

Modèles neuronaux non supervisés cartographiques pour le traitement des données documentaires multimédia : application à l'analyse du Web

THÈSE

présentée et soutenue publiquement le 28 Juin 2006

pour l'obtention du

Doctorat de l'université Henri Poincaré – Nancy 1

(spécialité informatique)

par

Shadi Al Shehabi

Composition du jury

<i>Président :</i>	Noëlle Carbonelle	Professeur, Université Henry Poincaré, Nancy
<i>Rapporteurs :</i>	Berand Dousset	Professeur, Université Paul Sabatier, Toulouse
	Wolfgang Glänzel	Professeur, Université de Leuven
<i>Directeurs :</i>	Frédéric Alexandre	Directeur de recherche, INRIA, Nancy
	Jean-Charles Lamirel	Maître de Conférence, Université Robert Schuman, Strasbourg



Mis en page avec la classe thloria.

Remerciements

Mes premiers remerciements vont à mon encadrant Jean-Charles Lamirel pour le temps qu'il m'a consacré à ce travail de thèse, son soutien, ses conseils scientifiques, sa disponibilité et son aide précieuse pour améliorer et aller jusqu'à bout pour terminer ce travail de thèse. Il m'a vraiment impressionné par ses qualités humaines et son esprit ouvert.

Je tiens à remercier mon directeur de thèse Frédéric Alexandre qui a accepté de m'encadrer durant ces années de thèse et pour son soutien.

J'exprime ma sincère reconnaissance aux membres du jury qui me font l'honneur de participer à la soutenance. En particulier, je remercie vivement Bernard Dousset et Wolfgang Glänzel d'avoir bien voulu être rapporteurs de ce travail de thèse ainsi que pour tout le temps qu'ils ont consacré à la lecture de ce document.

Je tiens à remercier les plus chers mon grand enseignant en Syrie Monsieur Rajab Deeb pour des conseils et son soutien.

Je tiens à remercier mon université de Syrie qui m'a envoyé pour faire une thèse à l'étranger.

Je tiens à remercier tous les membres de l'équipe CORTEX, Randa, Zhour, Claudio, Nizar, Tomas, Dominique, Maxime, Jérémy, Julien et Nicolas pour leur soutien et leur accueil et les membres de l'équipe ORPAILLEUR, Yanick, Laszlo, Hacène et Nizar.

Je tiens à remercier monsieur François Lamirel, le père de Jean-Charles, de son aide pour corriger le rapport de thèse et les papiers publiés.

Je tiens à remercier Claire François de l'INIST pour sa coopération.

Je tiens à remercier Mohammed Attik, qui a partagé avec moi le bureau durant les années de thèse, de sa coopération, sa discussion et son soutien.

Mes grands remerciements à ma grande famille en particulier mon père, à ma mère, à ma grande mère, à mes frères, à mes sœurs, à mon épouse et à mon frère et meilleur ami Mohammad nasani pour leur encouragement, leurs prières pour réussir dans ma vie professionnelle.

Je tiens particulièrement à remercier celui qui me donne d'énergie et de capacité de travailler et de vivre tous le temps dès mon enfance jusqu'à alors.

Enfin, Je tiens à remercier mes amis Houssam Murad, Johan, Abdessammar, Ahmad Hashoume, Rim, Hala, Housseine, Szalrd, Hatem, Imad, Moustafa, Kassem, Omar, Abdelaziz, Fadi, Hager, Mohammad Taha, Jamal, Riad, Abdeljabar, Hazem, ...

Table des matières

Introduction Générale

Partie I Traitement des données multidimensionnelles **7**

Chapitre 1

Représentation et exploitation des données documentaires

1.1	Introduction	9
1.2	L'indexation des documents	10
1.2.1	L'analyse de surface des documents	10
1.2.2	Modèles de représentation des documents	12
1.2.3	Notion de point de vue	12
1.3	Système de recherche documentaire	13
1.3.1	Les stratégies de recherche des documents	13
1.3.2	Réseaux bayésiens pour la recherche documentaire	15
1.3.3	Pondération par jugements de pertinence	19
1.3.4	Exploitation des relations sémantiques	20
1.3.5	Évaluation des systèmes de recherche documentaire	22
1.4	Extraction de connaissances dans les bases documentaires	24
1.4.1	Le processus d'extraction de connaissances	24
1.4.2	Fouille de données	24
1.4.3	Treillis de concepts et les motifs fréquents	25
1.4.4	Extraction de règles d'association	29
1.4.5	Inconvénients des modèles symboliques	32
1.5	Conclusion	33

Chapitre 2

Classification numérique non supervisée
--

2.1	Introduction	35
2.2	Méthodes de classification classiques	36
2.2.1	Méthodes de partitionnement (directes)	37
2.2.2	Méthodes hiérarchiques (indirectes)	39
2.2.3	Méthodes hybrides	41
2.2.4	Méthodes orientées par la densité	42
2.3	Approches neuromimétiques pour la classification	43
2.3.1	Principes généraux	43
2.3.2	Carte auto-organisatrice de Kohonen (SOM)	45
2.3.3	Neural gas (NG)	47
2.3.4	Neural gas plus competitive Hebbian learning	49
2.3.5	Growing neural gas (GNG)	49
2.3.6	SOM Incrémental Hiérarchique (GHSOM)	49
2.3.7	Le modèle MultiSOM	52
2.4	Évaluation de la qualité de classification	55
2.4.1	Qualité basée sur la distance (critères inertiels)	55
2.4.2	Qualité basée sur la similarité	56
2.4.3	Qualité basée sur la méthodologie et la métrique expérimentales	57
2.4.4	Qualité basée sur l'algèbre linéaire	58
2.5	Conclusion	59

Chapitre 3

Visualisation d'information

3.1	Introduction	61
3.2	Dimension intrinsèque des données multidimensionnelles	62
3.3	Visualisation par projection cartographique linéaire	63
3.3.1	L'analyse en composantes principales	63
3.3.2	La projection aléatoire (Random Mapping)	64
3.4	Visualisation par projection cartographique non linéaire	65
3.4.1	Méthodes de projection	65
3.4.2	Exemples applicatifs de visualisation dans le domaine documentaire	68
3.5	Visualisation hyperbolique	74
3.5.1	L'espace hyperbolique	75
3.5.2	Arbre hyperbolique (Hypertree)	75
3.5.3	Modèles neuronaux hyperboliques de la carte de Kohonen	76
3.6	Conclusion	79

Chapitre 4

Nouvelles mesures de qualité du clustering

4.1	Introduction	83
4.2	Qualité et analyse de classification non supervisée pour les données basées sur leur espace de description	84
4.2.1	Nouvelles mesures de la qualité basées sur la distribution des des- cripteurs de données	85
4.2.2	Équivalence au treillis de Galois	86
4.2.3	Sélection de modèle optimal	88
4.2.4	Expérimentations	88
4.3	Qualité et analyse de classification non supervisée pour les données avec des étiquettes multiples	97
4.3.1	Nouvelles mesures de qualité basées sur la distribution d'étiquettes .	99
4.3.2	Sélection de modèle optimal	102
4.3.3	Détermination du critère d'arrêt d'apprentissage	104
4.3.4	Définitions des types d'étiquettes	105
4.3.5	Définitions de pertinence d'étiquette et de classe	105
4.3.6	Analyse de classification	106
4.3.7	Expérimentations	107
4.4	Conclusion	116

Chapitre 5

**Généralisation du modèle multi-topographique : application à l'extraction
de connaissances**

5.1	Introduction	119
5.2	Mécanisme de communication	120
5.2.1	Modèle de réseau bayésien pour la communication inter-cartes . . .	121
5.2.2	Comparaison entre le mode probabiliste et le modèle bayésien	123
5.3	Mécanisme de comparaison des vues	125
5.3.1	Évaluation de la cohérence de propagation	125
5.3.2	Expérimentation webométrique de comparaison de vues	127
5.4	Mécanisme de généralisation	130
5.4.1	Mécanisme de généralisation de Neural Gas	130

5.4.2	Évaluation du mécanisme de généralisation	131
5.4.3	Expérimentation	132
5.5	Extraction des règles d'association par des modèles neuronaux non supervi- sés	136
5.5.1	Critères principaux	137
5.5.2	Extraction par le mécanisme de généralisation	137
5.5.3	Extraction par le mécanisme de communication	138
5.5.4	Expérimentation	139
5.6	Conclusion	146

Chapitre 6

Optimisation des stratégies de visualisation des résultats de classification

6.1	Introduction	149
6.2	Validation de la communication inter-cartes	150
6.2.1	Description des cartes	150
6.2.2	Analyse locale de la communication inter-cartes	151
6.2.3	Validation croisée des approches : MultiSOM et BibTechMon	154
6.3	Classification hiérarchique multi-niveaux orientée par la densité (DBHC)	157
6.3.1	Principe de DBHC	157
6.3.2	Visualisation hypertexte de DBHC	157
6.3.3	Application à la webométrie	160
6.4	Conclusion	161

Chapitre 7

Synthèse

Conclusion générale

Annexe A

Description des données

A.1	Introduction	173
A.2	Analyse des brevets	173
A.2.1	Définition des points de vue	173
A.2.2	Multi-indexation des brevets	174
A.3	Analyse du Web	176

A.3.1	Analyse du site web du Laboratoire d'Informatique de l'Université de Cambridge	176
A.3.2	L'analyse de sites des laboratoires de recherche européens	177
Liste des figures		179
Liste des tableaux		187
Liste des algorithmes		189
Liste des publications		191
Bibliographie		193



Introduction Générale

Problématique

La croissance exponentielle de l'information accessible en ligne et par l'intermédiaire des bases de données documentaires représente à la fois un défi et un challenge pour les analystes de l'information. En effet, ces derniers se voient confrontés à des sources de plus en plus riches dont le contenu global devient très complexe à appréhender et dans lesquelles il est également de plus en plus difficile de s'orienter efficacement pour y puiser de la connaissance pertinente. Le volume des données, leur diversité, et, l'hétérogénéité de leurs sources empêchant toute opération manuelle d'extraction d'une telle connaissance, la nécessité s'est donc fait sentir de proposer de nouvelles méthodologies ainsi que des outils efficaces pour y parvenir. L'étude de ces outils et de ces méthodologies a donné naissance à un courant spécifique de recherche appelé extraction de connaissances dans les bases de données ou ECBD. Le processus type d'extraction de connaissances fait généralement intervenir les connaissances d'un analyste, expert du domaine relatif aux données concernées, dont le rôle est de déterminer les étapes de l'analyse et les outils qui doivent être employés pour chacune d'entre elles. Ce dernier a aussi la responsabilité de valider, par le biais d'interfaces de visualisation des résultats intermédiaires et finaux, les résultats obtenus par les méthodes de fouille de données. Les méthodes de fouilles de données, qui interviennent de manière centrale dans le processus d'ECBD, font elle-même principalement appel aux techniques de l'analyse de données, et plus particulièrement, aux techniques de classification numériques et symboliques.

Chaque technique d'analyse des données a des avantages et des inconvénients lors de telles applications. Un problème majeur des techniques de classification symboliques est le temps de calcul lié aux problèmes combinatoires. Cette question peut être résolue par les techniques numériques et particulièrement par les techniques numériques neuronales utilisant la notion d'apprentissage et dont les facultés de synthèse d'information sont considérables. Un autre des problèmes importants des techniques de classification symboliques est la visualisation des résultats obtenus, particulièrement dans le cas de l'analyse de données fortement multidimensionnelles. Ce problème peut partiellement être résolu par les techniques numériques, soit de manière directe, par celles qui permettent à la fois une classification et une projection des résultats, soit de manière indirecte, en combinant les techniques de classification avec des techniques complémentaires de projection.

D'un autre côté, le défaut principal des techniques d'analyse numériques est que les solutions proposées sont variées, ce qui nécessite de valider leurs résultats en évaluant leur capacité de synthèse relativement aux données à traiter. Les méthodes d'évaluation existantes restent cependant très générales et très instables. Elles ne permettent pas réellement de comparer des méthodes, et s'avèrent encore moins convaincantes dans des cas spécifiques, comme celui de l'analyse des données documentaires. En effet, dans ce dernier cas, il s'avère indispensable de

prendre en compte des critères de qualité spécifiques liés à l'intelligibilité du résultat par les analystes. De plus, malgré leur puissance de synthèse, les techniques d'analyse numérique classiques ne peuvent pas réellement s'affranchir des problèmes de bruit liés à leur appréhension globale du contenu des données. De manière complémentaire, elles ne permettent pas de gérer l'instrumentalisation de la comparaison d'analyses différentes, en provenance de sources multiples ou construites à partir de descriptions multiples, ou locales. Cette dernière limitation nuit fortement à la qualité et la granularité et la portée des analyses susceptibles d'être produites.

La visualisation des résultats fournis par les méthodes d'analyses numériques reste également un problème relativement ouvert, malgré l'importance qu'il peut jouer dans la compréhension desdits résultats. Les techniques de projection les plus performantes, comme les techniques de projection non linéaires, avouent rapidement leurs limites dans le cas où les données à visualiser sont initialement représentés dans un espace fortement multidimensionnel. De plus, la prise en compte de relations entre ces données pose des problèmes supplémentaires de surcharge cognitive propres à la représentation des graphes.

Contribution

Le but de ce travail de thèse est celui de proposer un cadre général qui permette de résoudre de manière globale les problèmes liés à l'analyse et à la fouille de données que nous venons de mentionner. D'une manière plus synthétique, ces problèmes se rapportent à :

- l'intervention humaine dans chaque étape d'analyse et de fouille de données. A titre d'exemple, un des tâches de l'analyste utilisant une méthode de classification numérique reviendra la plupart du temps à :
 - déterminer subjectivement le nombre de classes à utiliser,
 - estimer empiriquement la qualité de la méthode proprement dite,
 - supprimer des informations inutiles qu'elle produit.
- l'exploitation des méthodes de classification numériques et de leurs résultats pour des tâches telle que l'extraction de connaissances.
- l'analyse basée sur une description globale des données, étant donné que ce type d'analyse produit à la fois du bruit et des résultats complexes, voir impossibles à interpréter.
- l'interaction globale entre les informations susceptibles de provenir d'analyses locales basées sur des critères, ou des points de vue, différents sur les données. A titre d'exemple, si une méthode d'analyse de données traite des données web, et que celles-ci se rapportent à la fois à des villes, des universités et des domaines de recherche, ce type d'interaction devra permettre de fournir :
 - des informations générales sur le comportement de la recherche et sur les stratégies de recherche dans les universités
 - des informations sur les types de coopération entre les universités, etc.,
- l'interaction purement locale entre les informations. Dans le cas de l'exemple précédent, ce type d'interaction devra permettre d'obtenir des informations spécifiques sur la distribution propre à chaque domaine de recherche dans les villes, ou inversement, sur la couverture thématique de chaque ville en terme de domaines de recherche etc.,
- l'interprétation des résultats obtenus par l'interaction locale et par l'interaction globale, ainsi que la comparaison de ces résultats, qui doivent être objectives.
- la visualisation des données, ainsi que des résultats de leur analyse en préservant les relations importantes entre les informations. Toujours dans le cas de l'exemple précédent,

la méthode de visualisation devra permettre d'illustrer le comportement du de référencement entre les universités par l'intermédiaire de la distribution des liens, entrant ou sortant, entre les sites associés, figurer la répartition géographique des domaines de recherche, etc.

L'approche NOMAD [Lamirel, 1995] représente une approche novatrice dans le domaine de la recherche d'information. Elle est la première à introduire la représentation des données documentaires selon des vues multiples et la communication entre les vues. Les auteurs de cette approche ont démontré son intérêt spécifique pour la recherche d'information. Ils en ont également développé une première extension, nommée MultiSOM, basée sur les cartes auto-organisatrices de Kohonen et dédiée spécifiquement à l'analyse des données documentaires. Nous reviendrons longuement sur ces travaux tout au long de notre propre travail de thèse, car ils représentent une base très solide pour définir un nouveau paradigme général d'analyse de données basée sur les points de vue.

Pour mettre en place ce paradigme, et comme il est naturellement «difficile de construire la maison sur des sables mouvants» nous nous intéressons donc prioritairement à consolider les mécanismes d'évaluation des résultats des analyses de données numériques. Pour cela, nous mettons donc en place de nouveaux mécanismes d'évaluation originaux. Notre but principal sera ensuite, grâce à ces mécanismes, de valider objectivement la supériorité de l'approche d'analyse de données numérique orientée par les points de vue par rapport à l'approche d'analyse de données numérique globale, bien que cette supériorité, notamment en terme de flexibilité, ait déjà été mentionnée subjectivement par des experts dans le cas du modèle MultiSOM. Ces mécanismes nous permettront également de faire plusieurs comparaisons entre différentes méthodes de classification de données numériques, de manière à sélectionner les méthodes les plus pertinentes. Ils nous permettront finalement de corriger les défauts, ainsi que d'optimiser le mode opératoire, des méthodes les plus intéressantes.

Nous nous attacherons ensuite à proposer un modèle générique de communication entre les vues pouvant être appliqué à des méthodes de classification différentes. Les mécanismes-noyaux de ce modèle devront permettre de gérer à la fois les interactions locales et les interactions globales entre les vues, de manière à pouvoir mener parallèlement des analyses générales et des analyses ciblées. De manière connexe, nous attacherons à définir des mécanismes de généralisation des vues, à la fois compatibles avec ce modèle de communication et calculables en ligne, c'est-à-dire, sans faire intervenir une nouvelle phase d'adaptation aux données.

Ces premiers travaux nous permettront ensuite de proposer un nouveau cadre pour la fouille de données à partir des méthodes de classification numériques. Nous mettrons notamment donc une nouvelle méthodologie d'extraction de règles d'association, fonctionnant à la fois sur des vues, ou classifications, élémentaires, mais permettant également de tirer parti des mécanismes de généralisation des vues élémentaires et de communication entre des vues multiples.

Nous nous attacherons finalement à corriger les défauts des méthodes de visualisation existantes. Nous proposerons pour cela une approche originale permettant de visualiser, sans surcharge cognitive, les interactions entre les informations ou les données analysées, ainsi que les différents degrés de cette interaction.

Plan du mémoire

La première partie de notre mémoire présente l'état de l'art relatif au traitement des données multidimensionnelles, et en particulier des données documentaires. Dans le premier chapitre,

nous commençons donc par détailler les méthodes d'analyse de surface utilisées pour représenter les gros volumes de données documentaires selon des modèles facilement exploitables en machine. Dans ce cadre, nous présenterons également les stratégies d'exploitation et d'optimisation de tels modèles pour l'analyse. Ensuite, nous présentons le processus d'extraction de connaissances dans les bases de données, ainsi que les approches symboliques utilisées dans ce processus. Puis, nous présentons un panorama des principes de la classification numérique non supervisée (clustering), couvrant les méthodes de classification classiques et les méthodes neuronales, en insistant sur les avantages spécifiques de ces dernières, et en particulier sur ceux des méthodes neuronales topographiques. Nous abordons après cela les techniques classiques utilisées pour évaluer ce type de méthodes. Nous terminons en présentant dans le troisième chapitre le problème de la visualisation des données multidimensionnelles. Nous présentons alors plusieurs techniques connues utilisées pour visualiser ce type de données, à savoir la projection linéaire, la projection non linéaire, et la visualisation hyperbolique. Nous présentons également quelques exemples d'application de ces méthodes.

La deuxième partie de notre mémoire présente notre contribution de thèse, il s'agit de mettre en place un nouveau paradigme pour l'analyse et la fouille de données orientée par les points de vue. Donc, nous commençons, dans le quatrième chapitre, par proposer de nouveaux mécanismes d'évaluation et d'analyse des méthodes de classification numériques non supervisées. Nous définissons d'abord une première série de mesures en nous inspirant de celles utilisées en recherche d'information et en nous basant également sur l'analogie entre les méthodes numériques et les méthodes symboliques. Nous abordons ensuite le problème de manière plus générale en nous intéressant à l'évaluation des données associées à des étiquettes multiples. Ceci nous amène alors à mettre en place de nouvelles mesures, inspirées des premières, mais basées sur une approche probabiliste. Il est possible de considérer que ce chapitre représente le fondement de notre travail de thèse. Dans le cinquième chapitre, nous proposons un modèle générique pour l'analyse et la fouille de données orientée par les points de vue, ceci en généralisant et en complétant les mécanismes proposés précédemment par le modèle MultiSOM. Nous montrons ainsi comment étendre le mécanisme de communication entre les vues en l'assimilant à un modèle d'inférence bayésien, et celui de généralisation, en utilisant des techniques de triangulation. Nous introduisons également un nouveau mécanisme important de comparaison des vues. En utilisant l'ensemble de nos propositions précédentes, nous proposons également une nouvelle méthodologie pour l'extraction de règles d'association basée sur l'exploitation des résultats d'un modèle de classification numérique et permettant l'optimisation du calcul des règles et la sélection des règles. Nous proposons ensuite une nouvelle méthode de visualisation hyperbolique basée sur la combinaison d'une classification hiérarchique originale, orientée par la densité, et une méthode de navigation interactive, de type focus+contexte, par arbre hyperbolique. Nous terminons cette partie par la synthèse de notre travail en présentant le schéma de notre paradigme d'analyse et de fouille de données orientée par les points de vue (cf. figure 1).

Nous utiliserons deux types de corpus différents dans nos expérimentations, à savoir un corpus de notice de brevets, et deux corpus différents des données web décrites à la fois par un contenu textuel et par des profils de liens hypertextes. Une description détaillée de ces corpus est fournie en annexe.

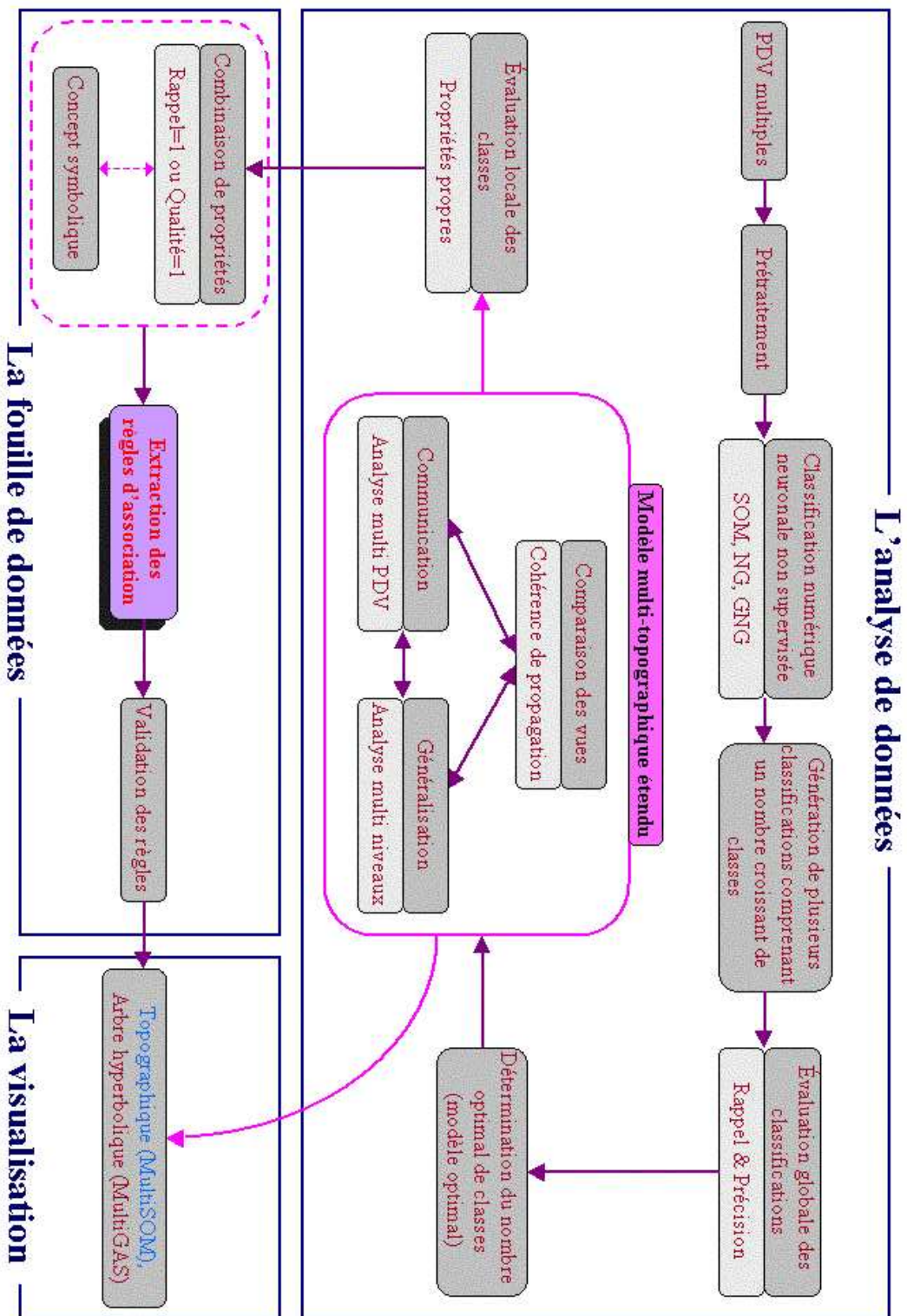


FIG. 1 – Paradigme de traitement de l'information orienté par les points de vue (MVDA).



Première partie

Traitement des données
multidimensionnelles

Chapitre 1

Représentation et exploitation des données documentaires

Sommaire

1.1	Introduction	9
1.2	L'indexation des documents	10
1.2.1	L'analyse de surface des documents	10
1.2.2	Modèles de représentation des documents	12
1.2.3	Notion de point de vue	12
1.3	Système de recherche documentaire	13
1.3.1	Les stratégies de recherche des documents	13
1.3.2	Réseaux bayésiens pour la recherche documentaire	15
1.3.3	Pondération par jugements de pertinence	19
1.3.4	Exploitation des relations sémantiques	20
1.3.5	Évaluation des systèmes de recherche documentaire	22
1.4	Extraction de connaissances dans les bases documentaires . .	24
1.4.1	Le processus d'extraction de connaissances	24
1.4.2	Fouille de données	24
1.4.3	Treillis de concepts et les motifs fréquents	25
1.4.4	Extraction de règles d'association	29
1.4.5	Inconvénients des modèles symboliques	32
1.5	Conclusion	33

1.1 Introduction

Le but de la recherche documentaire est celui d'isoler, dans un ensemble de documents souvent conséquent, un sous-ensemble de documents pertinents vis à vis du besoin d'un utilisateur, formulé par une requête généralement élémentaire. Selon ce processus, chaque document peut donc être considéré comme une unité d'information dont le système de recherche documentaire doit également, pour des raisons d'efficacité opérationnelle et d'adéquation avec les requêtes, manipuler une description élémentaire. D'un autre côté, la fouille de documents envisage une base de documents comme une vaste mine d'information brute dans laquelle il faut creuser et fouiller pour en trouver son «or». Cet or se matérialise comme des pépites de connaissances

[Simon, 2000] extraites des relations existantes entre les documents, mais surtout des concepts issus de leurs descriptions. L'extraction, ou la synthèse de l'information véhiculée par chaque document représente donc la première des problématiques à aborder que l'on se place dans le cadre de la recherche documentaire ou dans celui de la fouille de données.

Nous décrivons, dans le premier temps, les phases préliminaires utiles à la capture de cette information et à sa représentation selon un modèle de description global. Nous présentons également plusieurs des stratégies utilisées pour la recherche, la comparaison et l'adaptation des informations documentaires. Enfin, nous présentons de manière générale le processus d'extraction de connaissances dans les bases documentaires, en nous focalisant plus spécifiquement sur les techniques de fouille symboliques qui sont les techniques les plus couramment utilisées dans ce processus.

1.2 L'indexation des documents

L'indexation d'un document a pour but l'extraction de termes les plus importants et sa représentation sous forme d'un modèle conceptuel. Cette opération d'indexation, qui passe par une phase d'analyse et d'interprétation du contenu des documents, peut s'avérer plus ou moins complexe en fonction du modèle conceptuel retenu pour représenter les documents. Les contraintes d'efficacité, et les difficultés liées à l'interprétation du contenu des documents originaux, amènent une grande partie des systèmes de recherche documentaire (SRD) et des systèmes d'extraction de connaissances dans les bases documentaires (ECBD) à utiliser un modèle conceptuel simplifié et à procéder à une analyse de surface plutôt qu'à une analyse en profondeur des documents. Nous privilégierons donc ci-après la description de ce premier type d'analyse

1.2.1 L'analyse de surface des documents

Les analyses de surface ne permettent d'isoler que des caractéristiques élémentaires des documents. Ces analyses peuvent être basées sur un **langage contrôlé** ou sur un **langage libre** [Lamirel, 1995].

- Dans les analyses basées sur un langage contrôlé, des listes de candidats descripteurs susceptibles de représenter le contenu des documents sont construites a priori. La synthèse de contenu d'un document consiste alors à rechercher dans ces listes les candidats décrivant le mieux le contenu du document.
- Dans les analyses basées sur un langage libre, les descripteurs sont choisis a posteriori à partir de l'analyse du contenu des documents. La forme la plus simple de ce type d'analyse est l'extraction de termes d'index simples, ou unitermes. Cette méthode consiste à construire la liste des descripteurs d'un document en sélectionnant dans le contenu du document les mots ne relevant pas d'une liste de mots-outils (articles, prépositions, pronoms, ...). L'extraction d'unitermes a cependant pour inconvénient de produire des listes de descripteurs de taille importante. De plus, l'émiettement d'un texte en unités minimales a pour effet d'entraîner une polysémie maximale [Lamirel, 1995]. Des méthodes d'analyse plus sophistiquées permettent de conserver en partie l'intégrité des concepts. En effet, un concept est souvent exprimé comme une association de formes simples (par exemple, un cerf-volant, une base de données). A titre d'exemple, dans l'approche LEXINET, Chartron [Chartron, 1988] propose de repérer des groupes de termes fortement associés en utilisant des méthodes statistiques basées sur la variance. L'ensemble des

groupes de termes possibles est construit de manière incrémentale en partant de l'ensemble des unitermes adjacents, aux termes outils près.

Sélection des descripteurs

La sélection de descripteurs comprend habituellement plusieurs étapes de prétraitement de type d'analyse lexicale (élimination des suffixes, détection des racines équivalentes, etc.), dans le but de normaliser les termes extraits des documents. Elle fait également appel à des mécanismes de pondération, eux-mêmes basés sur la fréquence des termes d'index. Selon un premier objectif, la pondération pourra être utilisée comme une stratégie de sélection : seuls les termes les plus importants (c'est-à-dire les termes de plus forts poids) seront retenus parmi une liste de termes candidats. Selon un second objectif, qui peut être complémentaire du premier, la pondération servira à identifier l'importance de termes à la fois, les uns vis à vis des autres (pondération inter documents), et, vis à vis de chaque document (pondération intra document). Nous décrivons ci-après les deux mécanismes de pondération les plus couramment utilisés.

Pondération par $TF * IDF$: La fonction de pondération IDF (Inverse Document Frequency weighting) [Salton, 1971; Salton, 1989] s'inspire directement de l'approche de Sparck Jones en la simplifiant (cf. section 1.3.3). Elle consiste à attribuer les plus grands poids de pertinence aux termes d'index apparaissant dans un faible nombre de documents, ces derniers termes étant jugés porter le plus fort pouvoir d'information. La fonction IDF est définie comme :

$$IDF(t) = \log\left(\frac{N}{df(t)}\right) \quad (1.1)$$

Tel que t est un terme d'index, N est le nombre total de documents dans la collection, et $df(t)$ est le nombre de documents qui contiennent le terme d'index t . Le poids du terme d'index t dans un document d est alors défini comme :

$$W(d, t) = TF(d, t).IDF(t) \quad (1.2)$$

Tel que $TF(d, t) = \log(f(t, d))$ ¹, et $f(t, d)$ représente la fréquence d'apparition du terme d'index t dans le document d . Ainsi, un terme qui a une valeur de TF/IDF élevée doit être à la fois important dans le document dont ce terme est associé, et doit apparaître peu souvent dans les autres documents.

Pondération par l'entropie : La méthode de pondération par l'entropie est une méthode de pondération inter documents. Elle fonctionne selon un principe analogue à la méthode de pondération $TF * IDF$. Elle s'inspire de la fonction d'entropie de Shannon pour donner les poids les plus importants aux termes qui discriminent le mieux différentes zones de sujet. Elle est définie de la façon suivante [Kohonen *et al.*, 1996] :

Soit $r_d(t)$ la fréquence d'occurrence du terme t dans le document d de n dimensions $d =$

¹La linéarisation de la fréquence permet de répondre à loi de Zipf. Cette loi, qui se vérifie sur l'ensemble de données documentaire, stipule que plus un terme est fréquent dans un document, plus il a tendance à y être exponentiellement fréquent.

$\{t_1, t_2, \dots, t_n\}$ et $P_d(t)$ la probabilité que le terme t appartienne au document d . L'entropie H de ce terme est définie comme :

$$H(t) = - \sum_d P_d(t) \log(P_d(t)) \approx - \sum_d \frac{r_d(t)}{\sum_{d'} r_{d'}(t)} \log \frac{r_d(t)}{\sum_{d'} r_{d'}(t)} \quad (1.3)$$

Et le poids de pertinence $W(t)$ d'un terme t est défini comme :

$$W(t) = H_{max} - H(t) \quad (1.4)$$

où : $H_{max} = \log(n)$

1.2.2 Modèles de représentation des documents

Les modèles de représentation élémentaire de documents sont principalement deux types :

- **Modèle booléen** : Ce modèle a été initialement proposé par Salton [Salton, 1971]. Un document y est représenté par une liste de termes associés, tous les termes de la liste étant considérés comme des propriétés vraies, au sens booléen, pour le document. Ce principe rend cependant inexploitable la pondération des termes dans la représentation des documents.
- **Modèle vectoriel** : Ce modèle a également été proposé par Salton [Salton, 1971]. Un document y est représenté par un vecteur de termes pondérés dont la dimension correspond au nombre total de termes utilisés pour représenter les documents. Ainsi, un document d sera représenté par le vecteur de termes $d = (t_1, t_2, \dots, t_n)$ de dimension n .

1.2.3 Notion de point de vue

Afin d'augmenter la qualité et la granularité de l'analyse de données documentaires, Lamirel [Lamirel, 1995] a introduit la notion de points de vue multiples. Le principe de construction des points de vue est celui de séparer l'espace de description des documents en différents sous-espaces rattachés à des sous-ensembles de termes descriptifs différents. L'ensemble V des différents points de vue issus du sous-espace de description D d'un ensemble de documents peut être défini comme :

$$V = v_1, v_2, \dots, v_n, \quad v_i \in P(D), \quad \text{avec} \quad \bigcup_{i=1}^n v_i = D$$

Où chaque v_i représente un point de vue et $P(D)$ représente l'ensemble des parties de l'espace de description D des documents ; l'union des points de vue recompose alors l'espace initial D de description des documents.

Les différents points de vue correspondront alors aux différents sous-ensembles d'index rattachés aux différents champs de la description des documents : ainsi, si les champs « termes », « titre » et « auteurs » sont utilisés comme champs descripteurs d'un ensemble de documents, ils peuvent représenter trois points de vue différents sur cet ensemble de documents.

Cependant, Lamirel [Lamirel, 1995] a montré qu'il existe de nombreuses autres manières de définir des points de vue, comme par exemple, d'utiliser des relations non descriptives entre les données, telles que des liens de citations ou des relations de proximité.

Le principe général des points de vue s'applique particulièrement bien au modèle vectoriel, et permet, comme nous le montrons ci-après, d'en améliorer le comportement.

1.3 Système de recherche documentaire

Le processus central de la recherche documentaire est la stratégie de recherche proprement dite. Cette opération, également appelée mise en correspondance, consiste à évaluer la pertinence des documents du fonds vis à vis d'une requête formulée par un utilisateur [Van Rijsbergen, 1979].

1.3.1 Les stratégies de recherche des documents

Nous illustrerons ci-après quatre stratégies de recherche classiques telles que la stratégies de recherche booléenne, vectorielle, par classes et probabiliste.

Stratégie de recherche booléenne

Les requêtes booléennes sont exprimées sous forme de termes d'index qui peuvent être combinés par les opérateurs logiques usuels AND, OR, et NOT. La recherche booléenne recherchera donc les documents qui sont «vrais» pour une requête donnée en se basant sur des opérations ensemblistes élémentaires relatives aux documents associés au termes de la requête. Par exemple, le résultat de la requête $Q = (K_1 \text{ AND } K_2) \text{ OR } (K_3 \text{ AND } (\text{NOT } K_4))$ sera la liste de documents L tel que :

$$L = (L_{K_1} \cap L_{K_2}) \cup (L_{K_3} - L_{K_4})$$

où L_{K_i} correspond à l'ensemble de documents indexés par le terme K_i .

La syntaxe des requêtes booléennes rend cependant celles-ci difficiles à maîtriser pour un débutant. De plus, la taille du résultat de recherche n'est pas facile à contrôler. Ce problème est compliqué par le fait que ces résultats ne peuvent pas être classés selon un jugement de pertinence.

Stratégie de recherche vectorielle

Selon le modèle vectoriel, les requêtes et les documents sont représentés de manière homogène par des vecteurs de termes. La mise en correspondance vectorielle peut donc faire appel à de nombreuses fonctions de similarité. La fonction la plus utilisée est la fonction de corrélation Cosinus de Salton, ou mesure cosinus, soit :

$$Sim(Q, D) = \frac{\sum_{i=1}^t q_i d_i}{(\sum_{i=1}^t (q_i)^2 \sum_{i=1}^t (d_i)^2)^{\frac{1}{2}}} \quad (1.5)$$

$$Q = (q_1, q_2, \dots, q_t), D = (d_1, d_2, \dots, d_t),$$

q_i, d_i : les poids numériques associés au terme i .

L'équation 1.5 mesure le cosinus de l'angle formé par les vecteurs de termes de la requête Q et celui du document D . Une valeur de 1 signifie que les deux vecteurs sont identiques (les deux vecteurs ont les mêmes termes, avec les mêmes poids), une valeur de 0 signifie que il y a aucun terme partagé par les deux vecteurs.

La stratégie de recherche vectorielle admet un seuil de pertinence T . Seuls les documents qui ont une similarité supérieure à ce seuil sont conservés, ce qui définit l'ensemble résultat B ,

$$B = \{D_i \mid Sim(Q, D_i) > T\}; \quad T : \text{seuil}$$

Les documents de B sont finalement rangés dans l'ordre croissant ou décroissant de la valeur de la fonction : $Sim(Q, D_i)$.

Du fait qu'elle utilise l'espace de description global de documents, la mise en correspondance vectorielle classique génère du bruit issu des informations inutiles vis à vis du besoin d'utilisateur contenues dans les descriptions des documents. Pour résoudre ce problème Lamirel [Lamirel, 1995] a proposé une extension de la mise en correspondance vectorielle à une mise en correspondance partitionnée en domaines actifs, basée sur les points de vue. Cette stratégie, consiste à ne mettre à contribution, lors de la mise en correspondance, que les points de vue dont les termes apparaissent à la fois dans la requête et dans les documents. Selon le mode opératoire choisi, elle peut éventuellement impliquer une phase supplémentaire de synthèse de résultats issus des différents points de vue sélectionnés pour produire un classement global de pertinence à partir des classements partiels. Cette phase s'opère par le calcul des rangs moyens pondérés des documents D pour une requête Q , qui s'exprime sous la forme suivante :

$$r(Q, D) = \frac{\sum_{v_i \in V} \alpha_{v_i} r_{v_i}(Q, D)}{\sum_{v_i \in V} \alpha_{v_i}} \quad (1.6)$$

où α_{v_i} représente le poids de pertinence du domaine v_i lors de la mise en correspondance, $r_{v_i}(Q, D)$ est le rang du document D pour la requête Q selon la mise en correspondance partielle associée au domaine actif v_i .

Stratégie de recherche orientée par les classes

Cette démarche permet de simplifier la mise en correspondance. Elle repose sur la vérification par la base documentaire interrogée de l'hypothèse suivante, dite hypothèse de clusterisation [Van Rijsbergen, 1979] :

«Les documents fortement associés entre eux tendent à être pertinents pour les mêmes requêtes.»

La recherche orientée par les classes, elle repose sur l'utilisation préalable d'une méthode de classification (cf. chapitre 2). Le système SMART [Salton, 1971; Salton, 1983] est un exemple caractéristique de cette stratégie. Ainsi, dans ce système, le processus de mise en correspondance entre une requête et les documents s'opère en deux phases :

- Une première phase de mise en correspondance est opérée entre la requête et l'ensemble des classes représentant le contenu de la base. Cette phase a pour rôle de sélectionner la classe la plus pertinente.
- La seconde phase de la mise en correspondance est opérée entre la requête et les documents appartenant à la classe sélectionnée lors de la première phase. Cette phase a pour rôle de fournir un classement final de pertinence de ces documents.

Afin d'éviter le risque de produire des classes trop générales pour être exploitables par la mise en correspondance, cette stratégie implique l'utilisation d'une méthode de classification robuste. Cependant, comme le montre Lamirel [Lamirel, 1995], le choix de la méthode ne s'avère pas suffisant pour éliminer le bruit inhérent à une approche de classification globale. Ce dernier propose donc également d'appliquer le principe de partitionnement en points de vue à la mise en correspondance à base de classes, ce qui revient à définir une classification des documents différente pour chaque point de vue défini sur les documents. L'avantage de cette partition est

qu'elle permet d'obtenir des classes plus spécifiques qu'une méthode globale. Elle est de plus compatible avec la notion de domaine actifs définie précédemment.

Stratégie de recherche probabiliste

Le modèle probabiliste constitue un outil puissant pour les systèmes de recherche documentaire, car ils permettent de traiter d'une manière efficace l'incertitude intrinsèque au processus de recherche documentaire. Le principe de base du modèle de mise en correspondance probabiliste élémentaire est d'estimer la probabilité de pertinence des documents qui satisfont une requête donnée [Van Rijsbergen, 1975]. La règle de décision fondamentale de ce modèle est qu'un document d ne peut être classé comme pertinent pour une requête Q donnée que s'il est possible de vérifier l'inégalité stricte suivante :

$$P(Pert|d) > P(NonPert|d) \quad (1.7)$$

où *Pert* représente l'événement de pertinence et *NonPert* l'événement complémentaire de non pertinence, $P(Pert|d)$ représente la probabilité de pertinence conditionnelle au choix de d , et, $P(NonPert|d)$, la probabilité de non pertinence conditionnelle au choix de d .

Le modèle de mise en correspondance probabiliste élémentaire a été étendu et amélioré en utilisant le formalisme des réseaux bayésiens [Turtle and Croft, 1991]. L'avantage apporté par les réseaux bayésiens vis à vis de l'ensemble de modèles existants, est de permettre de combiner de manière homogène des informations provenant de différentes sources (ou points de vue) pour restituer les documents qui seraient les plus pertinents pour une requête donnée. De plus, ils permettent plus spécifiquement d'optimiser le fonctionnement du modèle de mise en correspondance probabiliste élémentaire. En effet, les calculs de probabilité a posteriori de documents pertinents, pour une requête donnée, peuvent être optimisés grâce à la topologie spécifique de ce type de réseau, ceci en considérant le tableau des probabilités conditionnelles qu'il est possible d'associer à cette topologie.

Nous présentons ci-après les bases de ce formalisme que nous utiliserons également par la suite pour généraliser l'interaction entre des points de vue multiples sur les documents.

1.3.2 Réseaux bayésiens pour la recherche documentaire

Les réseaux bayésiens introduisent une combinaison entre la théorie des graphes et celle des probabilités, ce qui fait d'eux des outils intuitifs et naturels pour traiter les données complexes et incertaines.

Principes des réseaux bayésiens

Les réseaux bayésiens sont des graphes dirigés et acycliques (DAG), où les nœuds V représentent les variables du problème à résoudre, et les arcs dirigés E représentent des liens de dépendance directe exprimant des relations de causalité, sous forme de dépendances conditionnelles, entre les variables du réseau $G = (V, E)$ [Pearl, 1988; Jensen, 1996]. Un réseau bayésien peut donc présenter deux aspects :

- Qualitatif, formulé par la structure du graphe G , qui exprime les relations entre les différentes variables du réseau (en illustrant les dépendances/indépendances entre les variables).

- Quantitatif, qui permet de spécifier la distribution de probabilité conditionnelle de chaque variable et qui quantifie ainsi les relations entre les nœuds.

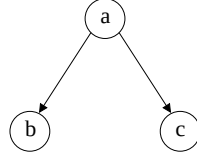


FIG. 1.1 – Exemple de réseau bayésien élémentaire représentant les dépendances conditionnelles entre trois variables

La distribution de probabilité conditionnelle d'une variable donnée $X_i \in V$ peut s'écrire comme :

$$P(X_i|X_{i-1}, \dots, X_1) = P(X_i|\text{parents}(X_i)) \quad (1.8)$$

où $\text{parents}(X_i)$ représente l'ensemble des parents de la variable X_i dans G . Ainsi, un arc allant d'un nœud X_a à un nœud X_b exprime le fait que X_b dépend directement de X_a (lien de cause à effet). L'absence d'arc ne renseigne alors que sur la non-existence d'une dépendance directe.

La distribution de probabilité jointe de l'ensemble des variables $X_1, X_2, \dots, X_n$ d'un réseau bayésien peut, quant à elle, s'écrire comme :

$$P(X_1, X_2, \dots, X_n) = \prod_i P(X_i|\text{parents}(X_i)) \quad (1.9)$$

Cette formule permet de simplifier considérablement le calcul de la probabilité jointe de l'ensemble $X_1, X_2, \dots, X_n$ en supposant que toute variable X_i ne dépend que de ses prédécesseurs directs (ses parents). En effet, la connaissance de toute autre variable constitue une information redondante et inutile.

La figure 1.1 illustre le principe de ce réseau bayésien dans un cas simple. L'ensemble des arcs, qui constitue un DAG, est donné comme suit : $E = \{(X_a, X_b), (X_a, X_c)\}$, soit, $\text{parents}(X_b) = \text{parents}(X_c) = \{X_a\}$. La distribution de probabilité jointe des variables $\{X_a, X_b, X_c\}$ est alors calculée comme suit :

$$P(X_a, X_b, X_c) = P(X_b|X_a) \cdot P(X_a) \cdot P(X_c|X_a)$$

La construction d'un réseau bayésien implique de prendre en compte trois problèmes principaux :

1. **La structure** : détermination du graphe causal qui permet de modéliser le problème considéré.
2. **L'inférence** : calcul des probabilités des différentes variables. Pour ce faire, il est nécessaire de connaître les valeurs de probabilités de certaines variables appelées «variables observées», les autres étant «cachées» étant donné que l'on ne connaît pas encore leurs valeurs de probabilités. Dans le cas où il n'existe pas de variables observées, des connaissances a priori sont indispensables. Le processus d'inférence a pour but de propager l'information disponible sur l'ensemble du réseau jusqu'à obtenir les valeurs des probabilités relatives à toutes les variables du réseau. Les algorithmes d'inférence les plus connus sont ceux de Pearl, connu sous le nom de «Passage de Message» [Pearl, 1988], et, de Jensen, connu sous le nom d'«Arbre de Jonction» [Jensen *et al.*, 1990].

