. Cette technique offre à l'utilisateur une représentation claire et lisible de l'information, initialement difficile d'accès.
Nous proposons un outil de visualisation de données relationnelles et temporelles, nommé VisuGraph. Cet outil permet la représentation de données issues de matrices symétriques ou asymétriques, évolutives ou non.

Ce questionnaire a pour objectif d'évaluer cet outil, en ciblant plusieurs axes :

- L'interface : convivialité, facilité d'accès aux données, facilité de manipulation.
- L'opérabilité : l'exploitation des résultats, la qualité des résultats.
- L'interactivité : la possibilité pour l'utilisateur de maîtriser pleinement son application.

Pour cela, nous vous proposons de tester notre outil sur plusieurs jeux de données. Nous utiliserons les notations suivantes :
TD=Très Difficile
D = Difficile
M = Moyen
F = Facile
TF = Très Facile

Le serveur peut être à l'origine de lenteur dans l'exécution de certaines applications. Nous ne ciblons pas ce critère de validation dans cette enquête.

Renseignements

- Niveau en informatique
  Débutant ☐ intermédiaire ☐ Expert ☐
- Expérience dans le décisionnel
  Débutant ☐ intermédiaire ☐ Expert ☐
- " Avez-vous déjà utilisé des outils de visualisation de données ? " Oui ☐ Non ☐

  Si oui, citez les outils que vous avez déjà utilisés :

  ....
  ....
  ....
  ....
  ....
- Niveau d'étude
  <Bac ☐ Bac+1 ☐ Bac+2 ☐ Bac+3 ☐ Bac+5 ☐ Bac+8 ☐
  Autre .....
- Votre domaine d'activité ?
  ……..…….....
- Profession actuelle

Etudiant(e)	☐	Cadre et profession libérale	☐
Agriculteur	☐	Inactif ou femme au foyer	☐
Indépendant	☐	Profession intermédiaire	☐
Employé	☐	Chômeur	☐
Ouvrier	☐	Retraité	☐

Jeu de données de type « **matrice symétrique non évolutive** »

	Réponse	Niveau de difficulté				
		TD	D	M	F	TF
• Lancez la représentation graphique de la matrice *AL-AL* du répertoire *Enquete*.						
• Combien de sommets composent ce graphe ?						
• Affichez les noms des nœuds.						
• Affichez le poids des arêtes						
• Agrandissez la taille de la police de caractère.						
• Eclaircissez la police de caractère.						
• Changez la couleur du fond d'écran.						
• « Masquez » les nœuds non liés.						
• Déplacez un nœud dans la fenêtre de représentation.						
• Citez deux auteurs importants de ce graphe.						
• « Filtrez » le graphe en ne conservant que les liens les plus forts.						
• Quels sont les auteurs ayant le plus collaboré ?						
• Revenir au monde sans filtrage (filtrage nul).						
• A l'aide du KCore, indiquez combien d'équipes se distinguent par leur taille imposante. Pour cela appliquer le KCore maximal ; puis appliquer les forces d'attraction via « Force paramètre attraction ».						
• Quel est le seuil du KCore atteint ?						
• Revenir au graphe initial						
• Via le menu, retrouvez l'auteur « Lopez m».						
o Quelle valeur est associée à cet auteur ?						
o Combien de voisins directs a-t-il (passer par la transitivité) ?						
o Retrouvez ses collaborations par transitivité.						
• Revenir au graphe initial.						
• Effectuez un clustering.						
o Combien de classes trouvez-vous ?						
o Extrayez l'ensemble des classes dans un fichier texte. A quelle classe appartient « Dousset » ?						
• Effectuez une représentation circulaire des données.						

Jeu de données de type « **matrice symétrique évolutive** »

	Réponse	Niveau de difficulté				
		TD	D	M	F	TF
• Lancez la représentation graphique de la matrice *AC-AC-DP* du répertoire *Enquete*.						
• Combien de périodes sont représentées dans ce graphe ?						
o Citez-les						
• De quelle couleur sont représentés :						
o Les sommets appartenant à la *première* période.						

○ Les sommets appartenant à la *deuxième* période.						
○ Les sommets appartenant à la *troisième* période.						
○ Les sommets appartenant à la *quatrième* période.						
○ Les sommets appartenant à la *cinquième* période.						
….						
○ Les sommets appartenant à la *n-ième* période						
• Citez un sommet spécifique de la première période						
○ de la *deuxième* période						
○ de la *troisième* période						
○ de la *quatrième* période						
…						
○ de la *n-ième* période						
• Citez un sommet présent durant toutes les périodes.						
• Quelle spécificité ont les sommets situés au centre ?						
• Quelle spécificité ont les sommets situés en haut à gauche de l'écran ?						
• Quels sont les auteurs émergents ?						
• Affichez uniquement les auteurs présents durant						
○ la *première* période.						
○ durant la *deuxième* période.						
○ durant la *troisième* période.						
○ durant la *quatrième* période.						
…						
○ durant la *n-ième* période.						
• Revenir au graphe global.						
• Trouvez un auteur de faible valeur mais présent durant toutes les périodes.						
• Appliquez les forces d'attraction et de répulsion :						
○ semi-paramétrées						
▪ Faites varier la température : quel effet cela a-t-il ?						
▪ Faites varier la distance d'attraction						
▪ Faites varier la distance de répulsions						
○ paramétrées						
▪ Faites varier la température : quel effet cela a-t-il ?						
▪ Faites varier la distance d'attraction						
▪ Faites varier la distance de répulsions						
○ paramétrées attraction						
▪ Faites varier la température : quel effet cela a-t-il ?						
▪ Faites varier la distance d'attraction						
▪ Faites varier la distance de répulsions						

• Voyez-vous une différence entre les différentes forces d'attraction et de répulsion ?	Oui ☐ Non ☐					
Si oui, veuillez préciser o Semi-paramétrée o Paramétrée o Paramétrée attraction						
• Obtenez un graphe révélant les différentes équipes.						