3. **L'apprentissage** : estimation de la structure du réseau et/ou de ses distributions de probabilités conditionnelles. Ce problème constitue l'un des problèmes les plus compliqués à résoudre (pour plus de détails voir [Stephenson, 2000; Jamoussi, 2004]).

En résumé, utiliser un réseau bayésien consiste à déterminer tout d'abord sa structure et à l'employer ensuite pour l'inférence en partant des données initialement observées.

Modèle de réseau bayésien à deux niveaux pour la recherche documentaire

Deux ensembles différents de variables (nœuds dans le graphe) peuvent être distingués dans un système de recherche documentaire (SRD) [de Campos *et al.*, 2002] :

- L'ensemble T de M termes T_i , dans une collection donnée,
- L'ensemble D des N documents D_j qui composent cette collection.

Chaque terme T_i , consiste en une variable aléatoire binaire prenant des valeurs dans l'ensemble $\{\bar{t}_i, t_i\}$, où $\bar{t}_i$ signifie que «le terme T_i n'est pas pertinent», et t_i signifie que «le terme T_i est pertinent».

De la même façon, chaque document D_j consiste en une variable aléatoire binaire prenant des valeurs dans l'ensemble $\{\bar{d}_j, d_j\}$, où $\bar{d}_j$ signifie que «le document D_j n'est pas pertinent pour une requête donnée», et d_j signifie que «le document D_j est pertinent cette même requête».

La structure du réseau bayésien à deux niveaux et sa topologie peuvent donc être déterminées en plusieurs phases. La première phase consiste à faire les hypothèses suivantes :

- Pour chaque terme associé à un document, il existe un lien entre le nœud représentant ce terme et chaque nœud représentant un document qui contient ce terme.
- Les relations entre les documents se produisent uniquement par les termes associés à ces documents.
- Les documents sont conditionnellement indépendants étant donné leurs termes associés. Ainsi, si les valeurs de pertinence (ou non pertinence) pour tous les termes associés à un document D_i sont connues, l'inférence de la pertinence de D_i ne sera pas affectée par la pertinence (ou non pertinence) d'un autre document D_j .

A l'issue de cette première phase, la structure du réseau peut être déterminée partiellement, comme suit :

- Les liens joignant les termes et les documents doivent être dirigés des nœuds termes vers des nœuds documents.
- Il n'y a pas de liens entre les nœuds documents.

Par conséquent, l'ensemble des parents d'un nœud document D_j , $Pa(D_j)$ représentera l'ensemble des nœuds termes qui appartiennent à D_j , c'est-à-dire, $Pa(D_j) = \{T_i \in T | T_i \in D_j\}$.

Pour déterminer complètement la topologie du réseau, l'hypothèse additionnelle suivante est opérée : les termes sont marginalement indépendants, ce qui implique qu'il n'y a pas de liens entre les nœuds termes (tous les nœuds termes sont des nœuds racines). Dans ce cas, nous obtenons un réseau composé de deux niveaux simples avec des liens des nœuds termes vers les nœuds documents (cf. figure 1.2).

Pour spécifier complètement le réseau, la phase finale consiste à estimer les distributions de probabilité associées à chaque nœud. Deux cas différents doivent être considérés :

- Nœuds termes : dans ce cas, les distributions marginales sont définies comme suit :

$$p(t_i) = \frac{1}{M} \text{ et } p(\bar{t}_i) = \frac{M-1}{M} \quad (1.10)$$

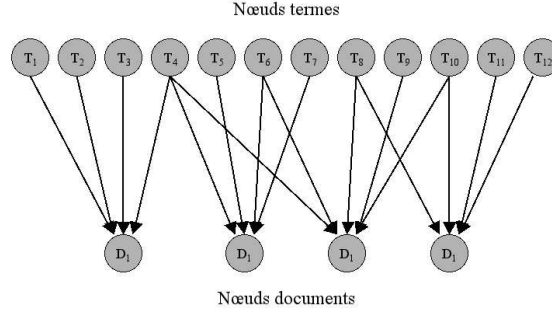


FIG. 1.2 – Réseau Bayésien à deux niveaux

- Nœuds documents : dans ce cas, un nouveau modèle canonique général permettant d'évaluer la probabilité conditionnelle de la pertinence de D_j comme suit :

$$p(d_j|Pa(D_j)) = \sum_{\substack{T_i \in D_j \\ t_i \in pa(D_j)}} w_{ij} \quad (1.11)$$

où les poids w_{ij} doivent vérifier $0 \leq w_{ij}$ et $\sum_{T_i \in D_j} w_{ij} \leq 1$.

Selon ce modèle, l'augmentation du nombre de termes pertinents dans $Pa(D_j)$ augmentera donc corrélativement la probabilité de pertinence de D_j .

Étant donné une requête Q , la stratégie de recherche basée sur ce réseau consiste à placer les évidences (les termes T_Q appartenant à Q) dans les nœuds termes du réseau, c'est-à-dire mettre leurs états à la valeur t_Q (valeur de pertinence). A l'issue du processus d'inférence, on obtient la probabilité de pertinence de chaque document, sachant que les termes de la requête sont également pertinents $p(d_j|Q)$. Pour finaliser le processus d'évaluation, les documents sont triés selon leur valeur de probabilité résultante.

Pour résoudre le problème de l'estimation complexe de probabilité conditionnelle du au grand nombre de parents dans le réseau, une méthode d'inférence spécifique a été proposée par [de Campos *et al.*, 2002]. Cette méthode prend en considération la topologie du réseau et la fonction de probabilité utilisée pour les nœuds documents. Le processus de propagation, défini à l'équation 1.11, est alors substitué par une seule évaluation pour chaque nœud document :

$$p(d_j|Q) = \sum_{T_i \in D_j} w_{ij} p(t_i|Q) \quad (1.12)$$

Étant donné que les termes sont marginalement indépendants, en utilisant l'équation 1.10, on obtient :

$$p(t_i|Q) = \begin{cases} 1 & \text{if } T_i \in Q \\ \frac{1}{M} & \text{if } T_i \notin Q \end{cases} \quad (1.13)$$

Ainsi, le calcul de $p(d_j|Q)$ peut être effectué comme suit :

$$p(d_j|Q) = \sum_{T_i \in D_j \cap Q} w_{ij} + \frac{1}{M} \sum_{T_i \in D_j \setminus Q} w_{ij} \quad (1.14)$$

Nous avons donné un exemple de l'utilisation des réseaux bayésiens dans le domaine de la recherche documentaire. Cependant, leur puissance expressive permet de les exploiter dans de nombreux autres domaines tels que, la recherche d'information (RI) [Ribeiro and Muntz, 1996; de Campos *et al.*, 2000; de Campos *et al.*, 2002], la fouille de données sur le Web [Tsirikas and Lalmas, 2002], la reconnaissance de la parole [Zweig and Russell, 1998], etc.

La mise en correspondance vectorielle d'après le réseau d'inférence

Dans le système de recherche documentaire (SRD) les documents et les requêtes sont représentés par des vecteurs de termes d'index (cf. section 1.2.2). La similarité entre un document D et une requête Q est donc usuellement estimée à partir de la valeur de cosinus entre les deux vecteurs (cf. équation 1.5).

Cependant, comme le montre Ribeiro et Muntz [Ribeiro and Muntz, 1996], la fonction de similarité entre un document D et une requête Q peut également être calculée en cherchant une relation de type $P(D|Q)$. En effet, en utilisant la théorème de Bayes, on peut écrire $P(D|Q) = \frac{P(D \wedge Q)}{P(Q)}$. Puisque $P(Q)$ est constante pour tous les documents dans la collection, on peut écrire $P(D|Q) \propto P(D \wedge Q)$ ce qui veut dire que la similarité affectée à un document D pour une requête Q est directement proportionnel à $P(D \wedge Q)$. Par conséquent, Ribeiro et Muntz dans [Ribeiro and Muntz, 1996] ont prouvé que :

$$P(D|Q) = K \times \text{Sim}(D, Q) \quad (1.15)$$

où K est constant.

Le rang des documents pertinents sachant une requête peut donc être défini par le rang de $\text{Sim}(D, Q)$ (cf. équation 1.15).

1.3.3 Pondération par jugements de pertinence

Sparck Jones a prouvé expérimentalement que si un terme d'index apparaît dans m documents dans une base comprenant N documents, alors le poids $\log(N/m) + 1$ mène à une recherche plus pertinente que si le terme était utilisé sans poids.

Cette approche s'inspire de modèle probabiliste, elle présuppose qu'un ensemble de jugements de pertinence a pu être opéré sur un ensemble de requêtes types [Robertson and Jones, 1976; Jones, 1979].

Soit un terme d'index t et une requête Q donnés, on définit alors les variables suivantes :

- N : Nombre de documents dans la collection,
- R : Nombre de documents pertinents pour Q ,
- m : Nombre de documents indexés par t ,
- r : Nombre de documents pertinents indexés par t ,

La table d'éventualité de la distribution de t pour un document donné est présentée dans le tableau 1.1.

Cette table permet de définir de nombreux types de fonctions de pondération tels que :

$$w^1 = \log\left(\frac{\frac{r}{R}}{\frac{m}{N}}\right) \quad F1$$

		Pertinence de document		
Indexation de document		Pertinent	Non-pertinent	
	$X_t = 1$	r	$m - r$	m
	$X_t = 0$	$R - r$	$N - m - R + r$	$N - m$
		R	$N - R$	N

TAB. 1.1 – La table d'éventualité de la distribution de document pour t .

$$w^2 = \log\left(\frac{\binom{r}{R}}{\binom{m-r}{N-R}}\right) \quad F2$$

La fonction $F1$ représente le rapport de la proportion de documents pertinents dans lesquels t apparaît, relativement à son apparition dans la collection entière.

La fonction $F2$ représente le rapport de la proportion de documents pertinents dans lesquels t apparaît, relativement à la proportion de documents non pertinents dans lesquels t apparaît également.

1.3.4 Exploitation des relations sémantiques

La plupart des méthodes de recherche documentaire dépendent des correspondances exactes entre les termes de requêtes d'utilisateurs et les termes des documents. Une raison de ceci est que les modèles de représentation des documents traitent les termes comme s'ils étaient orthogonaux ou indépendants, alors qu'ils ne le sont pas. Prenons comme exemple, les termes «automobile», «voiture», «conducteur», et «éléphant». Les termes «automobile» et «voiture» sont des synonymes, le terme «conducteur» représente un concept approchant, mais le terme «éléphant» est indépendant de ceux-ci. Si une requête traitant des éléphants ne doit ramener que les documents concernant les éléphants, il serait préférable qu'une requête traitant des automobiles ramène également des documents concernant les voitures, et à un degré moindre des documents concernant les conducteurs.

Plusieurs méthodes permettent d'exploiter les corrélations entre les termes pour améliorer l'exhaustivité de la recherche documentaire. Le principe de ces méthodes est d'encoder les termes de telle sorte que les termes similaires soient représentés par des codes similaires. Le but de ces méthodes est principalement d'améliorer l'exhaustivité du résultat lors d'une recherche de document. Les méthodes LSI et MatchPlus que nous présentons, ci-après, en illustrent ce principe fondateur :

Indexation sémantique latente (LSI) : Le modèle LSI a été décrit en détail par [Deerwester *et al.*, 1990; Dumais, 1991]. LSI représente une extension de la méthode de recherche vectorielle [Salton, 1983] dans laquelle les dépendances entre les termes sont explicitement prises en considération dans la représentation et exploitées au moment de la recherche. LSI est basé sur le principe qu'il existe une structure implicite ou «latente» dans le modèle d'un terme qui est utilisé dans les documents, et que des techniques statistiques peuvent être utilisées pour estimer cette structure.

LSI utilise la technique de décomposition en valeurs singulières (SVD) [Golub and Loan, 1991] pour modéliser les associations entre les termes et les documents comme suit :

Soit X une matrice de dimension $n \times N$ de rang $r \leq \min(n, N)$, qui représente la matrice termes-documents, où n est le nombre de terme, N est le nombre de documents. Il existe alors un triplet de matrices uniques U , V et Σ telles que

$$X_{n \times N} = U_{n \times r} \Sigma_{r \times r} V_{r \times N}^T \quad (1.16)$$

La matrice U de dimension $n \times r$ contient les vecteurs singuliers gauches et est formée des vecteurs propres de $X^T X$; la matrice V de dimension $r \times N$ contient les vecteurs singuliers droites et est formée des vecteurs propres de XX^T ; $\Sigma = \text{diag}(\alpha_i)$ contient les valeurs singulières de X sur la diagonale, qui sont elles mêmes les racines de ses valeurs propres.

S'il n'existe aucune association entre les termes, alors il n'y aura aucune possibilité de réduire la matrice Σ , et par conséquent, d'exploiter la matrice termes-documents pour améliorer la recherche documentaire.

S'il existe par contre des associations entre les termes, la matrice Σ pourra être réduite à une matrice de dimension $k \times k$ avec $k < r$. Cette matrice sera ensuite utilisée pour générer une matrice termes-documents transformée et réduite.

- Plutôt que d'être représentés par des vecteurs indépendants, les documents seront représentés par des valeurs continues dans la matrice termes-documents transformée, ou matrice de pseudo-documents.
- Si deux termes sont utilisés dans des documents similaires, alors, ils auront des vecteurs similaires dans la matrice de pseudo-documents.
- Toute requête Q soumise sera préalablement transformée en un pseudo-document d_Q par la formule suivante [Deerwester *et al.*, 1990] :

$$d_Q = X_Q^T U \Sigma^{-1} \quad (1.17)$$

- Finalement, la similarité entre deux pseudo-documents pourra être calculée comme suit :

$$\text{Sim}(d, d') = d \Sigma^2 d' \quad (1.18)$$

MatchPlus : Contrairement à la méthode LSI, la méthode MatchPlus [Gallant, 1991; Gallant *et al.*, 1992; Gallant *et al.*, 1993; Caid *et al.*, 1995] utilise des données d'apprentissage pour créer des vecteurs de contexte représentant les termes, les documents et les requêtes. Ainsi, des techniques d'apprentissage de type Bootstrapping sont utilisées pour ajuster les vecteurs de termes de telle manière que les termes qui existent dans des contextes similaires possèdent des vecteurs associés décrivant des directions similaires. Les vecteurs de contexte liés aux documents ou aux requêtes sont créés à partir de réseaux de neurones non supervisés eux-mêmes construits à partir des vecteurs de contexte des termes précédemment appris. La représentation résultante peut être utilisée pour la recherche documentaire et pour la classification des documents.

Dans les deux méthodes mentionnées ci-dessus, les documents sont cependant encodés de manière globale comme une somme des représentations de leurs termes. Quand des documents sont encodés de telle manière leur similarité n'est plus directement liée aux termes qu'ils contiennent, mais plutôt au degré de corrélation entre les termes eux-mêmes. Ce phénomène introduit un biais important lors de la mise en correspondance.

Il est à noter que les thésaurus peuvent également être utilisés pour trouver des relations sémantiques entre les termes d'index [Kohonen *et al.*, 2000; Lagus *et al.*, 1999].

Les méthodes d'exploitation des relations sémantiques précitées s'avèrent cependant souvent inefficaces du fait que le contenu des documents est généralement redondant. Ainsi, leur similarité ne dépend plus que très partiellement des relations sémantiques que les termes d'index élémentaires entretiennent entre eux.

1.3.5 Évaluation des systèmes de recherche documentaire

L'objectif d'un SRD est de restituer à un utilisateur l'information qu'il recherche. L'évaluation est une étape importante, puisqu'elle va permettre de caractériser les performances d'un système donné, et donc de fournir des éléments de comparaison entre systèmes.

On pourra dire qu'un système de recherche documentaire est efficace s'il permet de restituer un maximum d'informations pertinentes et un minimum d'informations non pertinentes. La première caractéristique se réfère au taux de rappel, la seconde au taux de précision [Van Rijsbergen, 1979].

Le taux de rappel permet de connaître le pourcentage de documents pertinents qui ont été retrouvés ; le taux de précision évalue le pourcentage de documents restitués qui sont effectivement pertinents (cf. figure 1.3) :

$$\text{Taux de Rappel} = TR = \frac{DPR}{DP} \quad (1.19)$$

$$\text{Taux de Précision} = TP = \frac{DPR}{DR} \quad (1.20)$$

où DPR est le nombre de documents pertinents restitués, DP est le nombre de documents pertinents, et DR est le nombre de documents restitués.

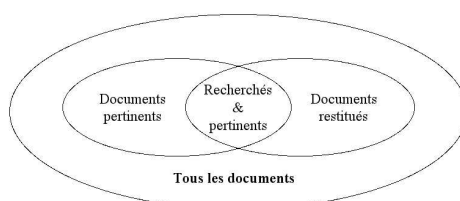


FIG. 1.3 – Evaluation de SRD

Ces deux taux évoluent généralement en sens inverse. À la vue des formules de calcul, on comprend que le taux de rappel puisse augmenter avec le nombre de documents restitués (DP est constant et DPR est une fonction croissante de DR), mais également que dans ce cas, le risque de voir le taux de précision diminuer soit important.

Le calcul d'une paire (taux de rappel, taux de précision) est réalisé à partir de la liste ordonnée des documents restitués. On parcourt cette liste à partir du document le plus ressemblant à la requête, jusqu'à obtenir le nombre de documents pertinents nécessaires pour arriver au taux de rappel recherché. À partir de cette sous-liste, le taux de précision associé est calculé à partir de la sous-liste obtenue.

Supposons par exemple que pour une requête Q , l'ensemble des documents pertinents soient constitué par les documents : [1], [3], [5], et [6], soit $DP = 4$. Dans le tableau (a) de la figure 1.4,

les documents rendus par le système sont rangés par ordre décroissant de ressemblance avec la requête. Le symbole X signifie que le document est pertinent. Le tableau (b), de la même figure, décrit les différentes valeurs possibles du taux de rappel et les valeurs correspondantes du taux de précision.

A titre d'exemple, obtenir un taux de rappel de $2/4$, soit $DPR = 2$, n'est possible que considérant les trois premiers documents restitués. Le taux de précision est alors de $DPR/DR = 2/3$.

Documents restitués		Rappel	Précision
[1]	X	1/4 = .25	1/1 = 1
[2]			
[3]	X	2/4 = .50	2/3 = .66
[4]			
[5]	X	3/4 = .75	3/5 = .6
[6]	X	4/4 = 1	4/6 = .66
[7]			
[8]			

(a)
(b)

FIG. 1.4 – Exemple de taux de rappel et de précision

La F-mesure constitue un compromis entre le rappel et la précision, faisant intervenir les deux termes selon la formule suivante :

$$F = \frac{2(R \times P)}{(R + P)} \quad (1.21)$$

Afin de calculer des valeurs moyennes pour une série de requêtes, les mesures de micro précision/rappel ou les mesures de macro précision/rappel peuvent être utilisées [Rocchio, 1971; Salton, 1983]. Les mesures de micro précision/rappel (cf. équation 1.22) privilégient les performances du système de recherche documentaire en fusionnant les résultats obtenus pour les différentes requêtes. Ces mesures permettent donc d'atténuer les erreurs dues aux requêtes fournissant naturellement peu des documents pertinents. Les mesures de macro précision/rappel (cf. équation 1.23) privilégient la qualité globale des résultats fournis à l'utilisateur en moyennant les résultats obtenus pour chaque requête.

Ces différentes mesures sont définies comme suit :

$$\text{micro-précision} = \frac{\sum_Q DPR_Q}{\sum_Q DR_Q} \quad \text{et} \quad \text{micro-rappel} = \frac{\sum_Q DPR_Q}{\sum_Q DP_Q} \quad (1.22)$$

où DPR_Q est le nombre de documents pertinents restitués par une requête Q , DR_Q est le nombre de documents restitués par la même requête Q et DP_Q est le nombre de documents pertinents également pour la requête Q .

$$\text{macro-précision} = \frac{\sum_Q TP_Q}{N_Q} \quad \text{et} \quad \text{macro-rappel} = \frac{\sum_Q TR_Q}{N_Q} \quad (1.23)$$

où N_Q le nombre de requêtes utilisées par l'utilisateur, TP_Q est le taux de précision pour la requête Q et TR_Q est le taux de rappel pour la même requête Q .

Dans la suite de ce travail, nous nous intéressons plus particulièrement aux mesures de macro précision/rappel.

1.4 Extraction de connaissances dans les bases documentaires

L'extraction de connaissances dans les bases de données (ECBD) (ou knowledge discovery in databases (KDD)) est un processus non trivial de découverte d'information implicites, précédemment inconnues et potentiellement utiles concernant les grandes bases de données [Fayyad *et al.*, 1996; Han *et al.*, 2001]. Nous en décrivons les grands principes ci-après.

1.4.1 Le processus d'extraction de connaissances

L'ECBD est un processus itératif qui se décompose en plusieurs étapes. Chaque étape de ce processus nécessite l'intervention de l'utilisateur qui est amené à prendre des décisions et à effectuer des choix (processus interactif). La découverte de nouvelles connaissances n'implique pas nécessairement l'arrêt du processus, mais conduit généralement à reprendre celui-ci sur de nouvelles données (processus incrémental) [Simon, 2000]. La figure 1.5 illustre les différents étapes de ce processus dans lequel il faut, dans l'ordre :

1. Sélectionner les données à analyser.
2. Prétraiter les données sélectionnées (analyse de surface).
3. Transformer les données prétraitées sous forme d'un modèle de représentation pour l'application de méthodes de fouille.
4. Appliquer la méthode de fouille en fonction des buts recherchés (classification, association, etc.).
5. Valider et interpréter les résultats issus de la fouille. Les informations extraites peuvent être visualisées ; celles sans intérêt ou redondantes supprimées ; et les résultats réécrits sous forme plus compréhensible. L'interprétation est susceptible de remettre en cause tout ou partie du processus.
6. Utiliser la connaissance découverte : celles-ci peuvent être incorporées en système d'information en vérifiant qu'elles sont cohérentes avec les connaissances existantes. Elles peuvent également être analysées et transmises à l'utilisateur, ou bien servir à prendre des décisions.

1.4.2 Fouille de données

La fouille de données (FD) ou data mining (DM) [Fayyad *et al.*, 1996; Han *et al.*, 2001], qui représente l'étape principal du processus d'ECBD, consiste de manière générale à «résoudre des problèmes par l'analyse des données couramment présentes dans les bases de données» [Witten and Frank, 2000]. D'une manière plus précise, cette opération revient à trouver, à partir des données, des modèles (ou patterns) qui sont significatifs et valides, selon des mesures d'importance et de validité définies par l'utilisateur.

Tâches de la fouille de données

La fouille de données comporte plusieurs tâches importantes qui sont nécessaires pour avoir une bonne connaissance sur la base de données traitée, telles que :

- La description : il s'agit de résumer les données en en donnant une description abstraite. Les techniques utilisées sont des méthodes de classification non supervisées numériques, dont l'objectif est de construire un ensemble de classes à partir d'un ensemble de données

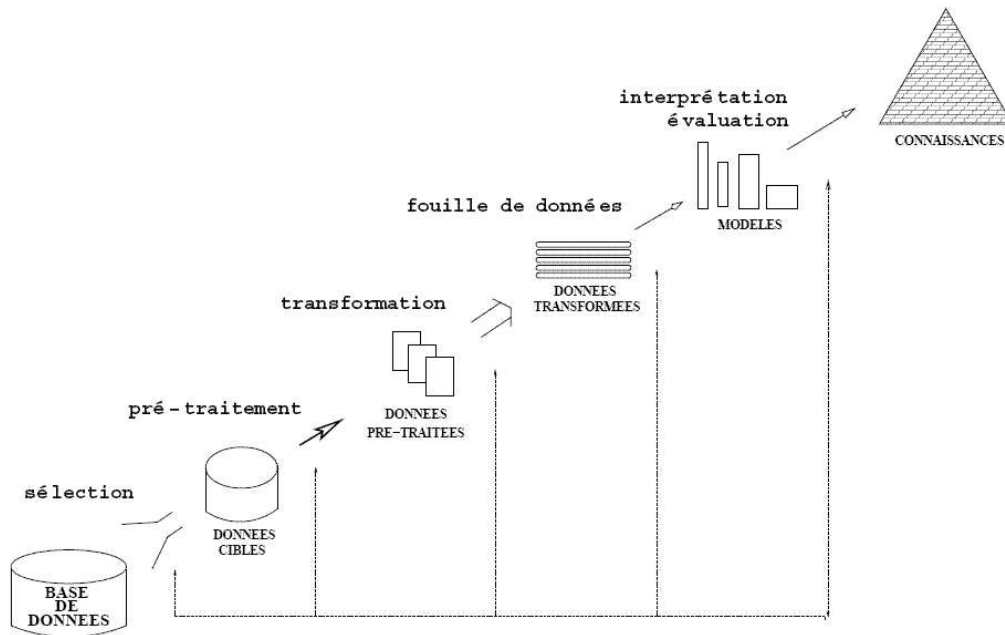


FIG. 1.5 – Les étapes du processus d'extraction de connaissances

non structuré (cf. chapitre 2), et des méthodes symboliques, telles que les treillis de concept (cf. section 1.4.3)

- La prédiction : le problème consiste à savoir classer correctement de nouveaux exemples grâce aux règles construites à partir du jeu d'essai. Ces méthodes sont basées sur une classification supervisée (les classes sont déjà prédéfinies par l'utilisateur, et décrites par des exemples). Ce sont des techniques d'induction. Parmi celles-ci on peut citer, du côté symbolique, les systèmes d'induction de règles et les arbres de décision, et, du côté numérique les réseaux de neurones supervisés, tel que le perceptron. Pour ce problème, les méthodes symboliques présente cependant l'avantage de produire des modèles facilement compréhensibles pour un utilisateur et de permettre d'intégrer les connaissances du domaine.
- L'analyse de corrélation : cette technique permet de mettre en évidence les liens existents entre des objets donnés, les attributs ou les objets et les attributs. Elle détermine des associations de la forme : « Si cet ensemble de faits est vérifié, alors, dans $x\%$ de cas, cet autre ensemble de faits l'est également ». Les méthodes utilisées sont, du côté numérique, les régressions, et, du côté symbolique, la recherche de règles d'association ou de motifs fréquents (cf. sections 1.4.3 et section 1.4.4).

1.4.3 Treillis de concepts et les motifs fréquents

La notion de treillis de Galois d'une relation (ou treillis de concepts) est à la base d'une famille de méthodes de classification conceptuelle introduite par Barbut et Monjardet [Barbut and Monjardet, 1970]. Les treillis de Galois ont d'abord été utilisés comme des outils mathématique dans le cadre de l'algèbre [Birkoff, 1967]. Ceux-ci ont, par la suite, joué un rôle très important en classification [Wille, 1982; ?]. En effet, un treillis de Galois permet d'obtenir des

classes structurées qui font ressortir les liens entre les données, entre les attributs, et également entre les données et les attributs. De fait, dans le cas de tableaux binaires de type données-propriétés, de nombreux auteurs utilisent cette approche pour interpréter, analyser, organiser les données [Wille, 1982; Duquenne, 1996; Godin *et al.*, 1995].

La classification par treillis

Nous donnerons premièrement quelques définitions préliminaires. Nous détaillerons ensuite, en donnant un exemple, le processus de construction d'un treillis de Galois d'une relation binaire. Puis, nous donnerons finalement quelques définitions liées aux motifs.

Définitions

– Ordre :

Soit E un ensemble et $\leq$ une relation binaire sur $E \times E$. La relation $\leq$ est une relation d'ordre partiel si elle est transitive, antisymétrique. On appelle *ordre*, ou *ensemble ordonné*, le couple $P = (E, \leq)$. Soient x et y deux éléments de X , on note $x < y$, si et seulement si, $x \leq y$ et $x \neq y$.

– Éléments minimaux et maximaux :

Soit A une partie arbitraire de l'ordre $P = (E, \leq)$, un élément x de A est *minimal* s'il n'existe pas d'élément y de A tel que $y < x$. De même, un élément x de A est *maximal* s'il n'existe pas d'élément y de A tel que $y > x$. Les ensembles d'éléments minimaux et maximaux de A sont donc respectivement donnés par :

$$\min A = \{x \in A \mid \nexists y \in A, y < x\} \quad \text{et} \quad \max A = \{x \in A \mid \nexists y \in A, y > x\}$$

– Ensembles des minorants et majorants :

Soit A une partie arbitraire de P , l'ensemble des *minorants* de A , noté $l(A)$, est l'ensemble des éléments de P plus petits que tout élément de A . De manière dual, l'ensemble des majorants de A , noté $u(A)$, est l'ensemble des éléments de P plus grands que tout élément de A . Plus formellement :

$$l(A) = \{x \in P \mid \forall y \in A, x \leq y\} \quad \text{et} \quad u(A) = \{x \in P \mid \forall y \in A, x \geq y\}$$

– Bornes d'un ensemble ordonné :

Si $l(A)$ a un plus grand élément unique, cet élément est appelé *borne inférieure* de A ou $\text{Meet}(A)$, ou infimum, $(\wedge)$ ou encore bottom $(\perp_A)$. De même, si $u(A)$ existe et a un plus petit élément unique, cet élément est appelé *borne supérieure* de A , ou $\text{Join}(A)$, ou supremum $(\vee)$, ou encore top $(\top_A)$.

– Treillis :

Un ensemble ordonné $(E, \leq)$ est un treillis, si et seulement si, tout couple (x, y) de E^2 possède une borne inférieure, notée $x \wedge y$, et une borne supérieure, notée $x \vee y$. Un treillis est généralement représenté par son diagramme de Hasse [Davey and Priestley, 1994].

– Contexte formel :

Soient $\mathcal{O}$ un ensemble fini d'objets, $\mathcal{P}$ un ensemble fini ordonné d'items, et, $\mathcal{R}$ une relation binaire entre ces deux ensembles, le *contexte formel*, ou la *base de données*, est le

triplet $\mathcal{CF} = (\mathcal{O}, \mathcal{P}, \mathcal{R})$. Les éléments $\mathcal{O}$ sont appelés *données* et ceux de $\mathcal{P}$ sont appelés *propriétés* ou *attributs*. Quand un individu o est en relation avec une propriété p , on dit que l'individu o possède la propriété p .

– **Concept formel :**

Un concept formel C d'un contexte formel $(\mathcal{O}, \mathcal{P}, \mathcal{R})$ est défini comme un couple $(\text{Extension}(C), \text{Intension}(C))$ où $\text{Extension}(C)$, notée $\text{Extent}(C)$, est l'ensemble des données de $\mathcal{O}$ possédant les propriétés de $\text{Intension}(C)$, et, $\text{Intension}(C)$, notée $\text{Intent}(C)$, l'ensemble des propriétés communes aux données de $\text{Extension}(C)$. Un concept formel rend donc compte d'un regroupement naturel de données possédant les mêmes propriétés. Il peut donc être décrit en extension et en intension.

– **Treillis de concepts ou treillis de Galois :**

Soient L l'ensemble des concepts formels d'un contexte, et, une relation $\leq$ sur L définie comme suit :

$$\forall C_1, C_2 \in L, C_1 \leq C_2 \iff \begin{aligned} \text{Extent}(C_1) &\subseteq \text{Extent}(C_2) \\ \text{Intent}(C_1) &\supseteq \text{Intent}(C_2) \end{aligned}$$

La relation $\leq$ est une relation d'ordre partiel sur L ; et $\mathcal{L} = (L, \leq)$ est un treillis complet appelé *treillis de concepts* ou *treillis de Galois*.

– **Opérateur de fermeture :**

Un opérateur de *fermeture* h sur un ensemble ordonné E est une application $h : E \rightarrow E$, vérifiant les propriétés suivantes :

- monotone croissante : $\forall x, y \in E, x \geq y \Rightarrow h(x) \geq h(y)$,
- extensive : $\forall x \in E, x \leq h(x)$,
- idempotente : $\forall x \in E, h \circ h(x) = h(h(x)) = h(x)$.

– **Fermé :**

Un sous-ensemble x de E est dit fermé pour l'opérateur de fermeture h , si et seulement si, $x = h(x)$.

Construction du treillis de Galois d'une relation binaire

Soient deux relations binaires f et g entre $\mathcal{O}$, l'ensemble d'objets, et $\mathcal{P}$, l'ensemble d'attributs, respectivement, dans le contexte $\mathcal{CF}$, telles que :

$$\begin{aligned} f : \mathcal{P} &\rightarrow \mathcal{O} \\ P &\mapsto f(P) = \{d \in \mathcal{O} \mid d \text{ contient } P\} \\ g : \mathcal{O} &\rightarrow \mathcal{P} \\ O &\mapsto g(O) = \{t \in \mathcal{P} \mid \forall d \in O, (d, t) \in \mathcal{R}\} \end{aligned}$$

Les compositions de ces deux fonctions $g \circ f$ et $f \circ g$ sont des opérateurs de fermeture. C'est sur la base de ces deux derniers opérateurs qu'est construit le treillis de Galois de la relation $\mathcal{R}$.

	d_1	d_2	d_3	d_4	d_5	d_6
WEB	x	x	x	x		
Recherche d'information (RI)	x	x	x		x	x
Moteurs de recherche (MR)	x	x	x		x	x
Annuaire (A)	x			x		
Yahoo ! (Y)	x					
Alta Vista (AV)					x	
Google (G)					x	

TAB. 1.2 – Indexation des documents.

Exemple : Le tableau 1.2 représente la relation binaire entre les documents et les propriétés associées, où $\mathcal{O} = \{d_1, d_2, d_3, d_4, d_5, d_6\}$ est un ensemble de documents et $\mathcal{P} = \{\text{WEB, Recherche d'information, Moteurs de recherche, Annuaire, Yahoo !, Alta Vista, Google}\}$ est un ensemble de termes associés à ces documents. Le treillis de Galois qui est généré à partir de tableau 1.2 est donné par la figure 1.6.

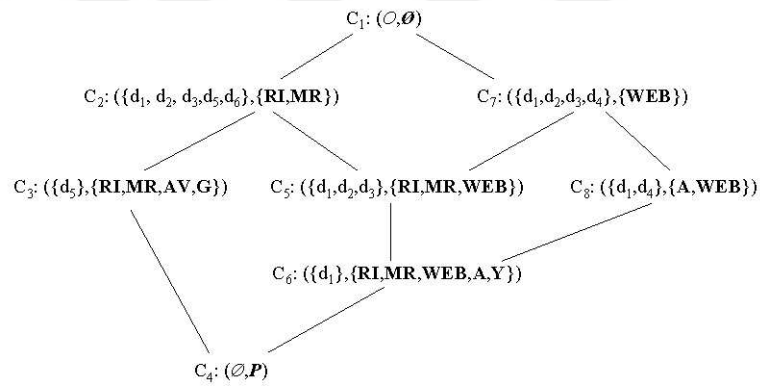


FIG. 1.6 – Représentation du treillis de Galois $\mathcal{L} = (\mathcal{O}, \mathcal{P})$

Motifs fréquents

Définitions

– Motif :

Un motif est un sous-ensemble de $\mathcal{P}$. On dit qu'un motif P est inclus dans l'objet o (ou o contient P) si P et o sont en relation, soit : $\forall p \in P, (o, p) \in \mathcal{R}$. Un motif de taille k (comprenant k items) est appelé un k -motif. Dans l'exemple précédent, d_2 et d_3 contiennent le 3-motif $\{\text{WEB, RI, MR}\}$.

– Support d'un motif :

Le support d'un motif P est défini par la fréquence d'apparition du motif dans l'ensemble $\mathcal{O}$. Dans l'exemple précédent, le support du motif $\{\text{WEB, RI, MR}\}$ est égal à 3.

– Motif fréquent :

Un motif P est dit fréquent s'il apparaît un nombre de fois supérieur ou égale à un seuil,

minsup, dans l'ensemble d'objets $\mathcal{O}$. Dans l'exemple précédent, si minsup est égale à 2, alors le motif $P = \{RI, MR\}$ sera considéré comme fréquent car $\text{support}(P) = 5 \geq \text{minsup}$.

– **Motif fermé :**

Un motif P est dit fermé, si et seulement si, $h(P) = P$. Calculer le motif fermé P revient à calculer l'intersection des «intensions» de tous les objets qui possèdent ce motif, utilisant la formule :

$$h(P) = \bigcap_{P \subseteq g(o)} g(o) \quad (1.24)$$

Dans l'exemple précédent, $\{A, WEB\}$ est un motif fermé, car c'est le motif maximal commun aux objets $\{d_1, d_4\}$.

En effet, $h(\{A, WEB\}) = \bigcap_{i \in \{d_1, d_4\}} g(d_i) = \{A, WEB, RI, MR, Y\} \cap \{A, WEB\} = \{A, WEB\}$

– **Classe d'équivalence :**

Deux motifs P_1 et P_2 sont dits équivalents ($P_1 \cong P_2$), si et seulement si, $f(P_1) = f(P_2)$. L'ensemble des motifs qui sont équivalents à un motif P_1 (appelé la classe d'équivalence de P) est dénoté par $[P_1] = \{P_2 \subseteq \mathcal{P} \mid P_1 \cong P_2\}$

– **Générateur minimal :**

Un motif générateur G_k d'un motif fermé F_k de taille k est un motif minimal (c'est-à-dire, le plus petit motif par rapport à l'inclusion) dont la fermeture est égale à F_k . Plusieurs motifs peuvent générer le même motif fermé. Ces motifs font partie d'une classe d'équivalence. Autrement dit, G_k est un motif générateur d'un fermé F_k , si et seulement si, $\nexists k' < k, G_{k'} \subsetneq G_k$ tels que $h(G_{k'}) = F_k$.

1.4.4 Extraction de règles d'association

L'extraction de règles d'association à partir de grandes bases de données [Agrawal and Srikant, 1994] est une tâche primordiale de la fouille de données. Elle revient à extraire d'une base de données les corrélations entre les attributs (propriétés) qui sont associés aux données. Ces corrélations entre attributs sont exprimées par des règles du type $A \rightarrow B$. De telles règles peuvent être pondérées par des mesures de validité. Elles sont intéressantes à plusieurs titres dans le cadre de la fouille :

- Elles sont faciles à lire.
- Elles sont généralement composées de peu d'attributs, ce qui facilite le travail d'interprétation en permettant d'identifier facilement les relations entre les attributs.

Définitions

– **Règles d'association :**

Une règle d'association est une règle d'implication conditionnelle permettant de trouver des corrélations entre des éléments qui sont liés par une relation R . Les règles d'association sont interprétées comme une cooccurrence d'attributs impliquant la présence

d'autres attributs dans les objets.

$R : A \rightarrow B$ est une règle d'association entre A et B qui vérifie les propriétés suivantes :

- A, B sont deux ensemble d'attributs,
- A est appelé antécédent, prémisses ou partie gauche,
- B est appelé conséquent, conclusion ou partie droite.

L'interprétation intuitive de la règle R est : si des données contiennent l'ensemble d'attribut A alors ces données ont tendance à contenir également l'ensemble d'attributs B avec une probabilité donnée.

– **Support d'une règle :**

Le support d'une règle est égal au support de l'union de sa prémisses et de sa conclusion.

$$\text{support}(R : A \rightarrow B) = \text{support}(A \cup B) \quad (1.25)$$

– **Confiance d'une règle :**

La confiance d'une règle représente le rapport entre le support de la règle et le support de sa prémisses (A).

$$\text{confiance}(R : A \rightarrow B) = \frac{\text{support}(A \cup B)}{\text{support}(A)} \in [0, 1] \quad (1.26)$$

La confiance d'une règle d'association est la mesure la plus utilisée pour évaluer la validité d'une règle d'association. Il existe cependant de nombreuses autres mesures telles que, la mesure de laplace, le gain, la conviction, etc. Pour une description plus précise de ces dernières mesures, on pourra se reporter à [Roberto J. Bayardo and Agrawal, 1999; Cherfi, 2004]

– **Règle d'association valide :**

Une règle d'association est dite «valide» si son support et sa confiance sont supérieurs ou égaux à deux seuils, minsup et minconf, respectivement.

– **Règle d'association exacte :**

Une règle d'association est dite «exacte» si sa confiance est de 100%; autrement dit, si les support de A et $A \cup B$ sont égaux.

– **Règle d'association approximative :**

Une règle d'association est dite «approximative» si sa confiance est inférieure à 100%.

– **Règle d'association informative :**

Une règle d'association est dite informative si :

- sa partie gauche, A , est minimale,
- sa partie droite, B , est maximale.

– **Règle d'association redondante :**

Une règle d'association est dite «redondante» par rapport à d'autres règles d'association, si l'information qu'elle apporte est présente dans d'autres règles. Une telle règle est donc naturellement moins informative que les autres règles, voir inutile.

Construction des règles d'association

La construction de règles d'association se décompose en deux étapes :

- détermination de l'ensemble des motifs fréquents (supérieurs à minsup), ainsi que de leur support,
- génération, pour chaque motif fréquent, de toutes les règles d'association dont la confiance est supérieure ou égale à minconf.

Cette méthode présente cependant l'inconvénient principal de nécessiter l'intervention d'un expert pour définir les valeurs de minsup et de minconf. De plus, la génération de toutes les combinaisons possibles à partir des motifs fréquents génère des règles redondantes. Ce dernier inconvénient peut être contourner par l'utilisation d'un treillis de Galois comme nous le montrons ci-après.

Extraction de règles d'association d'un treillis de Galois

La structure d'un treillis de Galois $\mathcal{L} = (\mathcal{O}, \mathcal{P})$ peut être exploitée afin d'extraire des règles d'association valides de la forme $A \rightarrow B$.

La formalisation mathématique de l'extraction d'un ensemble de règles d'implication à partir d'un treillis de Galois est due à Guigues et Duquenne [Guigues and Duquenne, 1986]. Des algorithmes d'extraction ont été proposés par Godin et Missaoui [Godin and Missaoui, 1994]. Simmon et Napoli [Simon and Napoli, 1999] ont également montré qu'un sous-ensemble minimal de règles peut être obtenu à partir d'un treillis de Galois. De plus, de manière à faciliter l'extraction des règles, Toussaint et Simon [Toussaint *et al.*, 2000; Simon, 2000] distinguent, pour chaque concept formel $\mathcal{CF}$ de treillis de Galois, d'une part, l'ensemble P des termes qui n'appartiennent pas aux intensions des concepts les plus généraux (au sens de $\leq$), aussi appelées **propriétés propres**, et, d'autre part, l'ensemble H des termes qui appartiennent aux intensions des concepts plus généraux, aussi appelée **propriétés héritées**. Les propriétés propres sont de plus mutuellement équivalentes (du point de vue de leurs occurrences dans les données).

Chaque concept $\mathcal{CF}$ permet alors de générer les règles d'implication $t_i \rightarrow H$, où $t_i \in P$, et P est l'ensemble de termes propres.

Dans l'exemple de la figure 1.6, le concept formel C_3 du treillis d'héritage possède deux propriétés propres $P = \{AV, G\}$ et deux propriétés héritées $H = \{RI, MR\}$. Ainsi, C_3 permet de générer les règles suivantes :

$$R_1 : AV \rightarrow RI \wedge MR$$

$$R_2 : G \rightarrow RI \wedge MR$$

$$R_3 : AV \rightarrow G$$

La règle R_3 exprime l'équivalence mutuelle des termes propres $\{AV, G\}$, qui s'interprète comme suit : tous les documents qui possèdent AV possèdent G . Les règles R_1 et R_2 s'interprètent de la façon suivante : l'emploi des termes AV ou G est toujours associé à celui des termes RI et MR .

Une méthode d'extraction alternative, permettant de générer des règles d'association informatives a été proposé par Szathmary et al. [Szathmary *et al.*, 2005]. Celle-ci se base sur

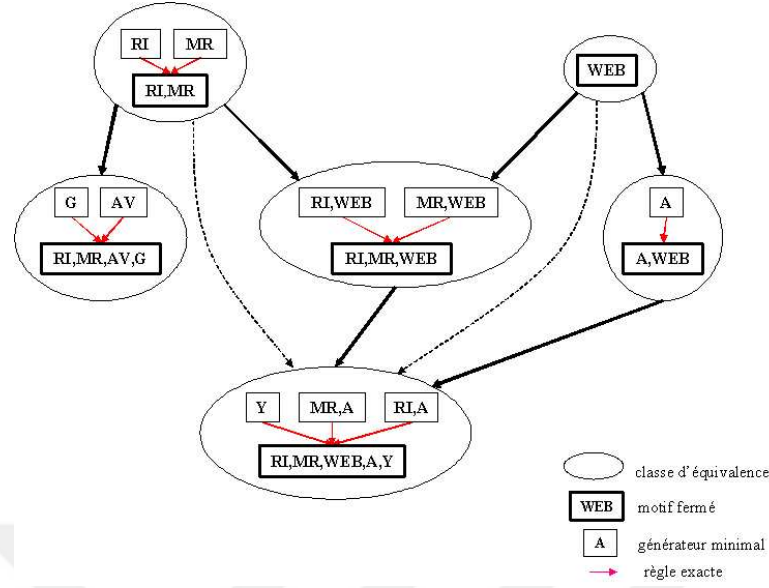


FIG. 1.7 – Représentation du treillis de Galois $\mathcal{L} = (\mathcal{O}, \mathcal{P})$ avec ses classes d'équivalence

les générateurs minimaux qui peuvent être trouvés à partir de motifs fermés fréquents. L'algorithme d'extraction, appelé Zart, vise à réduire les règles d'association redondantes sans perdre d'information (cf. figure 1.7). Il génère deux types de règles informatives :

- des règles exactes, qui peuvent être générées à l'intérieur des classes d'équivalence à partir des générateurs minimaux et de leur fermeture : pour chaque générateur minimal P_1 d'une classe d'équivalence et de sa fermeture P_2 , on génère une règle du type : $R : P_1 \rightarrow P_2 \setminus P_1$.
- des règles approximatives, qui peuvent être générées entre deux classes d'équivalence (c'est-à-dire, entre les générateurs minimaux d'une première classe d'équivalence C_1 et le motif fermé d'une deuxième classe d'équivalence C_2 , telles que $C_1 \subseteq C_2$).

1.4.5 Inconvénients des modèles symboliques

La génération de règles d'association par les méthodes symboliques est un processus qui reste coûteux en temps d'exécution, même dans le cas de l'utilisation d'un treillis de Galois. De plus, ces méthodes restent sujettes à la génération de règles redondantes, surtout si l'on travaille sur des données fortement multidimensionnelles. De même, la structure des treillis générés directement sur de tels types de données resterait très difficile à exploiter, tant le nombre de concepts ou de classes serait important.

Une solution pour réduire le nombre de règles redondantes, ainsi que pour diminuer la complexité des treillis générés, serait celle de diviser l'espace de description des données en plusieurs sous-espaces, ou points de vue (cf. section 1.2.3), et de baser la stratégie d'extraction sur des treillis de Galois propres à chacun des points de vue. Cette stratégie ferait cependant perdre des règles informatives du fait de l'absence de communication possible entre les treillis associés aux différents points de vue.

Un autre inconvénient général de méthodes symboliques est qu'elles nécessitent l'utilisation systématique de paramètres prédéfinis, tels que le seuil d'occurrence sur les motifs, pour trouver les motifs fréquents, et le seuil de validité sur les règles, pour éliminer les règles redondantes.

Finalement, ces méthodes permettent de ne traiter que des données binaires, sans ainsi prendre en considération des poids éventuels associés aux propriétés. Ce dernier point représente un handicap important pour un certain nombre de données, telles que les données documentaires, dans la description desquelles ces poids jouent un rôle déterminant.

1.5 Conclusion

Dans ce chapitre nous avons présenté les principes généraux de l'analyse et de la fouille des données documentaires en faisant spécifiquement état des analogies entre les deux domaines. Dans ce cadre nous avons décrit les principes conjoints de l'analyse de contenu et des modèles de représentation des documents, ainsi que les stratégies et de mise en correspondance associées à ces représentations. Nous avons montré que les stratégies d'analyse de contenu les plus couramment utilisées, à savoir l'analyse de surface, avait tendance à générer, si l'on considère le modèle de représentation des documents le plus couramment utilisé, à savoir le modèle vectoriel, des espaces de description de documents de taille importante.

A ce titre, nous avons montré que les techniques de partitionnement de l'espace de description global des documents en plusieurs sous-espaces locaux, ou point de vue, présentaient l'avantage de simplifier et d'enrichir la manipulation et l'analyse de données décrites dans un tel espace, tout en réduisant les phénomènes de bruit. Ces techniques s'harmonisent particulièrement bien avec l'utilisation des réseaux bayésiens pour la manipulation des données documentaires. En effet, ceux-ci permettent de combiner de manière homogène des informations issues de plusieurs sources.

Les méthodes de fouille les plus répandues sont les méthodes symboliques. Nous avons plus particulièrement présenté celles basées sur les treillis de concept. L'avantage de telles méthodes est qu'elles permettent d'identifier, en se basant sur une structure explicative, des relations entre les termes des documents sous forme de règles d'association. Ces méthodes restent cependant difficiles à exploiter dans les cas concrets. En effet, elles présentent l'inconvénient de procéder de manière combinatoire et de générer des règles d'association redondantes. De plus, leur structure explicative complexe ne peut pas être directement exploitée par les analystes.

Malgré les techniques d'identification de relations sémantiques entre les termes mentionnées dans ce chapitre, les données documentaires, une fois représentées en machine, conservent des caractéristiques descriptives particulières. Ainsi, chaque document a tendance à rester individuellement défini dans un sous-espace de description de petite taille, dont le recouvrement avec sous-espaces de description des documents voisins reste faible. Cette situation amène l'espace de description global à conserver une taille importante mais une faible densité, ce qui correspond à un espace ultra creux. Cette situation survient de manière encore plus prononcée dans le cas des données web, en particulier quand les liens sont pris en compte dans la description des documents. Elle complexifie singulièrement les processus d'analyse et de fouille de données, et va même jusqu'à invalider l'utilisation à grande échelle de certaines méthodes, comme les méthodes symboliques que nous avons décrites.

Les méthodes de classification numérique non supervisée jouent alors un rôle important pour l'analyse et la fouille des données documentaires du fait qu'elles traitent n'importe quel type de données et produisent de bonnes synthèses, accessibles aux analystes. Nous illustrerons donc dans le chapitre suivant le fonctionnement de ces méthodes et nous en présenterons plusieurs types, à savoir des méthodes classiques et des méthodes neuronales. Ces dernières constitueront l'essentiel de notre travail de thèse.

Parmi les domaines de recherche liés à l'analyse et à la fouille de données documentaires que nous allons aborder dans la suite de notre travail figurent, l'analyse de données documentaires, l'analyse des informations du web, et, la visualisation. Ce dernier domaine concerne aussi bien la visualisation directe des données documentaires que la visualisation des résultats d'analyse ou de fouille sur ces mêmes données.



Chapitre 2

Classification numérique non supervisée

Sommaire

2.1	Introduction	35
2.2	Méthodes de classification classiques	36
2.2.1	Méthodes de partitionnement (directes)	37
2.2.2	Méthodes hiérarchiques (indirectes)	39
2.2.3	Méthodes hybrides	41
2.2.4	Méthodes orientées par la densité	42
2.3	Approches neuromimétiques pour la classification	43
2.3.1	Principes généraux	43
2.3.2	Carte auto-organisatrice de Kohonen (SOM)	45
2.3.3	Neural gas (NG)	47
2.3.4	Neural gas plus compétitive Hebbian learning	49
2.3.5	Growing neural gas (GNG)	49
2.3.6	SOM Incrémental Hiérarchique (GHSOM)	49
2.3.7	Le modèle MultiSOM	52
2.4	Évaluation de la qualité de classification	55
2.4.1	Qualité basée sur la distance (critères inertiels)	55
2.4.2	Qualité basée sur la similarité	56
2.4.3	Qualité basée sur la méthodologie et la métrique expérimentales	57
2.4.4	Qualité basée sur l'algèbre linéaire	58
2.5	Conclusion	59

2.1 Introduction

Le clustering, ou classification numérique non supervisée, représente une tâche centrale de la fouille de données. C'est un processus de regroupement des données en un ensemble de classes, appelées également clusters. Ce regroupement s'opère en respectant deux contraintes interdépendantes : 1) les données regroupées dans les mêmes classes doivent être le plus similaires possible, 2) les données appartenant à des classes séparées doivent être le moins similaires possible. Le processus de classification non supervisée, est un processus de regroupement qui

ne dépend pas des classes prédéfinies mais uniquement des propriétés intrinsèques des données d'apprentissage [Jiang *et al.*, 2002].

Ce processus présente de nombreux avantages par rapport aux méthodes de regroupement symboliques. En effet, il permet de classer différents types de données, binaires ou non binaires, et de produire des synthèses efficaces qui prennent en compte les relations spatiales entre les données.

A titre d'exemple, dans le cas des données documentaires, les méthodes de clustering sont susceptibles de produire deux types de classes :

1. Classes de mots-clés (termes d'index) : ce type de classes permet l'analyse thématique du contenu d'un fonds donné.
2. Classes de documents : ce type de classes peut servir de support à l'extraction précise sur le fond, aussi bien qu'à l'optimisation de la recherche documentaire sur ce dernier (cf. chapitre 1). Ce type de classes permet de regrouper les documents similaires dans la même classe, d'une part ; il facilite le processus de la mise en correspondance pour une requête donnée dans le système de recherche documentaire ; d'autre part, il permet la fouille de données, et par conséquent l'extraction de connaissances, à partir de ce type de classes.

De nombreuses méthodes de classification non supervisées ont été proposées dans la littérature. Elles peuvent se diviser en deux grands types :

1. **Classification recouvrante** (ou soft clustering) : dans ce type de classification, il est possible qu'une donnée soit associée à plusieurs classes. Ce type de classification est également appelé *fuzzy clustering*,
2. **Classification disjointe** (ou hard clustering) : dans ce type de classification, chaque donnée est associée à une seule classe.

Il s'avère également intéressant de diviser les méthodes de classification non supervisées en deux sous-types, à savoir les méthodes classiques, et, les méthodes neuronales. En effet, les méthodes de classification neuronales présentent des avantages spécifiques qui sont liées à la fois au choix des fonctions d'apprentissage et à celui des topographies de classes qu'elles proposent. Nous présentons donc successivement ces deux types de méthodes, classiques et neuronales.

L'exploitation d'un modèle neuronal peut ouvrir de nouvelles voies dans le domaine du clustering en introduisant la possibilité de transformer un processus global en de multiples processus locaux coopérants. Cette technique, qui permet notamment de réduire sensiblement le bruit intrinsèquement lié à l'approche globale, représente une des bases de notre travail ultérieur et sera également présentée dans ce chapitre.

Pour pouvoir comparer les méthodes de classification entre elles, ainsi que pour permettre d'optimiser leur comportement individuel, il est nécessaire de comparer objectivement leurs résultats. Nous présenterons finalement quatre types de méthodes d'évaluations des résultats.

2.2 Méthodes de classification classiques

Les méthodes de classification classiques peuvent par la suite être classées en quatre catégories, à savoir les méthodes **directes**, les méthodes **indirectes**, les méthodes **hybrides** et les méthodes **orientées par la densité**.

2.2.1 Méthodes de partitionnement (directes)

Elles consistent à partitionner les objets en différents groupes en établissant des noyaux de classes, puis en regroupant les données à classer autour de ces noyaux. La méthode K-means, ainsi que ses dérivées (K-medioïde et K-means axiale), représentent des exemples types de méthodes directes.

K-means

L'algorithme K-means utilise un principe heuristique, à savoir celui de la minimisation de l'erreur quadratique [Hartigan and Wong, 1979; England and Beynon, 1981]. Il requiert une partition initiale et un nombre de classes K à construire. Cette méthode, encore appelée méthode des centres mobiles [Lebart *et al.*, 1998], permet de classer des objets en K groupes en les agrégeant autour de centroïdes. Elle vise à optimiser la fonction E définie par :

$$E = \sum_{i=1}^k \sum_{D \in C_i} |D - M_i|^2 \quad (2.1)$$

où D est un point qui représente un objet donné, et, M_i est le point central (profil moyen) de la classe C_i .

L'algorithme K-means est le suivant :

Algorithme 1 Algorithme de K-means

- 1: Choisir arbitrairement K objets initiaux, qui représenteront les centroïdes des K groupes, dans l'espace descriptif initial des objets à classer.
 - 2: Assigner chaque objet D au groupe qui a le centroïde le plus proche.
 - 3: Quand tous les objets ont été assignés, recalculer les positions des K centroïdes.
 - 4: Répéter les étapes 2 et 3 jusqu'à ce que centroïdes ne se déplacent plus (ce qui revient également à dire que les objets ne changent plus de groupe).
-

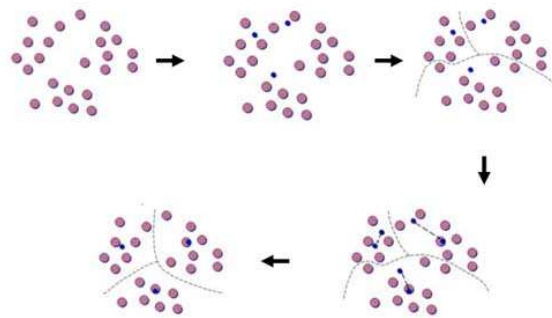


FIG. 2.1 – Méthode de K-means.

La figure 2.1 donne une illustration de K-means dans le plan. Il est prouvé mathématiquement que l'algorithme K-means converge vers un minima local qui dépend des centres initiaux [Bottou and Bengio, 1995]. Cette convergence est effective car la variance à l'intérieur d'une classe ne peut que décroître ou rester stationnaire [Lebart *et al.*, 1998].

Méthode K-medioïd

La méthode K-medioïd [Kaufman and Rousseeuw, 1990a] diffère de la méthode K-means par le fait qu'elle utilise la donnée la plus centrale (medioïde) dans une classe pour représenter le centre (profil) de classe, au lieu de prendre le profil moyen de données de la classe. Cet algorithme minimise la fonction 2.1. Mais avec M_i représente le medioïde de la classe C_i .

Grâce à cette différence, la méthode K-medioïd est moins sensible au bruit et aux données marginales que la méthode K-means [Han *et al.*, 2001].

K-means axiale (KMA)

La méthode K-means axiale [Lelu and Georgel, 1990; Lelu, 1994], de type soft clustering, a été spécifiquement développée pour traiter les données documentaires. Elle fait appel aux deux grandes familles de méthodes d'analyse de données, à savoir les méthodes de classification et les méthodes factorielles [Lebart *et al.*, 1982]. En fait, cette méthode représente une synthèse entre la méthode de classification K-means et la méthode d'analyse factorielle sphérique.

Cette méthode considère l'ensemble des documents comme un nuage de point plongé dans un espace géométrique où chaque dimension correspond à un mot-clé. Elle est caractérisée par une représentation des classes par des vecteurs pointant vers les zones de forte densité du nuage de points [Polanco and François, 2000].

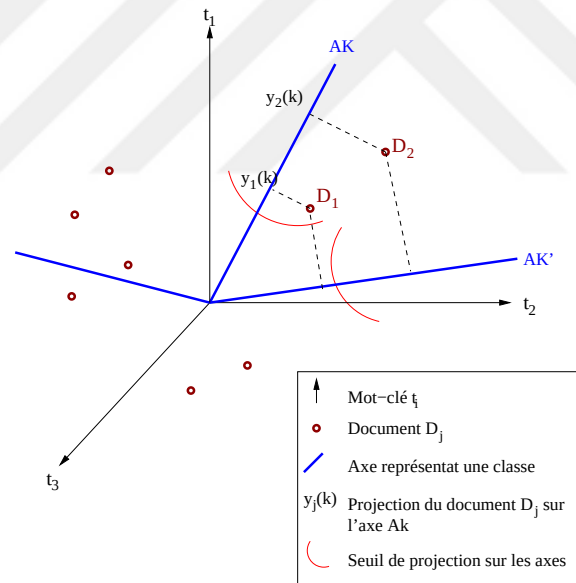


FIG. 2.2 – KMA : exemple d'un corpus de documents indexés par les 3 mots-clés t_1 , t_2 , et t_3 . Ces mots-clés définissent l'espace $\mathbb{R}^3$.

Les K-means axiales définissent les K classes recherchées par K demi-axes passant par l'origine de l'espace géométrique, ou K vecteurs unitaires pointant dans la direction de ses demi-axes. La position des K demi-axes est initialisée au hasard ou par les K premiers documents. Les projections orthogonales $y_j(k)$ de chaque document D_j normé sur les K demi-axes ainsi définis (cf. figure 2.2) sont ensuite calculées, en effectuant les produits scalaires entre le document D normé et les valeurs unitaire des K demi-axes. Chaque document est affecté à la classe K où sa projection $y(k)$ sur l'axe A_k est maximale et la position de l'axe est mise à jour

pour prendre en compte cette affectation. Par itérations successives, les axes se positionnent puis se stabilisent dans les zones de forte densité du nuage de documents, effectuant ainsi une classification stricte de documents. Pour obtenir des classes recouvrantes, un « seuil de typicité » est ensuite défini : un document appartient à la classe si sa valeur de projection sur l'axe représentant la classe est supérieure au seuil. Un document peut donc appartenir aux plusieurs classes si ses valeurs de projection sur les axes correspondants sont supérieures au seuil. Ensuite, les documents, appartenant à une classe, sont ordonnés selon la valeur de leur projection sur l'axe représentant la classe.

Cette méthode impose cependant, par construction, de travailler avec des données normalisées. Dans le cas contraire, le risque est en effet celui d'obtenir des classes non homogènes. La détermination d'une valeur pertinente de seuil de typicité s'avère également particulièrement complexe.

2.2.2 Méthodes hiérarchiques (indirectes)

La méthode hiérarchique est une méthode indirecte qui nécessitent le calcul de la matrice de similarité ou dissimilarité (distance) entre les données. Elle consiste à créer une hiérarchie de classes en formant un dendrogramme, les données les plus similaires étant regroupées dans les classes de plus bas niveau, tandis que les données les moins similaires le sont dans des classes de plus haut niveau [Van Rijsbergen, 1979]. Le dendrogramme peut être formé en deux manières (cf. figure 2.3) :

- de bas en haut : méthode ascendante,
- de haut en bas : méthode descendante.

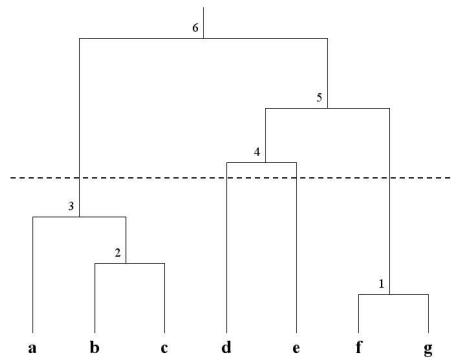


FIG. 2.3 – Exemple de classification hiérarchique

Méthodes hiérarchiques ascendantes

L'approche ascendante [Young and Calvert, 1974], est également appelée approche « agglomérative ». Chaque objet constitue une classe séparée. Les classes sont ensuite fusionnées par étapes successives en utilisant des mesures de similarité, ou encore d'agglomération, et ceci, jusqu'à ce que toutes les classes soient fusionnées dans une seule classe (le niveau le plus élevé de la hiérarchie), ou encore, jusqu'à atteindre une condition d'arrêt [Manning and Schütze, 2000].

Citons quelques exemples d'indices d'agrégations :

- Indice d'agrégation du lien minimum entre deux classes c_1 et c_2 :

$$\delta(c_1, c_2) = \min_{d_i \in c_1, d_j \in c_2} \text{dist}(d_i, d_j)$$

où $\text{dist}(d_i, d_j)$ est la distance entre l'objet d_i et l'objet d_j .

- Indice d'agrégation du lien maximum :

$$\delta(c_1, c_2) = \max_{d_i \in c_1, d_j \in c_2} \text{dist}(d_i, d_j)$$

- Indice des moyennes des distances :

$$\delta(c_1, c_2) = \frac{1}{|c_1||c_2|} \sum_{d_i \in c_1, d_j \in c_2} \text{dist}(d_i, d_j)$$

- Indice des centres de gravité :

$$\delta(c_1, c_2) = \text{dist}(G(c_1), G(c_2))$$

où $G(c_i)$ représente le centre de gravité de c_i .

Alors, il existe plusieurs techniques de classification hiérarchique basées sur l'agglomération. La différence entre les méthodes de classification hiérarchiques ascendantes est donc principalement liée à la stratégie d'agrégation utilisée. Les méthodes les plus souvent exploitées dans le domaine documentaire sont les suivantes :

Lien simple. La méthode du simple lien [Van Rijsbergen, 1979] calcule la matrice de dissimilarité entre les données. Les classes initiales sont représentées par les données élémentaires. Deux classes quelconques sont fusionnées, à un niveau donné, si leur dissimilarité est au plus égale au seuil associé à ce niveau, ce qui correspond à une stratégie d'agrégation de type lien minimum. Cette méthode tend à constituer de grosses classes et, de part le fait, tend à mal discriminer les classes.

Lien complet. La méthode du lien complet [Defays, 1977] agrège à chaque étape les deux classes ayant la plus petite distance entre leurs données les plus éloignées, ce qui correspond à une stratégie d'agrégation de type lien maximum. À l'inverse de la précédente, cette méthode tend à constituer des petites classes compactes.

Lien moyen. La méthode du lien moyen [Murtagh, 1983; Murtagh, 1985] agrège à chaque étape les deux classes dont les moyennes de distances entre les données sont les plus faibles. Cette méthode produit des classes dont la taille est intermédiaire entre celles des classes construites par les deux méthodes précédentes.

Voisins réciproques. La méthode des voisins réciproques [Rham, 1980] comprend deux phases distinctes. La première consiste à identifier toutes les paires d'objets (données initiales aux classes) les plus similaires. Ces couples sont appelés voisins réciproques. La deuxième phase agglomère tous les couples de voisins réciproques et forme une nouvelle classe pour chacun des couples. Les deux phases sont exécutées alternativement jusqu'à ce que toutes les données soient réunies en une seule classe. L'inconvénient de cette méthode est qu'elle tend à regrouper des données qui ne sont pas nécessairement similaires aux mêmes niveaux de représentation.

Méthodes hiérarchiques descendantes

Cette approche, également appelée approche «divisive», commence par une classe qui contient toutes les données. Ensuite, son principe est, à chaque itération, de chercher à singulariser la classe qui a la plus faible valeur de cohérence. La fonction de similarité utilisée dans l'approche descendante est donc généralement une mesure de cohérence de classe. La division s'arrête soit, dans le cas où chaque donnée est affectée à une classe séparée, soit, si une condition d'arrêt a été atteinte [Manning and Schütze, 2000]. Cette approche est moins utilisée que l'approche ascendante.

Avantages et limitations des méthodes hiérarchiques et des méthodes de partitionnement

Les méthodes de partitionnement nécessitent de fixer arbitrairement le nombre de classes désirées. Elles sont de plus sensibles aux conditions initiales et à l'ordre de prise en compte des données à classer, ainsi qu'aux données marginales et bruitées.

Les méthodes hiérarchiques nécessitent le calcul d'une matrice de distance ou matrice de similarité. La complexité en temps de calcul résultante rend ces méthodes difficilement utilisables pour des classification portant sur un grand volume de données. Il est de plus difficile, voire impossible, de contrôler le nombre de classes obtenues par ces méthodes. En effet, le seuil de coupure ne correspond lui-même qu'à un niveau d'arrêt dans le niveau de hiérarchisation et ne fixe donc pas le nombre de classes obtenues.

2.2.3 Méthodes hybrides

Les algorithmes de classification hybrides sont souvent identifiés comme des algorithmes multiphasés étant donné qu'ils combinent les techniques hiérarchiques et les techniques de partitionnement [André Pantel, 2003] tels que BIRCH [Zhang *et al.*, 1996], CURE [Guha *et al.*, 1998], ROCK [Guha *et al.*, 1999] et CHAMELEON [Karypis *et al.*, 1999]. Nous présenterons plus particulièrement l'algorithme CHAMELEON.

CHAMELEON est un algorithme de classification hiérarchique agglomérative qui utilise une modélisation dynamique pour l'agrégation de classes. À travers cet aspect dynamique, le but est de pouvoir retrouver des classes avec des formes irrégulières et des tailles différentes. CHAMELEON combine, selon deux étapes, un algorithme de partitionnement de graphe et un algorithme hiérarchique ascendant original.

- La première étape consiste à créer un ensemble de petites classes à partir des données initiales. Pour cette phase, une matrice de similarité est utilisée. Celle-ci regroupe toutes les similarités entre les données prises deux à deux. À partir de cette matrice, un graphe des k plus proches voisins est, dans un premier temps, construit. Dans ce graphe, il existera un lien entre une donnée d_i et une donnée d_j si d_j fait partie des k plus proches voisins de d_i et si d_i fait partie des k plus proches voisins de d_j . La valeur k est un paramètre de l'algorithme. Une fois le graphe des k plus proches voisins créé, les auteurs appliquent dans un second temps un algorithme de partitionnement de graphes appelé *hMETIS* [Karypis and Kumar, 1998] pour créer un ensemble de classes initiales.
- La seconde phase de CHAMELEON combine itérativement cet ensemble de petites classes initiales en utilisant un algorithme de classification hiérarchique. Cet algorithme repose sur deux mesures pour regrouper deux classes : l'interconnectivité relative (Relative Interconnectivity ou *RI*) et la proximité relative (Relative Closeness ou *RC*). Ainsi deux

classes sont regroupées si la valeur de leur RI et de leur RC est élevée ; c'est-à-dire, si les deux classes sont très connectées entre elles et si elles sont très proches.

Interconnectivité relative. L'interconnectivité relative entre deux classes est la valeur absolue de l'interconnectivité entre ces deux classes, normalisée par la moyenne arithmétique de l'interconnectivité de chaque classe. Ainsi l'interconnectivité entre deux classes C_i et C_j est définie de la façon suivante :

$$RI(C_i, C_j) = \frac{|EC(C_i, C_j)|}{\frac{1}{2}(EC(C_i) + EC(C_j))} \quad (2.2)$$

où $EC(C_i, C_j)$ est la somme des arcs du graphe des k plus proches voisins regroupant les classes C_i et C_j , $EC(C)$ est la somme minimum des arcs nécessaires pour diviser en deux la classe C .

Proximité relative. Les concepts évoqués pour la proximité relative sont analogues à ceux définis pour l'interconnectivité relative. La proximité absolue entre deux classes C_i et C_j est la moyenne pondérée des arcs (alors que l'interconnectivité absolue est la somme des arcs) qui connectent un objet de C_i à un objet de C_j .

$$RC(C_i, C_j) = \frac{\bar{S}_{EC(C_i, C_j)}}{\frac{|C_i|}{|C_i| + |C_j|} \bar{S}_{EC(C_i)} + \frac{|C_j|}{|C_i| + |C_j|} \bar{S}_{EC(C_j)}} \quad (2.3)$$

où $\bar{S}_{EC(C_i, C_j)}$ est la moyenne pondérée des arcs qui connectent des objets de C_i à des objets de C_j , $\bar{S}_{EC(C)}$ est la moyenne pondérée des arcs dans C pour diviser C en deux et $|C|$ est le nombre d'objets de la classe C .

Pour regrouper deux classes en utilisant ces deux mesures, plusieurs approches sont possibles :

- La première est d'utiliser deux seuils T_{RI} et T_{RC} respectivement pour RI et RC . Ainsi, pour une classe C_i , on calcule les deux mesures pour chaque classe C_j . Si, pour une classe C_j , la valeur de chaque mesure excède le seuil correspondant, alors les deux classes C_i et C_j sont regroupées. Si cette condition est remplie par plusieurs classes C_j , alors la classe C_i sera regroupée avec la classe C_j qui lui est la plus interconnectée ; c'est-à-dire, celle qui aura la plus grande valeur de $EC(C_i, C_j)$.
- La seconde est de maximiser une fonction combinant les deux mesures :

$$RI(C_i, C_j) \cdot RC(C_i, C_j) \quad (2.4)$$

L'avantage de la méthode de CAMELEON est qu'elle permet de trouver des classes de formes irrégulières, de taille et de densité différentes. Mais elle exige cependant de déterminer le paramètre de nombre de plus proches voisins (k), ainsi les deux seuils T_{RI} et T_{RC} . Dans la pratique ces trois paramètres s'avèrent difficiles à déterminer si l'on travaille sur des données fortement multidimensionnelles telles que les données documentaires.

2.2.4 Méthodes orientées par la densité

Les méthodes de classification basées sur la densité reposent sur un concept simple : les classes peuvent être considérées comme des régions denses dans l'espace de données qui sont

séparées par des régions intermédiaire de densité inférieure. La construction des classes s'opère ainsi de manière incrémentale en agglomérant les données dans les classes tant que la densité résultante de celle-ci excède un certain seuil. Les méthodes de cette catégorie s'avèrent très appropriées pour filtrer les données marginales et de découvrir des classes de formes arbitraires. Nous présentons ci-après l'algorithme DBSCAN qui fait partie de cette catégorie de méthode [Ester *et al.*, 1996].

L'algorithme DBSCAN est un algorithme qui a été plus spécifiquement développé pour l'analyse des données spatiale. Cet algorithme utilise deux paramètres ϵ et *MinPts*. Le voisinage d'une donnée dans un rayon ϵ est appelée ϵ -voisinage. Une donnée possédant au moins *MinPts* données dans son ϵ -voisinage est considérée comme une donnée noyau. La classification de DBSCAN obéit par la suite aux règles ci-après :

1. Une donnée ne peut appartenir à une classe que s'elle appartient au ϵ -voisinage d'au moins une donnée noyau.
2. Une donnée noyau o , qui appartient au ϵ -voisinage d'une autre donnée noyau p , doit appartenir à la même classe que p .
3. Une donnée non noyau q , qui appartient au ϵ -voisinage des données noyaux $p_1, \dots, p_i$, $i > 0$, doit appartenir à la même classe qu'au moins une des données noyaux $p_1, \dots, p_i$.
4. Une donnée non noyau r , qui ne se trouve au ϵ -voisinage d'aucune donnée noyau, est considéré comme une donnée bruitée.

Pour créer les classes à partir de données brutes, DBSCAN vérifie le ϵ -voisinage de chaque point de la base de données. Si le ϵ -voisinage d'un point p possède plus que *MinPts*, alors un nouveau classe avec p comme une donnée noyau est créée.

L'inconvénient d'une telle méthode est qu'elle est sensible au choix des paramètres ϵ -voisinage et *MinPts* qui doivent être fixés à l'avance. Une telle stratégie s'avère donc inadaptée aux cas des distributions complexes de données.

2.3 Approches neuromimétiques pour la classification

Dans cette section, nous aborderons les principes généraux des approches neuronales, puis nous nous focaliserons plus particulièrement aux modèles neuronaux non supervisés qui seront employés dans les données documentaires pour les tâches d'analyse de données et d'extraction de connaissances.

2.3.1 Principes généraux

Les réseaux de neurones représentent une transposition simplifiée du fonctionnement de cerveau humain. Alors, ils peuvent être considérés comme une organisation cohérente d'unités de traitement élémentaires. L'organisation des unités élémentaires en réseau induit l'émergence de propriétés nouvelles, analogues à celles que l'on attribue habituellement à l'intelligence humaine. Un réseau neuromimétique se caractérise par les types de ses neurones constitutants, par son comportement, par son architecture, et par sa dynamique. Les réseaux de neurones apprennent sur une population d'origine et sont par la suite capables d'exprimer des résultats sur des données inconnues. Ils sont utilisés dans la prédiction et la classification, dans le cadre de découverte dirigée des connaissances. Certaines variantes, comme les réseaux de Kohonen,

permettent également l'exploration des séries temporelles et des analyses non dirigées. Cependant, on leur reproche souvent d'être des «boîtes noires» : il est difficile de savoir comment les résultats sont produits, ce qui rend les explications délicates, même si les résultats sont bons.

Les réseaux de neurones se divisent en deux classes principales, à savoir, les réseaux à apprentissage supervisés (supervised learning) et les réseaux à apprentissage non supervisés (unsupervised learning).

- Pour les réseaux à **apprentissage supervisés** (Perceptron [Rosenblatt, 1958], Hopfield [Hopfield, 1982], etc.), on présente au réseau des entrées en parallèle avec les sorties attendues pour ces entrées. Le réseau doit alors se reconfigurer, c'est-à-dire adapter les poids des connexions entre ses unités élémentaires (neurones), afin que les sorties effectives qu'il fournit, correspondent bien aux sorties attendues. La règle générale d'apprentissage qui est appliquée dans le cas de l'apprentissage supervisé peut donc être considérée comme une règle de minimisation d'une fonction de coût [Widrow and Hoff, 1960; Rosenblatt, 1958], ou encore d'une fonction d'énergie [Hopfield, 1982].
- Pour les réseaux à **apprentissage non supervisé** (ART [Grossberg, 1976], Kohonen [Kohonen, 1984], Neural Gas [Martinetz, Thomas and Schulten, Klaus, 1991], Growing Neural Gas [Fritzke, 1995; Fritzke, 1997], etc.), on présente différentes entrées au réseau et on le laisse évoluer librement jusqu'à ce qu'il se stabilise. Les poids des connexions entre les neurones du réseau sont alors modifiés de manière à ce que les exemples possédant les mêmes caractéristiques produisent les même sorties. La règle fondamentale qui gouverne ce type d'apprentissage est la règle de Hebb [Hebb, 1949]. Cette règle, d'inspiration biologique, est une des premières règles d'apprentissage qui ait été formulée. Le principe de l'apprentissage non supervisé reste cependant très proche de celui employé dans les méthodes d'analyse de données.

Les réseaux à **apprentissage non supervisé** sont classés en deux catégories selon la manière de modification des poids d'un ou plusieurs neurones à chaque itération. Par la suite, la liste des poids des connexions d'entrée d'un neurone donné sera considérée comme un vecteur de référence :

- Pour les réseaux de type **Winner Take All**, l'apprentissage agit sur la modification du vecteur de référence d'un seul neurone également appelé **neurone gagnant**, à chaque itération. L'algorithme LBG [Linde *et al.*, 1980] est un exemple d'un tel type de réseau.
- Pour les réseaux de type **Winner Take Most**, l'apprentissage agit sur la modification des vecteurs de référence de plusieurs neurones. Les algorithmes de Kohonen, Neural Gas et Growing Neural Gas sont des exemples d'un tel type de réseau. Nous nous intéresserons plus particulièrement à ces trois derniers algorithmes dans le cadre de cette thèse car l'apprentissage de type «Winner Take Most» s'avère plus robuste que celui de type «Winner Take All». En effet, dans ce dernier cas un neurone mal initialisé ne peut jamais être le gagnant et ainsi ne jamais contribuer à la solution du problème². Ce cas se produit peu lors d'un apprentissage de type «Winner Take Most». Le fait de modifier plusieurs neurones à chaque itération rend donc l'apprentissage plus performant.

Dans la suite de notre travail nous considérons qu'un réseau de neurones peut être décrit de la manière simplifiée suivante :

²Ce problème est en tout point similaire à celui qui se pose lors de l'initialisation de la méthode K-means, qui pourrait être assimilée à une méthode neuronale de type Winner Take All (cf. section 2.2.1)

- le réseau est constitué d'un ensemble A de N neurones (ou nœuds) où $A = \{c_1, c_2, \dots, c_N\}$.
- chaque neurone a un vecteur de référence associé w_c décrit dans l'espace n –dimensionnel, soit $w_c \in \mathbb{R}^n$
- il existe un ensemble de liens $\mathcal{C}$ entre les neurones³, avec $\mathcal{C} \subset A \times A$. Ces liens matérialisent des connexion de voisinage et ne sont pas pondérés.
- le réseau est alimenté par un ensemble de M données d'entrée n –dimensionnelles $D = \{x_1, x_2, \dots, x_M\}$, $x_i \in \mathbb{R}^n$ (chaque donnée d'entrée x_i est représentée par le vecteur de poids $x_i = \{x_{i_1}, x_{i_2}, \dots, x_{i_n}\}$).

Nous considérons deux concepts géométriques importants qui peuvent être associés à l'organisation spatiale des neurones d'un réseau, par l'intermédiaire de leurs vecteurs de référence :

1. **Région de Voronoi** : Soit, un ensemble de vecteurs de référence $w_1, w_2, \dots, w_N \in \mathbb{R}^n$ où $w_i = \{w_{i_1}, w_{i_2}, \dots, w_{i_n}\}$ (cf. figure 2.4 a), la Région de Voronoi V_i d'un vecteur w_i est définie comme l'ensemble de tous les points de $\mathbb{R}^n$ pour lesquels w_i est le plus proche vecteur :

$$V_i = \{x \in \mathbb{R}^n | i = \arg \min_{j \in \{1, \dots, N\}} \|x - w_j\|\} \quad (2.5)$$

La partition de $\mathbb{R}^n$ formée par tous les polygones de Voronoi est appelée la «**Tessellation de Voronoi**» (cf. figure 2.4 b).

2. **Triangulation de Delaunay** : elle est formée par les connexions de tous les points (neurones) pour lesquels leurs Régions de Voronoi partagent le même bord (cf. figure 2.4 c).

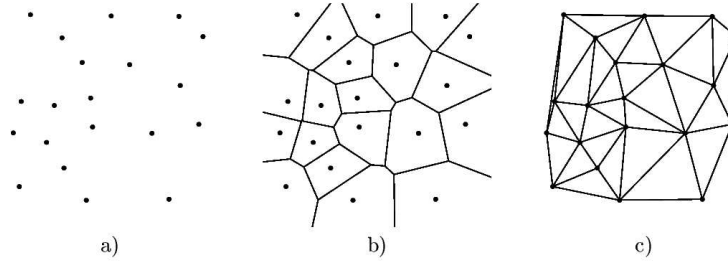


FIG. 2.4 – a) Un ensemble de points dans $\mathbb{R}^2$, b) la tessellation de Voronoi, c) la triangulation de Delaunay.

2.3.2 Carte auto-organisatrice de Kohonen (SOM)

L'hypothèse de base défendue par le modèle SOM est celle que la construction des connaissances opérée par le cortex repose sur des algorithmes d'apprentissage qui privilégient l'auto-organisation spatiale. Le modèle SOM proposé par Kohonen simule ce mécanisme sous la forme d'un processus de classification topographique [Kohonen, 1997; Kohonen, 2001]. Selon ce processus, les données d'entrée, représentables dans le cas général comme des vecteurs n –dimensionnels, sont donc ramenées à des classes qui s'auto-organisent selon une structure

³Cet ensemble peut être un ensemble vide comme le cas de Neural Gas.

de nœuds, ou neurones, sur laquelle les relations de voisinage sont prédéfinies (cf. figure 2.6). Le processus de classification topographique du modèle SOM combine donc une étape de classification avec une étape de projection des données (cf. figure 2.5).

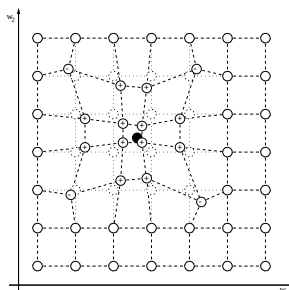


FIG. 2.5 – Modification de la topologie initiale dans le modèle SOM suite à la présentation d'une donnée (●).

Dans le modèle SOM, la structure de voisinage entre les neurones peut obéir à de nombreuses variantes différentes : structure en ficelle, bidimensionnelle rectangulaire ou hexagonale ou même multidimensionnelle (cf. figure 2.6). La structure bidimensionnelle étant la plus utilisée, on parle alors le plus souvent des cartes SOM. Nous nous intéressons donc par la suite à ce dernier cas.

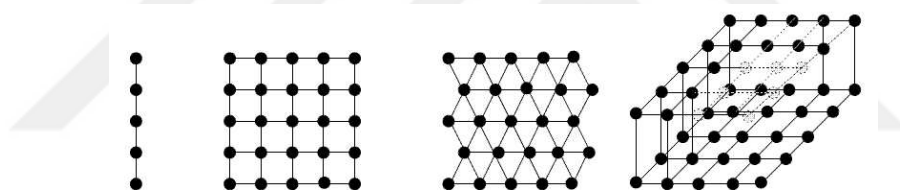


FIG. 2.6 – Options de voisinage du modèle SOM : ficelle (1D), rectangulaire (2D), hexagonale (2D) et tridimensionnelle (3D)

Les cartes SOM appartiennent à la catégorie des modèles neuronaux dit à apprentissage compétitif non supervisé. L'apprentissage compétitif se traduit par une concurrence d'activité des neurones (classes) d'une carte vis-à-vis des données d'entrée, si aucune information sur les classes résultantes susceptibles d'être obtenues n'est fournie au réseau. Une carte SOM est utilisée pour identifier des classes associées à des données d'entrée, et a posteriori, pour classifier des données inconnues. En produisant une cartographie d'un espace $\mathbb{R}^n$, pouvant s'avérer fortement multidimensionnel, vers un espace bidimensionnel ordonné $\mathbb{R}^2$, elle permet de visualiser de manière simplifiée des relations de nature métrique entre les données d'entrée. Le principe d'apprentissage d'une carte SOM est présenté de manière détaillée dans l'algorithme 2.

A l'issue de l'apprentissage, les données d'entrée peuvent être reprojetées sur la carte obtenue. Le neurone d'affectation d'une donnée représente alors celui dont le vecteur de poids est le plus proche du vecteur descriptif de la donnée. Il peut donc être sélectionné en réutilisant la seconde étape de l'algorithme 2 susmentionnée.

Les cartes SOM de Kohonen ont été utilisées avec succès dans l'évaluation qualitative des sciences. Elles ont également fourni des résultats significatifs pour d'autres types de tâches

Algorithme 2 Algorithme d'apprentissage des cartes auto-organisatrices de Kohonen (SOM)

- 1: **Initialisation** : initialiser les vecteurs de référence aléatoirement.
- 2: **Sélection du neurone gagnant** : présenter aléatoirement une entrée (x_k) au réseau et déterminer le neurone gagnant ($s_{x_k}(t)$) qui lui correspond au temps t , tel que :

$$s_{x_k}(t) = \arg \min_{c \in A} \|x_k(t) - w_c(t)\| \quad (2.6)$$

- 3: **Apprentissage non supervisé et sélection du voisinage** : après la sélection du neurone gagnant s , le vecteur de référence associé à ce neurone, ainsi que les vecteurs de référence associés aux neurones se trouvant dans un voisinage donné du neurone s (par exemple, les neurones se trouvant dans un carré, ou dans cercle, centrés autour de s) sont ajustés de manière à ce que leur profil se rapproche de celui de la donnée d'entrée. Cet ajustement de poids, qui caractérise l'apprentissage non supervisé du modèle, peut être décrit par l'équation :

$$w_{c_i}(t+1) = w_{c_i}(t) + \alpha(t) h(t) [x_{k_i}(t) - w_{c_i}(t)]; \quad 1 \leq i \leq n \quad (2.7)$$

où $\alpha(t)$ est un terme de gain ($0 \leq \alpha(t) \leq 1$) décroissant en fonction du temps et convergeant vers 0, et $h(t)$ est une fonction de voisinage qui peut être gaussienne, chapeau mexicain, bubble, etc.

- 4: Retourner à l'étape 2 jusqu'à ce que le réseau soit stable.
-

de traitement de l'information comme la classification de données socio-économiques et la cartographie du contenu des bases documentaires.

Leur principal limitation vient de fait qu'elles ne s'adaptent pas nécessairement à la structure des données originales en raison de leur structure topologique préfixée. De plus, les résultats peuvent être imprécis en raison de l'influence du voisinage sur les classes lors de l'apprentissage. Sur une carte SOM les neurones se trouvant sur le bord de la carte au moins de voisins que ceux se situant à l'intérieur de la carte ce qui amène souvent un phénomène supplémentaire de «effet du bord» [Kohonen, 2001]. En effet, les vecteurs de référence des neurones du bord de la carte sont sur influencés par les neurones centraux et donc par les données de ces derniers, ce qui a tendance à réduire la couverture de la carte vis à vis des données marginales. Il est à noter que la combinaison de l'effet de bord et de l'influence du voisinage lors de l'apprentissage peut générer, en retour, une imprécision supplémentaire sur les profils des classes de centre de la carte, si les données marginales sont nombreuses. Nous montrerons explicitement ce phénomène par la suite (cf. section 5.4.2).

2.3.3 Neural gas (NG)

Dans l'algorithme de neural gas ou gaz neuronal [Martinetz, Thomas and Schulten, Klaus, 1991], les vecteurs de référence des neurones sont adaptés sans aucun arrangement topologique fixe de ces derniers dans le réseau (contrairement aux cartes SOM). Au lieu de cela, cet algorithme utilise un classement des neurones voisins pour des données d'entrée. Les changements de poids ne sont pas déterminés par les distances relatives entre les neurones dans un treillis topologiquement pré-structuré, mais par la distance relative entre les neurones dans l'espace d'entrée, par conséquent le réseau nommé «gaz neuronal» (cf. figure 2.7).

La liberté topologique inhérente à cette méthode représente un de ses avantages majeurs.

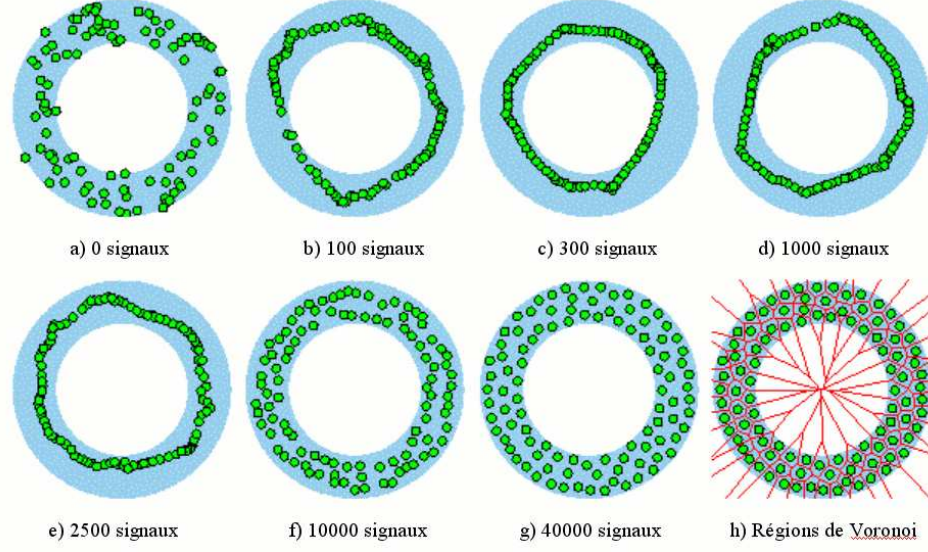


FIG. 2.7 – Classification de Neural Gas pour une forme de bague.

Algorithme 3 Algorithme d'apprentissage du modèle Neural Gas (NG)

- 1: Initialiser les vecteurs de références $w_i \in \mathbb{R}^n$ (avec $i \in [1, N]$ et N le nombre de classes) suivant $p(x)$, où $x \in \mathbb{R}^n$ est une forme d'entrée choisie aléatoirement dans l'espace d'entrée $\mathbb{R}^n$ avec la fonction de probabilité de densité de sélection $p(x)$.
- 2: Initialiser le paramètre temporel $t : t = 0$.
- 3: Choisir aléatoirement un vecteur d'entrée x selon $p(x)$.
- 4: Déterminer le numéro d'ordre k_i ($0 \leq k_i < N$) de chaque vecteur de référence $w_{i(k)}$ d'après la distance euclidienne entre w_j et x :

$$\|x - w_{i(0)}\| \leq \|x - w_{i(1)}\| \leq \dots \leq \|x - w_{i(k)}\| \leq \dots \leq \|x - w_{i(N-1)}\|, \quad j = i(k).$$

- 5: Modifier le vecteur de référence w_i :

$$w_i(t+1) = w_i(t) + \epsilon(t) * \exp\left(-\frac{k_i}{\lambda(t)}\right) * (x - w_i(t))$$

tel que :

$$\lambda(t) = \lambda_i(\lambda_f/\lambda_i)^{t/t_{max}},$$

$$\epsilon(t) = \epsilon_i(\epsilon_f/\epsilon_i)^{t/t_{max}},$$

où $\lambda_i = 10$, $\lambda_f = 0,01$, $\epsilon_i = 0,5$, $\epsilon_f = 0,005$ et t_{max} est l'itération maximale fixée par l'utilisateur

- 6: Incrémenter le paramètre $t : t = t + 1$.
 - 7: Retourner à l'étape 3, si $t < t_{max}$.
-

Elle lui permet en effet de s'adapter facilement au traitement des distributions multidimensionnelles complexes. Elle s'avère donc être une candidate idéale pour le traitement de données documentaires.

Ses inconvénients principaux résident dans le fait qu'elle impose de fixer a priori certains paramètres, tels que le nombre de neurones et le nombre d'itérations d'apprentissage. L'impossibilité de visualiser directement les classes (neurones) générées, étant donné que celles-ci restent définies dans l'espace originel de la description des données, représente également un inconvénient important. Des solutions qui répondent à l'ensemble de ces problèmes seront cependant proposées dans le cadre de notre thèse.