Jeu de données de type « **matrice asymétrique non évolutive**»

	Réponse	Niveau de difficulté				
		TD	D	M	F	TF
• Lancez la représentation graphique de la matrice *CO-PA* du répertoire *Enquete*						
• De quelle couleur sont représentés les pays ?						
• Combien de pays sont affichées ? Les citer						
• De quelle couleur sont représentées les compagnies?						
• Affichez les libellés longs des nœuds.						
• Citez une compagnie importante.						
• Citez une compagnie moins importante.						
• Représentez les données sous forme de nuance.						
• Représentez les données sous forme de barre.						
• Revenez à une représentation des données sous forme de cercle.						
• Retrouvez la France (son nom est FR). Après sélection de ce sommet, cliquez sur le graphe pour l'afficher.						

Jeu de données de type « **matrice asymétrique évolutive** »

	Réponse	Niveau de difficulté				
		TD	D	M	F	TF
• Lancez la représentation graphique de la matrice *AL-PA-DP* du répertoire *Enquete*						
• De quelle couleur sont représentés les pays ?						
• Appliquer la force paramètre attraction. Combien de pays sont affichées ? Les citer.						
• Appliquer les forces morphing. Faites varier la force repère temporels. Quel effet cela a t'il sur le graphe ?						
• De quelle couleur sont représentés les auteurs? (bleu ou multicolore ?)						
• Quel est le pays le plus présent sur l'ensemble de toutes les périodes ?						
• Appliquer le morphing. Vous pouvez augmenter la vitesse de morphing (de l'application). Quels sont les auteurs émergents ? quelles sont leur caractéristiques ?						
• Quels sont les pays émergents ? (citer les principaux ?) • Que peut-on dire du domaine étudié sur l'ensemble des périodes : Est-ce un domaine en pleine progression ? Est-ce un domaine en régression ? Est-ce un domaine en constante activité ?						

De manière générale…

- Comment qualifiiez-vous cet outil ?
  Très utile ☐ Assez utile ☐ Peu utile ☐ Inutile ☐ Sans avis ☐

- La représentation des données est
  Adaptée ☐ Non adaptée ☐ Sans avis ☐

- L'outil facilite t'il l'exploration des données
  Oui ☐ Non ☐ sans avis ☐

- La détection des différentes tendances, des signaux forts est elle facilité ?
  Oui ☐ Non ☐ sans avis ☐

- Qualifieriez-vous l'outil d'interactif ?
  Oui ☐ Non ☐ sans avis ☐

- Qualifieriez-vous l'outil d'ergonomique ?
  Oui ☐ Non ☐ sans avis ☐

- L'accès aux différentes fonctionnalités est-il difficile ?
  Oui ☐ Non ☐ sans avis ☐

- Avez-vous déjà effectué une analyse de données avec la plate-forme « Tétralogie » ?
  Oui ☐ Non ☐ sans avis ☐

- Si oui, que pensez-vous de l'ajout de ce module à cette plate-forme ?
  Indispensable ☐ Très utile ☐ Utile ☐ Sans grand intérêt ☐ Inutile ☐ Sans avis ☐

- Quels conseils donneriez-vous pour l'améliorer ?

  .....
  .....
  .....
  .....
  .....
  .....
  .....
  .....
  .....
  .....

- Si votre domaine d'activité n'est pas directement lié à la visualisation de données, pensez-vous que cet outil pourrait vous être utile professionnellement ?
  Oui ☐ Non ☐ sans avis ☐

Graphe symétrique

La figure suivante est issue d'une matrice symétrique, croisant des inventeurs de brevets entre eux. Une grosse équipe centrale se distingue des autres petites équipes, composées de deux ou trois inventeurs.