2.3.4 Neural gas plus competitive Hebbian learning

Cette méthode [Martinetz, Thomas and Schulten, Klaus, 1991; Martinetz and Schulten, 1994] est une superposition directe de la méthode neural gas et de l'apprentissage hebbien compétitif. On lui donne parfois le qualificatif de «construction de réseau dirigée par la topologie» [Martinetz and Schulten, 1994]. Son principe est d'adapter les vecteurs de référence des neurones selon la méthode neural gas et de créer une connexion spécifique entre le neurone gagnant et le second gagnant, ce qui représente l'apprentissage hebbien compétitif. Un mécanisme d'élimination des connexions qui ne sont pas régulièrement activées est également mis en place.

Le qualificatif de «construction de réseau dirigée par la topologie» est également applicable au modèle growing neural gas décrit ci-après.

2.3.5 Growing neural gas (GNG)

Cette méthode [Fritzke, 1995; Fritzke, 1997] diffère des méthodes précédentes par le fait que le nombre de neurones qu'elle utilise change périodiquement, en partant de deux neurones aléatoires. Ainsi, un neurone est ajouté périodiquement entre le neurone présentant le cumul d'erreur le plus important et son voisin ayant le cumul d'erreur le plus important. Seuls le neurone le plus proche de la donnée d'entrée et ses voisins ont leurs poids modifiés par l'apprentissage, suivant le même type de loi de celle de l'algorithme de Kohonen. Si une unité se retrouve sans connexion, elle est supprimée. Le nombre de neurones augmente jusqu'à un nombre maximum prédéfini ou plus généralement jusqu'à ce que le critère d'arrêt soit atteint (cf. figure 2.8).

Cette méthode s'est prouvée la plus efficace pour le traitement de données bidimensionnelles ou tridimensionnelles. Les paramètres d'apprentissage qui lui sont associées ont d'ailleurs été spécifiquement adaptés à ce type de données. Du fait du choix des critères d'arrêt (nombre de neurones minimal, performance minimale), elle ne présente cependant pas de garantie de terminaison. Ce problème peut plus spécifiquement survenir dans le cas de distributions multidimensionnelles complexes ou ultra-croisées.

2.3.6 SOM Incrémental Hiérarchique (GHSOM)

Rauber et al. ont proposé un nouvel algorithme, appelé SOM incrémental hiérarchique (GHSOM) [Rauber *et al.*, 2002], qui adresse deux limitations de SOM, à savoir sa structure statique, et ses capacités limitées à représenter des relations hiérarchiques entre les données. GHSOM donc est basé sur deux principes :

- une stratégie de croissance incrémentale du nombre de neurones, qui permet a priori de surmonter la détermination de taille de carte,

Algorithme 4 Algorithme d'apprentissage du modèle Growing Neural Gas

- 1: L'ensemble A des vecteurs de référence est initialisé avec deux neurones aléatoires a et b .
L'ensemble C des connexions et l'ensemble des voisins de a et de b sont vides.
- 2: Choisir un vecteur d'entrée x de l'ensemble d'apprentissage.
- 3: Déterminer les deux vecteurs de référence de A , s_1 et s_2 , les plus proches de x :

$$\|x - w_{s_1}\| \leq \|x - w_{s_2}\| \leq \|x - w_i\|, \quad \forall i \in A$$

- 4: Créer, si elle n'existait pas, une connexion entre s_1 et s_2
Si $(s_1, s_2) \in C$, alors $C = C \cup \{(s_1, s_2)\}$ et $N_{s_1} = N_{s_1} \cup \{s_2\}$
où N_{s_1} est l'ensemble des voisins topologiques de s_1 .
- 5: Mettre l'âge de la connexion à zéro : $age_{s_1, s_2} = 0$
- 6: Mettre à jour l'erreur locale du vecteur de référence gagnant :

$$E_{s_1}(t+1) = E_{s_1}(t) + \|x - w_{s_1}\|^2$$

- 7: Mettre à jour le vecteur de référence gagnant et ses voisins :

$$w_{s_1}(t+1) = w_{s_1}(t) + \epsilon_b(x - w_{s_1})$$

$$w_i(t+1) = w_i(t) + \epsilon_n(x - w_i), \quad \forall i \in N_{s_1}$$
- 8: Incrémenter la durée de vie de tous les voisins du vecteur de référence gagnant i.e.
si $s_1, j \in C$ alors $age_{s_1, j}(t+1) = age_{s_1, j}(t) + 1 \quad \forall i \in N_{s_1}$
- 9: Supprimer toutes les connexions du vecteur de référence gagnant qui excèdent la durée de vie ie
si $age_{s_1, j} > age_{max}$ et $s_1, j \in C$ alors $C = C \setminus s_1, j$
Si un vecteur de référence n'est plus connecté, i.e. n'a plus de voisins, alors il est supprimé.
- 10: Si le nombre d'exemples présentés est un multiple de la période λ , alors créer un vecteur de référence :

1. insérer un nouveau vecteur de référence r entre le vecteur de référence q dont l'erreur locale est la plus importante et son voisin f ayant l'erreur locale la plus forte.

$$A = A \cup \{r\}$$

$$w_r = 0.5(w_q - w_f)$$

2. connecter r à q et f : $C = C \cup \{(r, q), (r, f)\}$
et supprimer la connexion entre q et f : $C = C \setminus (q, f)$

3. décroître l'erreur locale de q et f

$$E_q(t+1) = E_q(t) - \alpha E_q(t) \text{ et } E_f(t+1) = E_f(t) - \alpha E_f(t)$$
interpoler l'erreur locale de r : $E_r = 0.5(E_q + E_f)$

- 11: Décroître l'erreur locale de tous les vecteurs de référence :

$$E_i(t+1) = E_i(t) - \beta E_i(t) \quad \forall i \in A$$

- 12: Si le critère d'arrêt (taille du réseau ou mesure de performance) n'est pas atteint, alors aller en 2.

L'initialisation des paramètres expérimentaux précédents est donnée d'après Fritzke [Fritzke, 1997] comme suit : $\lambda = 300$ $\epsilon_b = 0,05$ $\epsilon_n = 0,0006$ $\alpha = 0,5$ $\beta = 0,0005$ et $age_{max} = 88$.

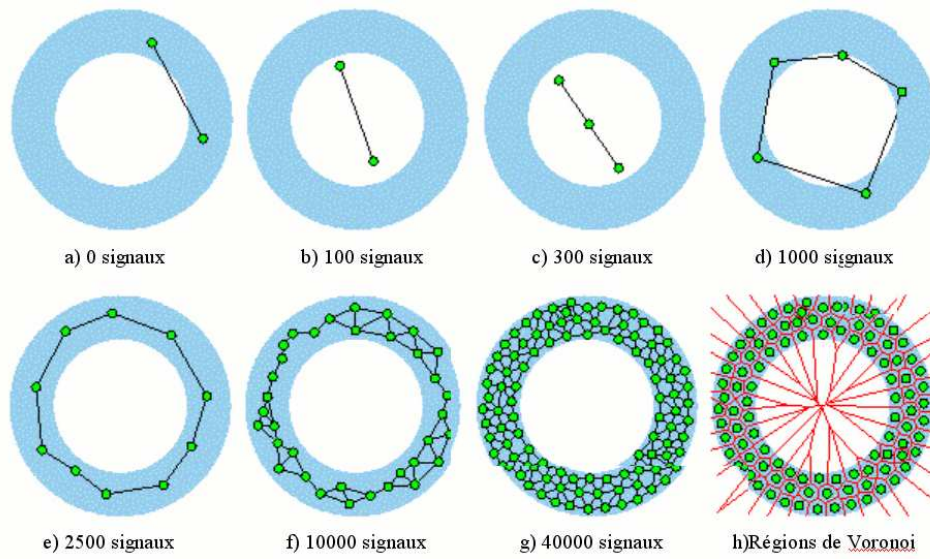


FIG. 2.8 – Classification de Growing Neural Gas d'une distribution en «bague».

– une évolution hiérarchique des structures qui reflète la structure hiérarchique des données d'entrée (cf. figure 2.9).

Pour plus de détails voir l'algorithme 5.

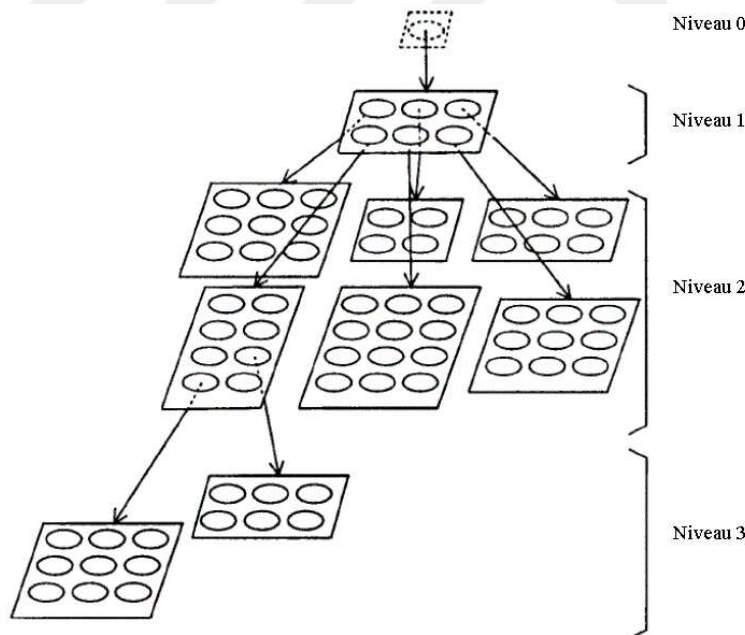


FIG. 2.9 – Exemple de structure générée par GHSOM.

Du fait du paramètre fixe qui définit le nombre maximal de données associées à chaque neurone, cette méthode a cependant tendance à sur-générer des neurones dans les zones très denses et à en sous-générer dans les zones moins denses. De plus, elle isole les cartes se trouvant

Algorithme 5 Algorithme de SOM incrémental hiérarchique (GHSOM)

- 1: Le niveau 0 contient seulement un neurone, son vecteur de référence représente le profil moyen de toutes les données d'entrée. Son erreur quadratique MQE_0^* est l'erreur quadratique moyenne de l'ensemble de données entier :

$$MQE_0^* = \frac{1}{M} \sum_{x_k} \|x_k - w_0\|^2$$

- 2: Le niveau 1 contient une carte SOM de 2×2 neurones.
- 3: Commencer le processus d'apprentissage de SOM.
- 4: Décision :
 - Insérer une nouvelle ligne ou colonne pour améliorer la qualité de la représentation dans ce niveau si $MMQE > \tau_1 MQE_0$, puis, aller à 4, où τ_1 est un paramètre défini par l'utilisateur, MQE_0 est l'erreur quadratique d'un neurone parent du niveau supérieur, $MMQE$ est la moyenne de la somme de l'erreur quadratique moyenne de chaque neurone i (MQE_i) :

$$MQE_i = \frac{1}{|c_i|} \sum_{x_k \in c_i} \|x_k - w_i\|^2$$

- Étendre les unités au prochain niveau hiérarchique pour améliorer la qualité de représentation dans ce niveau hiérarchique si $MQE_i > \tau_2 MQE_0^*$, ou, si $|c_i| > S$ avec $S = 10$, par exemple, puis aller à 6, où τ_1 est un paramètre prédéfini par l'utilisateur, $1 > \tau_1 > \tau_2 > 0$, et $|c_i|$ représente le nombre de données associées au neurone c_i .
 - Si la structure résultante représente bien les données, alors fin de l'algorithme.
- 5: Insérer une nouvelle ligne ou colonne, puis, aller à 3.
 - 6: Pour chaque neurone à étendre : créer une nouvelle carte SOM de 2×2 neurones, en préservant l'orientation des neurones parents, puis, aller à 3, en utilisant seulement le sous-ensemble des données qui sont déjà projetées sur le neurone parent étendu.
-

aux niveaux les plus bas, sans préserver leurs liens potentiels.

2.3.7 Le modèle MultiSOM

Toutes les méthodes de classification précédemment décrites opèrent une classification globale, c'est-à-dire, qu'elles utilisent l'espace de description global de données pour l'étape de classification. L'efficacité, en terme de réduction du bruit du principe de partitionnement de l'espace descriptif des données en points de vue avait déjà été démontrée par Lamirel [Lamirel, 1995] dans le domaine de la recherche documentaire pure. Lamirel a ensuite proposé de transposer ce principe au domaine de classification en développant le modèle MultiSOM qui représente une extension de carte SOM de Kohonen à un contexte multivues. Dans le modèle MultiSOM, chaque point de vue est représenté par un ensemble spécifique de cartes comprenant lui-même une carte de base et plusieurs généralisations. La restitution de la vue globale de l'analyse est opérée par l'utilisation de ce mécanisme de communication intercartes. Ce dernier mécanisme introduit également le principe de dynamique dans l'analyse des données.

Mécanisme de communication intercartes.

Cette communication est basée sur l'utilisation des données qui ont été re-projetées sur les cartes comme neurones intermédiaires de transmission entre les différentes cartes. Les cartes se comportent alors comme un système de croisement entre des thématiques (classes) appartenant à des points de vue différents, ou entre des vues de différents niveaux de généralité. La communication entre les cartes peut être décomposée en deux étapes successives (cf. figure 2.10) :

1. Génération d'une activité thématique, ciblée ou diffuse, sur une carte source,
2. Transmission de l'activité vers des cartes destination.

L'activité thématique de la carte source peut être générée de manière interactive ou automatisée, soit sur des neurones de la carte, soit plus globalement sur des aires de celle-ci. Cette activité se matérialise sous la forme d'un ensemble de décisions thématiques modales (choix ou rejet), associées à des activités neuronales prédéfinies. Les thématiques actives (positivement ou négativement) représenteront les thématiques originelles du croisement thématique.

La transmission d'activité aux cartes destinations s'opère elle-même en deux étapes (cf. figure 2.10) :

1. Une transmission descendante de la carte source vers les neurones-documents,
2. Une transmission ascendante des neurones-documents vers les cartes destinations.

L'activité d'une classe i d'une carte destination T dérivant de l'activité d'une carte source S peut être calculée comme :

$$A_i^T = f_{n \in i}(g(A_n)), \quad A_n = g(A_{j_n}^S)$$

Où n représente un neurone associé à une donnée, j_n sa classe d'appartenance sur la carte source, f est une fonction évaluant la force de croisement thématique telle que celle décrites ci-après g est une fonction de biais décrite ci-après.

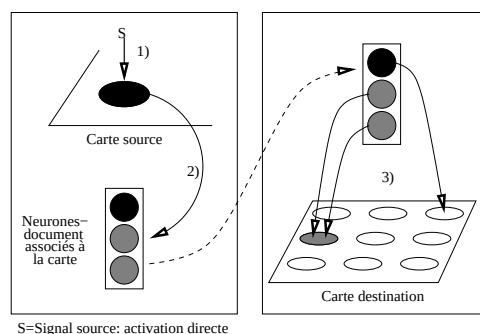


FIG. 2.10 – Mécanisme de communication intercartes

Les paramètres du mécanisme de communication entre les cartes susceptibles d'être choisis par l'analyste correspondent aux modes de calcul du croisement thématique matérialisé par la fonction f de l'équation de propagation précédemment décrite. Ce mode peut être soit probabiliste, soit possibiliste :

- En mode possibiliste : chaque classe d'une carte destination hérite de l'activité retransmise par ses données les plus actives. La fonction f peut alors être décrite comme :

$$f = \text{Max}_{n \in i}(A_n^+) + \text{Max}_{n \in i}(A_n^-)$$

A^+ représente une valeur d'activité positive (matérialisant un choix), et A^- une valeur d'activité négative (matérialisant un rejet). Cette approche peut permettre de détecter des croisements faibles, ou signaux faibles, entre des thèmes appartenant à des points de vue différents.

- En mode probabiliste : chaque classe hérite de l'activité moyenne retransmise par l'ensemble des données associées. La fonction f peut alors être décrite comme :

$$f = \frac{1}{\|i\|} \sum_{n \in i} A_n \quad (2.8)$$

$\|i\|$ représente le nombre de données associées à la classe i . Contrairement au calcul possibiliste, le calcul probabiliste donne une indication plus fiable de la force de croisement thématique. Il peut donc être utilisé pour distinguer entre les croisements forts et les croisements faibles.

Fonction de biais : le rôle de la fonction g , qui peut être utilisée d'une manière optionnelle, est celui de moduler la transmission d'activité entre les cartes en tenant en compte des degrés de similarité entre les classes et les données.

Cependant, la fonction f (cf. équation 2.8) ne permet pas de mesurer précisément la distribution des modalités d'activité générées par la carte source sur la carte destination. Nous proposons pour cela, dans le chapitre 5, une modification du comportement de la communication intercarte de MultiSOM basée sur l'analogie avec un modèle de réseau bayésien.

Mécanisme de généralisation.

Une fonctionnalité importante pour assister à l'analyse des informations fournies par une carte consiste à résumer le contenu de cette dernière sous la forme d'un ensemble de classes plus génériques par l'intermédiaire d'un processus de généralisation en ligne. La méthode réduit le nombre de classes (i.e. le nombre de neurones) d'une carte de manière cohérente. Comme les cartes originales ont été construites sur la base d'une structure de voisinage carrée entre les neurones, la transition d'un niveau de généralisation vers un autre s'opère alors en construisant un nouvel ensemble de neurones, dans lequel le profil de chaque neurone représente la composition moyenne des profils de quatre voisins immédiats sur le niveau original. La figure 2.11 illustre le fonctionnement du processus de généralisation.

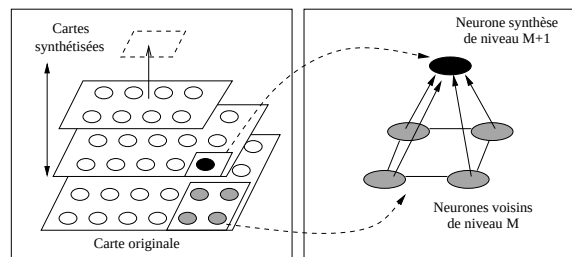


FIG. 2.11 – Mécanisme de généralisation

La méthode de généralisation proposée peut être vue comme la détermination de l'ensemble des centres de gravité d'un niveau source pour construire un nouveau niveau. Soit

$n \times m$ ($n, m \geq 2$) les dimensions en nombre de neurones d'une carte d'un niveau donné, le processus de généralisation produira alors à l'étape suivante une carte de niveau immédiatement plus général de dimensions $(n - 1) \times (m - 1)$. L'avantage de cette méthode est celui de préserver la structure originale de voisinage sur l'ensemble des niveaux générés. Elle permet de plus d'assurer la conservation des propriétés topographiques des vecteurs de référence de neurones [Lamirel *et al.*, 2000]. Cette méthode peut finalement s'apparenter à une forme implicite et distribuée de classification hiérarchique ascendante basée sur le voisinage réciproque (cf. section 2.2.2).

La méthode de généralisation peut être plus spécifiquement utilisée pour mesurer la cohérence des classes qui ont été déterminées sur la carte originale. En effet, du fait des propriétés topographiques inhérentes aux cartes (SOM), la cohérence d'une thématique donnée est directement liée à la stabilité, durant les différentes étapes de la généralisation, du nombre de classes qui la représente.

2.4 Évaluation de la qualité de classification

Beaucoup de méthodes de classification sont proposées dans la littérature, ce qui amène naturellement à évaluer les solutions qu'elles proposent. Nous présentons donc ci-après quatre types de critères permettant de mener à bien une telle évaluation.

2.4.1 Qualité basée sur la distance (critères inertiels)

Les mesures les plus répandues pour évaluer la qualité de la méthode de classification sont basées sur les critères inertiels, à savoir l'inertie intra-classe et l'inertie inter-classes [Lebart *et al.*, 1982].

- L'inertie intra-classe peut être définie comme suit :

$$\frac{1}{|C|} \sum_{c \in C} \frac{1}{|c|} \sum_{d \in c} \|p_c - p_d\|^2 \quad (2.9)$$

- L'inertie inter-classes peut être définie comme suit :

$$\frac{1}{|C|^2 - |C|} \sum_{c_i \in C} \sum_{c_j \in C, c_j \neq c_i} \|p_{c_i} - p_{c_j}\|^2 \quad (2.10)$$

où $|C|$ représente le nombre de classes non vides associées à une classification, $|c|$ représente le cardinal d'une classe c (c'est-à-dire, le nombre de données associées à la classe c), d représente une donnée de la classe, p_x représente le vecteur (ou le profil) associé à la donnée x et $\|\cdot\|$ représente la distance euclidienne.

L'inertie intra-classes évalue l'homogénéité de données présentes à l'intérieur des classes alors que l'inertie inter-classe évalue l'hétérogénéité entre les différentes classes.

Des valeurs faibles d'inertie intra-classe désignent des classes homogènes vis à vis des données qui la composent et des valeurs fortes d'inertie inter-classes désignent des classes hétérogènes entre elles. Ces grandeurs constituent donc des critères de qualité des classes, et par conséquent de la classification proprement dite.

Cependant, ces mesures sont souvent fortement biaisées, car les dimensions intrinsèques des profils de classes (nombre de composants différents de zéro dans les profils) ne sont pas du

même ordre de grandeur que les dimensions intrinsèques des profils de données. Ceci est spécialement vrai dans le domaine documentaire où le nombre d'index présent dans les documents est souvent extrêmement faible relativement à la dimension de leur espace global de description. Ce phénomène cause une augmentation anormale de l'inertie intra-classe. La normalisation des profils de classes pendant la phase de construction de classes, telle que celle proposée par les modèles de classification sphérique comme K-means axial (cf. section 2.2.1), peut cependant indirectement aider à résoudre ce problème.

Un second biais, propre à la mesure intra-classe, vient du fait qu'elle ne permet pas de distinguer correctement les classes cohérentes de celles qui ne le sont pas. En effet, pour la même valeur de l'inertie intra, les éléments pourraient être écartés autour du profil de classe dans les classes incohérentes, comme ils pourraient être groupés ensemble à une distance donnée du profil de classe dans les classes cohérentes. Pour traiter ce problème, une autre mesure, appelée l'inertie moyenne, a été proposée par Yahya [Yahya, 1997]. Cette mesure utilise le profil des données de la classe, et est définie comme suit :

$$\frac{1}{N} \sum_{i \in C, i=1}^{n-1} \sum_{j \in C, j=i+1}^n \|p_{d_i} - p_{d_j}\|^2, \quad N = \sum_{c \in C} \frac{|c|(|c| - 1)}{2} \quad (2.11)$$

Cette mesure est plus stable que l'inertie intra-classe étant donné qu'elle est moins sensible au nombre de données classées ensembles. De plus, elle ne dépend que des profils de données associées à la classe. Malheureusement, elle ne peut que donner une évaluation moyenne de la cohérence, du fait qu'elle gère les profils de données de classe d'une manière globale.

Il est également à noter que, dans le cas d'une classification de Kohonen, le phénomène de biais est accentué par le comportement de la méthode qui a tendance à rapprocher les profils des classes de manière à conserver les propriétés topographiques sur la carte, faisant par conséquent diminuer artificiellement la valeur de l'inertie inter-classes.

2.4.2 Qualité basée sur la similarité

Une autre manière d'appréhender la qualité d'une méthode de classification consiste à se baser sur des mesures de corrélation. Dans [Zhao and Karypis, 2001; Zhao and Karypis, 2002], les auteurs proposent d'utiliser la similarité cosinus (cf. équation 1.5). Cette dernière se base à la définition de trois types de critère :

1. Les **critères internes** ($\mathcal{I}_1$ et $\mathcal{I}_2$) :

- Le critère $\mathcal{I}_1$ maximise la somme des similarités moyennes entre les données associées à chaque classe :

$$\text{maximize } \mathcal{I}_1 = \sum_{c \in C} \frac{1}{|c|} \sum_{d_i, d_j \in c} \cos(d_i, d_j) \quad (2.12)$$

- Le critère $\mathcal{I}_2$ maximise la similarité entre chaque donnée et le centroïde (ou profil) de la classe à laquelle est associé :

$$\text{maximise } \mathcal{I}_2 = \sum_{c \in C} \sum_{d_i \in c} \cos(d_i, c) \quad (2.13)$$

2. Le **critère externe** (ε_1) minimise la similarité entre le centroïde de chaque classe (c) et le centroïde de toutes les classes (C) :

$$\text{minimise } \varepsilon_1 = \sum_{c \in C} |c| \cos(c, C) \quad (2.14)$$

3. Les **critères hybrides** ($\mathcal{H}_1$ et $\mathcal{H}_2$) combinent les critères $\mathcal{I}_1$ avec ε_1 et $\mathcal{I}_2$ avec ε_1 , respectivement, soit :

$$\text{maximise } \mathcal{H}_1 = \frac{\mathcal{I}_1}{\varepsilon_1} \quad (2.15)$$

$$\text{maximise } \mathcal{H}_2 = \frac{\mathcal{I}_2}{\varepsilon_1} \quad (2.16)$$

Ces critères présentent cependant les mêmes défauts que les critères inertiels décrits précédemment.

2.4.3 Qualité basée sur la méthodologie et la métrique expérimentales

Selon cette approche, la qualité d'une méthode de classification est évaluée par comparaison de ses résultats avec ceux fournis par une classification de référence. Cette qualité est mesurée en utilisant trois métriques différentes, le FScore, l'entropie et la pureté [Zhao and Karypis, 2001; Zhao and Karypis, 2003] :

- Le **FScore** : cette mesure combine les valeurs de rappel et de précision (cf. section 1.3.5). L'auteur traite chaque cluster comme s'il était le résultat d'une requête et chaque classe comme s'elle était un ensemble de données attendues pour cette requête. Le rappel et la précision de ce cluster pour chaque classe donnée sont calculés ensuite. Pour un cluster c et une classe l , il vient :

$$\text{Rappel}(l, c) = R(l, c) = \frac{n_{lc}}{|l|}$$

$$\text{Précision}(l, c) = P(l, c) = \frac{n_{lc}}{|c|}$$

où n_{lc} représente le nombre de données associées au cluster c appartenant à la classe l , $|l|$ représente le cardinal de la classe l et $|c|$ représente le cardinal du cluster c .

Donc, la F – mesure de cluster c et une classe l est donnée comme :

$$F(l, c) = \frac{2 * R(l, c) * P(l, c)}{R(l, c) + P(l, c)} \quad (2.17)$$

Pour une classification hiérarchique (arbre hiérarchique) la F – mesure d'une classe est la valeur maximale qui atteint n'importe quel nœud dans l'arbre, par conséquent, la F – mesure est calculée comme :

$$FScore = \sum_{l \in L} \frac{|l|}{n} FScore(l) \quad (2.18)$$

où

$$FScore(l) = \max_{c \in T} F(l, c)$$

tel que L représente l'ensemble total de classes, T représente un arbre hiérarchique et n est le nombre total de données dans la base de données.

La classification est dit optimale dans le cas où chaque classe a un cluster correspondant qui contient le même ensemble de données dans le résultant de classification, dans ce cas le $FScore$ sera un. En général, plus les valeurs de $FScore$ sont hautes, plus la solution

de classification est meilleure.

- l'**entropie** : d'une manière différente du *FScore*, qui évalue la qualité d'un arbre hiérarchique en utilisant seulement un sous-ensemble de ses nœuds (clusters), l'entropie prend en compte de la distribution des données dans tous les nœuds (clusters) d'un arbre hiérarchique (classification non hiérarchique).

$$E(T) = \sum_{c \in T} \frac{1}{|T|} E(c) \quad (2.19)$$

$$E(c) = -\frac{1}{\log |L|} \sum_{l \in L} \frac{n_c^l}{|c|} \log \frac{n_c^l}{|c|}$$

où n_c^l est le nombre de données de classe l qui sont assignées au nœud (cluster) c .

La classification est dite optimale dans le cas où chaque cluster contient des données d'une classe seulement. En général, plus la valeur de l'entropie est basse plus la solution de classification est meilleure.

- la **pureté** : elle mesure à quel moment le cluster contient des données d'une classe :

$$\text{Pureté} = \sum_{c \in C} \frac{|c|}{n} P(c) \quad (2.20)$$

$$P(c) = \frac{1}{|c|} \max_l (n_c^l)$$

En général, plus forte est la valeur de la pureté, meilleure sera la solution de classification.

Malheureusement, ces critères nécessitent une connaissance a priori des classes ainsi la distribution de données associés à ces dernières. Par conséquent nous ne pouvons pas les utiliser pour une évaluation non supervisée.

2.4.4 Qualité basée sur l'algèbre linéaire

Mather [Mather, 2000] a défini une méthode de mesure de qualité de classification basée sur l'algèbre linéaire. Il en propose une application spécifique aux données documentaires. Le principe de cette méthode est de considérer qu'un ensemble de documents est idéalement classable si chaque classe peut faire intervenir un ensemble de termes disjoints. Elle revient à comparer les valeurs propres de la matrice terme-document A à celle d'une matrice bloc-diagonale (i.e. de classes disjointes) B , dérivée de la matrice A par un processus de bloc-sérialisation. La comparaison est effectuée en utilisant la métrique η définie comme suit :

$$\eta = \frac{s_1^T s_2}{||s_1|| ||s_2||} \quad (2.21)$$

où s_1 représente l'ensemble de valeurs singulières obtenues par SVD de la matrice originale, s_2 représente l'ensemble de valeurs singulières obtenues de la matrice bloc-diagonale dérivée.

La métrique η repose sur une propriété importante de la décomposition en valeurs singulières (SVD) (cf. section 1.3.4) qui stipule que si une matrice A est bloc-diagonalisable, sa décomposition en valeurs singulières sera équivalente à celle de sa transformation bloc-diagonale, ce qui revient à dire que les deux matrices posséderont le même ensemble de valeurs propres.

Ainsi, la valeur de η sera maximale, dans le cas où les données sont idéalement classables. En effet, dans ce dernier cas, la matrice A sera une matrice bloc-diagonalisable et la matrice B représentera alors sa transformation bloc-diagonale effective.

L'inconvénient de cette mesure repose cependant sur le fait qu'elle se focalise sur les descripteurs de données originales, en évaluant leur capacité à être classifiés ensemble. Par conséquent, elle ne peut pas être employée a posteriori, ni pour comparer les résultats de différentes méthodes de classification, ni pour optimiser les résultats d'une méthode de classification donnée.

2.5 Conclusion

Nous avons présenté dans ce chapitre deux approches de classification non supervisée des données, à savoir l'approche classique et l'approche neuromimétique. Concernant l'approche classique, les méthodes directes nécessitent de fixer arbitrairement le nombre de classes désirées, et sont particulièrement sensibles aux conditions initiales. Les méthodes hiérarchiques, quant à elles, ne permettent pas de contrôler le nombre de classes obtenues et leurs résultats sont difficiles à exploiter et à interpréter. Les méthodes hybrides et les méthodes orientées par la densité sont utiles pour trouver des classes de taille et de densité différentes, mais elles sont difficiles à employer sur des distributions complexes.

Concernant l'approche neuronale, nous nous sommes focalisés sur les méthodes topographiques non supervisées. Les cartes SOM, ainsi que leurs variantes, telles que MultiSOM, qui reposent sur une topologie fixe, implémentent en une seule étape les tâches de classification et de projection cartographique non linéaire des données. Cette famille de méthodes a tendance à fournir des résultats de classification imprécis du fait de l'usage d'une structure topographique préfixée et, dans certains cas, à être sujette à des effets de bord supplémentaires dus à celui d'une structure discontinue. De leur côté, les méthodes neuronales NG et GNG reposent sur une topologie libre, ce qui leur permet de bien s'adapter au traitement de données multidimensionnelles complexes, comme les données documentaires. Ces méthodes conservent néanmoins quelques défauts. La méthode NG impose de fixer le nombre de neurones et de déterminer le nombre d'itérations d'apprentissage et la méthode GNG ne garantit pas la terminaison de l'apprentissage pour les données fortement multidimensionnelles.

Enfin, nous avons présenté plusieurs critères permettant d'évaluer d'une manière objective l'efficacité des méthodes de classification. Nous avons montré que ces critères posaient beaucoup de problèmes d'utilisation et d'interprétation, ce qui les rend difficilement applicables à la comparaison de méthodes de classification. Nous mettrons donc en place dans le chapitre 4 des mécanismes d'évaluation originaux faisant à la fois office de mécanismes de mesure de qualité de classification et de mécanismes d'optimisation du nombre de classes. Nous montrerons également dans ce chapitre que ces mécanismes permettent de traiter l'ensemble des défauts importants de la méthode NG.

A travers l'exemple du modèle MultiSOM, nous avons également mis en évidence l'intérêt de l'utilisation de modèles de classification basés sur les points de vue en terme de précision de résultat et d'adaptabilité à des types d'analyse impliquant des facteurs multiples. Nous reviendrons donc sur l'intérêt de la généralisation d'un tel modèle à l'ensemble des méthodes de classification non supervisées dans le chapitre 5.

En dehors des méthodes de famille de SOM, les méthodes de classification décrites précédemment ne proposent pas directement de visualisation synthétique de leurs résultats dans le cas du traitement de données multidimensionnelles. Ces méthodes nécessitent donc de faire

appel à des techniques spécifiques de visualisation. Nous présenterons les principes de ces techniques au chapitre suivant. Nous proposons également notre propre méthodologie de visualisation dans le chapitre 6.



Chapitre 3

Visualisation d'information

Sommaire

3.1	Introduction	61
3.2	Dimension intrinsèque des données multidimensionnelles . . .	62
3.3	Visualisation par projection cartographique linéaire	63
3.3.1	L'analyse en composantes principales	63
3.3.2	La projection aléatoire (Random Mapping)	64
3.4	Visualisation par projection cartographique non linéaire . . .	65
3.4.1	Méthodes de projection	65
3.4.2	Exemples applicatifs de visualisation dans le domaine documentaire	68
3.5	Visualisation hyperbolique	74
3.5.1	L'espace hyperbolique	75
3.5.2	Arbre hyperbolique (Hypertree)	75
3.5.3	Modèles neuronaux hyperboliques de la carte de Kohonen	76
3.6	Conclusion	79

3.1 Introduction

Nous avons décrit dans le chapitre précédent différentes méthodes de classification. La plupart du temps les classes issues de ces méthodes sont représentées dans le même espace de description que celui des données sources, qui peut s'avérer fortement multidimensionnel. La représentation synthétique des résultats de ces méthodes se ramène donc à un problème de visualisation sur un plan de représentation. Le domaine des techniques de visualisation est un domaine très riche. Différentes techniques peuvent être appliquées en fonction du but recherché et des propriétés des données elles-mêmes. Les techniques les plus couramment utilisées sont la projection sur un plan bidimensionnel, l'organisation en arbres, et l'organisation en réseau ou en graphe.

Les techniques de projection sur un plan bidimensionnel, à savoir les méthodes linéaires et non linéaires, ont pour but d'optimiser le dépliage en deux dimensions de la structure de données originale. Les techniques de visualisation de données en arbre, en réseau ou en graphe sont, quant à elles, spécialisées dans la représentation conjointe des données et de leurs relations potentielles.

Plusieurs facteurs importants, liés à l'ergonomie, sont également à prendre en compte dans le choix d'une technique de visualisation. En effet, une technique de visualisation appropriée doit, dans la mesure du possible, répondre aux contraintes suivantes [Bouthier, 2004] :

1. avoir une vue globale : permet de se faire une idée de l'ensemble des données,
2. zoomer : permet d'accéder, dans le contexte, à une donnée spécifique,
3. avoir une vue en détail : permet d'accéder à tous les détails d'une donnée spécifique,
4. voir les relations : permet de se concentrer sur les relations entre les données et non pas sur les données elles-mêmes.

L'opération d'étiquetage des données est également une opération importante pour l'interprétation intuitive et synthétique d'un résultat de visualisation. Elle s'avère même cruciale dans le cadre spécifique de la visualisation de résultats d'analyse et de fouille de données documentaires (cf. chapitre 2).

Nous présenterons plus spécifiquement dans ce chapitre les problèmes liés à la visualisation des données fortement multidimensionnelles et les solutions ou les techniques proposées pour visualiser ce type de données. Nous présenterons ci-après trois familles importantes de techniques, à savoir la visualisation par la projection linéaire, la visualisation par la projection non linéaire, et la visualisation hyperbolique.

3.2 Dimension intrinsèque des données multidimensionnelles

Dans cette section nous abordons succinctement quelques problèmes de fond posés par la projection des données multidimensionnelles, qui peut être considérée comme un processus central pour la visualisation.

Avant toute opération de projection de données multidimensionnelles, il est naturellement important d'évaluer la dimension idéale de l'espace cible pour représenter ces données. En effet, si la dimension de l'espace cible est trop faible, une partie de l'information sera perdue lors de la projection. Si elle est trop grande, l'intérêt de la projection sera en partie perdu. Le concept de dimension intrinsèque est le plus souvent employé pour évaluer cette dimension. La dimension intrinsèque correspond au nombre effectif de degrés de liberté d'un ensemble des données [Lendasse *et al.*, 2001]. Ce concept est présenté ici en prenant comme exemple la distribution bien connue du fer à cheval (cf. figure 3.1) : pour cette distribution de données, la dimension intrinsèque est égale à deux, car deux degrés de liberté sont suffisants pour déterminer de manière univoque n'importe quelles données appartenant à la distribution, bien que ces données soient initialement définies dans $\mathbb{R}^3$. La détermination de la dimension intrinsèque reste cependant très problématique dans le cas des données fortement multidimensionnelles.

Un phénomène important est également à prendre en compte lors du traitement des données multidimensionnelles. Ce phénomène est appelé «phénomène de l'espace vide» [Verleysen, 2003]. Considérons par exemple une hyper-sphère en n dimensions. Le volume de cette hyper-sphère en fonction du rayon r vaut :

$$V(n) = \frac{\pi^{n/2} r^n}{\Gamma(1 + n/2)} = V(n-2) \frac{\pi r^2}{d-2} \quad (3.1)$$

avec $V(1) = 2$ $V(n) = \pi$. L'observation de l'évolution de $V(n)$ où $r = 1$. L'observation de la figure 3.2 montre que le volume décroît vers 0 quand la dimension n s'accroît, c'est-à-dire que plus la dimension d'un espace est grande, plus celui-ci semble vide. Ce qui revient

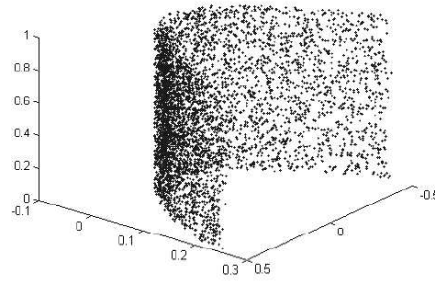


FIG. 3.1 – Distribution du «fer à cheval».

corrélativement à dire que, dans l'espace multidimensionnel, les données ont naturellement tendance à migrer vers un hypercentre virtuel et que, de ce fait, les relations entre celles-ci ont tendance à se complexifier plus que prévu. Ce type de comportement représente un problème supplémentaire pour la visualisation.

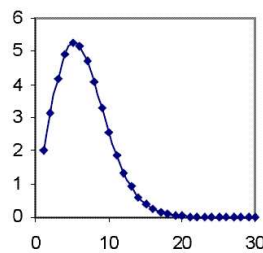


FIG. 3.2 – Évolution du volume de sphère en fonction de nombre de dimensions.

Les remarques précédentes montrent que la projection des données fortement multidimensionnelles représente un problème très difficile à traiter directement. Cependant, toutes les variables ou tous les descripteurs de l'espace de description ne sont pas d'égale importance pour comprendre la structure intrinsèque qui sous-tend une distribution de données multidimensionnelles. Une manière de contourner les problèmes posés par la projection d'un tel type de distribution consiste donc à éliminer ou atténuer l'effet des variables ou des descripteurs marginaux. Ce principe est appliqué dans l'ensemble des méthodes de projection que nous décrivons ci-après.

3.3 Visualisation par projection cartographique linéaire

3.3.1 L'analyse en composantes principales

L'analyse en composantes principales (ACP) est une méthode de réduction de dimension linéaire [Jolliffe, 1986]. Son principe est la description des données aux combinaisons linéaires orthogonales (les CP) des variables originales possédant la plus grande variance. La première CP, est la combinaison linéaire avec la plus grande variance. La deuxième CP est la combinaison linéaire avec la deuxième plus grande variance, orthogonale à la première CP, et ainsi de suite. Le nombre de CP est égale au nombre de dimensions de l'espace de description des données

originales. Pour beaucoup d'ensembles de données, les premiers CP expliquent la partie majeure de la variance, donc, le reste peut être négligé au prix d'une perte d'information minimale.

Étant donnée la matrice de covariance :

$$\Sigma_{n \times n} = \frac{1}{N} X X^T$$

où $X = \{x_{i,j} : 1 \leq i \leq n, 1 \leq j \leq N\}$ représente la matrice variable-donnée.

nous pouvons employer le théorème de décomposition spectral pour écrire Σ comme :

$$\Sigma = U \Lambda U^T$$

où $\Lambda = \text{diag}(\lambda_1, \dots, \lambda_n)$ est la matrice diagonale des valeurs propres ordonnées $\lambda_1 \leq \dots \leq \lambda_n$ et U est une matrice diagonale $n \times n$ qui contient des vecteurs propres. Les CP sont données par les n lignes de la matrice S de $n \times N$ [Imola K. Fodor, 2002], où :

$$S = U^T X \quad (3.2)$$

Le nombre de CP à conserver peut être choisi, soit en fixant un nombre k tel que $k < n$, soit en fixant un seuil λ_0 , en suite, soit en ne gardant que les vecteurs propres dont les valeurs propres sont supérieures à λ_0 . Cette dernière méthode est préférée par [Jolliffe, 1972].

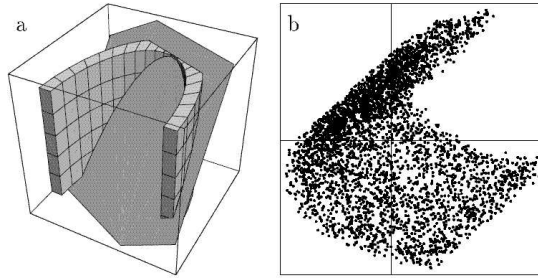


FIG. 3.3 – Distribution en «fer à cheval» : (a) Distribution et plan principal trouvé par l'ACP . (b) Projection de la distribution sur ce plan.

Cette méthode de projection prend en compte uniquement les dépendances linéaire entre les variables. Donc, elle n'est pas capable de projeter correctement une distribution non linéaire, comme la distribution en «fer à cheval» (cf. figure 3.3). La figure montre que cette méthode est incapable de restituer une information pertinente sur cette distribution en considérant sa projection selon les deux premiers CP dans ce type d'exemple.

3.3.2 La projection aléatoire (Random Mapping)

La méthode linéaire la plus simple est celle du Random Mapping (projection aléatoire) [Kaski, 1998; Lagus *et al.*, 1999]. Le Random Mapping consiste simplement à multiplier le vecteur descriptif initial d'un document par une matrice aléatoire pour produire un vecteur descriptif transformé de plus faible dimension, soit :

$$y = R \times x \quad (3.3)$$

où $x \in \mathbb{R}^n$ est un vecteur de document. R est une matrice aléatoire des valeurs de dimension $k \times n$, et la longueur euclidienne de chaque colonne de cette matrice a été normalisée à l'unité. $y \in \mathbb{R}^k$ est le vecteur résultat de cette multiplication, $k < n$.

L'interprétation de projection aléatoire doit considérer ce qui arrive à chaque dimension de l'espace original $\mathbb{R}^n$ dans l'espace de projection. Si la $i^{\text{ème}}$ colonne de R est dénotée par r_i l'opération de la projection aléatoire (cf. équation 3.3) peut être exprimée comme :

$$y = \sum_i x_i r_i \quad (3.4)$$

Il a été démontré que la méthode du Random Mapping produit des résultats similaires à l'ACP, alors que cette dernière méthode s'avère plus coûteuse, surtout si la dimension originale des données est très grande. En dehors de son efficacité opérationnelle cette méthode souffre cependant des mêmes limitations que l'ACP.

Cette méthode est principalement utilisée avant l'étape de classification dans le cadre de l'application WEBSOM (cf. section 3.4.2).

3.4 Visualisation par projection cartographique non linéaire

Comme nous avons mentionné précédemment que les méthodes linéaires prennent uniquement en compte les dépendances linéaires entre les variables, elles ne sont donc pas capable de projeter correctement une distribution non linéaire. Plusieurs méthodes de projection non linéaires ont été proposées pour palier aux problèmes des méthodes linéaires.

3.4.1 Méthodes de projection

De nombreuses méthodes de projection non linéaires des données ont été proposées dans la littérature. Parmi celles-ci, on peut citer : les cartes de Kohonen (SOMs) (cf. section 2.3.2) utilisées pour projeter les données sur une ou deux dimensions, les méthodes basées sur la conservation de la distance (telles que le Multidimensional Scaling (MDS) [Shepard, 1962], la projection de Sammon (NLM) [Sammon, 1969], l'analyse en composantes curvilignes (CCA) [Demartines and Hérault, 1997] et ses extensions, telle que l'analyse par distance curviligne (CDA) [Lee *et al.*, 2000], et Isomap [Tenenbaum *et al.*, 2000]). Ces méthodes sont basées sur le principe général suivant : considérons n points dans l'espace de dimension d , il s'agit alors d'essayer de placer ces points dans un espace de faible dimension m ($m < d$) en préservant les distances mutuelles entre les points issus de l'espace original et les points correspondants dans l'espace de projection. Cette dernière condition est naturellement impossible à réaliser dans l'absolu, du fait qu'il y a $n(n-1)$ conditions à satisfaire avec nm degrés de liberté [Verleysen, 2003]. Ces méthodes cherchent donc en priorité à conserver, lors de la projection, l'intégrité des distances les plus courtes de manière à préserver au mieux la topologie originale.

Multidimensional Scaling (MDS). Dans le cas du Multidimensional Scaling (MDS), où la fonction-objectif est simplement basée sur la minimisation de la somme des rapports des distances entre les données d'entrée et les données projetées.

Projection non linéaire de Sammon (NLM). D'une manière similaire de MDS, la projection de Sammon (NLM), est basé sur une fonction objectif qui est l'erreur quadratique des différence entre les distances des données d'entrée et des distances des données projetées. Cette méthode cherche donc à minimiser la fonction d'erreur suivante :

$$E = \frac{1}{c} \sum_i \sum_{j < i} (X_{ij} - Y_{ij})^2 F(X_{ij}) \quad (3.5)$$

où

$$F(X_{ij}) = \frac{1}{X_{ij}} \quad \text{et} \quad c = \sum_i \sum_{j < i} X_{ij}$$

Un inconvénient majeur des deux méthodes que nous avons présentées, est qu'elles souffrent d'une faible fiabilité de projection pour des données de grande dimension [Sammon, 1969]. Une autre de leurs limitations importantes a été démontrée par Démartines [Demartines, 1994]. Elle est liée au fait que ces méthodes effectuent une projection directe basée sur les distances mesurées dans l'espace d'entrée. Cette stratégie limite sévèrement les possibilités de dépliage de structures complexes offertes par ces méthodes.

Analyse en composantes curvilignes (CCA). L'analyse en composantes curvilignes (CCA) opère en deux phases distinctes. La première phase est une quantification vectorielle de l'espace de données d'entrée. Cette phase est importante pour utiliser un nombre k de points inférieur au nombre n de points original. La deuxième phase est une projection non linéaire des vecteurs d'entrée dans un espace de dimension $m < d$ qui minimise l'erreur quadratique entre les distances. Mais contrairement aux méthodes précédemment décrites, cette minimisation s'opère relativement aux distances dans l'espace de projection. Ces principes augmentent la qualité de projection, d'après Démartines [Demartines and Hérault, 1997]. Ainsi, la méthode CCA consiste à optimiser le critère suivant :

$$E = \sum_{i=1}^k \sum_{j=1}^k (X_{ij} - Y_{ij})^2 F(Y_{ij}, \lambda) \quad (3.6)$$

où X_{ij} est la distance entre les points x_i et x_j dans l'espace d'entrée, Y_{ij} est la distance entre les points correspondants y_i et y_j , $F(\cdot)$ est une fonction décroissante et monotone dont l'étendue est contrôlée par un paramètre λ ajustable. La figure 3.4 montre la projection de la distribution du «fer à cheval» par CCA. Cette méthode détecte donc le caractère uniforme de cette distribution en la dépliant. Un des inconvénients de cette méthode est le choix des paramètres. De plus, malgré les améliorations apportées par rapport aux méthodes précédemment décrites, elle continue cependant à souffrir d'une mauvaise capacité de dépliage des structures complexe.

Analyse par distance curviligne (CDA). Afin d'améliorer la méthode de CCA, Lee et al. [Lee et al., 2000] ont proposé une analyse par distance curviligne (CDA) en utilisant un type de «distance curviligne» δ_{ij} au lieu de la distance euclidienne standard pour X_{ij} (cf. figure 3.5). La distance curviligne est mesurée sur le chemin le plus court passant par des liens liant des

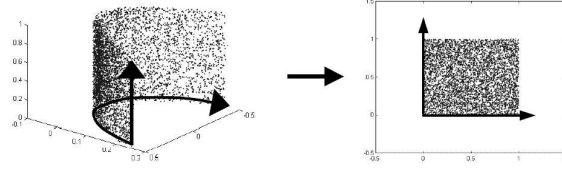


FIG. 3.4 – Projection faite par CCA de $\mathbb{R}^3$ à $\mathbb{R}^2$ de la distribution du «fer à cheval».

centroïdes adjacents (cf. figure 3.6). Les centroïdes sont eux-mêmes obtenus par un processus de quantification vectorielle [Kohonen, 2001] :

$$E = \sum_{i=1}^N \sum_{j=1}^N (\delta_{ij} - Y_{ij})^2 F(Y_{ij}) \quad (3.7)$$

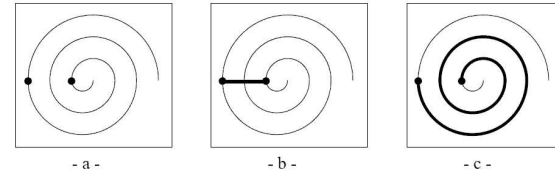


FIG. 3.5 – (a) deux points d'un spirale, (b) la distance euclidienne entre ces deux points et (c) la distance curviligne ou géodésique

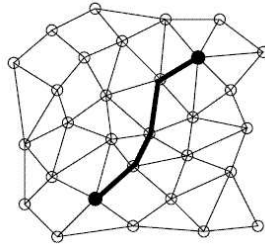


FIG. 3.6 – Approximation de la distance curviligne à l'aide du chemin le plus court par l'intermédiaire des liens entre les centroïdes (ici la distance entre les deux centroïdes noircis)

La figure 3.7 montre un exemple de «nœud de tresse» qui est un objet mono-dimensionnel inclus dans un espace tridimensionnel. Le CDA le déplie dans un espace mono-dimensionnel rapidement et automatiquement.

Isomap. Tenenbaum et al. [Tenenbaum *et al.*, 2000] ont proposé la méthode Isomap qui utilise, d'une façon analogue à CDA, la distance géodésique, à la place de la distance euclidienne, ainsi le voisinage de chaque point est calculé en utilisant, soit la notion de k plus proches voisins, soit le rayon fixé. Ensuite, après avoir établi le graphe de liens entre tous les points voisins, la distance géodésique entre deux points est approximée par le plus court chemin de la somme des longueurs des arcs liant les deux points. Le plus court chemin est calculé par l'algorithme de Dijkstra [Dijkstra, 1959].

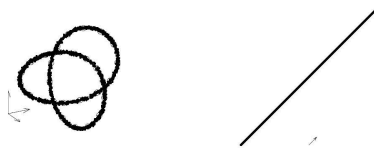


FIG. 3.7 – CDA : Projection non-linéaire d'un «nœud de tresse» (de dimension 3 à 1)

La figure 3.8 montre un exemple de la projection d'un «rouleau suisse» sur une carte par Isomap le CDA détecte le caractère uniforme de cette distribution en la dépliant, tandis que Isomap produit des trous étranges comme une tranche de fromage suisse! D'après Lee et al [Lee *et al.*, 2002], ce problème pourrait être résolu en ajoutant une étape de quantification vectorielle, comme dans de la CDA.

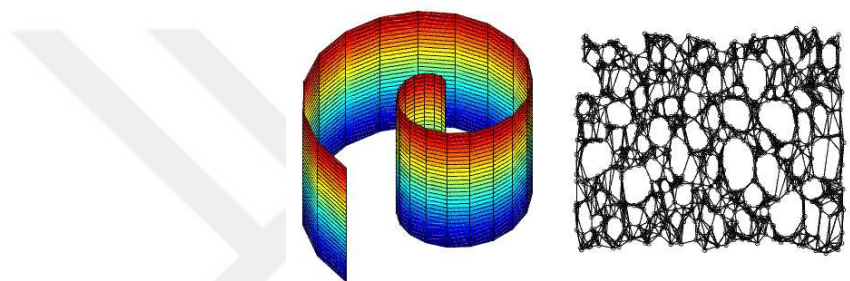


FIG. 3.8 – Isomap : Exemple du «rouleau suisse» (à droite) et de la projection de 20000 échantillons tirés du rouleau par Isomap.

3.4.2 Exemples applicatifs de visualisation dans le domaine documentaire

Nous illustrons dans cette section deux modèles neuronaux cartographiques non linéaires basés sur le modèle SOM (cf. section 2.3.2) ainsi qu'un modèle basé sur le graphe. Ces modèles sont particulièrement appliqués aux données documentaires. Les deux modèles neuronaux proposent une phase importante de l'assistance à l'analyse des données fournies par une carte de documents. Il s'agit de définir des noms de classes qui représentent au mieux le contenu de ces dernières lorsque la carte est présentée à l'utilisateur. En tenant compte du fait qu'il n'y a pas de solution absolue pour résoudre ce problème dit de «labélisation de classe (étiquetage de classes)», bien connu par les spécialistes de la classification.

Modèle WEBSOM

WEBSOM représente une application directe du modèle SOM aux données documentaires⁴ [Honkela *et al.*, 1996b; Honkela *et al.*, 1996a; Lagus *et al.*, 1996; Kaski *et al.*, 1998; Kohonen *et al.*, 1996]. Il est utilisé pour organiser de grandes collections de documents sur les cartes visuelles où les documents proches (respectivement similaires) apparaissant sur des zones géographiques proches (respectivement similaires). WEBSOM a été utilisé aussi bien pour la recherche d'informations que pour la fouille de données textuelles [Lagus *et al.*, 1999].

⁴WEBSOM utilise des collections des articles obtenus à partir de Usenet discussing groups

Codage des documents. WEBSOM travaille directement sur la représentation en plein texte des documents. Pour des raisons de facilité de calcul, la méthode considère cependant les documents comme des collections de mots, négligeant de fait l'ordre des mots [Lagus *et al.*, 1999]. La représentation vectorielle des documents ainsi obtenue est ensuite pondérée, en utilisant la méthode de l'entropie (cf. équation 1.4), puis réduite, en utilisant la méthode de projection aléatoire (cf. section 3.3.2), comme le montre la figure 3.9.

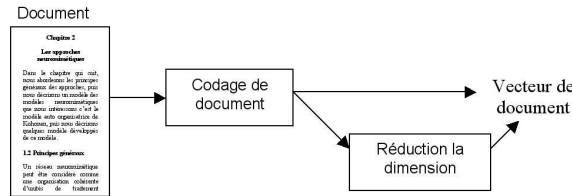


FIG. 3.9 – La base de construction du système WEBSOM. Le document est encodé en un vecteur numérique. La dimension de représentation peut être réduite pour des raisons de calcul.

Étiquetage et parcours de la carte de documents. L'algorithme SOM, utilisant en entrée les vecteurs fournis par la phase de codage précédemment décrite, est ensuite appliqué pour fournir en sortie la carte thématique bidimensionnelle de la collection de documents étudiée.

1. **L'étiquetage** : Les caractérisations peuvent par exemple être employées pour étiqueter des régions d'un image de carte de document. La méthode proposée par Lagus et al. [Lagus and Kaski, 1999; Lagus *et al.*, 2004] produit de telles caractérisations. Une classe de documents, les mots caractéristiques qui la décrivent peuvent être obtenus en utilisant la mesure de qualité suivante :

$$G(w, j) = F_j(w) \frac{F_j(w)}{\sum_i F_i(w)} \quad (3.8)$$

où $F_j(w)$ est le nombre d'occurrences du mot w dans le cluster j . Cette mesure compare le nombre relatif d'occurrences aux autres clusters.

Les étiquettes ont été donc automatiquement choisies et placées sur la carte de sorte qu'elles :

- (a) apparaissent souvent dans les documents de la zone qu'elles décrivent,
- (b) apparaissent relativement plus souvent dans ladite zone que dans des régions éloignées sur la carte.

Les étiquettes peuvent guider l'utilisateur vers des centres d'intérêt précis sur la carte tout en lui donnant une idée générale du contenu de la collection.

2. **Navigation basée sur la carte de documents** : L'algorithme SOM, utilisant en entrée les vecteurs fournis par la phase de codage précédemment décrite, est ensuite appliqué pour fournir en sortie la carte thématique bidimensionnelle de la collection de documents étudiée. L'espace des documents est organisé selon trois niveaux hiérarchiques distincts : la carte, les nœuds, et les différents documents. N'importe quel sous-domaine thématique de la carte peut ainsi être choisi et feuilleté d'un simple clic de souris.

Une information colorimétrique supplémentaire est donnée sur la carte, de manière à orienter la navigation de l'utilisateur en fonction des propriétés locales de voisinage sur celle-ci. Si les documents voisins sont proches, des nuances légères sont utilisées entre eux, sinon, une nuance foncée indique qu'il y a beaucoup de différence entre des documents voisins (cf. figure 3.10).

L'interface de la carte a été équipée d'une zone de saisie dans laquelle l'utilisateur peut taper des requêtes. Une requête utilisateur est prétraitée de manière à construire un vecteur descriptif de même forme que celui employé pour les documents [Honkela, 1997]. Le vecteur résultant est alors comparé aux vecteurs modèles de toutes les unités (classes) de la carte, puis les classes les plus pertinentes sont identifiées par des cercles sur l'affichage de la carte.

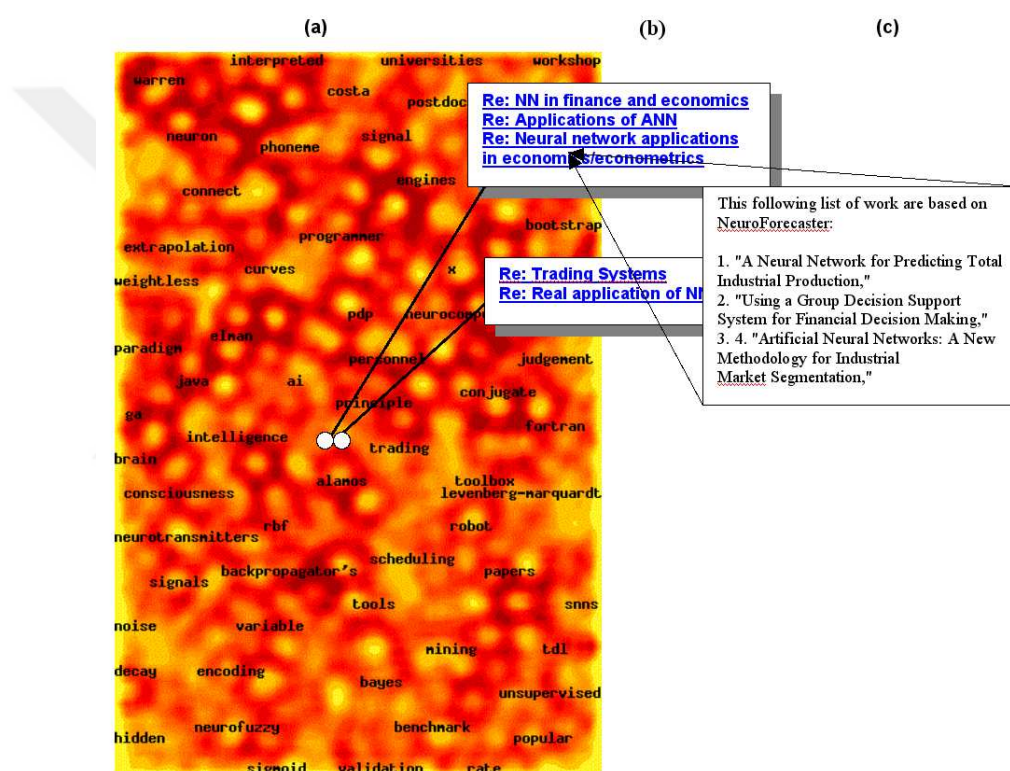


FIG. 3.10 – Exemple d'accès à un document à partir d'une carte WEBSOM. L'accès à un document fait intervenir 3 niveaux : (a) La carte WEBSOM de la collection. (b) Le contenu d'un nœud de la carte. (c) Le document proprement dit (ici un article de newsgroup).

Application. WEBSOM est applicable à n'importe quel type de collection de documents textuels (les documents sont alors organisés clairement selon leur contenu). Cette méthode est particulièrement adaptée aux tâches d'exploration de collections destinées aux utilisateurs ne connaissant que faiblement les domaines concernés par celles-ci. Une interface de navigation a été spécifiquement développée pour explorer les cartes à partir du World Wide Web [Kaski *et al.*, 1998].

Avantages et inconvénients de la méthode WEBSOM La méthode WEBSOM permet d'organiser automatiquement une collection de documents. Elle permet également de parcourir et d'explorer facilement cette collection en fonction de ses différentes orientations thématiques [Kaski *et al.*, 1998].

En contrepartie, WEBSOM ne permet de traiter qu'une seule classification à la fois. Cette stratégie globale génère un bruit de classification important qui nuit à l'interprétation du contenu de la collection (cf. chapitre 1). De plus, WEBSOM ne permet de traiter que des données textuelles pures, et non pas les données multimédias (textes, images, etc.) présentes dans de nombreuses collections.

Modèle MultiSOM

Dans le modèle MultiSOM (cf. section 2.3.7), l'exploitation coordonnée du principe des cartes topographiques est de celui du mécanisme de communication inter-cartes permet à l'analyste d'opérer à des analyses interactives et visuelles, à des niveaux de granularité variés, des résultats de classification. Ainsi, la présentation des résultats s'opère selon plusieurs principes :

Codage des documents. MultiSOM travaille, tout comme WEBSOM, directement sur la représentation vectorielle des documents. Mais, de manière différente de WEBSOM, cette représentation est pondérée en utilisant la méthode de $TF * IDF$ (cf. équation 1.2).

Construction des cartes élémentaires (étiquetage et découpage). La phase de construction des cartes élémentaires utilise l'algorithme SOM. Chaque carte ainsi obtenue est organisée de manière à être rendue lisible à l'utilisateur, en lui fournissant notamment une vue générale des résultats de la classification. Cette tâche, correspond dans la pratique à un étiquetage des nœuds de la carte, dont les profils contiennent la description des classes obtenues, ainsi qu'à un découpage de la carte en zones thématiques cohérentes, ou aires logiques.

1. **Étiquetage d'une carte** : Deux stratégies peuvent être indifféremment utilisées lors de la consultation des cartes :

- Les stratégies dirigées par les profils des classes consistent à attribuer à chaque classe un nom qui représente la composition des labels des composantes dominantes de son vecteur de profil. Ces stratégies sont particulièrement adaptées pour mettre en évidence les thématiques majeures isolées par le processus de classification.
- Les stratégies dirigées par les profils des données consistent à attribuer à chaque classe un nom qui représente la combinaison des labels des composantes dominantes, qui du vecteur de profil du membre le plus caractéristique de la classe, qui d'un vecteur de profil calculé représentant le profil moyen de l'ensemble des membres de la classe.

Ces stratégies sont utiles pour fournir des informations complémentaires pour l'interprétation des thématiques. En effet, les profils de classes représente une information indirecte, issue du processus de classification proprement dit, alors que les profils des documents représentent une information directe issue du contenu des documents.

2. **Découpage d'une carte en aires logiques** : cette phase consiste à regrouper figurativement les classes de la carte en groupe cohérents ou «aires logiques fermées». Chaque aire, qui peut être considérée comme une macro-thématique, fournira alors une information fiable sur l'importance relative des différentes thématiques de base (cf. figure 3.11). L'algorithme de calcul des aires est basé sur les propriétés topographiques des vecteurs de

profils associées aux nœuds d'une carte SOM. Cet algorithme représente une généralisation de celui proposé par Lin [Lin *et al.*, 1991]. Il est décrit en détail dans [Lamirel, 1995]. Les thématiques ou les aires se trouvant proches sur la carte, représentent des notions proches et/ou associées.

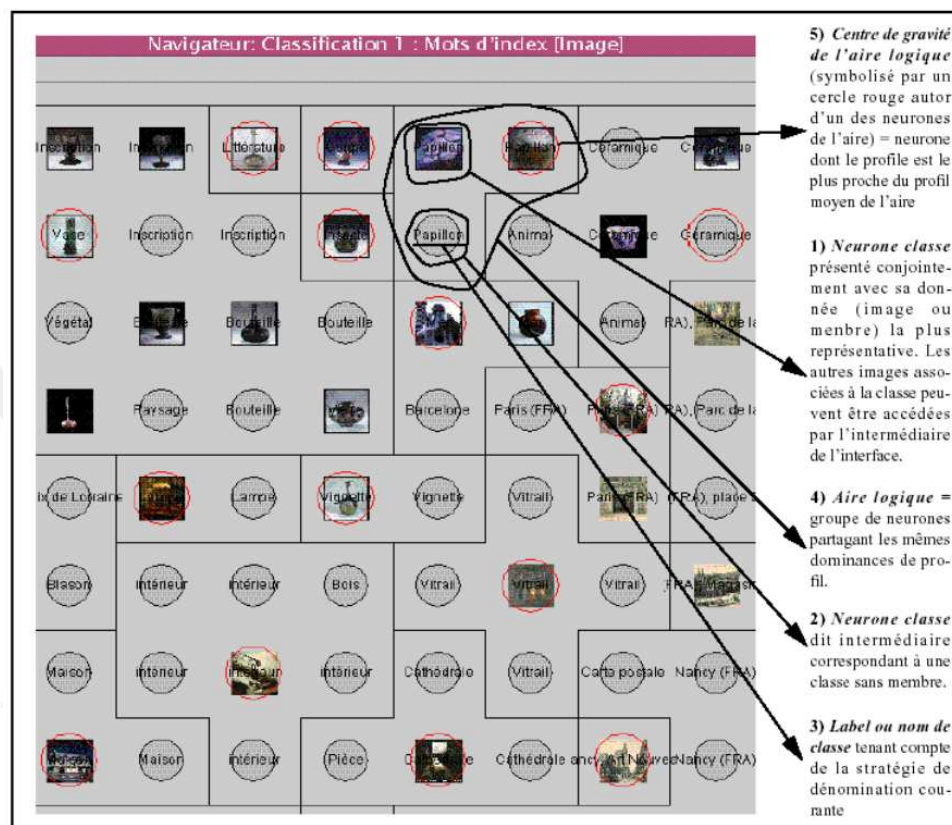


FIG. 3.11 – Vue partielle d'une carte de 12×12 neurones (classes). La carte est organisée comme une grille carrée $2D$ de neurones. Les profils de classe sont ensuite générés par l'intermédiaire d'un apprentissage compétitif non supervisé. Le point de vue choisi ici est celui des **Termes d'index**. Les noms des classes illustrent les thématiques qui ont été isolées par le processus d'apprentissage. Après l'apprentissage, les neurones globalement rattachés aux mêmes thématiques ont été regroupés en aires logiques en utilisant les propriétés topographiques de la carte.

Navigation visuelle par la communication inter-cartes. la démarche de ce mécanisme est illustrée à la section 2.3.7. En comparaison avec les méthodes de projection cartographiques statistiques précédemment décrites, la puissance du modèle MultiSOM est clairement représentée par la souplesse du mécanisme de cartographie multicartes combiné au processus de communication inter-cartes. Ces mécanismes peuvent être exploités de manière interactive et visuelle par un analyste pour opérer des analyses ciblées.

La figure 3.12 illustre une déduction thématique s'opérant entre quatre points de vue issus du domaine des huiles moteur (cf. annex A).

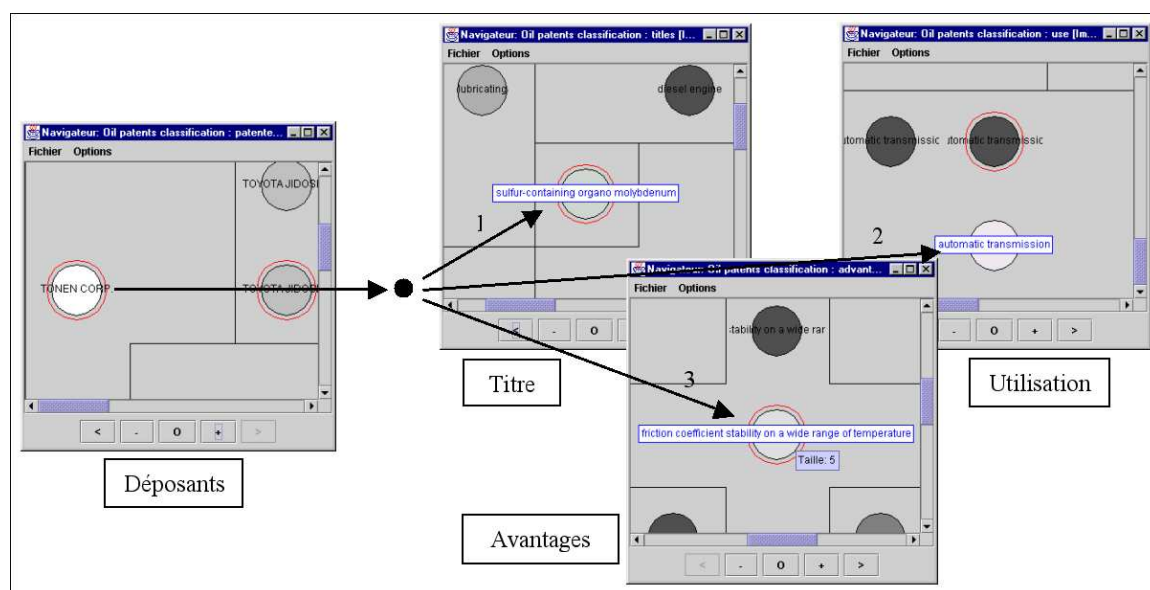


FIG. 3.12 – Exemple d'exploitation interactive du mécanisme de communication inter-carte pour la veille technologique. La décision de l'analyste d'activer l'aire correspondant au déposant TONEN CORP. sur la carte des **Déposants** et de propager cette activité vers les cartes thématiques associées aux points de vue **Utilisations**, **Avantages** et **Titres** correspond à une requête transthématique virtuelle dont la formulation explicite pourrait être la suivante : «Je voudrais connaître les domaines de compétence spécifiques du déposant TONEN CORP. (concernant l'utilisation des huiles, leur composition et les avantages attendus), si ceux-ci existent». L'application MULTISOM lui permet de trouver de manière interactive que le déposant TONEN CORP. est un spécialiste de la lubrification des transmissions automatiques [flèche n°2 sur la carte] et que ce déposant a adopté pour ce type de lubrification des composés à base d'organomolybdène [flèche n°1] dont les avantages principaux sont de fournir des huiles dont le coefficient de friction est stable sur une large gamme de températures [flèche n°3]. Dans ce cas de figure, une propagation inverse depuis les thématiques cibles aurait également dû être utilisée pour vérifier que ces thématiques sont parties exclusives des domaines de compétence du déposant (ce qui est le cas). Plus la couleur des nœuds des cartes se rapproche du blanc, plus forte est leur activité résultante.

Application. De nombreuses applications du modèle MultiSOM sont référencées, aussi bien pour les bases textuelles [Lamirel *et al.*, 2000] que pour les bases multimédia [Lamirel, 2001; Lamirel *et al.*, 2000]. Le modèle a également été choisi comme modèle de référence dans le projet européen EISCTES (cf. annex A), dont le but était de fournir une cartographie du web (cf. chapitres 6).

Visualisation par l'analyse de graphe

Nous présentons dans cette section la méthode BibTechMon (**B**ibliometric **T**echnology **M**onitoring) qui représente une méthode bibliométrique graphique [Kopcsa and Schiebel, 1998]. Cette méthode repose sur le calcul des co-occurrences des termes apparaissant dans les documents (Kostoff, 1993). Les termes représentent donc des données d'entrée. La relation entre

deux termes est considérée comme forte s'ils apparaissent souvent ensembles (i.e. co-occurent souvent) dans les documents. Les valeurs de co-occurrence de termes sont normalisées, par des indices tels que «Jaccard», «Cosinus», etc., pour travailler sur des valeurs relatives dans l'intervalle $[0, 1]$. Le réseau de relation est ensuite déterminé à partir des relations individuelles entre les termes, puis transcrit sous forme graphique en utilisant un «modèle de Spring» (cf. figure 3.13). Le «modèle de Spring» est un modèle d'équilibre de forces qui représente une alternative au Multidimensional Scaling (MDS) [Morrison and Chalmers, 2003]. Selon cette représentation, chacun de n termes de l'espace de description sera associé à un point fixe (ou nœud) $d_j \in \mathbb{R}^2$, $j = 1, \dots, n$. La taille de chaque nœud représentera le nombre d'occurrences de ce terme. Les positions des nœuds sur la carte seront calculées en basant sur leur relation avec tous les autres nœuds. Les nœuds les plus centraux de la représentation seront donc ceux dont les termes associés partagent les co-occurrences normalisées les plus fortes.

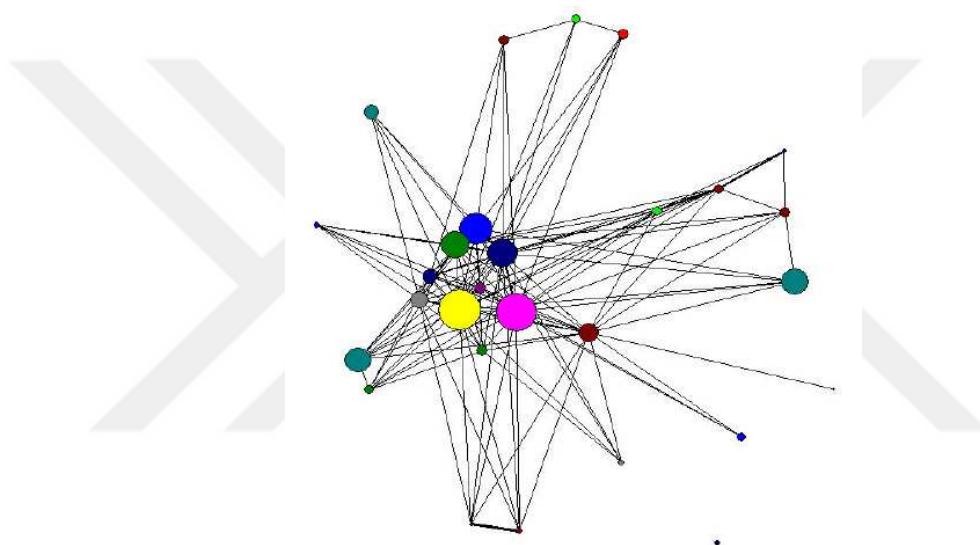


FIG. 3.13 – BibTechMon : réseau de mots basé sur les relations entre eux. Ce réseau contient 28 nœuds et 131 connexions

L'intérêt principal de cette approche est qu'elle fournit à l'analyste une vue globale de l'interaction entre les données, mais au prix d'une structure très complexe et surchargée. De fait, un filtrage sur les liens privilégiant seulement les liens les plus forts, ou d'autres critères spécifiques, est souvent nécessaire pour pouvoir effectuer une quelconque analyse des résultats. Ce type de méthodologie est cependant susceptible de faire perdre à l'analyste des informations utiles pour l'analyse.

3.5 Visualisation hyperbolique

La visualisation hyperbolique est connue pour ses capacités à traiter le problème de la surcharge cognitive produite par les approches basées sur les graphes ou sur la projection cartographique précédemment décrites. En effet, cette technique permet de visualiser des relations complexes entre les données par les mécanismes de focus et de contexte [Bergé and Bouthier, 2003].

3.5.1 L'espace hyperbolique

L'espace hyperbolique H^2 représente un élément important de la géométrie non-euclidienne. C'est un candidat idéal pour gérer de grandes structures hiérarchiques grâce à ses propriétés géométriques spécifiques [Lamping and Rao, 1994]. En effet, cet espace peut être projeté sur un plan bidimensionnel de l'espace euclidien sous la forme d'un disque unité également appelé disque de Poincaré [Henle, 2001; Ratcliffe, 1994]. Le disque Poincaré a un certain nombre de propriétés commodes pour la visualisation : premièrement, il préserve la forme originale de distribution de H^2 , deuxièmement, le modèle permet de matérialiser la translation de H^2 original d'une manière très élégante. Ainsi, la fovéa peut être déplacée vers l'ensemble des parties du disque hyperbolique infini. Ceci permet à l'utilisateur de se concentrer sur les parties intéressantes d'une distribution originale dans H^2 tout en gardant toujours une vue générale de son contexte.

3.5.2 Arbre hyperbolique (Hypertree)

La technique de visualisation en arbre hyperbolique, aussi appelée visualisation hypertree, a été inventée par Lamping et Rao [Lamping J. and P., 1995; Lamping and Rao, 1996] dans le but de représenter et de manipuler de larges structures hiérarchiques dans l'espace hyperbolique.

La visualisation hypertree permet de représenter des données structurées sous forme d'arbre. L'arbre est dessiné de manière radiale - la racine au centre, les nœuds fils se trouvant sur un cercle ayant pour centre leur père dans le plan hyperbolique. Puis, une fois chacun des nœuds de l'arbre placé dans l'espace hyperbolique, l'arbre tout entier est projeté sur le disque de Poincaré. La racine est projetée au centre du disque unité. Les branches de l'arbre deviennent des géodésiques du modèle de Poincaré, c'est-à-dire soit des diamètres, soit des arcs de cercle (cf. figure 3.14) [Bergé and Bouthier, 2003; Bouthier, 2004].

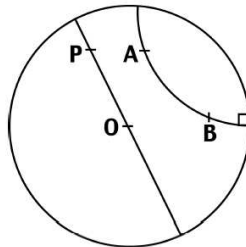
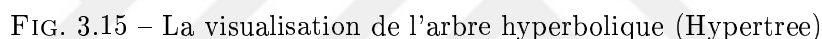


FIG. 3.14 – Deux types de géodésique : un diamètre passant par **O** et **P** et un arc de cercle **AB** orthogonal au cercle unité

La visualisation hypertree est une technique de visualisation « focus+ contexte » : elle donne une vue détaillée sur une petite surface (focus), tout en gardant une vue globale de la structure (contexte). Puisque l'ensemble du plan hyperbolique est projeté dans le disque unité, une distorsion de l'espace apparaît : plus on s'approche du bord du disque, plus les distances sont compressées. Ainsi, le centre du disque fournit une vue détaillée sur une partie des nœuds de l'arbre (focus), tandis que le bord du disque fournit une vue globale du reste de la structure de l'arbre (contexte).

La visualisation hypertree est une technique interactive. L'arbre peut être déplacé de façon à avoir une vue détaillée au centre du disque de n'importe laquelle de ses parties. Un nœud



- l'arbre est dessiné dans le disque unité, comme si l'arbre a été tiré dans l'espace hyperbolique, puis projeté sur le disque de Poincaré.
- l'arbre entier peut être dragué par la souris pour placer une partie de cet arbre au milieu du disque.
- un clic sur un nœud le déplace automatiquement au centre du disque.

3.5.3 Modèles neuronaux hyperboliques de la carte de Kohonen

76

polygones réguliers de telle manière que chaque point de la grille soit entouré par un nombre n de voisins fixé⁵ [Magnus, 1974]. La figure 3.16 montre deux exemples de tessellations (pour les valeurs de $n = 7$, et de $n = 10$, respectivement), visualisées sur le disque unité de Poincaré. Il est cependant à remarquer que, dans l'espace H^2 hyperbolique original, chaque triangle de tessellation aurait la même taille.

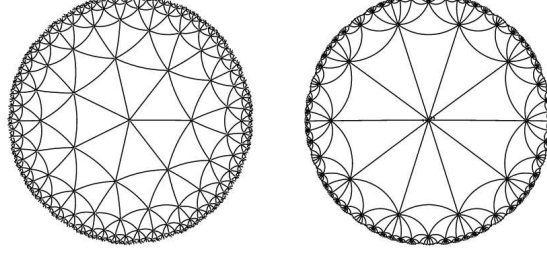


FIG. 3.16 – Tessellations régulières à base de triangles d'une carte hyperbolique, projetées sur le disque unité de Poincaré. La figure de gauche illustre le cas où le nombre n de triangles équilatéraux pour chaque sommet est de 7 (en dehors du bord) ; la figure de droite est construite avec $n = 10$.

Carte auto-organisatrice hyperbolique (*HSOM*)

Le modèle *HSOM* [Ritter, 1999] utilise une grille hyperbolique finie. Les nœuds (neurones) de la grille sont donc organisés en anneaux autour d'un nœud original (ce qui représente la manière la plus simple de construire une grille approximativement circulaire). Cependant, le nombre de nœuds d'une telle grille se développe très rapidement (asymptotiquement exponentiellement) avec le rayon choisi R . Le tableau 3.1 illustre l'évolution de nombre total N_R de nœuds de la grille hyperbolique en fonction de rayons R différents, variant de 1 jusqu'à 10, pour une valeur de référence de $n = 7$. Chaque nœud r de la grille porte un vecteur de référence $\vec{W}_r \in \mathbb{R}^D$, issu de l'espace D -dimensionnel. L'apprentissage se déroule de manière analogue à celui de *SOM*. Après la détermination du nœud gagnant s au temps t , l'ajustement de son voisinage $N(s, t)$ s'opère de manière radiale dans la grille selon la règle suivante :

$$\Delta \vec{W}_r = \eta h_{rs}(\vec{x} - \vec{W}_r) \quad (3.9)$$

où $r \in N(s, t)$, $h_{rs} = \exp(-d^2(r, s)/2\sigma^2)$ et σ représente le rayon de voisinage. Cependant, le voisinage $N(s, t)$ et la distance carrée de nœud $d^2(r, s)$ doivent être calculés selon la métrique naturelle héritée par la grille hyperbolique.

R	1	2	3	4	5	6	7	8	9	10
N_R	8	29	85	232	617	1625	4264	11173	29261	76616

TAB. 3.1 – Le nombre de nœuds N_R de grille hyperbolique pour des valeurs différentes du rayon R , et pour $n = 7$

⁵En dehors de neurones du bord de la grille, pour lesquels le nombre de voisins est inférieur, comme dans le cas des cartes *SOM*.

Un nombre complexe z_r est associé à chaque nœud r pour identifier sa position dans l'espace hyperbolique. La distance entre les nœuds sur le disque Poincaré est alors obtenue par :

$$d = 2 \operatorname{arctanh}\left(\left|\frac{z_r - z_s}{1 - \bar{z}_s \cdot z_r}\right|\right) \quad (3.10)$$

Il est montré dans [Ontrup and Ritter, 2001] que la taille exponentiellement croissante d'un voisinage autour d'un point dans l'espace hyperbolique fournit plus de liberté pour projeter une distribution de données complexe, et permet ainsi de mieux s'affranchir de la gestion des relations spatiales entre les données dans l'espace projeté. Le modèle *HSOM* a été appliqué avec succès, par ces mêmes auteurs, aux tâches de catégorisation de texte, et comparé favorablement aux autres méthodes de classification usuellement appliquées dans ce domaine.

Le principal problème du modèle *HSOM* est que l'effet de bord, produit sur le centre de la grille par les données marginales, y est beaucoup plus fort que dans le modèle SOM classique. En effet, le chemin vers le bord de la grille, en partant de n'importe quel point, est plus court que dans le cas de SOM.

SOM hyperbolique hiérarchique incrémental (H^2SOM)

Le modèle H^2SOM a été proposé par Ontrup et Ritter [Ontrup and Ritter, 2005]. Cependant, cette méthode a été proposée pour s'affranchir des limitations de visualisation de de SOM hiérarchique (cf. section 2.3.6). Il utilise le même type de grille que le modèle *HSOM*. Le réseau est construit en deux étapes successives :

1. Initialisation : un anneau de n_b triangles hyperboliques équilatéraux, centrés à l'origine de H^2 , (cf. figure 3.17), est construit avec $n_b + 1$ nœuds qui forment les triangles de premier niveau hiérarchique.
2. Croissance : chaque nœud dans la périphérie de ce réseau est augmenté en l'entourant par des nœuds issus de $n_b - 2$ triangles équilatéraux, ce qui revient finalement à former n_b -gones autour du nœud considéré (cf. figure 3.17 b), ce qui permet finalement de générer un nouveau niveau hiérarchique.

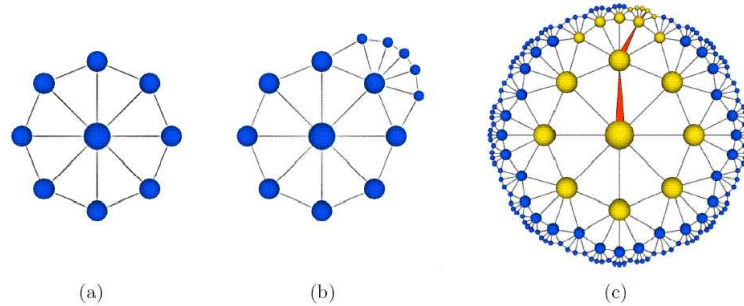


FIG. 3.17 – (a) Les nœuds aux sommets - dans ce cas $n_b = 8$ - des triangles équilatéraux forment le premier niveau dans la hiérarchie du H^2SOM . (b) Quand un nœud répond à un critère de croissance, un ensemble de $n_b - 3$ nœuds fils va être généré. Dans (c) la hiérarchie présentée est une hiérarchie complète à 3 niveaux, et un chemin virtuel correspondant à l'ensemble de nœuds visités par un utilisateur est accentué.

La phase d'apprentissage de ce modèle fonctionne d'une façon similaire à celle du modèle *HSOM* (cf. section précédente), en dehors le fait que le centre s et les n_b premiers nœuds sont initialisés à partir de profil moyen de données, et de petites variations de celui-ci.

Les modèles *HSOM* et *H²SOM* représentent des méthodes de visualisation efficaces qui combinent les vertus du modèle *SOM* avec un apprentissage extrêmement rapide et des erreurs de quantification et de classification moindres. Cependant, leur capacité d'adaptation aux distributions complexes restent moindre que celle de modèles tels que NG.

3.6 Conclusion

Nous avons présenté dans ce chapitre plusieurs types de techniques de visualisation de données. La visualisation par projection cartographique linéaire montre rapidement ses limites dans le cas des distributions complexes. Les méthodes de visualisation par projection non linéaires permettent de mieux traiter la visualisation de tels types de distribution, sans pour autant s'affranchir entièrement des problèmes liés au pliage et à la compression de l'information dans le cas de distributions de données fortement multidimensionnelles. De plus, ces méthodes doivent compenser l'effet de complexification supplémentaire dû au phénomène de l'espace vide (cf. section 3.2) quand le nombre de dimensions de l'espace de description devient très grand. Ils se ramènent également aux problèmes de surcharge cognitive propres à la représentation en graphes. Ces problèmes se compliquent encore si une représentation des liens entre les données est prise en compte.

La visualisation hyperbolique représente une alternative très intéressante car elle permet de s'affranchir de ces problèmes, au prix d'une faible perte d'information, ceci en utilisant un mécanisme spécifique de présentation par dépliage dynamique (focus + contexte). Nous avons présenté plusieurs types d'application de cette méthode : (1) l'arbre hyperbolique (hypertree), utilisé pour représenter des structures hiérarchiques de taille importante, (2) deux modèles neuronaux hyperboliques basés sur SOM, à savoir *HSOM* et *H²SOM*, proposés pour surmonter les limitations en terme de visualisation des modèles SOM, et SOM hiérarchique, respectivement.

L'exemple de la méthode MultiSOM suggère également l'intérêt de l'interaction visuelle dans le processus d'interprétation des résultats, par l'intermédiaire de la notion de plan cartographique multiples et communicants.

Nous reviendrons régulièrement dans la suite de notre thèse sur un certain nombre de ces points-clés, ainsi que sur le potentiel des méthodes que nous avons décrites dans ce chapitre dans le cadre de l'analyse et de la fouille des données documentaires. Nous nous intéresserons en particulier aux mécanismes d'organisation hyperboliques des résultats de classification et aux mécanismes de validation d'analyses et de fouille basés sur l'interaction entre des représentations visuelles différentes.



Deuxième partie

Nouveau paradigme pour l'analyse, la
fouille et la visualisation de données

Chapitre 4

Nouvelles mesures de qualité du clustering

Sommaire

4.1	Introduction	83
4.2	Qualité et analyse de classification non supervisée pour les données basées sur leur espace de description	84
4.2.1	Nouvelles mesures de la qualité basées sur la distribution des descripteurs de données	85
4.2.2	Équivalence au treillis de Galois	86
4.2.3	Sélection de modèle optimal	88
4.2.4	Expérimentations	88
4.3	Qualité et analyse de classification non supervisée pour les données avec des étiquettes multiples	97
4.3.1	Nouvelles mesures de qualité basées sur la distribution d'étiquettes	99
4.3.2	Sélection de modèle optimal	102
4.3.3	Détermination du critère d'arrêt d'apprentissage	104
4.3.4	Définitions des types d'étiquettes	105
4.3.5	Définitions de pertinence d'étiquette et de classe	105
4.3.6	Analyse de classification	106
4.3.7	Expérimentations	107
4.4	Conclusion	116

4.1 Introduction

La classification non supervisée est beaucoup plus difficile à mettre en œuvre que la classification supervisée. Ce problème vient principalement du fait que le nombre de classes nécessaires pour décrire les données peut être considéré comme le paramètre central inhérent à cette première approche. Plusieurs techniques ont été proposées pour déterminer la valeur optimale de ce paramètre. Nous résumons ci-après les principes généraux de ces techniques, qui se font toutes elles-mêmes appel aux mesures de qualité que nous avons présentées au chapitre précédent (cf. section 2.4) :

1. La première technique consiste à utiliser une fonction qui atteint son maximum pour le nombre optimal de classes, en cherchant typiquement à maximiser les distances inter-classes tout en minimisant les distances intra-classe. Il suffit donc de calculer cette fonction objectif pour chaque n classes tel que $n \in [1, N]$ et de choisir la valeur de n qui donne le résultat le plus élevé. Les fonctions dépendent parfois uniquement des distances intra-classe [Krzanowski and Lai, 1985; Hartigan, 1975], mais également, dans certains cas, des distances inter-classes [Calinski and Harabasz, 1974; Kaufman and Rousseeuw, 1990b].
2. La deuxième technique consiste à utiliser la méthode de rééchantillonnage [Dudoit and Fridlyand, 2002] par comparer les classifications obtenues sur les données réelles à des classifications «de référence», correspondant à l'absence de classes. Pour cela, des données sont simulées à partir de distributions uniformes, et des classifications de référence sont ainsi réalisées sur les échantillons obtenus. Le nombre de classes n pour lequel l'écart entre la classification réelle et la classification de référence est le plus grand est le nombre retenu de classes.

Les mesures d'inertie, tout comme les autres les mesures de similarité existantes, (cf. section 2.4) posent cependant de nombreux problèmes d'interprétation (cf. section 2.4.1), ce qui amène à également à réserver notre jugement quant au résultat obtenus par l'intermédiaire des techniques que nous venons de décrire. En dehors de leur instabilité, un des problèmes majeurs de ces mesures de qualité usuelles est qu'elles ne sont pas orientées-utilisateur, du fait qu'elles ne se focalisent pas sur la qualité de l'interprétation du contenu de classe par ces derniers. La qualité de l'interprétation du contenu des classes est principalement liée à l'homogénéité des descriptions des membres de chaque classe. Une mesure adaptée à l'utilisateur devra donc plus spécifiquement évaluer à quel point les membres d'une classe partagent un ensemble de propriétés communes, au lieu de se concentrer sur une distance globale entre les profils de membres.

L'ensemble des mesures que nous proposons dans ce chapitre est basé sur ce principe, qui, comme nous le montrerons par les nombreuses expérimentation que nous proposons, dans ce chapitre et dans les suivants, ouvre de nouvelles perspectives dans l'évaluation des classifications numériques. Ce principe permet en effet de traiter de manière stable et homogène les opérations d'évaluation de la qualité d'une classification, de recherche de modèle optimal, de validation du modèle obtenu, de gestion de l'apprentissage et d'extraction de connaissances. Il est naturellement adapté à la comparaison des méthodes de classification.

Nous utiliserons trois corpus de référence complémentaires dans nos expérimentations, à savoir, un corpus de notices de brevets et deux corpus de données web différents décrits en annexe A.