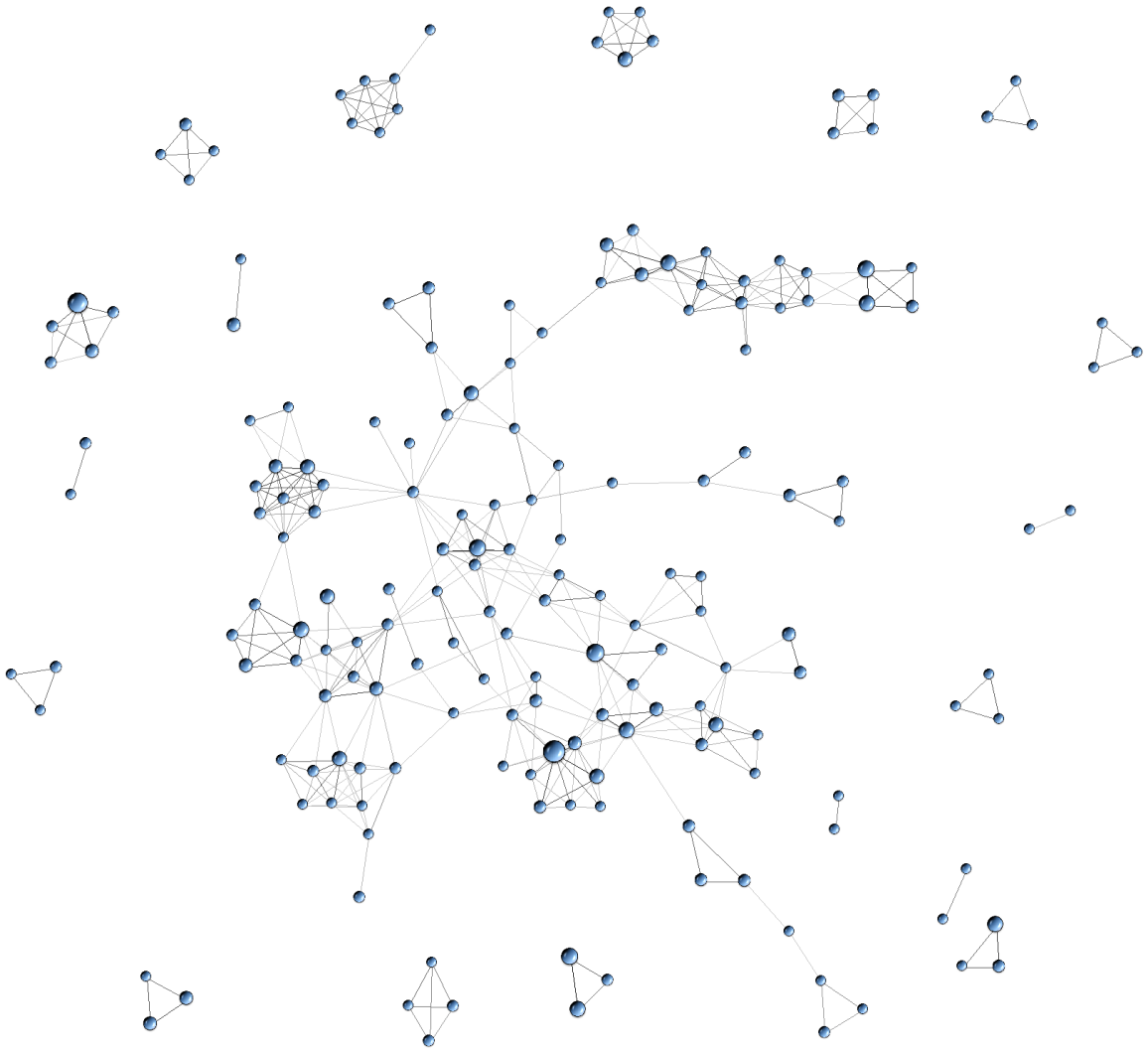


Figure 114. *Graphe symétrique croisant des compagnies pharmaceutiques.*

Graphe asymétrique orienté

Le graphe suivant est issu d'une matrice asymétrique croisant des sociétés liées au domaine de l'aéronautique sous-traitant une activité auprès de sociétés de services. Le prédécesseur est l'entreprise d'origine et le successeur est la société de service.

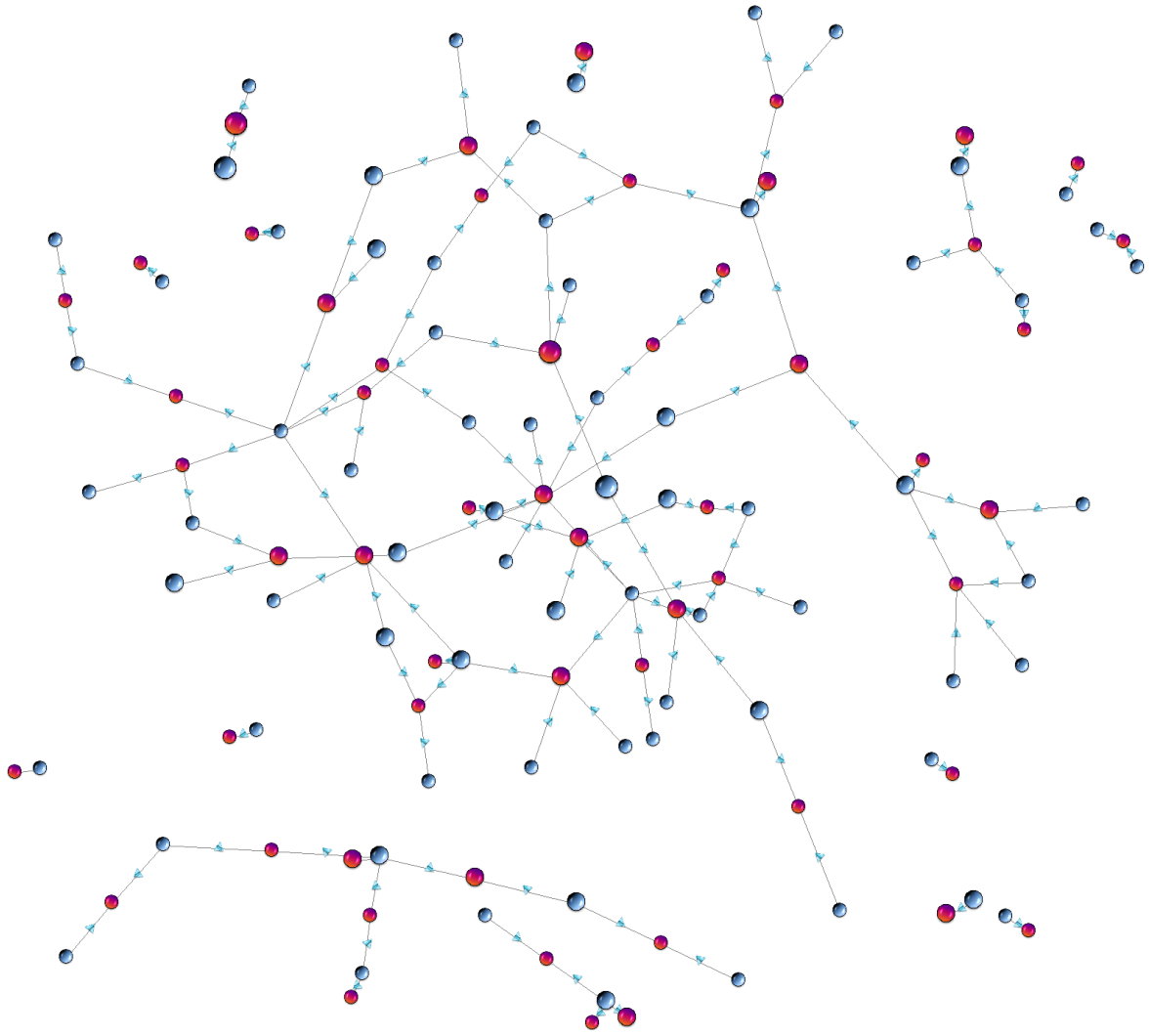


Figure 115. *Graphe asymétrique orienté.*

Graphe asymétrique, non orienté

Le graphe suivant est issu du croisement entre des compagnies pharmaceutiques achetant des licences (LI) et celles les accordant (LO). Ainsi, la caractéristique du licencié/ licencieur est traduit par une couleur spécifique. Les sommets bi-colores signifient leur appartenance aux deux catégories, ils accordent et achètent à la fois des licences.