4.2 Qualité et analyse de classification non supervisée pour les données basées sur leur espace de description

Nous aborderons dans cette section le problème lié à l'évaluation objective de la qualité de classification numérique non supervisée en proposant une solution par mettre en place des nouvelles mesures de qualité.

4.2.1 Nouvelles mesures de la qualité basées sur la distribution des descripteurs de données

Dans la recherche documentaire (cf. section 1.3), le rappel (R) représente le pourcentage de documents pertinents qui ont été retrouvés pour une requête donnée, et la précision (P) évalue le pourcentage des documents restitués qui sont effectivement pertinents pour cette même requête (cf. section 1.3.5). Ces estimateurs peuvent être adaptés à notre contexte, pour mesurer la capacité d'un système de classification numérique à se comporter comme un système de recherche documentaire optimal. Nous introduisons donc ci-après de nouvelles mesures de **Rappel** et de **Précision**. Ces mesures évaluent la qualité d'une méthode de classification en estimant la pertinence des contenus de ses classes résultantes (le contenu d'une classe est représenté par un sous-ensemble de données originales qui lui ont été associées par le processus de classification) en termes de propriétés (termes d'index) partagées. Dans nos descriptions, le contenu d'une classe est censé être représenté par les documents et les termes d'index et les documents sont censés être pondérés par des valeurs dans l'intervalle $[0, 1]$ [Lamirel *et al.*, 2004a; Lamirel *et al.*, 2003b; Lamirel *et al.*, 2003a].

Considérons C un ensemble de classes issues de l'application d'une méthode de classification appliquée sur un ensemble D de documents, la mesure de **Précision** (P) et la mesure de **Rappel** (R) peuvent être définies respectivement comme :

$$P = \frac{1}{|\overline{C}|} \sum_{c \in \overline{C}} \frac{1}{|S_c|} \sum_{p \in S_c} \frac{|c_p^*|}{|c|} \quad (4.1)$$

$$R = \frac{1}{|\overline{C}|} \sum_{c \in \overline{C}} \frac{1}{|S_c|} \sum_{p \in S_c} \frac{|c_p^*|}{|C_p^*|} \quad (4.2)$$

où S_c est un ensemble de propriétés qui sont propres à une classe c . S_c , qui est décrit comme :

$$S_c = \{p \in d, d \in c | \overline{W}_c^p = \text{Max}_{c' \in C} (\overline{W}_{c'}^p)\} \quad (4.3)$$

où $\overline{C}$ représente un ensemble propre de classes extraites à partir des classes de C , qui vérifie :

$$\overline{C} = \{c \in C | S_c \neq \phi\} \quad (4.4)$$

et

$$\overline{W}_c^p = \frac{\sum_{d \in c} W_d^p}{\sum_{c' \in C} \sum_{d \in c'} W_d^p} \quad (4.5)$$

où W_x^p représente le poids de la propriété p pour un élément x , et :

$$c_p^* = \{d \in c | W_d^p \neq 0\} \quad (4.6)$$

De façon similaire à la recherche documentaire (RD), la mesure F, ou F-mesure, (cf. équation 1.21) peut être utilisée pour combiner les résultats des mesures de rappel et de précision.

Précision : Le critère de Précision mesure l'homogénéité de contenu des classes générées par une méthode de classification. La valeur de précision augmentera si les intentions des documents appartenant aux mêmes classes sont similaires l'une à l'autre, ce qui rendra en conséquence l'interprétation du contenu de classe par un utilisateur plus facile.

Rappel : D'une manière complémentaire, le critère de Rappel mesure l'exhaustivité du contenu des classes, en évaluant à quel point les propriétés propres sont associées aux classes singulières. Le critère de rappel devrait être considéré comme une application spécifique du concept statistique de sensibilité, correspondant au taux de vrais positifs, aux propriétés d'une classe [Ahn, 1997].

Nous démontrons, ci-après dans la section 4.2.2, que si les deux valeurs de Rappel et de Précision atteignent la valeur unité pour un ensemble de propriétés propres associées à une classe donnée c , alors cet ensemble de propriétés propres représentera un motif fermé ou concept formel de treillis de Galois. Donc, la combinaison de ces deux mesures permet d'évaluer à quel point un modèle de classification numérique peut être considéré comme un classificateur élémentaire aveugle de type treillis de Galois. Cette propriété représente une autre illustration du fait que nos nouvelles mesures de Rappel et de Précision ont été construites afin d'être indépendantes des méthodes de classification. Ainsi elles sont bien adaptées pour comparer différentes méthodes de classification. En plus cette propriété va jouer un rôle important pour l'extraction de connaissances à partir de modèle numérique (cf. section 5.5) du fait que ces mesures permettent à lier le modèle numérique avec le modèle symbolique de treillis de Galois qui est utilisé pour extraire de connaissance (règles d'association) (cf. section 1.4.4).

D'un autre côté, du fait qu'elles sont liées à la recherche documentaire, ces mesures sont également appropriées à évaluer les possibilités de la classification numérique pour agir en tant qu'un SRD. Il est particulièrement important d'évaluer cette capacité si on considère qu'une classification peut servir d'interface à une base de données documentaire. En effet, une classification peut agir en tant que un filtre pour une estimation a priori de la pertinence des requêtes d'utilisateur relativement au contenu de la base de données [Lamirel *et al.*, 2001].

4.2.2 Équivalence au treillis de Galois

Étant donné un ensemble de documents D et un ensemble de propriétés P associées aux documents D .

Soit $f : P \rightarrow D$ une fonction qui associe à un ensemble de propriétés p , un ensemble de documents qui possèdent au moins toutes les propriétés p .

Soit $f : D \rightarrow P$ une fonction qui associe à un ensemble de documents d , un ensemble de propriétés qui sont au moins communes à tous les documents d .

$\forall c \in C, D_c = \{d_1, d_2, \dots, d_n\}$ est un ensemble de documents associés à la classe c , $\forall d_i \in D_c$, p_i est un ensemble de propriétés propres au document d_i et $P_c = p_1 \cup p_2 \cup \dots \cup p_n$ est un ensemble entier de propriétés propres de la classe c .

La classe $c \in C$ représente un concept formel (c'est-à-dire, c contient un motif fermé) de treillis de Galois s'elle vérifie la condition :

$$\boxed{h(P_c) = P_c}$$

où $h(P_c) = \text{gof}(P_c)$

Démonstration

- Nous démontrons ci-après, que $f(P_c) = D_c$ pour $R = 1$ et $P = 1$:

$$f(P_c) = f(p_1 \cup p_2 \cup \dots p_n) = f(p_1) \cap f(p_2) \cap \dots f(p_n) = \{d_1, d_2, \dots, d_n\} = D$$

Une valeur de Précision égale à 1 implique que chaque propriété propre p_i de la classe c est possédée, au moins, par tous les documents D_c , ainsi :

$$\forall p_i, f(p_i) \supseteq D_c \quad (4.7)$$

Une valeur de Rappel égale à 1 implique que tous les documents possédant une propriété propre p_i de la classe c sont inclus dans cette classe D_c , ainsi :

$$\forall p_i, D_c \supseteq f(p_i) \quad (4.8)$$

De l'équation 4.7 et l'équation 4.8 :

$$f(p_1) = f(p_2) = \dots f(p_n) = D_c$$

En conséquence :

$$D = D_c \quad (4.9)$$

- Nous démontrons ci-après, que $g(D_c) = P_c$ pour $R = 1$ et $P = 1$:

$$g(D_c) = g(d_1 \cup d_2 \cup \dots d_n) = g(d_1) \cap g(d_2) \cap \dots g(d_n) = \{p_1, p_2, \dots, p_n\} = P$$

Une valeur de Précision égale à 1 implique que chaque document d_i de la classe c possède, au moins, toutes les propriétés P_c , ainsi :

$$\forall d_i, g(d_i) \supseteq P_c \quad (4.10)$$

Une valeur de Rappel égale à 1 implique que chaque propriété p_i de la classe c est possédée exclusivement par les documents de cette classe, ainsi :

$$\forall d_i, P_c \supseteq g(d_i) \quad (4.11)$$

De l'équation 4.10 et l'équation 4.11 :

$$g(d_1) = g(d_2) = \dots g(d_n) = P_c$$

En conséquence :

$$P = P_c \quad (4.12)$$

Finalement :

$$\begin{aligned} gof(P_c) &= g(f(P_c)) \\ &= g(D_c) \text{ de l'équation 4.9} \\ &= P_c \text{ de l'équation 4.12} \end{aligned}$$

Conclusion : Des valeurs unités de Rappel et Précision pour un ensemble de classes propres d’une classification numérique impliquent que cet ensemble peut être assimilé à un sous-treillis de Galois.

4.2.3 Sélection de modèle optimal

Le principe de notre algorithme de sélection du modèle de classification optimal, qui est présenté dans l’algorithme 6, est de rechercher un point d’équilibre (c’est-à-dire, point d’intersection) entre le **Rappel** et la **Précision**. Cette technique est également utilisée en recherche documentaire pour évaluer la qualité d’un SRD [Salton, 1971]. Dans notre cas, la classification qui est la plus proche du point d’équilibre représente alors le meilleur compromis entre l’exhaustivité et la spécificité. Notre algorithme inclut également une phase d’amélioration, si un premier point d’équilibre a été trouvé. Le but de cette phase est d’augmenter la qualité globale du résultat (une meilleure valeur F) ou, si cette première opération échouait, de diminuer le nombre de classes pour faciliter l’interprétation globale de la classification. Quand aucun point d’équilibre ne peut être trouvé, notre algorithme recherche alors la classification qui réduit au minimum la différence entre le **Rappel** et la **Précision** [Lamirel *et al.*, 2004a; Lamirel *et al.*, 2003b].

4.2.4 Expérimentations

Comparaison des approches de la classification orientée par les points de vue et de la classification globale

Notre expérimentation [Lamirel *et al.*, 2004a] consiste à comparer l’approche de classification orientée par les points de vue avec l’approche de classification globale, en termes de qualité de classification. Dans ce but, nous utilisons les critères de qualité que nous avons proposés précédemment, soit les mesures de rappel (cf. équation 4.2) et de précision (cf. équation 4.1).

Notre base de test est un corpus de 1000 notices de brevets (section A.2). Pour chaque point de vue, une carte SOM de 10×10 neurones (classes) est générée. Un point de vue global, qui correspond à l’union des index de tous les points de vue spécifiques, est représenté par deux cartes SOM différentes. La première (**GlobMin**) a comme contrainte d’avoir le même nombre de classes que les classifications de points de vue (c’est-à-dire, 10×10 classes). La seconde (**GlobMax**) a comme contrainte d’avoir la somme du nombre de classes de toutes les classifications de points de vue (c’est-à-dire qu’elle représente une classification de 20×20 classes).

De plus, pour chaque point de vue spécifique, ainsi pour le point de vue global, 100 classes sont finalement générées en utilisant deux autres méthodes neuronales, soit NG (cf. section 2.3.3) et GNG (cf. section 2.3.5).

Cas de SOM. Les résultats d’évaluation de l’analyse d’inertie et de l’analyse de qualité sont présentés dans les tableaux 4.1 et 4.2 respectivement.

Critères inertiels : En considérant les mesures d’inertie (cf. tableau 4.1), la classification de **Déposants** est la seule dont l’inertie d’intra-classe est sensiblement inférieure à l’inertie inter-classes. Cela est partiellement lié au fait que la dimension de l’espace descriptif de ce point de vue est considérablement inférieure à celle des autres points de vue. La

4.2. Qualité et analyse de classification non supervisée pour les données basées sur leur espace de description

	Déposants	Titres	Utilisations (SOM)	Utilisations (KMA)	Avantages	GlobMin	GlobMax
I-intra	0.12	0.78	0.49	0.31	0.62	0.81	0.60
I-moyenne	-	-	0.54	0.74	-	-	-
I-inter	0.64	0.14	0.29	1.01	0.23	0.13	0.26

TAB. 4.1 – Résultats de l'évaluation de l'inertie, I-intra : inertie intra-classe, I-moyenne : inertie moyenne, I-inter : inertie inter-classes.

	Déposants	Titres	Utilisations (SOM)	Utilisations (KMA)	Avantages	Taux F	GlobMin	GlobMax
R	0.94	0.89	0.78	0.97	0.77		0.87	0.84
P	0.92	0.40	0.63	0.55	0.60		0.48	0.65
F	0.93	0.55	0.70	0.70	0.69	0.71	0.61	0.68

TAB. 4.2 – Résultats de l'évaluation du **Rappel**, de la **Précision** et de la F-mesure.

classification **GlobMax** semble être légèrement meilleure que la classification **GlobMin**, et la classification **Utilisations** légèrement meilleure que la classification **Avantages**. De plus, le tableau 4.1 montre que les classes générées par plusieurs cartes selon les différents point de vue sont en général plus homogènes vis à vis des individus qui les constituent et également plus hétérogènes entre elles que les classes de **GlobMin**. Les résultats globaux du tableau 4.1 pourraient également amener à conclure au manque de stabilité des critères inertiels si une comparaison fiable de méthode de classification doit être réalisée.

Critères de Rappel et Précision : L'examen de la mesure de **Précision** et de la mesure de **Rappel** (cf. tableau 4.2) montre que les résultats sont plus fiables et plus stables du fait que ces mesures sont indépendantes de la méthode de classification et de la taille de l'espace de description. La supériorité globale de l'approche orienté par les points de vue par rapport à une approche globale avec le même nombre de classe (**GlobMin**) est ainsi mise en évidence. Lorsque le nombre de classes est fortement augmenté dans l'approche globale (**GlobMax**), la **Précision** est alors simultanément augmentée, mais l'avantage de l'approche orientée par les points de vue reste évident du fait que la valeur moyenne sur tous les points de vue de F est plus élevée que la valeur F de **GlobMax**, ceci avec un nombre plus raisonnable de classes par cartes pour chaque point de vue. Le cas spécifique de la classification **Titres** doit être discuté ici. La mauvaise qualité de cette classification est due à la dispersion des termes d'indexation de ce champ⁶, ainsi qu'à un nombre inadéquat de classes, relativement de la taille de l'espace de description associé à ce point de vue. Une stratégie intéressante serait alors de se servir de notre critère de qualité afin de trouver le nombre optimal de classes pour cette classification. Finalement, un déséquilibre entre le **Rappel** et la **Précision** (en faveur du **Rappel**) peut être observé dans le cas des classifications les plus mauvaises (**GlobMin** et **Titres**). Un tel déséquilibre signifie que des documents qui possèdent différents ensembles de propriétés sont groupés dans les mêmes classes, ce qui mène au risque de confusion dans l'interprétation du contenu des classes par l'utilisateur.

⁶Cela peut être confirmé par les résultats de critère inertiels pour ce point de vue.

	Déposants	Titres	Utilisations	Avantages	GlobalMin
I-intra	0	0.50	0.20	0.31	0.52
I-inter	0.61	0.44	0.53	0.57	0.42
R	0.91	0.89	0.82	0.78	0.86
P	1	0.54	0.85	0.72	0.54

TAB. 4.3 – Résultats de l'évaluation de l'Inertie, du **Rappel** et de la **Précision** pour la méthode GNG.

	Déposants	Titres	Utilisations	Avantages	GlobalMin
I-intra	0.001	0.53	0.23	0.29	0.56
I-inter	0.87	0.83	0.76	0.77	0.83
R	0.95	0.88	0.75	0.78	0.86
P	1	0.83	0.80	0.80	0.83

TAB. 4.4 – Résultats de l'évaluation de l'Inertie, du **Rappel** et de la **Précision** pour la méthode de NG.

Cas de GNG. Dans ce cas, les mesures de **Rappel** et de **Précision** (cf. tableau 4.3) montrent la supériorité globale de l'approche orientée par les points de vue par rapport à l'approche globale. Nous pouvons également observer un déséquilibre entre le **Rappel** et la **Précision** dans le cas des classifications **GlobMin** et **Titres** [Al Shehabi and Lamirel, 2004b].

Cas de NG. Cette méthode fournit des résultats de qualité très élevés et très homogènes dans la plupart des cas, comme le montre les valeurs des mesures données par le tableau 4.4. Dans le cas de NG, la supériorité de l'approche orientée par les points de vue par rapport à l'approche globale ne peut pas être montrée directement par l'utilisation des mesures de qualité (**Rappel** et **Précision**). En effet, les mesures de qualité sont également bonnes pour les deux approches. Cependant, les mesures d'inertie fournissent des informations complémentaires intéressantes. Ces mesures montrent que les classes sont plus homogènes (inertie intra-classe inférieure) dans l'approche orientée par les points de vue que dans l'approche globale [Al Shehabi and Lamirel, 2004b].

Comparaison entre la méthode neuronale de SOM et K-means axiale

Pour tester l'efficacité pure de la méthode neuronale SOM, nous avons mené sa comparaison avec une méthode de classification non neuronale, à savoir la méthode KMA (cf. section 2.2.1). Pour cela, nous avons effectué une autre classification des notices de brevets, selon le point de vue de **Utilisations**, en utilisant la méthode KMA. La classification KMA a finalement produit 32 classes différentes (à comparer aux 100 classes produites par la méthode SOM) [Lamirel *et al.*, 2004a].

Critères inertiels : Le tableau 4.1 montre que les classes issues de la classification KMA selon le point de vue **Utilisations** sont plus homogènes vis à vis des données (inertie intra-classe (cf. équation 2.9)) qui la constituent et également plus hétérogènes entre elles (inertie inter-classes (cf. équation 2.10)) que celles de la classification SOM, pour le même point de vue. Par contre, l'inertie intra-classe basée sur le profil des membres (inertie moyenne (cf. équation 2.11)) rapporte une meilleure valeur pour SOM. Ces ré-

sultats pourraient confirmer l'hypothèse que les contraintes produites par le principe de construction de topographie proposée par la méthode de SOM ont une tendance à affecter d'une manière anormale les profils de classes, sans nécessairement avoir d'effet sur la cohérence du contenu des classes⁷. Cette différence peut partiellement s'expliquer par la différence entre les nombres de classes produites par les deux méthodes (32 classes pour KMA contre 100 pour SOM). Elle peut aussi partiellement s'expliquer par la différence des stratégies de projection des documents utilisées par les deux méthodes. En effet, KMA autorise l'affectation d'un document à plusieurs classes (cas de classification recouvrante) alors que SOM ne le propose pas par défaut (classification disjointe). Cependant, les critères inertiels restent toujours trop instables pour la réalisation d'une comparaison objective entre deux méthodes.

Critères de Rappel et Précision : Un déséquilibre entre le **Rappel** et la **Précision** en faveur du **Rappel** peut être observé dans le cas des classifications de KMA (cf. tableau 4.2). Un tel déséquilibre signifie que les documents comprenant des ensembles différents de propriétés sont groupés dans les mêmes classes. Ce déséquilibre pourrait naturellement être expliqué par la stratégie de projection de KMA précédemment mentionnée.

Optimisation des résultats de la classification de webométrie

Cette expérimentation [Lamirel *et al.*, 2004a] représente une expérimentation originale dans le domaine de la webométrie étant donné qu'elle combine pour la première fois la classification de contenu (l'information textuelle) et la classification de liens, pour analyser un grand ensemble non homogène de pages web. Cette expérimentation fait partie du projet EICSTES sur un corpus de pages du site de Cambridge (cf. section A.3.1), dont l'objectif global est de définir des outils et des méthodes pour cartographier l'Web. Elle utilise le modèle MultiSOM qui représente un des modèles de référence du projet. Par ailleurs, dans cette expérimentation, la classification de contenu et celle des liens, peuvent s'échanger de l'information l'une à l'autre en utilisant les principes de communication dynamique du modèle MultiSOM (cf. section 2.3.7).

L'optimisation du nombre de classes de différentes classifications est un problème particulièrement crucial dans le cas de l'analyse de pages web, principalement en raison du caractère épars inhérent aux données web. Deux ensembles de classifications différentes sont donc générés, le premier étant basé sur l'information de lien et le second étant basé sur l'information textuelle indexée (cf. section A.3.1). Pour les deux ensembles, différentes cartes SOM sont finalement générées de 6×6 neurones (classes) jusqu'à 24×24 neurones (classes).

Les figures 4.1 et 4.2 montrent l'évolution des différents critères (R , P , F – mesure, inertie intra-classe et inertie inter-classes), selon la croissance du nombre de classes.

La figure 4.1, montre que l'inertie intra-classe et l'inertie inter-classes sont maintenues aux valeurs basses, quelque soit le nombre de classes. Par contre, la mesure de **Rappel** diminue lentement de 0,8 à 0,7, la mesure de **Précision** augmentant conjointement de 0,3 à 0,7, avec un point d'inflexion à 400 classes. Notre algorithme de recherche de modèle optimal qui est basé sur les mesures R , P , F a déterminé avec précision ce dernier point en tant que un point optimal. Également, la figure 4.1 montre clairement que ce point est un point spécifique des courbes de R , de P , et de F , tandis que de tels points caractéristiques seront plus difficiles à

⁷ A cause du même principe d'apprentissage, notre étude sur la généralisation montrera cependant que, dans le cas d'une forte proportion de données marginales, l'influence de ces dernières peut influencer plus fortement les profils de classes et ainsi générer une redistribution de données homogènes dans des classes différentes

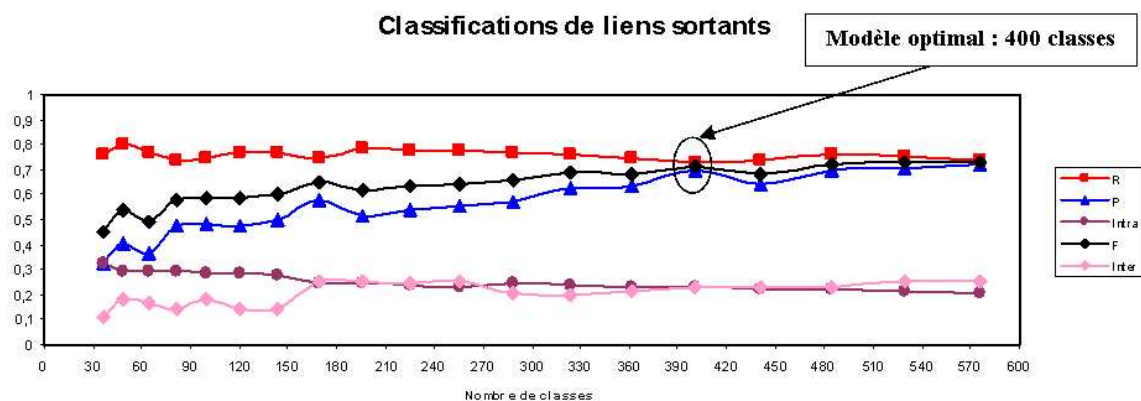


FIG. 4.1 – Évolution des mesures de qualité en considérant le nombre de classes pour le point de vue **liens sortants** (R = Rappel, P = Précision, Intra = inertie intra-classe, F = F-mesure, Inter = inertie inter-classes).

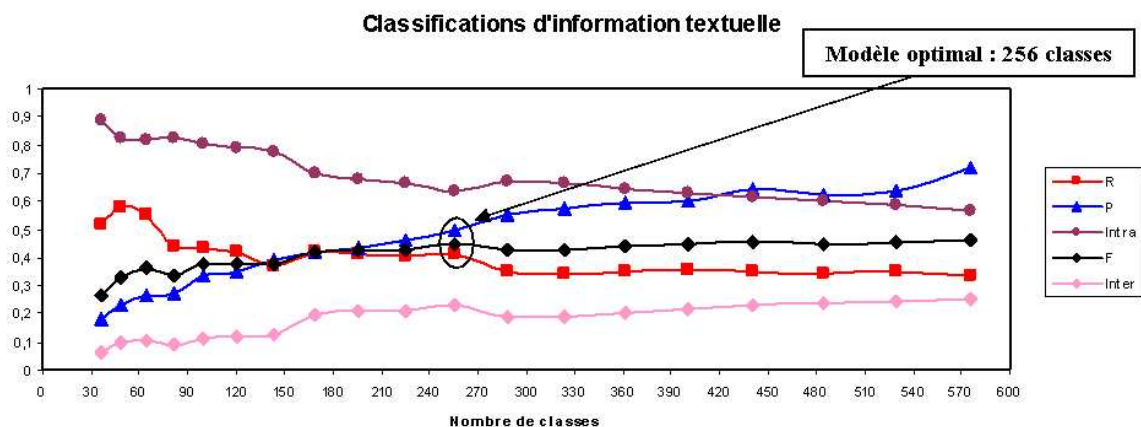


FIG. 4.2 – Évolution des mesures de qualité en considérant le nombre de classes pour le point de vue **information textuelle** (R = Rappel, P = Précision, Intra = inertie intra-classe, F = F-mesure, Inter = inertie inter-classes).

trouver automatiquement si on se base sur des courbes d'inertie, ceci en raison de nombreuses intersections entre ces courbes dans un intervalle constant.

La figure 4.2, montre que l'inertie intra-classe diminue lentement de 0,9 à 0,6 tandis que l'inertie inter-classes augmente lentement de 0,1 à 0,25, et qu'aucune intersection, ou point de compromis, ne peut ainsi être trouvée entre ces courbes. Par conséquent, le comportement des courbes d'inertie illustre mal la qualité des résultats de classification. Par contre, dans le même cas, les courbes de R , P , F présentent des points caractéristiques. Les courbes de R et de P intersectent à 144 classes, alors qu'un point optimal est trouvé par notre algorithme à 256 classes. Ces deux points caractéristiques sont également les points caractéristiques des courbes d'inertie, mais dans ce dernier cas il sera plus difficile les détecter automatiquement en raison des coefficients de variations insuffisantes des courbes d'inertie.

La consultation de courbes de F met en évidence une différence significative de qualité

entre les classifications textuelles et les classifications de liens, en faveur des classifications de liens. Ces différences peuvent être expliquées par l'analyse de la distribution des termes d'index textuels et de celle des liens dans les pages web. Pour cela, nous faisons trois analyses différentes basées sur les descripteurs communs (termes textuels ou liens) partagés par les documents (cf. figure 4.3) en faisant des intersection deux à deux les documents. Ces intersections génèrent des différentes tailles des motifs⁸ et chaque motif généré a un nombre de documents participés (ou support de motif (cf. section 1.4.3)). Dans toutes les analyses, l'axe des abscisses représente la taille des motifs de descripteurs en augmentant leurs tailles. La première analyse a) souligne le nombre de différents descripteurs inclus dans chaque taille de motif. La deuxième analyse b) se concentre sur le nombre de différentes combinaisons des descripteurs pour chaque taille de motif. La troisième analyse c) décrit le nombre de documents qui sont associés à chaque taille de motif.

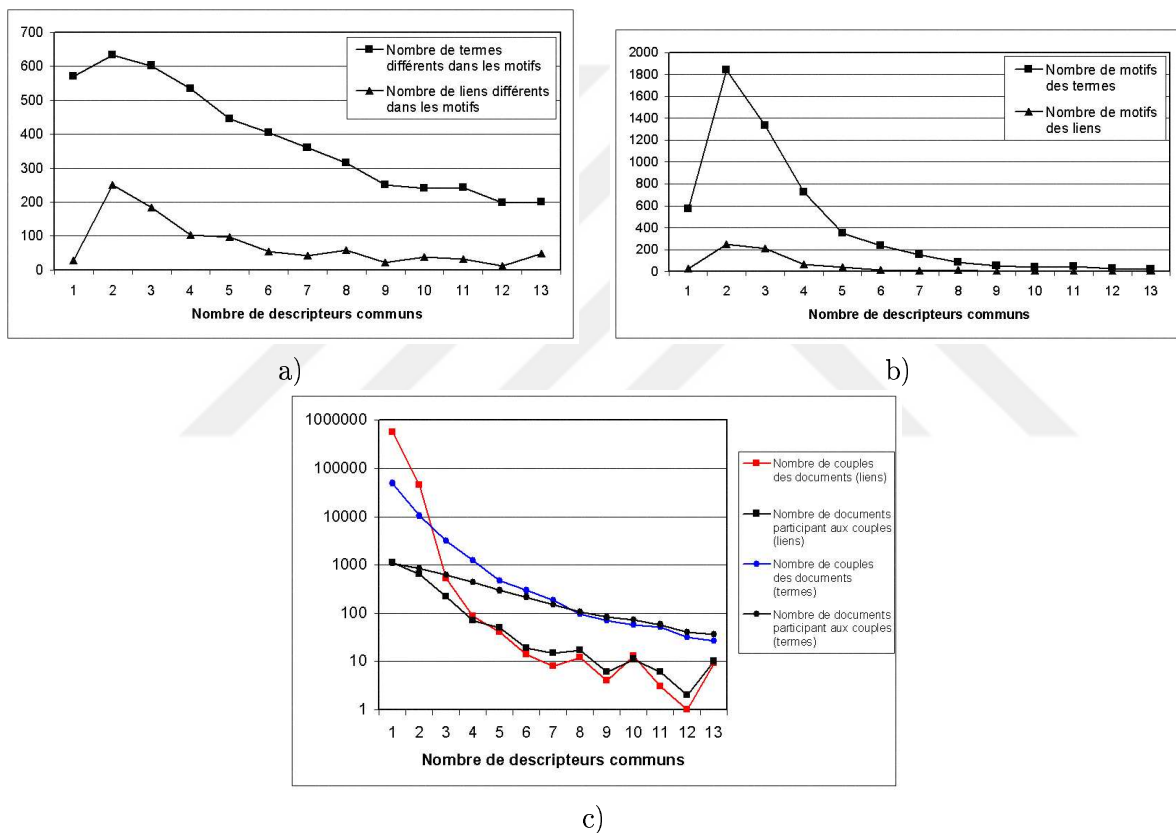


FIG. 4.3 – Distribution des termes et des liens dans les pages web : a) Nombre de descripteurs inclus dans des motifs de descripteur des documents, b) Nombre de différents motifs de descripteur inclus dans le motif de descripteur de documents, c) Nombre de couples des documents/nombre de documents indexés par motif.

Les deux courbes résultant de la première analyse a) ont la même forme pour les liens et pour les termes, avec un sommet important pour des motifs de la taille 2. Néanmoins, la dispersion des termes semble beaucoup plus importante que la dispersion des liens à cause du

⁸La définition du motif est décrite dans la section 1.4.3

nombre plus élevé de termes inclus dans les motifs. Par ailleurs, la deuxième analyse b) montre que la richesse des combinaisons des motifs est également beaucoup plus importante pour les termes que pour les liens. La troisième analyse c) montre que, avec une taille augmentée de motif, le nombre de documents impliqués est beaucoup plus stable avec des termes qu'avec des liens. Tous ces critères confirment que le recouvrement dans des descriptions de document est plus important quand les termes sont considérés. Ceux-ci pourraient expliquer le fait que la discrimination des documents dans les classes séparées est moins efficace dans la classification de termes, menant à faire baisser la qualité des classes.

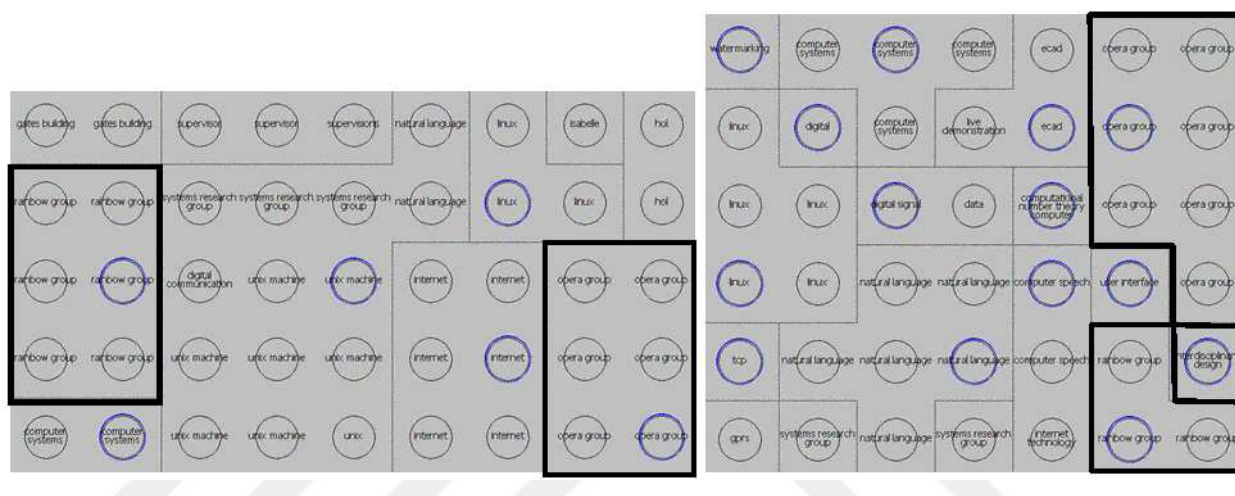


FIG. 4.4 – Comparaison entre la carte 12×12 à gauche et la carte 16×16 à droite qui représentent la classification d'**information textuelle**. Les aires **Opera Group** et **Rainbow Group**, qui représentent le focus de la comparaison, sont spécifiquement entourées par des lignes noires épaisses.

Pour évaluer le perfectionnement de notre algorithme d'optimisation de classification, nous utilisons le point de vue **information textuelle**, car il est plus facile à analyser. Nous avons ainsi décidé de comparer subjectivement la classification ($16 \times 16 = 256$ classes) qui est considérée en tant que optimale par notre algorithme, avec une autre classification ($12 \times 12 = 144$ classes) de mauvaise qualité. En regardant la figure 4.4, la carte 16×16 fournit à l'utilisateur des informations plus précises : de plus petits aires thématiques en moyenne avec des étiquettes plus précises. A titre d'exemple, la figure 4.4 illustre l'interaction entre deux équipes importantes du Cambridge Computer Laboratory, que sont le groupe Opera (**Opera Group**), travaillant sur les systèmes distribués (**Distributed Systems**), et le group Rainbow (**Rainbow Group**) travaillant sur l'interaction homme-humaine (**Human Computer Interaction**). La carte 16×16 illustre directement l'interaction entre ces groupes par leur proximité sur la carte et, par ailleurs, elle clarifie des thématiques d'interaction spécifiques par les classes intermédiaires (c'est-à-dire, les classes **User Interface**, **Interdisciplinary Design** et **Computer Speech/Internet Technology**) qui sont des classes voisines de l'intersection. Dans la carte 12×12 ces deux groupes sont éloignés et séparés par des aires de classes générales.

Optimisation du fonctionnement interne d'une méthode de classification

Dans cette expérimentation nous comparons la méthode Neural Gas (NG) standard à une autre optimisation de cette méthode nommée le gaz neuronal normalisé (NG-normalisé) [Yahya, 1997]. Dans le cas du gaz neuronal normalisé, les classes sont renormalisées à chaque étape de l'apprentissage. Les raisons pour une telle opération sont que, dès que les données documentaires de la collection seront habituellement normalisées, la normalisation des classes a tendance à réduire les différences potentielles qui pourraient apparaître entre les classes et les données pendant le processus de classification.

Nous avons basés cette expérimentation sur le corpus des 1000 notices de brevets (cf. section A.2) en sélectionnant un point de vue spécifique, à savoir le point de vue (**Utilisations**). A partir de ce point de vue nous avons générés de 36 jusqu'à 225 classes (neurones) en utilisant parallèlement des méthodes de NG-standard et de NG-normalisé [Al Shehabi and Lamirel, 2004b]. Les figures 4.5 et 4.6 montrent l'évolution des résultats des différents critères de qualité selon la croissance du nombre de classes.

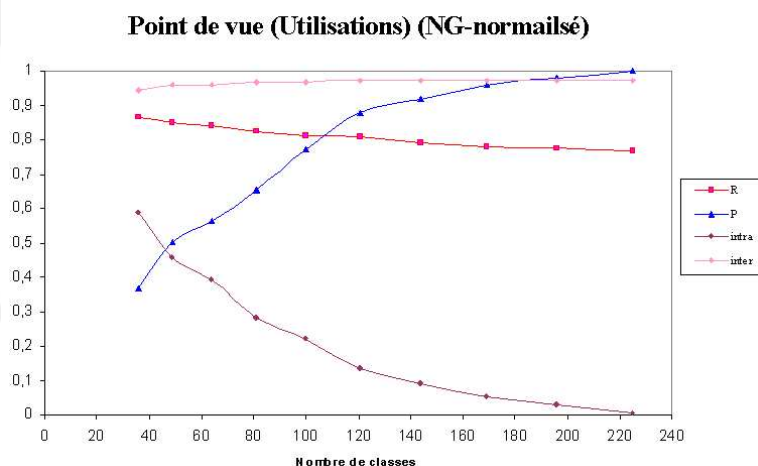


FIG. 4.5 – Évolution des critères de qualité selon le nombre de classes pour le point de vue **Utilisations** en utilisant la méthode de NG-normalisé.

L'examen des courbes de qualité montre que NG-normalisé a une mauvaise précision par rapport à NG-standard au début de la courbe (c'est-à-dire, pour un petit nombre de classes). L'examen de courbes met également en évidence que le point d'équilibre de NG-standard est vérifié avant celui de NG-normalisé, avec dans ce premier cas un meilleur compromis entre le **Rappel** et la **Précision**. Enfin, la mauvaise précision de NG-normalisé pour un plus petit nombre de classes est du fait que les projections des documents se produisent dans des classes fausses, à cause du processus de normalisation des classes. Ce phénomène est particulièrement important quand la distribution des données à classer est éparse, ou ultra-creuse, comme c'est souvent le cas dans le domaine documentaire. Il tend naturellement à disparaître dès que le nombre de classes a augmenté jusqu'à atteindre le nombre de documents.

Notre expérimentation démontre la supériorité de la méthode originale de NG-standard par rapport à NG-normalisé en terme de capacité de synthèse. La capacité de synthèse d'une méthode est liée à son aptitude à fournir la meilleure classification possible avec le nombre le plus faible possible de classes. Elle ne peut cependant être évaluée que de manière comparative.

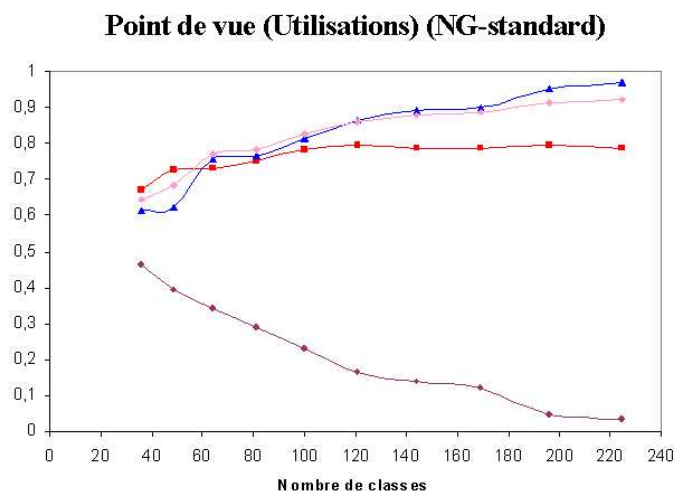


FIG. 4.6 – Evolution des critères de qualité selon le nombre de classes pour le point de vue **Utilisations** en utilisant la méthode de NG-standard.

Cette conclusion semble encore une fois être contraire à l'intuition, mais la raison de ce comportement apparaît plus claire quand nous réalisons une étude théorique. De nouveau, dans cette expérimentation notre mesure de qualité nous fournit une assistance précieuse pour mettre en évidence la différence de comportement et l'efficacité entre les méthodes de classification.

Comparaison entre différentes méthodes de classification neuronales

Cette expérimentation consiste à faire une comparaison entre les trois méthodes non supervisées (NG, GNG, et SOM) en utilisant le même corpus de pages du site de Cambridge (cf. section A.3.1). La comparaison est effectuée par l'analyse de l'évolution de nos critères de qualité en fonction d'évolution le nombre de classes [Al Shehabi and Lamirel, 2004b].

Une classification est alors générée en basant sur l'**information textuelle**. Pour ce point de vue, de 36 jusqu'à 225 neurones (classes) issues de NG, de GNG, et de SOM sont générés.

Les figures 4.7, 4.8 et 4.2 montrent l'évolution des différents critères, inertie et qualité, selon la croissance du nombre de classes. Ces figures montrent que les courbes de **Rappel** et de **Précision** de NG sont meilleures que celle de GNG et de SOM. De plus, les mêmes courbes sont bien meilleures pour GNG que pour SOM. Les mêmes remarques s'appliquent sur des courbes d'inertie. Finalement, l'examen de toutes les courbes met en évidence que le point d'équilibre (c'est-à-dire, la classification optimale) apparaît plus tôt (c'est-à-dire, avec un plus petit nombre de classes) dans l'ordre de NG que dans GNG et SOM respectivement.

Les mauvais résultats de la méthode GNG comparés à ceux de la méthode NG peuvent être surprenants dès lors que le principe de la méthode GNG est plus élaboré que celui de la méthode NG (cf. section 2.3.5). La raison de ces mauvais résultats est plutôt à chercher du côté des paramètres de GNG, tels que l'âge de connexion entre les neurones, qui sont très difficiles à contrôler. Cela est particulièrement sensible dans le cas des données fortement multidimensionnelles.

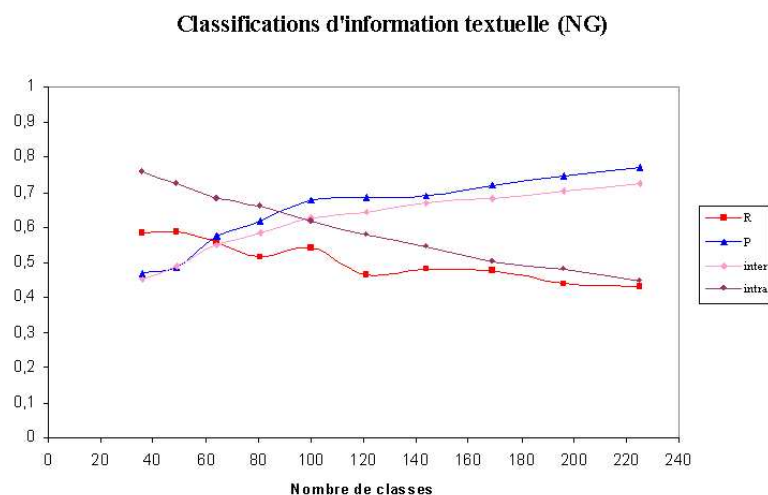


FIG. 4.7 – Evolution des critères de qualité selon le nombre de classes pour le point de vue **information textuelle** en utilisant la méthode de NG.

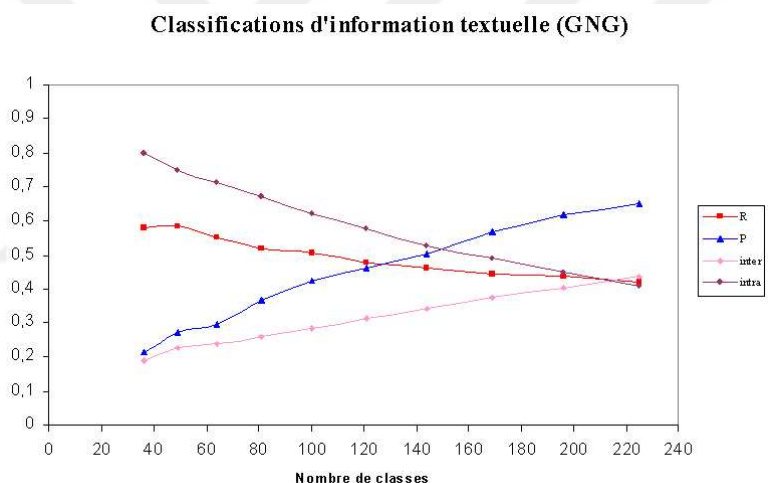


FIG. 4.8 – Évolution des critères de qualité selon le nombre de classes pour le point de vue **information textuelle** en utilisant la méthode GNG.

4.3 Qualité et analyse de classification non supervisée pour les données avec des étiquettes multiples

Nous aborderons dans cette section le problème de la classification de données lorsque ces données sont associées aux étiquettes multiples. Nous traitons particulièrement le cas où chaque étiquette n'a aucune étiquette antagoniste et l'absence d'une étiquette pour des données n'implique pas nécessairement que ces données ne peuvent pas avoir ladite étiquette, par exemple les substances associées à un emplacement géographique dans l'exploration minière, les termes d'index des pages web, ... Cette forme spécifique de problème de classification n'est pas résolu, même si elle est rencontrée dans de nombreux domaines tels que l'analyse de modèle, la recherche documentaire, le traitement d'image, le traitement de la parole, la catégorisation des

textes, la catégorisation de Web, le bioinformatique et le système d'information géographique (GIS).

D'une manière générale, deux approches principales sont considérées pour classifier des données avec des étiquettes multiples [Duda and Hart, 1973] : l'approche de classification supervisée, ou catégorisation, avec des multi-classes et l'approche de classification non supervisée. L'approche de classification supervisée exige des données avec des étiquettes positives et négatives. Selon le type précédent de données, une telle approche devient imprécise du fait que l'absence d'une étiquette pour un exemple ne signifie pas nécessairement que ledit exemple représente un exemple négatif. De leur côté, les méthodes de classification non supervisée (cf. chapitre 2), classifient ces données sans prendre en compte leurs étiquettes spécifiques même si ce genre d'information peut être considéré comme utile pour choisir un bon modèle de classification. Cette remarque se rapporte de celle que nous avons fait à la section précédente (cf. section 4.2) concernant la non prise en compte des propriétés des données associées aux classes pour les méthodes d'évaluation de la qualité des classifications non supervisées. Elle la généralise cependant en considérant les étiquettes associées aux données comme des propriétés externes au processus de classification. A titre d'exemple, une méthode couramment utilisée en scientométrie, ou évaluation de la science, est celle d'opérer une classification des auteurs en groupes de collaboration à partir de leur profil de co-publication ou de co-citation. Dans ce même domaine, le type de démarche que nous proposons peut s'avérer très utile pour identifier a posteriori les caractéristiques les plus représentatives de chaque groupe. Par exemple, en projetant sur chaque groupe d'auteurs des étiquettes correspondant à des thèmes des recherches couverts par chacun des auteurs du groupe, et en opérant une analyse de la répartition de ces étiquettes, il est possible de faire émerger les thèmes de recherche dominants des différents groupes. De plus, la gestion de l'information des étiquettes devrait être également utile pour des tâches précises d'extraction de connaissances. Ainsi, dans le contexte de l'étiquetage précédemment décrit, l'analyse fine des étiquettes associées à un groupe d'auteurs peut permettre d'identifier des associations pertinentes ou nouvelles entre des thèmes de recherche.

Nous proposons donc dans cette section des mesures de qualité de classification originales avec une nouvelle analyse de classification (CAMDL) tout en se basant sur les données avec des étiquettes multiples [Attik *et al.*, 2006b; Attik *et al.*, 2006a].

Les mesures de qualité et l'analyse de classification appelée (CAMDL) proposées sont basées sur l'utilisation de deux informations principales : la similarité entre les données regroupées par l'algorithme de classification et la distribution des étiquettes dans le modèle après la projection de ces étiquettes sur le modèle de classification. La méthode CAMDL se fonde sur deux types d'analyses qui sont conduites d'une manière parallèle : analyse d'étiquettes et analyse de classes. Elle est basée sur des mesures de qualité probabilistes pour mener à bien ces analyses.

Ces mesures de qualité sont utiles pour réaliser les choses suivantes :

1. une classification de données très précise,
2. une identification du modèle de classification optimal parmi plusieurs modèles,
3. la détermination du critère d'arrêt pendant la phase d'apprentissage d'un modèle de classification.
4. l'utilisation de critères complémentaires de validation pour la classification qui s'appliquent eux-mêmes pour :
 - (a) la validation de la stratégie d'étiquetage de données utilisée;

- (b) la détermination des différents types d'étiquettes associées aux données : implicite, bruitée, nouvelle et explicite ;
- (c) la détermination de la pertinence d'étiquette ;
- (d) la sélection des classes intéressantes (classes pertinentes) qui peuvent être utilisées pour une autre étape ultérieure d'extraction de connaissances.

4.3.1 Nouvelles mesures de qualité basées sur la distribution d'étiquettes

L'approche proposée dans la section 4.2 peut être adaptée aux données avec des étiquettes multiples si les propriétés ($p \in P$) sont remplacées par des étiquettes ($l \in L$). Mais son défaut principal est que la détermination du modèle optimal est basée sur l'utilisation d'un seuil qui doit être fixé par un expert. Au cours de nos expérimentations, nous avons pu nous apercevoir que ce seuil dépendait des propriétés de chaque ensemble de données spécifique à analyser.

Par conséquent, nous proposons ci-après des nouvelles mesures pour évaluer le modèle de classification et pour sélectionner un modèle optimal. Ces mesures étendent les mesures du **Rappel** et de la **Précision** décrites à la section 4.2. Elles évaluent la pertinence du contenu des classes en terme d'étiquettes partagées en se basant sur la théorie des probabilités.

Soit $X = \{x_1, x_2, \dots, x_N \mid x_i \in \mathbb{R}^D\}$ un ensemble de N données de dimension D , $L = \{l_1, l_2, \dots, l_M\}$ indique un ensemble d'étiquettes différentes, et $|L| = M$, $|\bar{L}|$ indique le nombre total d'étiquettes affectées aux données et :

$$\psi : x_i \mapsto L_i \text{ où } L_i \subseteq L$$

$\bigcup_i L_i = L$ et chaque L_i peut avoir une taille différente.

L'algorithme de classification A consiste à produire une partition de X en K classes, $C = \{c_1, c_2, \dots, c_K\}$, telles que :

$$A : x_i \mapsto c_j, \quad c_{k_1} \cap c_{k_2} = \emptyset \quad \forall k_1 \neq k_2 \quad \text{et} \quad X = \bigcup_k c_k$$

Soit ϕ et Φ deux fonctions telles que :

$$\begin{cases} \phi : (c_l, l_k) \mapsto \{L_i \mid l_k \in L_i \wedge \psi^{-1}(L_i) \in c_l\} \\ \Phi : l_k \mapsto \{L_i \mid l_k \in L_i \wedge L_i \subseteq L\} \end{cases}$$

où $x_i \in X$, $l_k \in L$, $c_l \in C$.

Définition 1

Le *Rappel de classification (CR)* R_{l_k, c_l} représente le rapport entre la fréquence de l'étiquette l_k dans la classe c_l et la fréquence de la même étiquette l_k dans tout l'ensemble de données :

$$R_{l_k, c_l} = \frac{|\phi(c_l, l_k)|}{|\Phi(l_k)|} \quad (4.13)$$

Le rappel fait une analyse globale (au niveau des toutes les classes). Il représente la probabilité conditionnelle $p(c_l | l_k)$. Donc, il vérifie :

$$\sum_{c_l \in \bar{C}} R_{l_k, c_l} = \sum_{c_l \in \bar{C}} \frac{|\phi(c_l, l_k)|}{|\Phi(l_k)|} = 1$$

Définition 2

La *Précision de Classification (CP)* P_{l_k, c_l} représente le rapport entre la fréquence d'étiquette l_k dans la classe c_l et la somme de toutes les fréquences des différentes étiquettes de cette même classe c_l

$$P_{l_k, c_l} = \frac{|\phi(c_l, l_k)|}{\sum_{L_i \in \phi(c_l, \cdot)} |L_i|} \quad (4.14)$$

La précision fournit une analyse locale (au niveau de la classe). Elle représente la probabilité conditionnelle $p(l_k|c_l)$. Donc, elle vérifie :

$$\sum_{l_k \in \bar{L}_{c_l}} P_{l_k, c_l} = \sum_{l_k \in \bar{L}_{c_l}} \frac{|\phi(c_l, l_k)|}{\sum_{L_i \in \phi(c_l, \cdot)} |L_i|} = 1$$

où $\bar{C}$ représente un ensemble de classes non vides extrait à partir des classes C c'est-à-dire qu'il vérifie : $\bar{C} = \{c_l \in C \mid |c_l| \geq 1\}$ et $\bar{L}_{c_l} = \{l_k \mid \exists L_i \text{ où } L_i \in \phi(c_l, l_k)\}$

La relation entre le rappel et la précision est donnée par la formule de Bayes :

$$R_{l_k, c_l} = p(c_l|l_k) = \frac{p(l_k|c_l) \times p(c_l)}{\sum_i p(l_k|c_i) \times p(c_i)}$$

Définition 3

L'*Entropie de Rappel d'Etiquette (LRE)* H_{l_k} est définie comme :

$$H_{l_k} = E(\log R_{l_k, c_l}) = - \sum_{c_l \in \bar{C}} R_{l_k, c_l} \log R_{l_k, c_l} \quad (4.15)$$

Deux cas particuliers de valeurs de l'Entropie de Rappel d'Etiquette (LRE) sont représentés par :

- $H_{l_k} \longrightarrow 0$: l'étiquette l_k existe seulement dans une classe.
- $H_{l_k} \longrightarrow \log(|\bar{C}|)$: les classes contiennent les mêmes fréquences d'étiquette l_k .

Définition 4

L'*Entropie de Précision de Classe (CPE)* H_{c_l} est définie comme :

$$H_{c_l} = - \sum_{l_k \in \bar{L}_{c_l}} P_{l_k, c_l} \log P_{l_k, c_l} \quad (4.16)$$

Deux cas particuliers de l'Entropie de valeurs de Précision d'Etiquette (CPE) sont représentés par :

- $H_{c_l} \longrightarrow 0$: la classe possède seulement une étiquette.
- $H_{c_l} \longrightarrow \log(|\bar{L}_{c_l}|)$: les étiquettes sont équiprobables dans la classe.

On peut montrer que LRE et CPE se comportent d'une manière antagoniste. Ainsi, le LRE augmente et le CPE diminue si le nombre de classes augmente.

Définition 5

De l'équation 4.13, le *Rappel de Classe (ClustR)* R_{c_l} est donné comme :

$$R_{c_l} = \frac{1}{|L_{c_l}|} \sum_{l_k \in \bar{L}_{c_l}} R_{l_k, c_l} = \frac{1}{|L_{c_l}|} \sum_{l_k \in \bar{L}_{c_l}} \frac{|\phi(c_l, l_k)|}{|\Phi(l_k)|} \quad (4.17)$$

Deux cas particuliers de valeurs de Rappel Classe (ClustR) sont représentés par :

- $R_{c_l} \longrightarrow 0$: les fréquences moyennes des étiquettes sont basses dans la classe c_l
- $R_{c_l} \longrightarrow 1$: les fréquences moyennes des étiquettes sont élevées dans la classe c_l

Définition 6

De l'équation 4.14 la *Précision d'étiquette (LP)* P_{l_k} est donnée comme :

$$P_{l_k} = \frac{1}{|\bar{C}_{l_k}|} \sum_{c_l \in \bar{C}} P_{l_k, c_l} = \frac{1}{|\bar{C}_{l_k}|} \sum_{c_l \in \bar{C}} \frac{|\phi(c_l, l_k)|}{\sum_{L_i \in \phi(c_l, \cdot)} |L_i|} \quad (4.18)$$

où $\bar{C}_{l_k}$ représente les classes qui contiennent l'étiquette l_k et $\bar{L}$ représente toutes les étiquettes.

Deux cas particuliers de valeurs de précision d'étiquette (LP) sont représentés par :

- $P_{l_k} \longrightarrow 0$: l'étiquette l_k a une faible représentation dans les classes
- $P_{l_k} \longrightarrow 1$: l'étiquette l_k a une forte représentation dans les classes

Définition 7

De l'équation 4.13, la *Moyenne de Rappel de Classification (MCR)* est donnée comme :

$$MCR = \frac{1}{|\bar{C}|} \sum_{c_l \in \bar{C}} \frac{1}{|L_{c_l}|} \sum_{l_k \in \bar{L}_{c_l}} R_{l_k, c_l} \quad (4.19)$$

alors,

$$MCR = \frac{1}{|\bar{C}|} \sum_{c_l \in \bar{C}} \frac{1}{|L_{c_l}|} \sum_{l_k \in \bar{L}_{c_l}} \frac{|\phi(c_l, l_k)|}{|\Phi(l_k)|}$$

Deux cas particuliers de valeurs de Moyenne de Rappel de Classification (MCR) sont représentés par :

- $MCR \longrightarrow 0$: chaque étiquette existe dans toutes les classes.
- $MCR \longrightarrow 1$: chaque étiquette existe seulement dans une classe.

Définition 8

De l'équation 4.14, la *Moyenne de Précision de Classification (MCP)* est donnée comme :

$$MCP = \frac{1}{|L|} \sum_{l_k \in \bar{L}} \frac{1}{|\bar{C}_{l_k}|} \sum_{c_l \in \bar{C}} P_{l_k, c_l} \quad (4.20)$$

alors,

$$MCP = \frac{1}{|L|} \sum_{l_k \in \bar{L}} \frac{1}{|\bar{C}_{l_k}|} \sum_{c_l \in \bar{C}} \frac{|\phi(c_l, l_k)|}{\sum_{L_i \in \phi(c_l, \cdot)} |L_i|}$$

où $\bar{C}_{l_k}$ représente les classes qui contiennent l'étiquette l_k .

Deux cas particuliers de valeurs de *Moyenne de Précision de Classification (MCP)* sont représentés par :

- $MCP \longrightarrow 0$: chaque classe contient des étiquettes équiprobables.
- $MCP \longrightarrow 1$: chaque classe contient seulement une étiquette.

Définition 9

La *Moyenne d'Entropie de Rappel de Classification (MCRE)* est définie comme :

$$MCRE = \frac{1}{|L|} \sum_{l_k \in L} H_{l_k} \quad (4.21)$$

ou

$$MCRE = -\frac{1}{|L|} \sum_{l_k \in L} \sum_{c_l \in \bar{C}} R_{l_k, c_l} \log R_{l_k, c_l}$$

Deux cas particuliers de valeurs de *Moyenne d'Entropie de Rappel de Classification (MCRE)* sont représentés par :

- Si l'étiquette existe seulement dans une classe alors $H_{l_k} = 0$.
- Si les classes contiennent les mêmes fréquences de la même étiquette alors $H_{l_k} = \log|\bar{C}|$.

Définition 10

La *Moyenne d'Entropie de Précision de Classification (MCPE)* est définie comme :

$$MCPE = \frac{1}{|\bar{C}|} \sum_{c_l \in \bar{C}} H_{c_l} \quad (4.22)$$

ou

$$MCPE = -\frac{1}{|\bar{C}|} \sum_{c_l \in \bar{C}} \sum_{l_k \in L_{c_l}} P_{l_k, c_l} \log P_{l_k, c_l}$$

Deux cas particuliers de valeurs de *Moyenne d'Entropie de Précision de Classification (MCPE)* sont représentés par :

- Si toutes les classes contiennent une seule étiquette alors $MCPE = 0$.
- Si toutes les classes représentent des étiquettes équiprobables (i.e. les étiquettes dans les classes ont les mêmes membres) alors $MCPE = \log|\bar{L}| - \log|\bar{C}|$.

4.3.2 Sélection de modèle optimal

Notre approche pour la sélection de modèle optimal parmi un ensemble de modèles différents peut être décrite par les étapes suivantes [Attik *et al.*, 2006b] :

1. **Génération d'un ensemble de modèles de classification** : lancer un algorithme de classification avec un nombre croissant de classes en utilisant seulement l'ensemble de données non étiquetées X .
2. **Projection des étiquettes** : dans cette étape les étiquettes sont projetées sur les classes en fonction de leurs données préalablement associées.
3. **Sélection d'un modèle optimal** : le modèle optimal est sélectionné par l'utilisation des approches d'analyse basées sur le compromis (décrites ci-après).

La technique basée sur l'analyse de compromis

Lorsque le MCRE augmente et le MCPE diminue d'une manière parallèle en augmentant le nombre de classes, les courbes de MCRE et de MCPE peuvent être obtenues en générant plusieurs modèles avec un nombre croissant de classes. Les techniques basées sur le compromis doivent alors être utilisées afin d'effectuer une analyse combinée de telles courbes. Nous proposons ci-après deux différentes approches basées sur le compromis :

- L'approche basée sur la normalisation : le but est de trouver un point où les courbes de MCRE et de MCPE sont presque égales (c'est-à-dire, un point d'équilibre). Cependant, une normalisation doit être réalisée dans une première étape pour trouver le compromis optimal entre MCRE et MCPE.
- L'approche basée sur la dérivation : le but est de trouver un point qui est caractéristique à partir de deux courbes MCRE et MCPE par l'utilisation de leurs premiers dérivés, cela sera respectivement appelé VMCRE et VMCPE. Un tel point peut être trouvé quand il représente un minimum local pour la courbe de VMCRE et un maximum local pour la courbe VMCPE.

Ces deux approches sont présentées comme suit :

L'approche basée sur la normalisation : Plusieurs méthodes de normalisation peuvent être utilisées dans cette approche. Nous présentons ci-après deux méthodes habituelles :

1. **Normalisation de Min-Max.** Elle représente une transformation linéaire de l'intervalle original de données dans un nouvel intervalle de données. L'avantage de cette méthode est qu'il préserve exactement les relations originales entre les données. En effet, elle n'introduit aucun biais potentiel dans les données. Dans le cas de $NMCRE_{y_i} \in [0, 1]$ nous avons :

$$NMCRE_{\{y_i\}} = \frac{MCRE_{\{y_i\}} - Min}{Max - Min} \quad (4.23)$$

où Min est la valeur minimale et Max est la valeur maximale de l'ensemble de valeurs de $MCRE$. Y représente l'ensemble de tous les modèles générés et $y_i \in Y$ est un modèle donné.

2. **Normalisation de z-score.** Les valeurs sont normalisées sur la base de la moyenne et de l'écart type. Cette méthode fonctionne bien quand le minimum et le maximum réels des données d'entrée ne sont pas connus. Pour les valeurs de MCRE nous avons :

$$ZMCRE_{\{y_i\}} = \frac{MCRE_{\{y_i\}} - Mean}{\sigma} \quad (4.24)$$

$$\sigma = \sqrt{\frac{1}{|Y|} \sum_{i=1}^{|Y|} (MCRE_{\{y_i\}} - Mean)^2}$$

et

$$Mean = \frac{1}{|Y|} \sum_{i=1}^{|Y|} MCRE_{\{y_i\}}$$

les valeurs de $ZMCPE$ correspondant aux valeurs de $MCPE$ peuvent être obtenues de la même manière.

Dans les deux cas, le modèle qui est le plus proche du point d'équilibre est considéré comme un modèle optimal. La première méthode de normalisation est sensible aux valeurs de Min et de Max et la deuxième méthode de normalisation est sensible à la valeur de $Mean$.

Dès lors qu'elle repose sur la prise en compte d'un seul point caractéristique sur les courbes, une approche basée sur la normalisation représente une manière simple d'obtenir une évaluation d'un modèle, en évitant un processus de comparaison directe entre les modèles. Cependant, son inconvénient principal est qu'elle exige de calculer beaucoup de modèles afin d'obtenir une stabilité dans les résultats. Ainsi, une telle stabilité est atteinte quand il n'y a plus de changements des valeurs de Min et Max (pour la première méthode de normalisation) et dans ceux de $Mean$ (pour la deuxième méthode de normalisation), ceci lorsque le nombre de classes augmente.

L'approche basée sur la dérivation : Les dérivées premières de $MCRE$ et $MCPE$ sont définies comme suit :

$$VMCRE_{\{y_i\}} = \frac{MCRE_{\{y_{i+1}\}} - MCRE_{\{y_i\}}}{\bar{E}_{y_{i+1}} - \bar{E}_{y_i}} \quad (4.25)$$

$$VMCPE_{\{y_i\}} = \frac{MCPE_{\{y_{i+1}\}} - MCPE_{\{y_i\}}}{\bar{E}_{y_{i+1}} - \bar{E}_{y_i}} \quad (4.26)$$

où $\bar{E}_{y_i}$ est le nombre de classes non vides issus du modèle y_i .

Dans ce cas, le modèle pour lequel les premiers minimum local de la courbe $VMCRE$ et maximum local de la courbe $VMCPE$ se produisent en même temps est considéré comme un modèle optimal.

L'avantage de cette approche est qu'elle exige seulement trois modèles pour examiner l'existence d'un compromis entre $MCRE$ et $MCPE$.

4.3.3 Détermination du critère d'arrêt d'apprentissage

Notre approche pour déterminer le critère d'arrêt de la phase d'apprentissage pour un modèle de classification peut être décrite par les étapes suivantes :

1. **Initialisation** : fixer le nombre de classes.
2. **Cycle d'apprentissage** : lancer l'algorithme de classification en utilisant seulement l'ensemble de données non étiquetées X pour seulement un cycle.
3. **Projection des étiquettes** : projeter les étiquettes sur les classes en fonction de leurs données préalablement associées.
4. **Évaluation** : dans le cas de l'approche basée sur la normalisation, l'arrêt de l'apprentissage est défini quand la phase de stabilité de $MCRE$ (cf. équation (4.21)) et $MCPE$ (cf. équation (4.22)) est atteinte.

Dans le cas de l'approche basée sur la dérivation, l'arrêt de l'apprentissage est défini quand $VMCE$ et $VMCPE$ sont égaux à zéro.

Si la condition d'arrêt n'est pas respectée, alors aller à l'étape de cycle d'apprentissage.

4.3.4 Définitions des types d'étiquettes

Nous définissons quatre types d'étiquettes pour l'analyse de classification [Attik *et al.*, 2006a] :

Définition 11

Une *étiquette nouvelle* est une étiquette qui apparaît dans un **faible** nombre de classes avec une représentation **faible**. Les étiquettes d'un tel type ne sont pas nécessairement attribuées d'une manière incorrecte. Elles doivent plutôt être considérées comme des informations cachées.

Définition 12

Une *étiquette bruitée* est une étiquette qui apparaît dans des classes d'une manière **globale** avec une représentation **faible**. Ce type d'étiquettes peut être éliminé afin d'améliorer la performance d'apprentissage et pour faciliter l'interprétation de données.

Définition 13

Une *étiquette explicite* est une étiquette qui apparaît dans des classes d'une manière **globale** avec une représentation **forte**. Ce type d'étiquettes fournit une information redondante dans le modèle. Il est donc possible de considérer qu'une étiquette globale qui doit être éliminée afin de faciliter l'interprétation de classes.

Définition 14

Une *étiquette implicite et spécifique* est une étiquette qui apparaît dans un **faible** nombre de classes avec une représentation **forte**.

Définition 15

Toutes les étiquettes qui sont différentes des étiquettes nouvelles, bruitées, explicites et spécifiques, sont considérées en tant qu'*étiquettes implicites et générales*.

Une étiquette implicite, qu'elle soit spécifique ou générale, peut être considérée comme une étiquette importante pour l'interprétation des données.

Selon les différentes définitions données ci-dessus :

- Le taux d'étiquettes nouvelles est défini par : $\frac{c}{c+d+e+f}$
- Le taux d'étiquettes bruitées est défini par : $\frac{e}{c+d+e+f}$
- Le taux d'étiquettes explicites est défini par : $\frac{f}{c+d+e+f}$
- Le taux d'étiquettes implicites est défini par : $\frac{d}{c+d+e+f}$

où c est le nombre d'étiquettes nouvelles, d est le nombre d'étiquettes implicites, e est le nombre d'étiquettes bruitées et f est le nombre d'étiquettes explicites.

4.3.5 Définitions de pertinence d'étiquette et de classe

Dans cette section nous sommes intéressés à définir la pertinence d'étiquette et la pertinence de classe à l'aide des mesures précédentes [Attik *et al.*, 2006a].

Définitions de la pertinence d'une étiquette

Afin d'évaluer la performance d'une stratégie d'étiquetage donnée, nous définissons deux notions, à savoir les *étiquettes pertinentes* et les *étiquettes non pertinentes* :

Définition 16

D'après les différentes définitions de type d'étiquette données ci-avant, une *étiquette pertinente* peut être définie comme une étiquette nouvelle ou implicite. Inversement, une *étiquette non pertinente* peut être définie comme une étiquette bruitée ou explicite.

- Le taux d'étiquettes pertinentes est défini par : $\frac{c+d}{c+d+e+f}$
- Le taux d'étiquettes non pertinentes est défini par : $\frac{e+f}{c+d+e+f}$

Définitions de la pertinence d'une classe

Nous proposons deux définitions différentes de la pertinence d'une classe : la première est basée sur le Rappel de Classe (*ClustR*) (cf. définition 5) et la seconde est basée sur la pertinence d'étiquette (cf. définition 16) :

Définition 17

Une classe est pertinente si sa valeur associée de Rappel de Classe (*ClustR*) est supérieure à un seuil donné. Une telle classe contiendra plusieurs étiquettes qui ont une bonne représentativité en ce qui concerne leur distribution globale.

Définition 18

Une classe est pertinente si elle contient des étiquettes pertinentes.

La définition 17 est une définition générale qui convient pour évaluer la qualité de la représentation d'un modèle. Donc, elle est basée sur la distribution d'étiquettes. La deuxième définition est plus précise. Elle peut spécifiquement être utilisée pour des tâches d'extraction de connaissances.

4.3.6 Analyse de classification

Dans cette section nous présentons trois types différents d'analyse : le premier est basé sur les étiquettes, le second est basé sur les classes, et le dernier est une combinaison des deux types d'analyse.

Analyse d'étiquettes

Dans cette étape nous étudions la distribution des étiquettes dans un modèle donné. Le but de cette analyse est de détecter : les étiquettes *nouvelles*, les étiquettes *implicites*, les étiquettes *explicites* et les étiquettes qui représentent du *bruit*. Notre approche pour l'analyse d'étiquettes peut être décrite par les étapes suivantes :

- **Apprentissage** : l'algorithme de classification est lancé en utilisant seulement les données non étiquetées X .
- **Projection des étiquettes** : dans cette étape les étiquettes sont projetées sur les classes en fonction de leurs données préalablement associées.
- **Étape d'analyse** : D'après les valeurs de H_{l_k} et P_{l_k} nous avons les résultats suivants :
 - $H_{l_k} \longrightarrow 0$ et $P_{l_k} \longrightarrow 0 \implies l_k$ est une étiquette nouvelle

- $H_{l_k} \longrightarrow 0$ et $P_{l_k} \longrightarrow 1 \implies l_k$ est une étiquette implicite
- $H_{l_k} \longrightarrow \log(|\bar{C}|)$ et $P_{l_k} \longrightarrow 0 \implies l_k$ est une étiquette bruitée
- $H_{l_k} \longrightarrow \log(|\bar{C}|)$ et $P_{l_k} \longrightarrow 1 \implies l_k$ est une étiquette explicite

Analyse de classes

Dans cette étape nous étudions la distribution des étiquettes dans le modèle pour chaque classe. Le but de cette analyse est de déterminer les classes *pertinentes* et les classes *non pertinentes* avec une ou plusieurs étiquettes.

- **Apprentissage** : l'algorithme de classification est lancé en utilisant seulement les données non étiquetées X .
- **Projection des étiquettes** : dans cette étape les étiquettes sont projetées sur les classes en fonction de leurs données préalablement associées.
- **Étape d'analyse** : D'après les valeurs de H_{c_l} et R_{c_l} nous avons les résultats suivants :
 - $H_{c_l} \longrightarrow 0$ et $R_{c_l} \longrightarrow 0 \implies c_l$ est une classe non pertinente ou bruitée
 - $H_{c_l} \longrightarrow 0$ et $R_{c_l} \longrightarrow 1 \implies c_l$ est une classe pertinente avec une seule étiquette
 - $H_{c_l} \longrightarrow \log(|\bar{L}_{l_k}|)$ et $R_{c_l} \longrightarrow 0 \implies c_l$ est une classe non pertinente ou bruitée
 - $H_{c_l} \longrightarrow \log(|\bar{L}_{l_k}|)$ et $R_{c_l} \longrightarrow 1 \implies c_l$ est une classe pertinente avec plusieurs étiquettes

Analyse hybride

Cette méthode représente un compromis entre l'analyse d'étiquettes et l'analyse de classes. Elle correspond à une stratégie de filtrage dans le but d'étudier la distribution de différents types d'étiquettes selon chaque classe, ainsi que la distribution des classes selon chaque type d'étiquette.

Cette méthode peut également être considérée comme une méthode d'analyse multi-objectifs. Donc, plusieurs alternatives peuvent en être proposées. Une alternative simple consiste à filtrer les classes pertinentes avec leurs étiquettes implicites. Une autre alternative consiste à opérer une analyse croisée, en filtrant les classes pertinentes ou les classes non pertinentes avec tous les types possibles d'étiquettes, à savoir : (1) étiquettes nouvelles (2) étiquettes implicites (3) étiquettes bruitées (4) étiquettes explicites.

4.3.7 Expérimentations

Notre base de test est la base de 1000 notices de brevets issus du domaine de l'ingénierie des huiles moteur (cf. section A.2).

Dans cette étude, nous nous sommes focalisées sur les descriptions des documents associés au point de vue **Avantages**. Les étiquettes des documents ont été ainsi établies par l'utilisation des termes de plus hautes fréquences extraits de ce point de vue spécifique. Cette stratégie d'étiquetage peut être considérée comme la plus simple possible. D'une telle manière, chaque liste d'étiquette L_i associé au document D_i peut avoir un nombre différent des étiquettes. Dans le cas du point de vue **Avantages**, chaque document possède un ensemble d'étiquettes L_i spécifique avec $|L_i| \in [1, 8]$. Le nombre total d'étiquettes manipulées est de 207.

Dans cette expérimentation nous avons utilisé l'algorithme de classification Neural Gas présenté dans la section 2.3.3.

Expérimentation pour la sélection de modèle

Notre première expérimentation consiste à générer, pour le point de vue **Avantages**, un ensemble de classifications de 3×3 neurones (classes) jusqu'à 11×11 neurones (classes) en utilisant la méthode de classification NG basée sur des vecteurs de description pondérés des documents. Le but de cette expérimentation est ainsi de comparer nos mesures de qualité à celle qui ont été décrites dans la section (4.2) et donc de mettre en évidence l'avantage de notre nouvelle méthode pour la sélection du modèle optimal. Ainsi, pour chaque modèle de classification, les MCR , MCP , $MCRE$, $MCPE$ et P sont calculés.

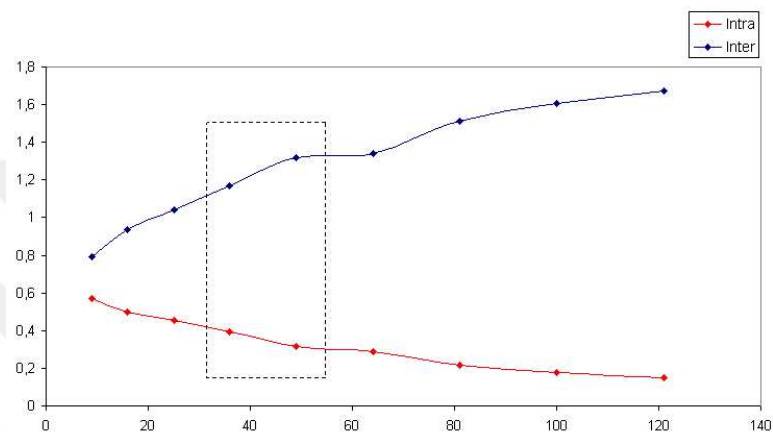


FIG. 4.9 – Mesures d'inertie intra-classe et inter-classes.

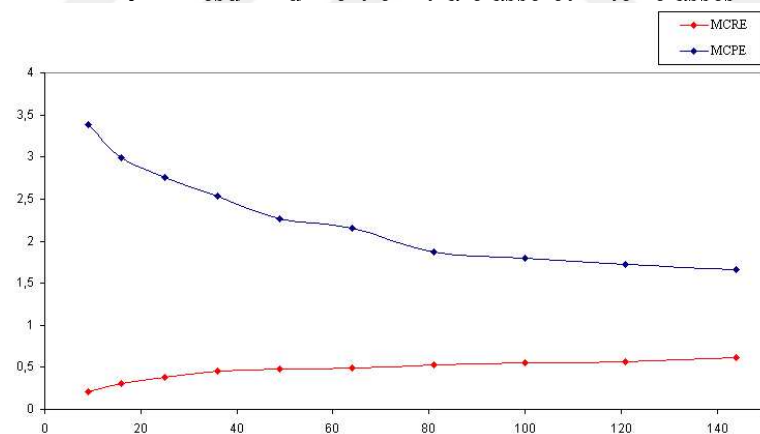


FIG. 4.10 – Mesures de MCRE et de MCPE.

Les résultats présentés dans les figures (4.10, 4.11, 4.12, 4.13, 4.14) illustrent l'évolution de différentes mesures MCR , MCP , P , $MCRE$, $MCPE$, $NMCRE$, $NMCPE$, $ZMCRE$, $ZMCPE$, $VMCRE$, $VMCPE$. L'évaluation de ces mesures est présentée relativement à la croissance du nombre de classes pour le point de vue **Avantages**.

La figure 4.10 montre que $MCRE$ et $MCPE$ se comportent d'une manière antagoniste. Le $MCRE$ augmente et le $MCPE$ diminue si le nombre de classes augmente.

Le modèle optimal a été trouvé expérimentalement à 49 classes. En effet, ce modèle a été considéré comme le meilleur compromis entre le nombre des classes et la qualité de ces dernières.

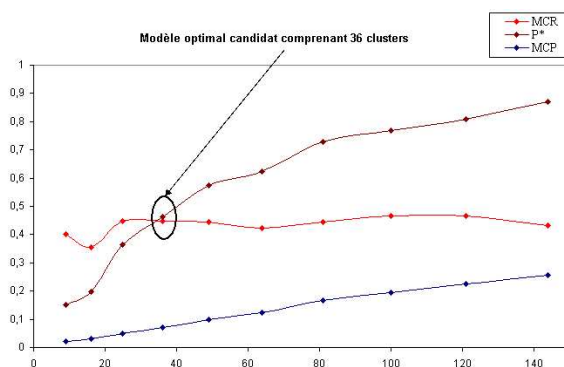


FIG. 4.11 – Mesures de MCR, de P^* et de MCP.

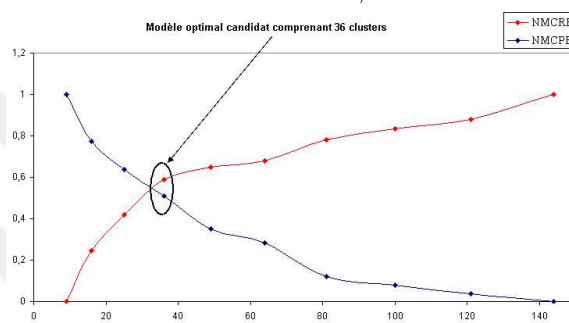


FIG. 4.12 – Mesures de MCRE et de MCPE : Approche basée sur la normalisation de min-max.

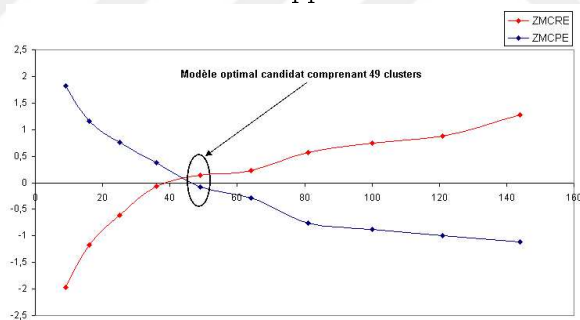


FIG. 4.13 – Mesures de MCRE et de MCPE : Approche basée sur la normalisation de z-score.

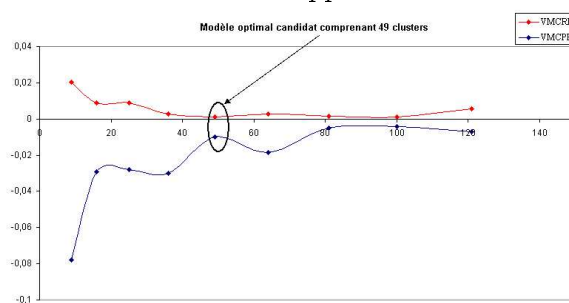


FIG. 4.14 – Mesures de MCRE et de MCPE : Approche basée sur la dérivation.

Même si ce modèle ne peut pas être trouvé d'une manière automatique par l'analyse des courbes d'inertie, on peut remarquer qu'il représente également une étape stable dans l'évolution de ces courbes (cf. figure 4.9).

La figure 4.11 illustre le cas où les points d'intersection sont utilisés entre MCR et P . Dans ce cas, le choix du modèle optimal est difficile quand on applique l'algorithme heuristique 6 décrit à la section 6. En effet, plusieurs paramètres, tels que le seuil entre MCR et P , doivent être fixés en tenant compte du nombre de documents à classer et de la taille de l'espace de description. Même après ces étapes spécifiques, seul un modèle sous-optimal de 36 classes peut être trouvé pour le point de vue **Avantages**.

La figure 4.12 et la figure 4.13 illustrent la détermination du modèle optimal en trouvant le modèle qui est le plus proche au point d'équilibre entre les courbes normalisées de $MCRE$ et de $MCPE$. On peut voir que le résultat diffère légèrement selon la méthode de normalisation choisie. Ainsi, le modèle sous-optimal à 36 classes est trouvé en tant qu'un modèle optimal dans le cas de la méthode de la normalisation min-max (cf. figure 4.12), et le modèle optimal à 49 classes est trouvé dans le cas de la méthode de normalisation de z-score (cf. figure 4.13).

Ces résultats illustrent le fait que les approches basées sur la normalisation peuvent être sensibles au nombre de modèles qui sont utilisés pendant le processus d'évaluation. Ainsi, d'avantage de modèles auraient été nécessaires pour trouver le modèle optimal dans le cas de l'approche de la normalisation Min-Max.

La figure 4.14 illustrent la détermination du modèle optimal en utilisant la première co-occurrence du premier minimum local de la courbe de $VMCRE$ et du premier maximum local de la courbe de $VMCPE$. Cette méthode peut être considérée comme la plus robuste. En effet, elle met directement en évidence le modèle optimal à 49 classes.

Expérimentation pour le critère d'arrêt

Notre deuxième expérimentation consiste en calculant ces mesures ($MCRE$ et $MCPE$) avec l'inertie intra-classe et l'inertie inter-classes, parce que chaque cycle d'apprentissage de NG (chaque cycle d'apprentissage correspond à un balayage complet de toutes les données) avec un nombre fixé de classes. Le but de cette expérimentation est de mettre en évidence les capacités de nos mesures de qualité à être utilisées en tant que critères d'arrêt pendant la phase d'apprentissage d'un modèle de classification.

Les résultats décrits dans les figures (4.17, 4.16, 4.15, 4.18) montrent l'évolution des différentes mesures relativement à la croissance des cycles d'apprentissage dans le cas d'un seul modèle de classification pour le point de vue **Avantages**. Le modèle choisi qui est utilisé dans cette expérimentation est le modèle optimal à 49 classes.

La figure (4.18) montre que le comportement des courbes d'inertie illustre mal l'information liée à l'évolution des cycles d'apprentissage, dès lors que l'inertie intra-classe diminue lentement et l'inertie inter-classes augmente lentement en considérant la croissance des cycles d'apprentissage. De plus, les deux courbes ne présentent pas de variations locales significatives pendant l'évolution de l'apprentissage.

D'autre part, les figures (4.15, 4.16, 4.17) montrent que les courbes de $MCRE$ et de $MCPE$ donnent une vue plus précise de l'évolution de l'apprentissage que les courbes d'inertie. Ainsi, l'examen de ces courbes montrent que la stabilité de l'apprentissage est atteinte après 36 cycles. L'examen des courbes d'inertie aurait amené à conclure que 91 cycles sont obligatoires pour atteindre cette stabilité même si les cycles d'apprentissage, complémentaires aux 36 premiers

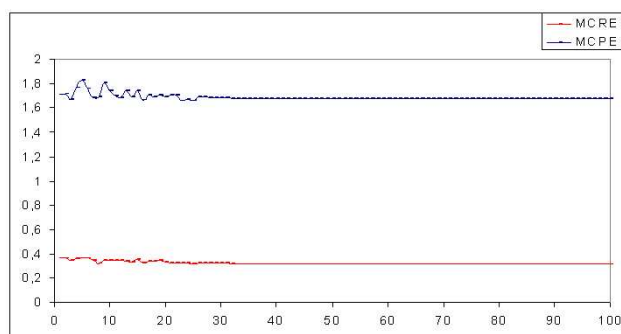


FIG. 4.15 – Mesures de MCRE et de MCPE.

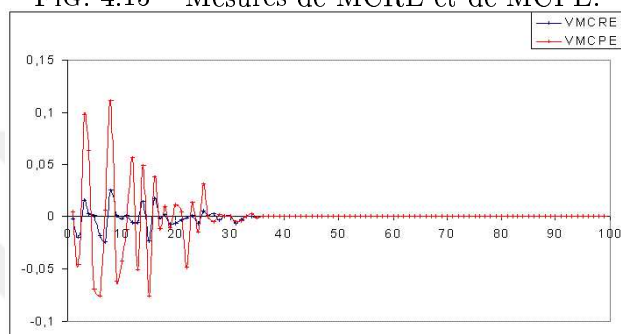


FIG. 4.16 – Mesures de MCRE et de MCPE : Approche basée sur la dérivation.

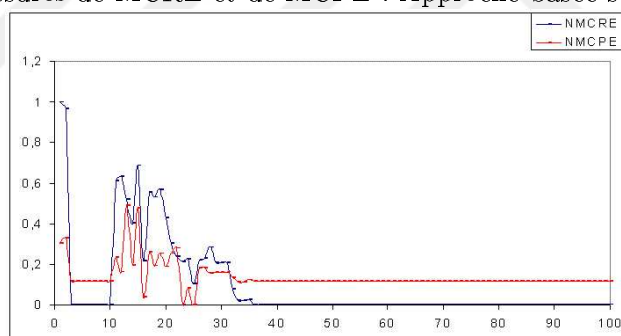


FIG. 4.17 – Mesures de MCRE et de MCPE : Approche basée sur la normalisation de min-max.

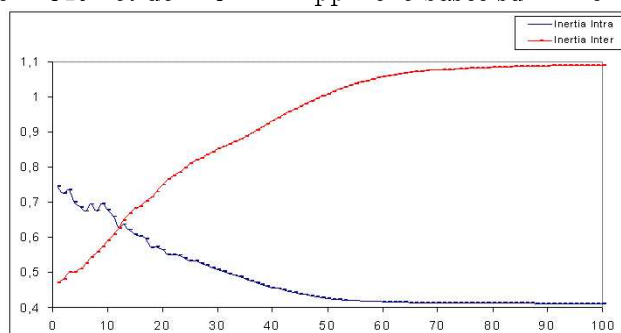


FIG. 4.18 – Mesures d'inertie intra-classe et inter-classes.

cycles considérés précédemment, ne s'avèrent en définitive pas nécessaires pour réaliser l'apprentissage. Cette expérimentation met en évidence que la stabilité conjointe des mesures *MCRE* et *MCPE* peut être utilisée comme un critère fiable d'arrêt d'apprentissage. En effet, cette stabilité exprime l'absence de changement dans la distribution des étiquettes dans les classes.

Expérimentation pour l'analyse des étiquettes

Cette expérimentation et les deux expérimentations suivantes consistent à générer une classification de 36 neurones (classes) pour le point de vue **Avantages** en utilisant la méthode de classification de NG basée sur des vecteurs de description pondérés des documents.

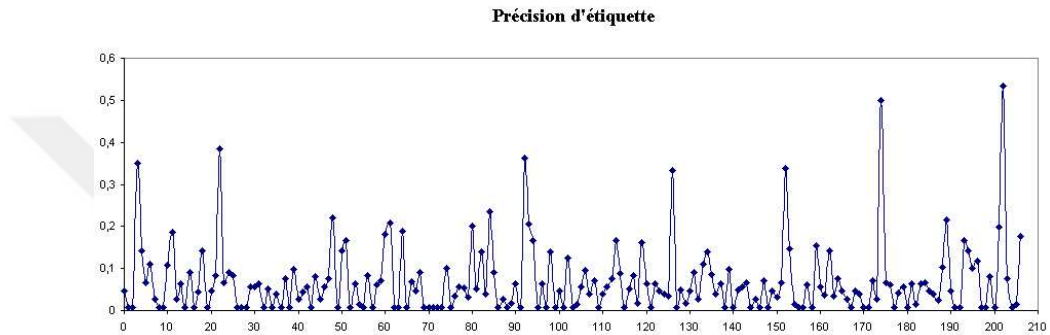


FIG. 4.19 – Évaluation de P_{l_k} pour chaque étiquette l_k dans le modèle de classification NG.

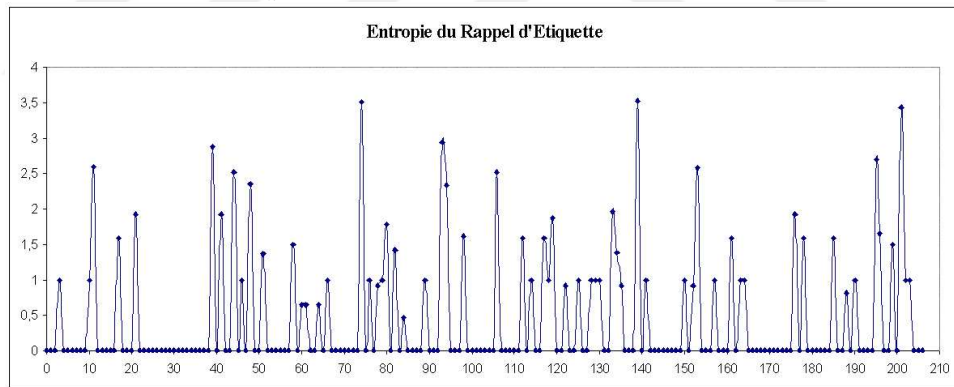


FIG. 4.20 – Évaluation de H_{l_k} pour chaque étiquette l_k dans le modèle de classification NG.

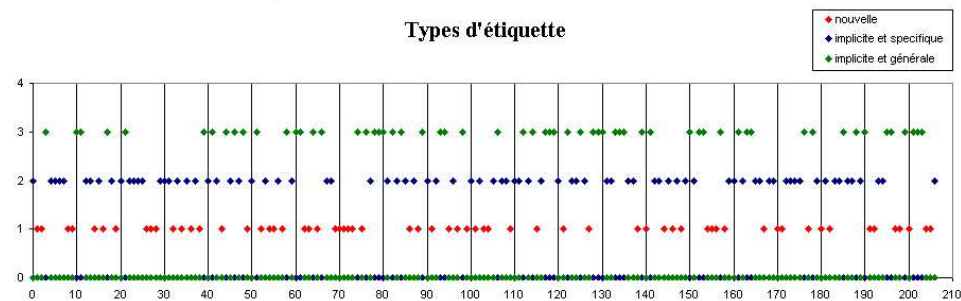


FIG. 4.21 – Caractérisation de la catégorie de chaque étiquette l_k dans le modèle de classification NG.

Pour chaque étiquette de ce modèle, les valeurs de H_{l_k} et de P_{l_k} sont calculées (cf. figures 4.19, 4.20 et 4.21). Pour chaque classe de ce modèle de classification, les valeurs de H_{c_l} et de R_{c_l} sont également calculées (cf. figures 4.22 et 4.23).

Les résultats présentés aux figures 4.19 et 4.20 illustrent la distribution des différentes mesures (P_{l_k}, H_{l_k}) parmi les étiquettes. Ces mesures peuvent être utilisées pour classer les étiquettes dans des catégories spécifiques, à savoir, implicite, explicite, bruitée et nouvelle et dans des catégories recouvrantes plus générales, à savoir, pertinentes et non pertinentes (cf. figure 4.21).

- La figure 4.19 présente les valeurs de la précision d'étiquette qui définit les degrés de représentativité des étiquettes. Par exemple, l'ensemble $\{l_{202}, l_{174}, l_{22}, l_{92}, l_3, l_{152}, l_{126}\}$ correspond aux étiquettes dont les valeurs de précision sont les plus élevées, donc aux étiquettes les plus représentatives dans le modèle. Inversement, l'ensemble $\{l_1, l_2, l_8, l_9, l_{14}, l_{16}, l_{19}\}$ correspond aux valeurs les plus basses de précision, donc aux étiquettes les moins représentatives dans le modèle.
- La figure 4.20 présente les valeurs de l'entropie de rappel d'étiquette qui définit la distribution des étiquettes dans les classes. Par exemple, l'ensemble $\{l_{139}, l_{74}, l_{201}, l_{93}, l_{39}, l_{195}, l_{11}\}$ correspond aux étiquettes avec des valeurs d'entropie de rappel les plus élevées, donc aux étiquettes appartenant à plusieurs classes. Inversement, l'ensemble $\{l_1, l_2, l_8, l_9, l_{14}, l_{16}, l_{19}\}$ correspond aux étiquettes avec des valeurs nulles d'entropie de rappel, donc aux étiquettes appartenant à une seule classe.
- La figure 4.21 résume les résultats de notre analyse d'étiquettes décrite dans la section 4.3.6. Cette figure met en évidence l'existence de seulement deux types d'étiquettes, à savoir, des étiquettes implicites et des étiquettes nouvelles. Donc, selon nos mesures, aucune étiquette bruitée ni explicite n'a été trouvée dans cet ensemble de données. Comme mentionné avant, les étiquettes implicites peuvent elles-mêmes être divisées en deux sous-groupes : *implicite spécifique* et *implicite générale*. Quand $H_{l_k} = 0$, l'étiquette l_k représente soit une étiquette nouvelle (cf. définition 11) ou une étiquette implicite et spécifique selon la valeur de la précision P_{l_k} . Par exemple, l'ensemble $\{l_{174}, l_{22}, l_{92}, l_{152}\}$ représente les étiquettes implicites et spécifiques, alors que l'ensemble $\{l_1, l_2, l_8, l_9, l_{14}, l_{16}, l_{19}\}$ représente les étiquettes nouvelles.