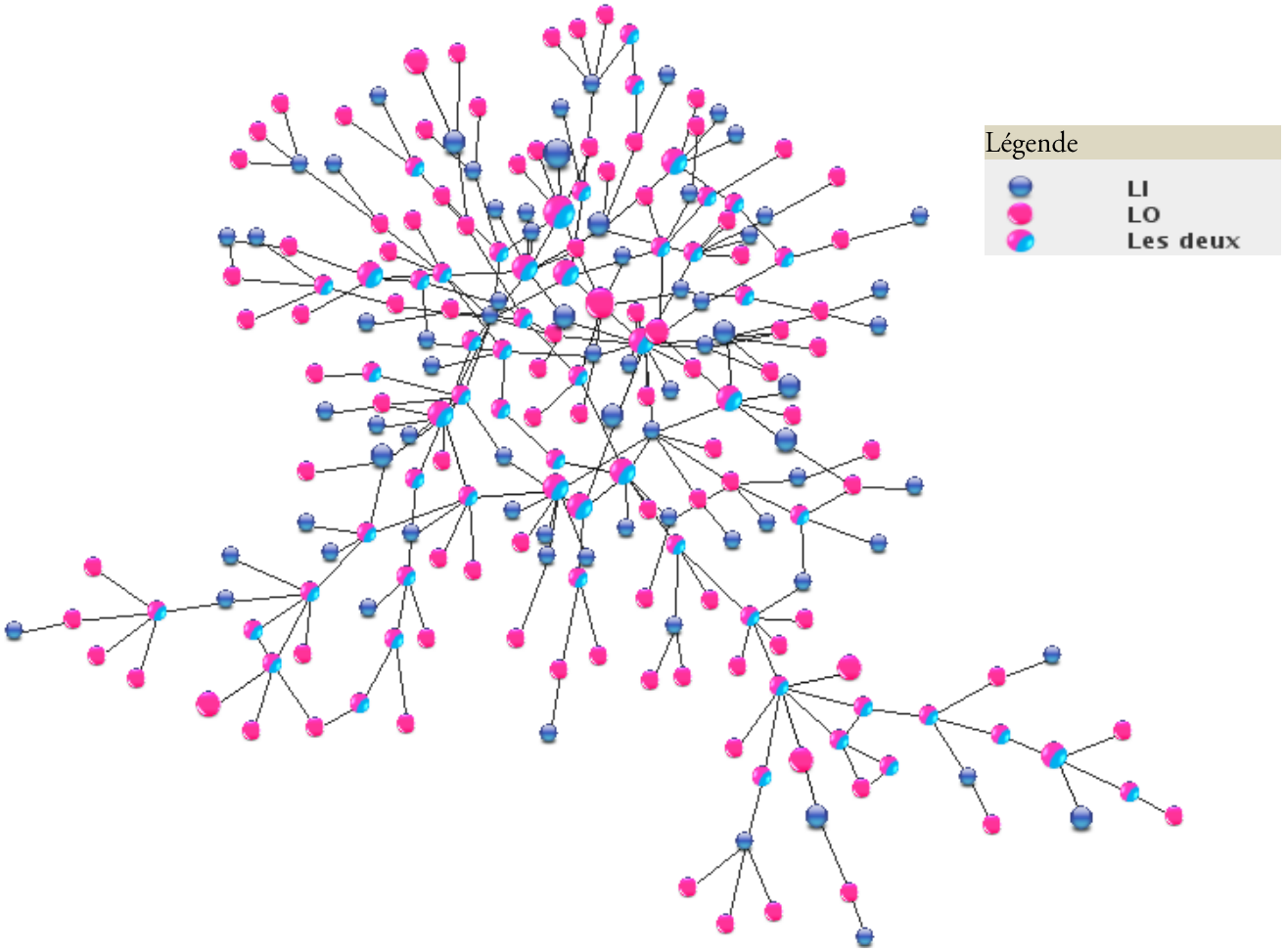


Figure 116. *Graphe asymétrique des licenciés/ licenseurs.*

Graphe asymétrique, orienté et issu du croisement de trois variables

Le graphe suivant est issu d'une matrice asymétrique croisant des compagnies achetant des licences (LI) et celles qui les accordent (LO). A ce croisement est ajouté celui des pays d'appartenance de ces sociétés. Ainsi, chaque pays est visible par un carré de couleur en dessous de chaque icône circulaire. Ce graphe est orienté afin de distinguer les acheteurs des vendeurs. La flèche à double sens indique l'accord d'une licence respective d'une compagnie à l'autre.

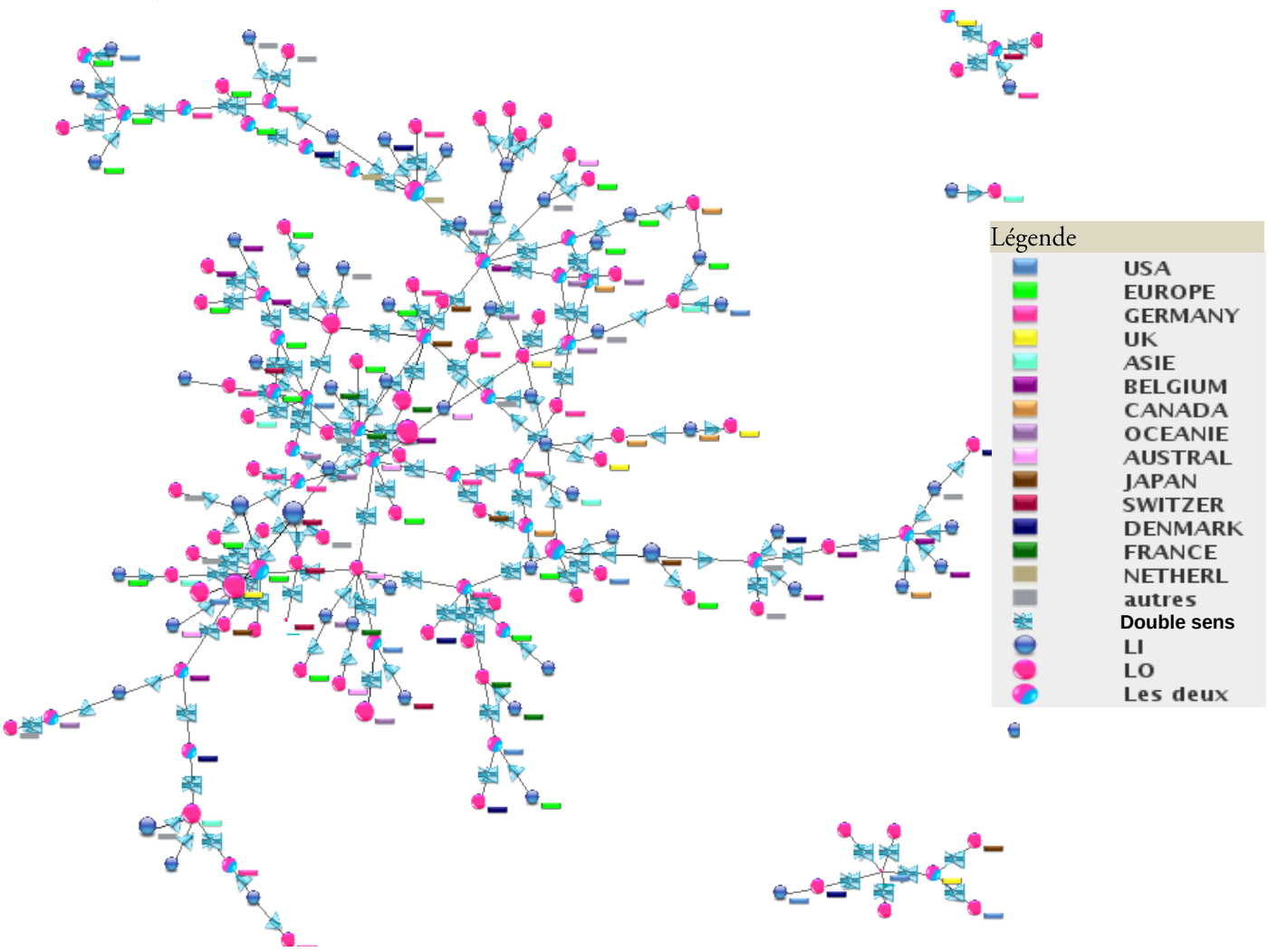


Figure 117. *Graphe asymétrique, orienté et issu du croisement de trois variables.*

Graphe temporel, orienté

Le graphe suivant est un graphe temporel sur deux périodes. Ce graphe se base sur le croisement de compagnies pharmaceutiques croisées entre elles. A ce résultat sont croisés les pays d'appartenance de ces sociétés. Ainsi, chaque pays est visible par un carré de couleur en dessous de chaque barre indiquant la caractéristique temporelle du sommet. Ce graphe est orienté afin de distinguer les sociétés pharmaceutiques faisant appel aux services d'une autre.

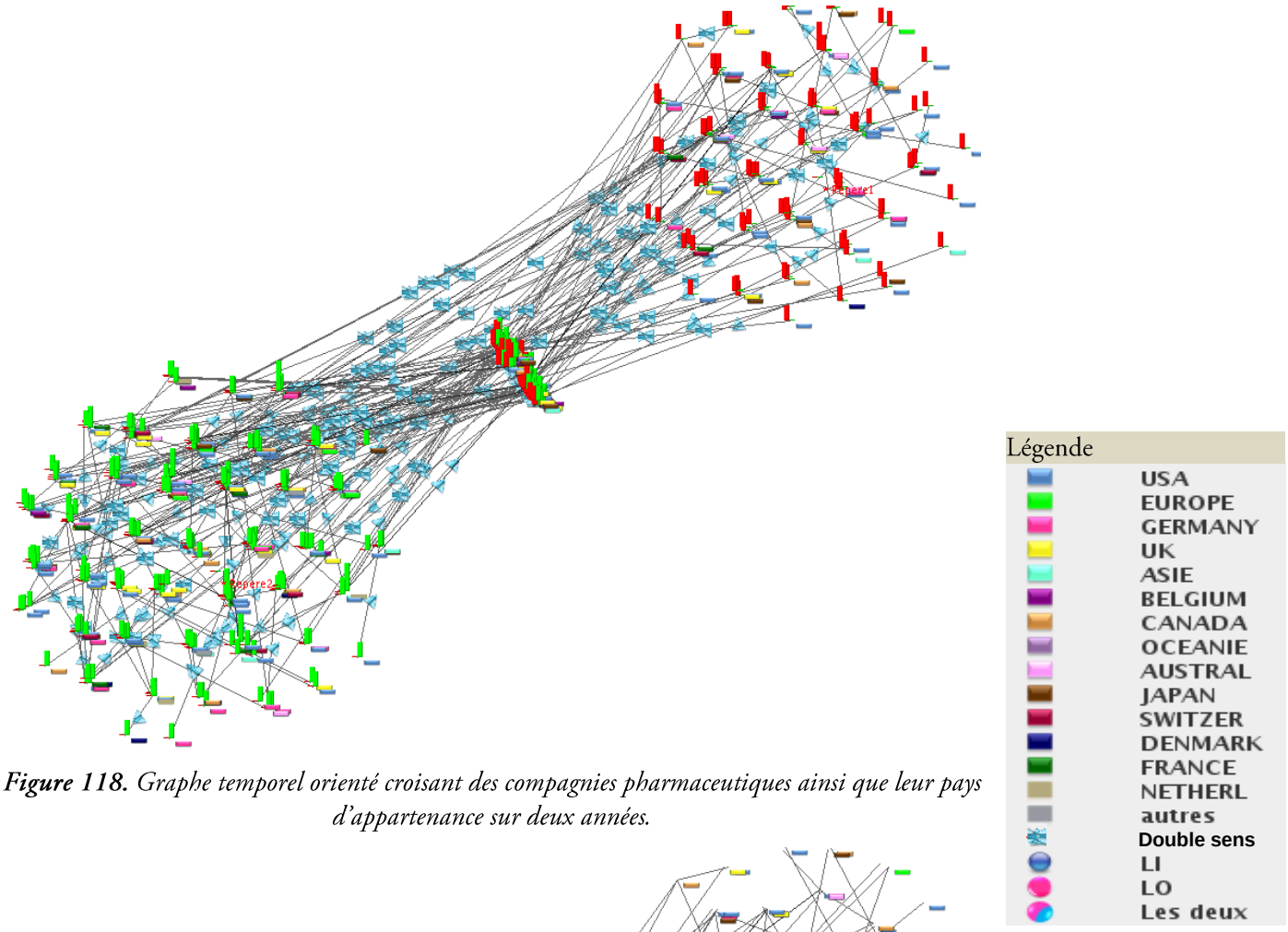


Figure 118. *Graphe temporel orienté croisant des compagnies pharmaceutiques ainsi que leur pays d'appartenance sur deux années.*

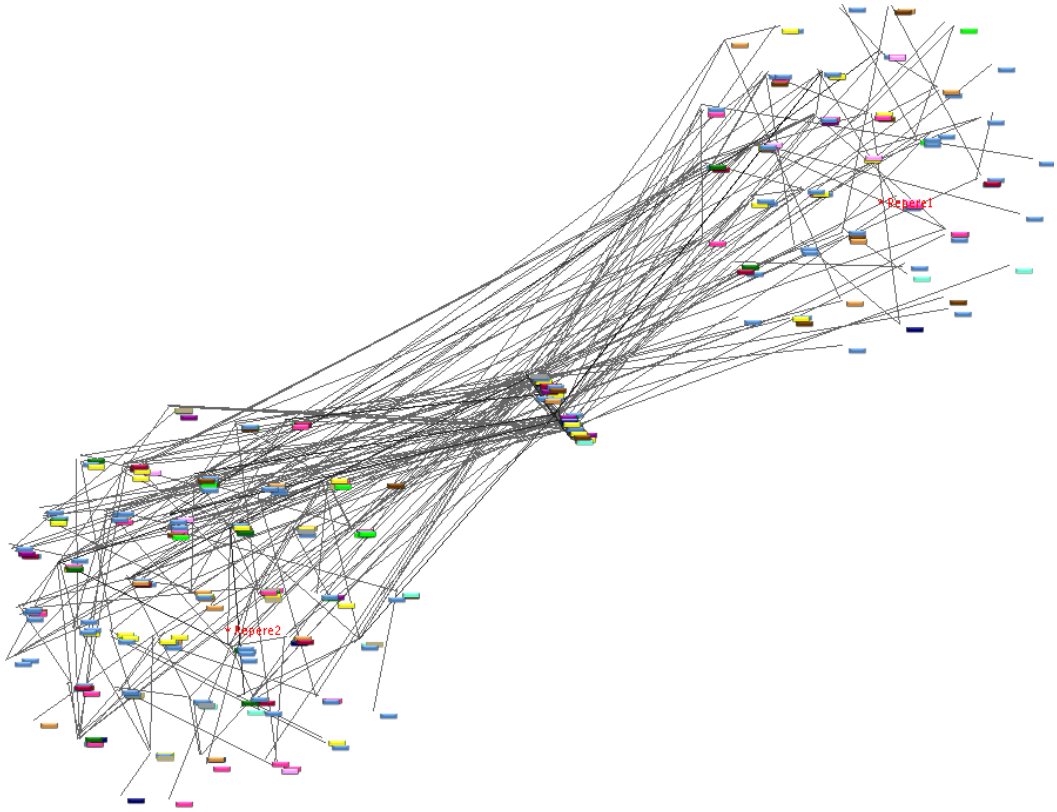


Figure 119. *Figure précédente dont l'orientation des arcs à été masquée, ainsi que les icônes traduisant la temporalité des sommets.*

Graphe symétrique des dépôts de brevets

Ce graphe représente les collaborations entre les entreprises américaines dans les dépôts de brevets en aéronautique. On constate la bipolarité Boeing - MacDouglas et un certain nombre de clusters indépendants.

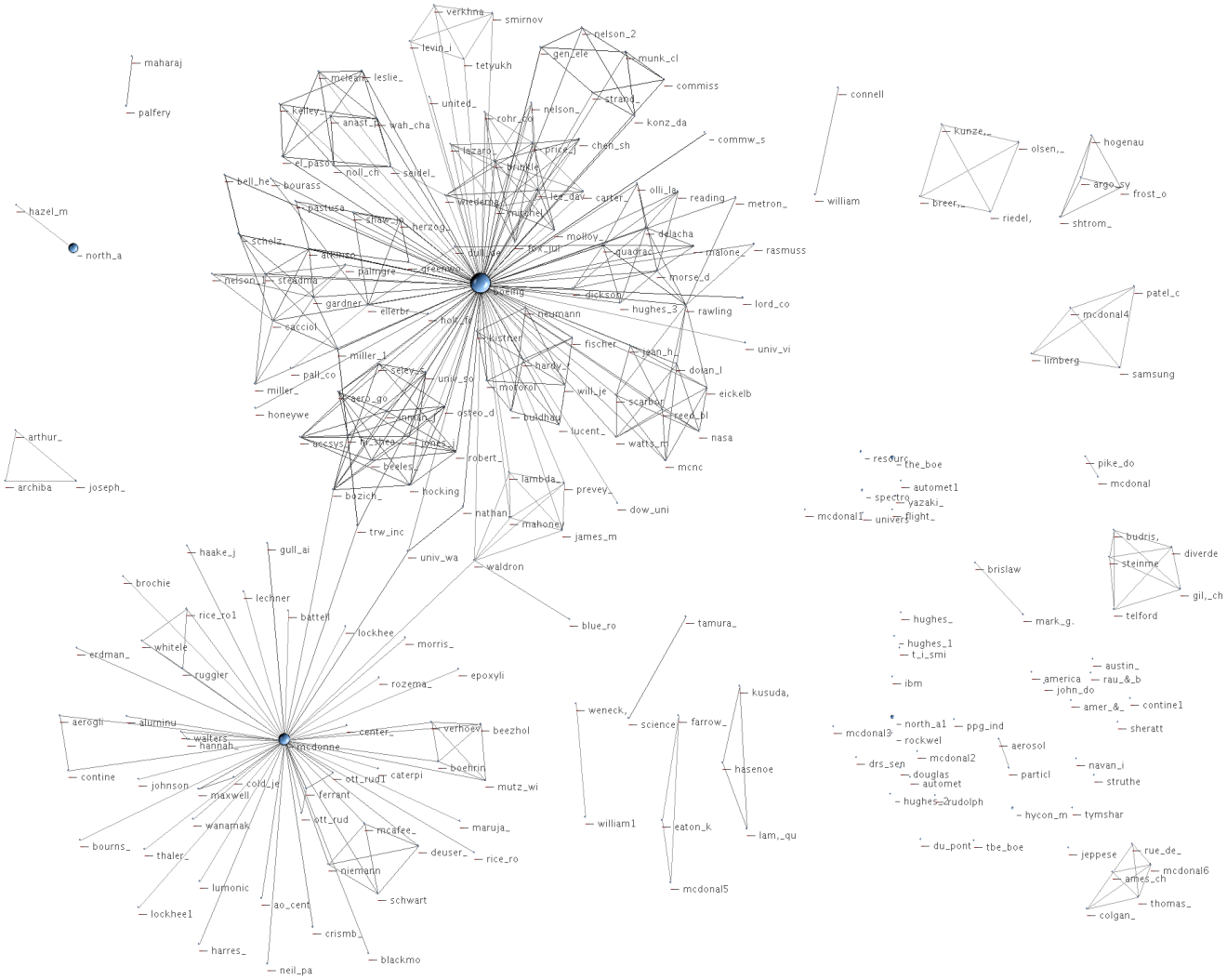


Figure 120. *Graphe symétrique des dépôts de brevet dans le domaine de l'aéronautique.*

Graphe symétrique dans le domaine des véhicules hybrides

Ce graphe représente me croisement des inventeurs de brevets japonais portant sur le domaine des véhicules hybrides. On constate l'importance de Yamaguc, Suzuki-4, Katsuda comme points d'articulations et Tuga-y, Sasaki, Takagi les inventeurs les plus importants.

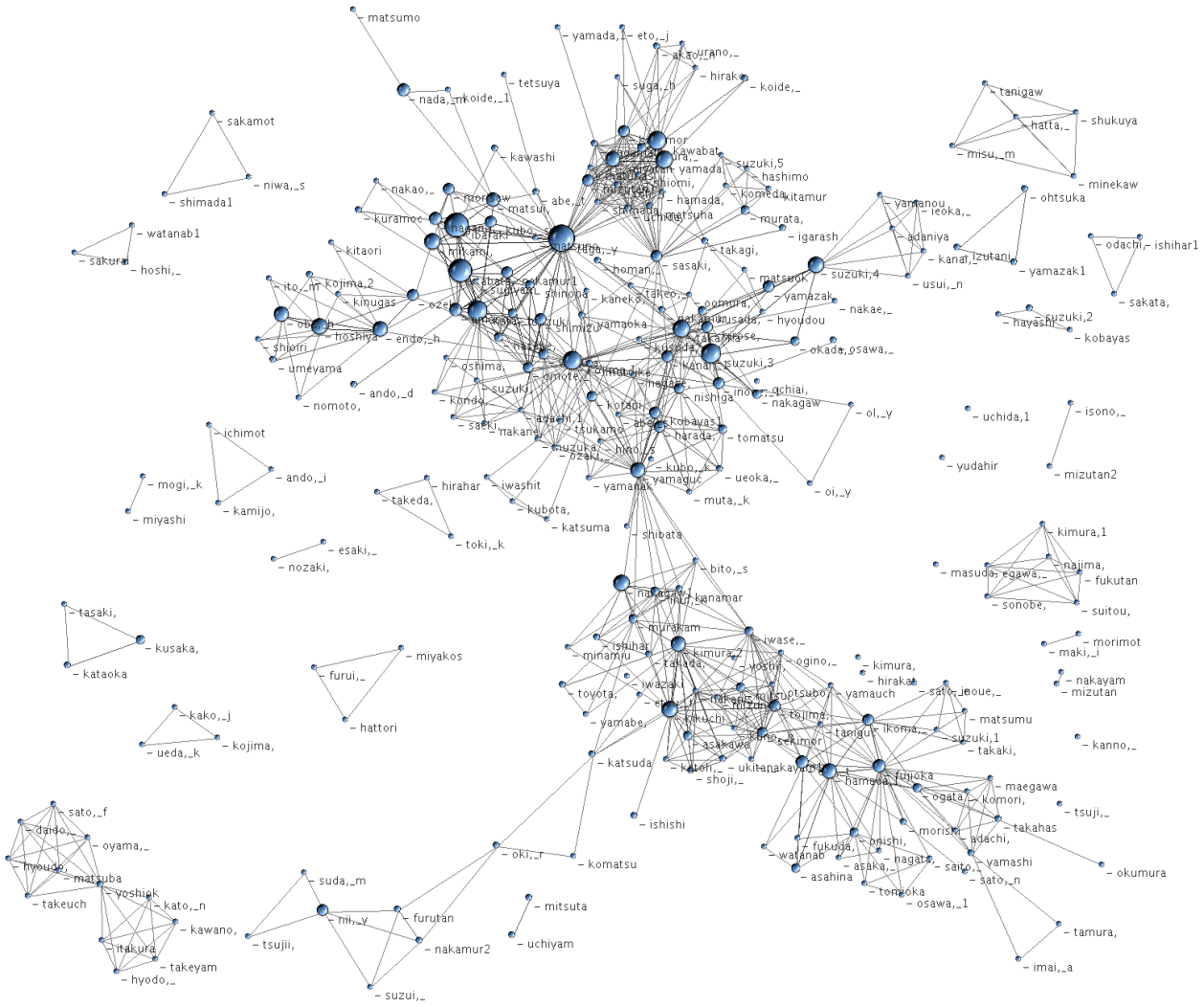


Figure 121. *Graphe symétrique des inventeurs de brevet dans le domaine des véhicules hybrides.*