Les résultats de l'analyse précédente sont résumés dans (le tableau 4.3.7) et les différents taux sont donnés ci-dessous :

- taux d'étiquettes nouvelles = 31%
- taux d'étiquettes implicites = 69%
- taux d'étiquettes bruitées = 0%
- taux d'étiquettes explicites = 0%
- taux d'étiquettes non pertinentes (explicites et bruitées) = 0%
- taux d'étiquettes pertinentes (nouvelles et implicites) = 100%

Ces résultats prouvent que la stratégie d'étiquetage se basant sur les termes de hautes fréquences est une stratégie pertinente et précise pour cet ensemble de données. En effet, elle ne crée aucune étiquette bruitée ou explicite.

Type d'étiquette	étiquettes pertinentes	étiquettes non pertinentes
étiquette nouvelle	64	-
étiquette implicite générale	82	-
étiquette implicite spécifique	61	-
étiquette bruitée	-	0
étiquette explicite	-	0

TAB. 4.5 – Résultats globaux de processus de catégorisation des étiquettes.

Expérimentation pour l'analyse de classes

Les résultats décrits dans les figures 4.22 et 4.23 montrent l'évolution des différentes mesures (R_{c_l} , H_{c_l}) dans le but de répartir les classes en deux catégories, à savoir pertinentes et non pertinentes.

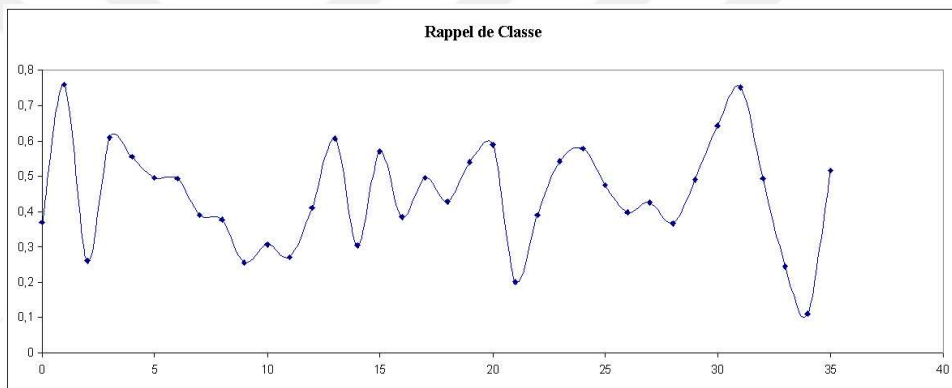


FIG. 4.22 – Évaluation de R_{c_l} pour chaque classe c_l dans le modèle de classification NG.

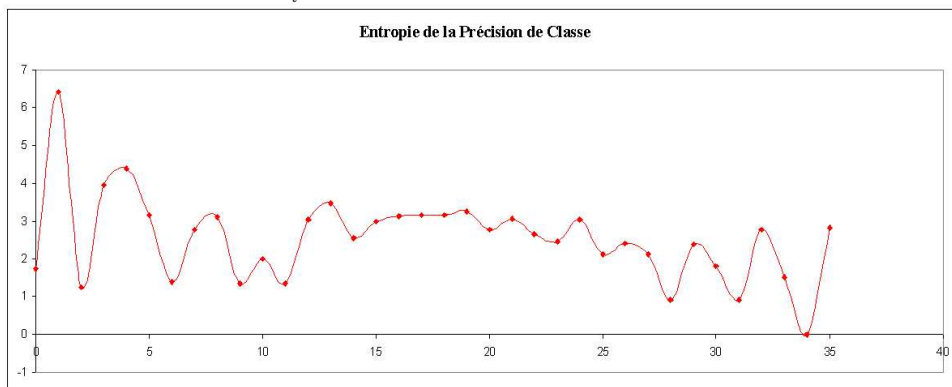


FIG. 4.23 – Évaluation de H_{c_l} pour chaque classe c_l dans le modèle de classification NG.

- La figure 4.22 présente les valeurs du rappel de classe qui définissent les degrés de représentativité de l'ensemble d'étiquettes dans les classes. Par exemple, la classe C_{34} rapporte une mauvaise représentation de ses étiquettes car les fréquences des étiquettes dans cette

classe sont basses. Inversement, la classe C_1 rapporte une bonne représentation de ses étiquettes car les fréquences de ces étiquettes sont hautes.

- La figure 4.23 présente les valeurs de l'entropie de précision de classe qui rapporte des informations sur le nombre d'étiquettes dans la classe. Par exemple, la classe C_{34} contient seulement une étiquette. Inversement, la classe C_1 contient de nombreuses étiquettes.

Selon notre approche d'analyse de classe décrite à la section 4.3.6 en considérant la figure 4.22 et la figure 4.23, les classes peuvent être ordonnées par degrés de pertinence. Par exemple, la classe C_{34} est une classe non pertinente parce qu'elle contient seulement une étiquette avec une faible représentation. Inversement, la classe C_1 contenant beaucoup d'étiquettes avec une représentation élevée et la classe C_{31} contenant quelques étiquettes avec une représentation élevée sont des classes pertinentes.

Finalement, deux taux différents peuvent être calculés afin d'obtenir des informations globales sur la pertinence des classes selon la définition 17 :

- taux de classes pertinentes = 97,2%
- taux de classes non pertinentes = 2,8%

Ces résultats prouvent que ce modèle de classification représente exactement la distribution des données avec des étiquettes multiples.

Expérimentation de l'analyse hybride

Ce type d'analyse nous permet d'étudier la distribution des différents types d'étiquettes selon chaque classe, ainsi que la distribution des classes selon chaque type d'étiquette.

La figure 4.24 montre la distribution des étiquettes nouvelles, des étiquettes implicites et générales et des étiquettes implicites et spécifiques. Le tableau 4.3.7 donne, quant à lui, le nombre de classes par type d'étiquette. Le croisement de ces deux informations permet de déterminer quel sous-groupe de classes choisir pour étudier des types spécifiques d'étiquettes. Par exemple, si un utilisateur est spécifiquement intéressé par des étiquettes nouvelles, il devra seulement étudier la classe C_1 .

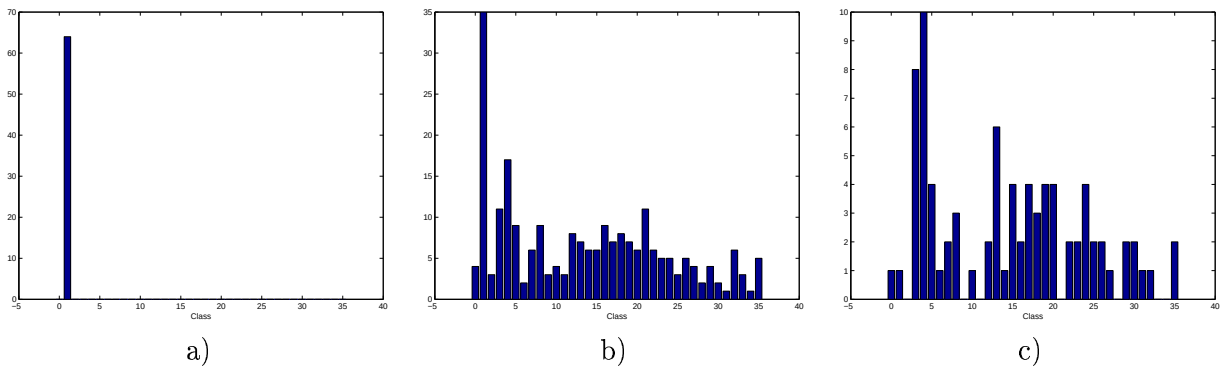


FIG. 4.24 – a) Distribution des étiquettes nouvelles. b) Distribution des étiquettes implicites générales (à l'exclusion des étiquettes implicites spécifiques). c) Distribution des étiquettes implicites spécifiques.

Types d'étiquette	Nombre de classes
nouvelle	1
spécifique implicite	29
générale implicite	36

TAB. 4.6 – Distribution des types d'étiquettes selon les classes

4.4 Conclusion

Nous avons proposé dans la première partie de ce chapitre de nouvelles mesures pour évaluer la qualité des classifications numériques. Ces mesures nous ont permis de mettre en évidence l'analogie existante entre les modèles numériques et les modèles de classification symbolique de type treillis de Galois. Elles présentent l'avantage majeur, contrairement aux mesures classiques d'inertie, ou assimilées, d'être indépendantes de la méthode de classification utilisée. Ces mesures ont donc un large champ d'application. Ainsi, elles permettent à la fois de comparer objectivement différentes méthodes de classification, mais également de comparer différents modes opératoires relatifs à une même méthode. Nous montrerons également par la suite qu'elle permettent de mener à bien des extractions de connaissances à partir d'un modèle de classification numérique (cf. section 5.5).

Dans le cadre de ce chapitre, nous avons mis à profit l'utilisation de nos mesures de nombreuses manières différentes. Premièrement, nous les avons utilisées pour comparer l'approche de classification orientée par les points de vue à une approche de classification globale, ce qui nous a permis de démontrer objectivement la supériorité de l'approche orientée par les points de vue, qui tend à réduire le bruit produit par un processus global de classification. Une deuxième expérimentation nous a permis de démontrer la supériorité, en terme de précision de résultats, d'une méthode de classification non recouvrante, telle que SOM, relativement à une méthode de classification recouvrante, telle que KMA. Une troisième expérimentation nous a permis de montrer comment ces mesures pouvaient être utilisées pour la sélection d'un modèle de classification optimal, en cherchant le point d'équilibre entre le Rappel et la Précision. Nous avons également présenté les résultats pratiques de la recherche d'un tel modèle à partir de l'analyse d'une série de cartes SOM. Cette même expérimentation nous a de plus permis de démontrer la supériorité de la méthode NG, en terme de pouvoir de synthèse, par rapport aux autres méthodes topographiques telles que GNG et SOM. Enfin, une dernière expérimentation nous a permis de démontrer le comportement plus homogène de la méthode originale NG par rapport à une de ses variantes, à savoir la méthode NG-normalisé.

Dans la seconde partie de ce chapitre, nous avons adressé le problème de classification des données associées aux étiquettes multiples. Ce type de problème, qui représente une généralisation du problème de l'évaluation des classifications, peut être considéré comme un nouveau champ fertile de recherche. Dans ce cadre, nous avons proposé une nouvelle approche qui se base sur la théorie des probabilités. Cette approche représente une généralisation de l'approche basée sur le Rappel et la Précision. Elle utilise des mesures de similarité entre les données classées par une méthode de classification et la distribution des étiquettes après la projection de ces dernières sur le modèle de classification.

Comme les premières mesures de Rappel et de Précision que nous avons proposées, ces nouvelles mesures s'avèrent utiles pour l'évaluation de la qualité de la classification et pour la sélection du modèle optimal, sans, dans ce dernier cas, en avoir les défauts concernant la

nécessité du choix de seuils. De plus, elles permettent de résoudre un autre des problèmes centraux de la classification non supervisée, à savoir le choix du critère d'arrêt de l'apprentissage. Enfin, elles ouvrent de nouvelles perspectives dans le domaine de la validation et de l'étiquetage des classes, étant donné qu'elles peuvent servir de base à des analyses portant sur les types d'étiquettes et sur les types de classes.

L'ensemble des résultats des expérimentations basées sur les mesures de qualité que nous avons proposé dans ce chapitre, va nous permettre, par la suite, d'opérer des choix stratégiques importants pour la validation et la généralisation du paradigme d'analyse basée sur les points de vue.



Algorithme 6 Algorithme de recherche du modèle de classification optimal

Entrée : Etant donné S une séquence d'index de classification. Soit $Suc(S, i)$ la fonction qui associe à un élément i , son successeur dans S . Soit $Pre(S, i)$ la fonction qui associe à un élément i , son prédécesseur dans S . S_0 est le premier élément d'une séquence S vérifiant $Pre(S_0, i) = \phi$. S_* est le dernier élément d'une séquence S vérifiant $Suc(S_*, i) = \phi$. Les fonctions $R(i)$, $P(i)$ et $F(i)$ calculent respectivement le **Rappel**, la **Précision** et la **F-mesure** d'une classification i .

Étant donné S' la séquence des index de classification vérifiant : $\forall i \in S', NbClass(Suc(i, S')) = NbClass(i) + 2i + 1$

Etant donné S'' la séquence des index de classification vérifiant : $\forall j \in S'', j \in S'$ and $R(j) \geq R(Suc(j, S''))$.

Trouver=Faux, $k = S_0$.

Sortie : Le modèle de classification optimal k_f .

tant que ($k \neq \phi$) **faire** /* rechercher le point d'équilibre entre R et P */

si $|R(k) - P(k)| \leq seuil_1$ **alors**

 Trouver=Vrai, $k_0 = k$, $k_f = k_0$, **Sortir** tant que

fin si

$k = Suc(S'', k)$,

fin tant que

si Trouver=Vrai **alors**

si $P(Suc(S', k_0)) > P(k_0)$ **alors** /* Phase de raffinement ascendant (pour améliorer le résultat global) */

$Max_F = F(k_0)$, $R_0 = R(k_0)$,

$k = Suc(S', k_0)$,

tant que $k \neq \phi$ **faire**

si $F(k) > Max_F$ **alors**

si $R(k) - R_0 \geq seuil_2$ **alors**

$Max_F = F(k)$, $k_f = k$

sinon **Sortir** tant que,

$k = Suc(S', k)$

fin tant que

sinon /* Phase descendante d'amélioration (pour obtenir un nombre de classes inférieur) */

 */

$F_0 = F(k_0)$, $Min_D = |R(k_0) - P(k_0)|$,

$k = Pre(S', k_0)$

tant que $k \neq \phi$ **faire**

si $|R(k) - P(k)| \leq Min_D$ et $F(k) - F_0 \leq seuil_2$ **alors**

$Min_D = |R(k) - P(k)|$, $k_f = k$

fin si

$k = Pre(S', k)$

fin tant que

fin sinon

sinon /* Résultat dégénéré : aucun point d'équilibre trouvé */

$k = S'_0$, $Min_D = |R(k_0) - P(k_0)|$

$k = Suc(S', k)$

tant que $k \neq \phi$ **faire**

si $|R(k) - P(k)| \leq Min_D$ **alors**

$Min_D = |R(k) - P(k)|$, $k_f = k$

118 **fin si**

$k = Suc(S', k)$

fin tant que

fin sinon

Chapitre 5

Généralisation du modèle multi-topographique : application à l'extraction de connaissances

Sommaire

5.1	Introduction	119
5.2	Mécanisme de communication	120
5.2.1	Modèle de réseau bayésien pour la communication inter-cartes	121
5.2.2	Comparaison entre le mode probabiliste et le modèle bayésien	123
5.3	Mécanisme de comparaison des vues	125
5.3.1	Évaluation de la cohérence de propagation	125
5.3.2	Expérimentation webométrique de comparaison de vues	127
5.4	Mécanisme de généralisation	130
5.4.1	Mécanisme de généralisation de Neural Gas	130
5.4.2	Évaluation du mécanisme de généralisation	131
5.4.3	Expérimentation	132
5.5	Extraction des règles d'association par des modèles neuro- naux non supervisés	136
5.5.1	Critères principaux	137
5.5.2	Extraction par le mécanisme de généralisation	137
5.5.3	Extraction par le mécanisme de communication	138
5.5.4	Expérimentation	139
5.6	Conclusion	146

5.1 Introduction

Les expérimentations que nous avons menées au chapitre précédent, à l'aide de mesures originales, nous ont permis de démontrer objectivement le fort potentiel de l'analyse orientée par les points de vue en termes de capacité d'interaction pour l'analyse et pour la fouille de données et de précision globale des résultats obtenus. Le but de ce nouveau chapitre est donc globalement celui de proposer des solutions permettant d'augmenter encore le potentiel et la portée de l'analyse orientée par les points de vue, en la rendant notamment plus générique.

En tenant compte des remarques que nous avons faites sur le modèle MultiSOM au chapitre 2 (cf. section 2.3.7), un des premiers objectifs des travaux de ce chapitre est celui de réviser les mécanismes de communication entre les points de vue associés à ce modèle de manière à les rendre plus génériques. L'approche que nous proposons consiste à établir sur l'analogie du mode de communication inter-vues avec un modèle de raisonnement bayésien (cf. section 1.3.2).

Afin d'augmenter la qualité de l'analyse locale orientée par les points de vue, l'approche que nous proposons est d'utiliser le modèle Neural Gas comme une nouvelle base pour étendre le modèle MultiSOM à un modèle MultiGAS [Al Shehabi and Lamirel, 2005b]. En effet, nous avons montré au chapitre précédent que le modèle NG s'avérait nettement plus efficace que le modèle SOM pour les tâches de classification où une visualisation des résultats de l'analyse de données n'est pas exigée. L'extension à un tel modèle de l'approche multi-vues passe cependant par la mise en place d'un nouveau mécanisme de généralisation qui s'affranchit de la contrainte de topographie fixe inhérente au modèle SOM. C'est exactement ce que nous proposons dans la deuxième partie de ce chapitre.

En parallèle, nous proposons une nouvelle mesure permettant d'évaluer la cohérence de propagation entre les topographies. Ce type de mesure est à la source de nouveaux mécanismes inédits et très puissants d'automatisation de la comparaison des vues. Ces mécanismes peuvent avoir de nombreuses applications originales. Nous le montrons dans la troisième partie de ce chapitre en proposant une expérimentation webométrique qui peut être considérée comme une première tentative pour effectuer une identification précise et automatique du comportement de liens dans un domaine scientifique par l'analyse de la communication des institutions académiques relatives à un pays sur le web.

L'ensemble de ces extensions étant applicables sur n'importe quel type de méthode de classification, elle nous permettent donc de définir un nouveau paradigme générique d'analyse de données orientée par les points de vue.

Pour compléter ce paradigme, et l'étendre à la fouille de données, nous proposons dans une quatrième partie de ce chapitre, d'y intégrer une nouvelle approche pour l'extraction de connaissances à partir de vues multiples [Al Shehabi and Lamirel, 2005a; Lamirel and Al Shehabi, 2005]. Cette consiste à utiliser le modèle multi-topographique pour l'extraction non supervisée des règles d'association. Un tel modèle représente en effet un candidat naturel pour régler les problèmes liés à l'inflation de règles et au choix des règles et à la complexité d'extraction qui sont inhérents aux méthodes symboliques, nous montrons en effet que ses capacités de synthèse peuvent être utilisées pour réduire le nombre de règles et pour extraire les plus significatives.

Il reste à noter que, nous utilisons deux corpus de référence dans nos expérimentations, à savoir un corpus de notices de brevets et un corpus de données web.

5.2 Mécanisme de communication

Dans cette section nous proposons une approche originale consiste en une modification du comportement de la communication inter-cartes de MultiSOM basée sur l'analogie avec un modèle de propagation bayésien [Al Shehabi and Lamirel, 2004a; Lamirel *et al.*, 2004b]. Cette approche est utilisée à la fois pour valider et généraliser les principes de communication inter-cartes du modèle MultiSOM et pour perfectionner le mode probabiliste de calcul de corrélation qui est déjà fourni par le modèle (cf. section 2.3.7). D'une manière générale, cette démarche nous a également amené à montrer qu'un modèle de classification multi-vues non supervisée peut se comporter comme un réseau bayésien d'inférence dans lequel l'évaluation des probabilités a

posteriori devient un processus simple, utilisant uniquement des mesures de similarité a priori. Le modèle de propagation bayésien que nous allons proposer est donc un modèle suffisamment générique pour que l'on puisse envisager de l'appliquer directement à n'importe quel type de méthode de classification non supervisée.

5.2.1 Modèle de réseau bayésien pour la communication inter-cartes

Le réseau d'inférence bayésien auquel le modèle de MultiSOM peut être assimilé est constitué de trois couches. Dans ce réseau les neurones peuvent être considérés comme des nœuds. La première couche du réseau contient M nœuds (neurones) associés à la carte source S . Ces nœuds représentent eux-mêmes les sources potentielles de l'activité initiale. La deuxième couche contient N nœuds intermédiaires D (c'est-à-dire, nœuds-données) qui sont associés aux nœuds de la carte source. La troisième couche contient L nœuds de la carte destination T qui peuvent être activés par les nœuds intermédiaires D qui sont associés à la topographie⁹ source [Al Shehabi and Lamirel, 2004a; Lamirel *et al.*, 2004b; Lamirel *et al.*, 2004c].

La structure du réseau peut être déterminée en considérant les directives suivantes (cf. figure 5.1) :

1. Les relations entre les nœuds classes dans les topographies source et destination se produisent seulement par les données inclus dans ces classes.
2. Les liens joignant des classes et des données doivent être dirigés des nœuds-classes source aux nœuds-données et des nœuds-données aux nœuds-classes destination.

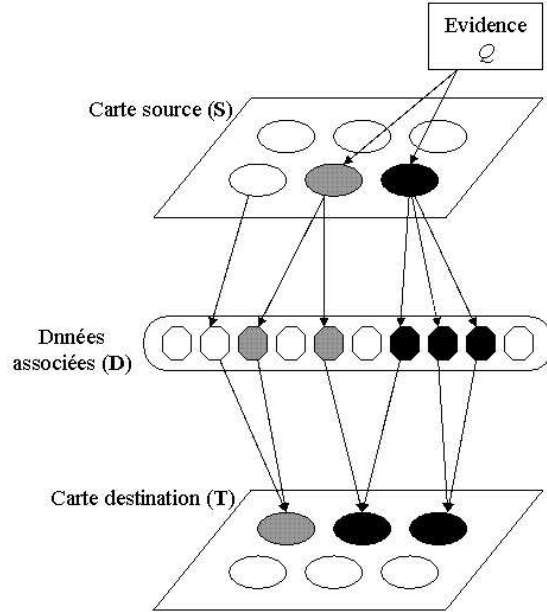


FIG. 5.1 – La structure de réseau bayésien pour la communication inter-topographies.

En tenant compte de la stratégie de la construction de carte (cf. section 2.3.7), chaque nœud-donnée $d \in D$ ne peut avoir qu'un seul parent dans la carte source, soit :

⁹Nous employons le terme topographie pour désigner, soit une carte, soit un gaz. Le mot gaz représente une classification effectuée par la méthode neural gas

$$Pa(d) = S_i \in S \quad ; \quad d \in S_i \quad (5.1)$$

La probabilité d'une classe S_i de la carte source, considérant l'évidence Q , peut donc elle-même être estimée comme :

$$P(S_i|Q) = \frac{1}{M} \quad (5.2)$$

L'ensemble des parents d'un nœud-classe T_j destination, est représenté par l'ensemble de nœuds-données qui appartiennent à T_j , c'est-à-dire, $Pa(T_j) = \{d \in D | d \in T_j\}$.

Dans le modèle original de MultiSOM, il n'existe pas de liens entre les nœuds-données ou entre les nœuds-classes, ce qui revient à dire que les nœuds-classes source peuvent être considérés comme les nœuds-racines.

L'algorithme de propagation de réseau d'inférence bayésien standard est utilisé pour calculer les probabilités a posteriori de nœuds-classes T_j , de la topographie destination qui ont hérité de l'activité (l'évidence Q) transmise par ses nœuds-données associés D . Ces calculs peuvent être effectués d'une manière efficace grâce à la topologie de réseau d'inférence bayésien spécifique qui peut être associée au modèle MultiSOM. Donc, il est possible de calculer la probabilité $P(act_m|T_j, Q)$ pour une activité de modalité act_m sur la classe T_j qui est héritée des activités générées sur la topographie source. Ce calcul est réalisé comme suit :

$$P(act_m|T_j, Q) = \frac{P(act_m \cap T_j|Q)}{P(T_j|Q)} \quad \text{Bayes} \quad (5.3)$$

$$P(act_m \cap T_j|Q) = \sum_{d \in act_m \cap T_j} P(d|Q) \quad (5.4)$$

$$P(T_j|Q) = \sum_{d \in T_j} P(d|Q) \quad (5.5)$$

$$P(d|Q) = \sum_{i=1}^M P(d|S_i)P(S_i|Q) \quad (5.6)$$

de l'équations 5.1 et l'équation 5.6, nous obtenons :

$$P(D|Q) = P(D|S_D)P(S_D|Q) \quad (5.7)$$

de l'équation 5.4, l'équation 5.5 et l'équation 5.7, nous obtenons :

$$P(act_m|T_j, Q) = \frac{\sum_{d \in act_m, T_j} P(d|S_i).P(S_i|Q)}{\sum_{d \in T_j} P(d|S_i).P(S_i|Q)} \quad (5.8)$$

$$P(d|S_i) = \frac{P(d \cap S_i)}{P(S_i)} \quad \text{Bayes} \quad (5.9)$$

A cette étape, étant donné que la probabilité $P(d \cap S_i)$ peut être considérée comme une mesure de corrélation entre les d et S_i , nous faisons donc l'hypothèse qu'elle peut être assimilée à la mesure de corrélation cosinus $Sim(d, S_i)$ (cf. équation 1.5). En effet, la mesure de corrélation cosinus prend ses valeurs dans l'intervalle $[0, 1]$. Plus les deux vecteurs associés à d et à S_i sont

(positivement) corrélés, plus cette mesure est proche de 1. Moins ils sont corrélés, plus cette mesure est proche de zéro ¹⁰. Ainsi :

$$P(d \cap S_i) = Sim(d, S_i) \quad (5.10)$$

$$0 \leq Sim(d, S_i) \leq 1$$

en combinant l'équation 5.8 avec l'équation 5.10

$$P(act_m|T_j, Q) = \frac{\sum_{d \in act_m, T_j} Sim(d, S_i)}{\sum_{d \in T_j} Sim(d, S_i)} \quad (5.11)$$

Finalement, les probabilités de toutes les activités, correspondant aux r différentes modalités, sur la classe T_j dans la topographie destination peuvent être représentées comme :

$$P(Act|T_j, Q) = arg \max_m \{P(act_m|T_j, Q)\} \quad \text{avec } m \in [1, r] \quad (5.12)$$

L'équation 5.12 permet de choisir quelle modalité a la meilleure probabilité pour la classe T_j dans la classification destination.

5.2.2 Comparaison entre le mode probabiliste et le modèle bayésien

Notre expérimentation consiste à comparer la performance de deux modèles de propagation d'activité, c'est-à-dire, notre modèle bayésien et celui de mode probabiliste original, qui est fourni par MultiSOM, et le modèle de réseau bayésien d'inférence que nous avons défini [Al Shehabi and Lamirel, 2004a; Lamirel *et al.*, 2004b]. Pour ce faire, nous avons à nouveau utilisé notre corpus de test de 1000 notices de brevets (cf. section A.2). Nous avons construit deux cartes SOM différentes qui représentent les classifications de deux points de vue (**Utilisations** et **Avantages**). La carte associée au point de vue **Utilisations** est considérée comme une carte source et la carte associée au point de vue **Avantages** est considérée comme une carte destination. De manière à comparer les deux mesures, le processus de test consiste, pour chacune d'entre elles, à lancer séparément sur toutes les classes de la carte source une activité correspondant à une modalité de choix catégorique, puis, à observer l'activité transmise aux classes de la carte destination, ceci par l'intermédiaire des valeurs de probabilité a posteriori des classes activées sur cette dernière carte. Les valeurs de probabilité a posteriori obtenues sont ensuite mises en correspondance avec le pourcentage de données activées dans ces dernières classes. Les résultats de ce processus sont présentés à la figure 5.2, pour la mise en correspondance probabiliste, et à la figure 5.3, pour la mise en correspondance bayésienne.

En observant ces résultats, nous pouvons conclure que le modèle bayésien est plus linéaire que le mode probabiliste. Ce qui revient à dire que pour le même nombre de données activées la déviation de la probabilité a posteriori est plus faible dans le cas de modèle bayésien (cf. figure 5.3) que dans celui du mode probabiliste (cf. figure 5.2). Ce qui signifie également que, dans le cas du modèle bayésien, les résultats sont les plus similaires entre eux, pour une proportion de

¹⁰ A notre avis, cette hypothèse est à la fois plus générale et plus précise que celle qui est choisie par [Ribeiro and Muntz, 1996] (cf. section 1.3.2). Dans cette dernière approche, la mesure de corrélation de cosinus est utilisée pour représenter la probabilité a posteriori d'un document relativement à une requête donnée $P(\text{document}|\text{requête})$.

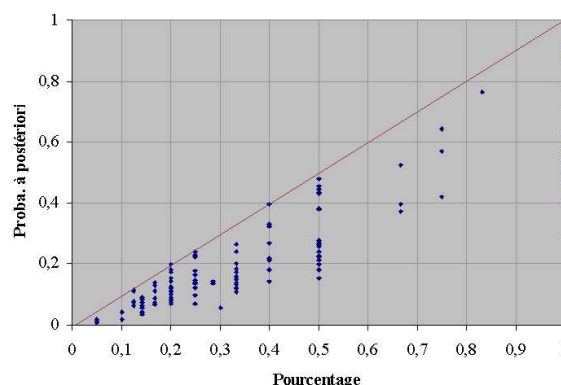


FIG. 5.2 – La figure illustre le comportement de mode probabiliste basé sur la relation entre le pourcentage de données activées dans les classes destination et la mesure de probabilité a posteriori générée par le processus de propagation.

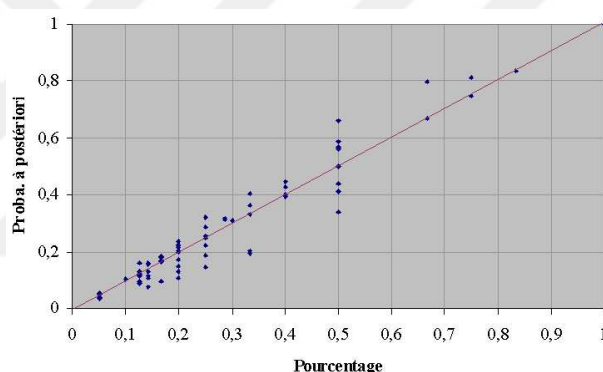


FIG. 5.3 – La figure illustre le comportement du modèle de réseau bayésien d'inférence basé sur la relation entre le pourcentage de données activées dans les classes destination et la mesure de probabilité a posteriori générée par le processus de propagation.

données actives fixées, et que la proportion de données actives dans les classes de la carte destination est plus conforme avec la probabilité de l'activité de ces classes. Ce dernier phénomène est particulièrement évident pour des valeurs de probabilité très faibles et très élevées.

Ces résultats nous amènent à retenir le modèle bayésien comme le modèle de propagation de référence dans la suite de notre travail. De manière plus générale, ce modèle permet de transformer un modèle de classification global en un modèle basé sur les points de vue. Nous en avons donné un premier exemple en l'appliquant au modèle MultiSOM de façon à optimiser et à généraliser son mode opératoire. Ce modèle permet également, sans aucune transformation, d'étendre le modèle neural gas à un modèle MultiGAS. Ce que nous faisons, de manière implicite, dans la suite de notre travail.

5.3 Mécanisme de comparaison des vues

5.3.1 Évaluation de la cohérence de propagation

En nous basant sur le modèle bayésien décrit précédemment, nous proposons dans cette section une mesure globale dont le but est d'évaluer la cohérence entre des classifications associées à des points de vue différents [Al Shehabi and Lamirel, 2005b]. Nous proposons une définition de cette mesure dans le cadre conjoint des modèles MultiSOM et MultiGAS. Cependant, comme il sera facile de le constater, son principe est suffisamment général pour la rendre applicable à toute autre méthode de classification orientée par les points de vue.

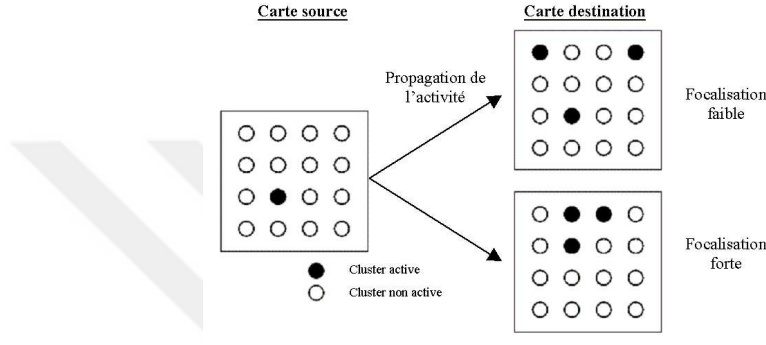


FIG. 5.4 – Types de profil d'activités sur des topographies destination.

La propagation d'activité d'une classe S_k sur une topographie source à un ensemble de classes T^k sur une topographie destination est définie par la fonction :

$$PRG : S_k \rightarrow T^k$$

Cependant, comme le montre la figure 5.4, plusieurs types de profil d'activité peuvent être générés sur la topographie destination. Dans le cas d'un profil focal, les classes activées T^k se trouvent dans les zones proches de la topographie destination. Dans le cas d'un profil diffus, les classes activées T^k se trouvent réparties sur une topographie. Une mesure de focalisation d'activité générée par une topographie source sur une topographie destination doit à la fois tenir compte la distance moyenne¹¹ entre les classes activées T^k , et la probabilité moyenne entre ces dernières. La formule que nous proposons est directement inspirée de la formule d'une force de gravitation. Selon ce principe, la mesure de focalisation d'activité CP_k d'une classe S_k d'une topographie source est inversement proportionnelle au carré de la distance moyenne entre les classes activées T^k sur la topographie destination, Soit :

$$CP_k = \frac{\sum_i P(act|T_{i_k})}{D_k + 1} \quad (5.13)$$

avec,

$$D_k = \frac{2 \sum_i \sum_{j=i+1} \|T_{i_k} - T_{j_k}\|}{|T^k| \cdot (|T^k| - 1)}$$

¹¹Définie selon une métrique dépendante du modèle

où D_k est une distance moyenne entre les classes activées T^k dans la topographie destination, $|T^k|$ est le nombre de classes activées dans la topographie destination.

La cohérence globale de propagation (CP) d'activité générée (act) par l'ensemble des classes de la topographie source sur une topographie destination est alors donnée par :

$$CP = \frac{1}{|\bar{S}|} \sum_{k; S_k \neq \emptyset} CP_k \quad (5.14)$$

où $P(act|T_{i_k})$ est la probabilité d'une activité (act) dans la classe T_{i_k} qui appartient à la topographie destination lors d'une activation de la classe S_k qui appartient à la topographie source.

et,

$$T_{i_k} = \begin{cases} (w_1^{i_k}, \dots, w_n^{i_k}) & \text{NG : la position originale de vecteur de l'espace } \mathbb{R}^n \\ (a^{i_k}, b^{i_k}) & \text{SOM : la coordonné sur la grille} \end{cases}$$

$\bar{S}$ représente l'ensemble particulier de classes extraites à partir de la topographie source de S , qui vérifie :

$$\bar{S} = \{S_k \in S | S_k \neq \emptyset, S_k \in act\}$$

$$T^k = \{T_{i_k} \in T | T_{i_k} \in PRG(S_k)\}$$

Cette mesure présente l'intérêt majeur de permettre d'automatiser la comparaison entre des classifications de différents types, dans différents contextes. Il est à noter qu'elle n'est naturellement pas symétrique sauf dans le cas où les deux classifications associées aux points de vue sont parfaitement identiques. Notre mécanisme de comparaison des vues est donné ci-après (cf. algorithme 7). Nous donnerons dans la suite de ce travail, de nombreux exemples de son utilisation (cf. sections 5.3.2, 5.4.2 et 6.2).

Algorithme 7 Algorithme de l'évaluation de la cohérence de propagation

Entrée : La classification source S , la classification destination T , le type d'activité act .

- 2: **pour chaque** classe source activée $S_k \in S$; $S_k \neq \emptyset$ **faire**
 - 3: /* Calculer la distance moyenne entre les classes activées */
 - 4: Soit $Dist(T^k)$ la somme de distances entre les classes qui appartiennent à T^k ,
 - 5: $D_k = \frac{2 * Dist(T^k)}{|T^k|(|T^k| - 1)}$
 - 6: **pour chaque** classe destination activée $T_{i_k} \in T^k$ **faire**
 - 7: $Sigma(P(act|T^k)) + = Fonction(act, T_{i_k}, S_k)$
 - 8: **fin pour**
 - 9: $CP_k + = \frac{Sigma(P(act|T^k))}{D_k + 1}$
 - 10: **fin pour**
 - 11: $CP = \frac{CP_k}{|S|}$
-

Algorithme 8 Fonction(act, T_{i_k}, S_k)

```

1:  $Sigma(d_{ST}) = Sigma(d_S) = 0$ 
2: pour chaque données  $d \in T_{i_k}, S_k$  faire
3:    $Sigma(d_{ST}) += Sim(d, S_k)$ 
4: fin pour
5: pour chaque donnée  $d \in T_{i_k}, S_d ; S_d \in S$  faire
6:    $Sigma(d_T) += Sim(d, S_d)$ 
7: fin pour
8: retourner ( $\frac{Sigma(d_{ST})}{Sigma(d_T)}$ )

```

5.3.2 Expérimentation webométrique de comparaison de vues

L'étude de l'utilisation des liens entre les sites web académiques pour créer un mode simple de communication scientifique est un nouveau champ prometteur de recherche. Des études préliminaires ont montré que des métriques basées sur le nombre de liens pouvaient être appliquées de nombreuses manières pour quantifier les échanges liées aux activités scientifiques [Wilkinson *et al.*, 2003]. D'autres analyses plus récentes [Chu, 2005] ont cependant montré que, pour assurer leur validité, les études de comportement de référencement ne devaient pas seulement considérer le nombre de liens, mais également la typologie de ces derniers. [Thelwall *et al.*, 2003] ont également prouvé que les comportements de référencement étaient fortement dépendant du domaine étudié. Ces derniers ont notamment démontré que les domaines des mathématiques et de l'informatique avaient des comportements de référencement spécifiques, l'informatique ayant tendance à référencer (liens sortants), et les mathématiques à être référencés (liens entrants), plus fortement que les autres domaines scientifiques. Plusieurs travaux récents sont allés encore plus loin, en démontrant qu'une analyse précise du comportement de liens devait tenir compte de nombreux facteurs complémentaires, tels que, par exemple, la discipline scientifique et la localisation géographique associées aux sites académiques étudiés [Thelwall, 2002]. L'exécution d'un tel type d'analyse à grande échelle proscrit néanmoins l'utilisation du mode manuel, tel que cela est proposé jusqu'à aujourd'hui.

Il est à noter que des méthodes basées sur les graphes [Björneborn and Ingwersen, 2004; Thelwall *et al.*, 2005] ou sur les réseaux de neurones [Dekker, 2001] ont été envisagées pour étudier automatiquement des structures de liens sur le web. Elles ne représentent cependant pas un support fiable pour exécuter une analyse précise, basée sur des facteurs multiples.

L'approche orientée par les points de vue (cf. section 2.3.7), associée au mécanisme de communication entre les vues et au mécanisme de comparaison des vues que nous avons présentés dans ce chapitre, représente une base prometteuse pour résoudre le problème de l'analyse automatique des comportements de référencement. Comme nous l'avons montré, elle peut être à la fois exploitée pour établir des relations entre des données provenant de différents types de source [Lamirel *et al.*, 2004c] et peut également être utilisée pour faire une analyse croisée entre les données appartenant au même ensemble de source, en se basant sur des facteurs multiples.

La méthode MultiSOM qui implémente l'approche que nous venons de décrire a été déjà utilisée avec succès pour comparer indirectement une analyse de contenu et une analyse de liens sortants en opérant sur le même ensemble de référence de pages web extraites d'un site web d'établissement (cf. section 4.2.4). Le rôle de cette étude est de franchir une étape supplémentaire. Elle représente une première tentative d'identification automatisée du comportement de référencement dans un domaine scientifique ciblé en utilisant des facteurs, ou points de

Points de vue	Nombre de classes
Cities	256
UCodes	81
OutLinks	144
InLinks	289

TAB. 5.1 – Nombre optimal de classes pour chaque point de vue.

vue multiples. Le domaine qui a été choisi pour cette étude est le domaine de l'informatique en Allemagne. Dans celle-ci [Francois *et al.*, 2004] le corpus utilisé est l'ensemble des sites du laboratoires de recherche européens, en prenant plus spécifiquement en compte les points de vue **UCodes**, **Cities**, **InLinks** et **OutLinks** (cf. annexe A.3.2).

Optimisation d'analyse de classification

Afin de définir la taille optimale de la carte SOM pour chaque point de vue, nous générons différentes cartes, pour chaque point de vue, de 9 neurones (3×3) jusqu'à 400 neurones (20×20). Le tableau 5.1 présente les résultats finaux de l'algorithme de sélection de la carte optimale (cf. section 4.2.3) et la figure 5.5 présente les valeurs de qualité basées sur le Rappel, la Précision et la F-mesure (cf. section 4.2).

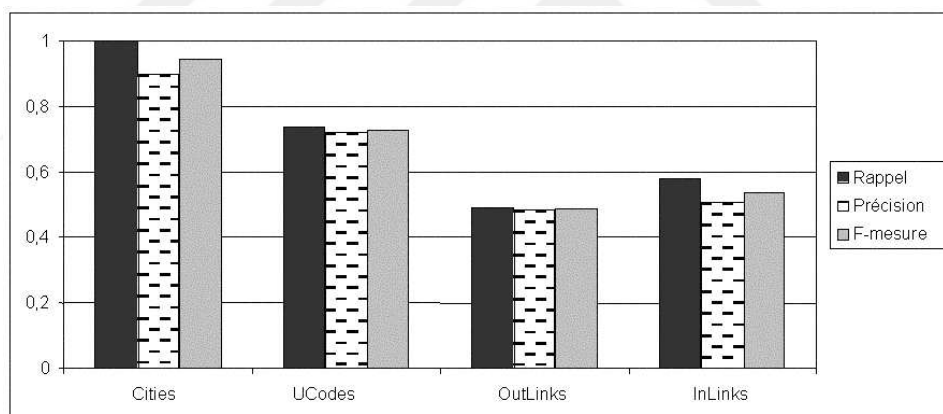


FIG. 5.5 – Qualité de la carte pour chaque point de vue.

La figure 5.5 met en évidence des valeurs de qualité très fortes pour les points de vue **Cities** et **UCodes**, et des valeurs de qualité plus modérées dans le cas des points de vue **OutLinks** et **InLinks**. Dans le cas des points de vue **Cities** et **UCodes**, les classes sont tout à fait homogènes et distinctes entre elles. Cette distribution s'opère facilement du fait que chaque site web est indexé par un faible nombre de propriétés recouvrantes¹². Tant que chaque site web présente un nombre relativement significatif de liens entrants et sortants, les chevauchements sont ainsi potentiellement beaucoup plus significatifs, ceci implique des valeurs relativement modérées de qualité pour les points de vue **OutLinks** et **InLinks**, même après le processus d'optimisation. Ces résultats préliminaires seront pris en considération dans le reste de notre étude.

¹²La propriété est dite recouvrante si elle partagée avec au moins une classe.

$\vec{r}$	Cities	UCodes	OutLinks	Out-Coupl	InLinks	In-Coupl
Cities	1	0.096	0.250	0.196	0.270	0.171
UCodes	0.216	1	0.554	0.359	0.416	0.288
OutLink	0.117	0.095	1	0.231	0.270	0.174
Out-Coupl	0.168	0.140	0.418	1	0.347	0.232
InLinks	0.080	0.055	0.130	0.110	1	0.150
In-Coupl	0.183	0.164	0.374	0.312	0.557	1

TAB. 5.2 – Valeurs de consistance entre les cartes issues de MultiSOM pour différents points de vue.

Évaluation de consistance entre les points de vue différents

La mesure de consistance de propagation (ou cohérence de propagation) (cf. équation 5.14) évalue la focalisation d'activité produite par une carte source sur une carte destination.

Une consistance élevée mesurée d'une carte source vers une carte destination signifie que les classes de la carte destination peuvent être correctement expliquées par les classes de la carte source. Ainsi, si la consistance inverse, mesurée d'une carte destination vers une carte source, est également élevée, ceci peut amener à considérer que les deux cartes sont similaires. Par contre, si la consistance inverse est faible, alors il n'y a pas d'explication claire concernant l'explication inverse fournie par la carte destination sur la carte source.

Donc, à partir de tableau 5.2, présente les valeurs de consistance entre les différents points de vue susmentionnés, nous étudions le comportement de liens dans le domaine webométrique en donnant quelques exemples d'interprétation des résultats de ce tableau ci-après.

Comportement de référencement selon les sous-domaines de recherche. La valeur la plus élevée de la consistance est **UCodes** (i.e. **Sous-domaines** de recherche) vers **OutLinks** (0,554). Celle-ci peut être interprétée comme un indicateur de cohérence du comportement de référencement sortants de laboratoires appartenant aux sous-domaines de recherche communs. La valeur de consistance de **Sous-domaines** vers **InLinks** est significativement élevée (0,416), tandis qu'elle est plus faible que la valeur précédente. Celle-ci indique une bonne identification des sous-domaines de recherche des laboratoires, bien qu'ils n'aient pas le contrôle sur des liens qu'ils reçoivent.

D'une manière inverse, la valeur de consistance d'**OutLinks** vers **Sous-domaines** est particulièrement faible (0,095), et la valeur de consistance d'**InLinks** vers **Sous-domaines** (0,055). Ces valeurs indiquent que le comportement global de référencement des laboratoires de recherche outrepassé en grande partie leurs sous-domaines de recherches.

Distribution géographique des sous-domaines scientifiques. La valeur de consistance de **Sous-domaines** vers **Cities** (ou villes) (0,216) est modérément élevée, indiquant une certaine cohérence dans la distribution géographique des sous-domaines de recherche : les sous-domaines de recherches tendent à être regroupés dans des villes spécifiques. Inversement, la valeur de consistance de **Cities** vers **Sous-domaines** est faible (0,096), indiquant que beaucoup de sous-domaines de recherche co-existent dans les mêmes villes.

Comparaison globale des comportements de référencement entrants et sortants. Le tableau 5.2 montre que les valeurs de consistance d'**OutLinks** sont plus élevées que celles

d'**InLinks**, ceci vers tous les points de vue. Cela indique que les classes de la carte **OutLinks** sont plus cohérentes que celles de la carte **InLinks**. Nous pouvons généraliser cette observation en mettant l'hypothèse que celui des laboratoire qui le référence. Cependant, cette hypothèse doit être considérée avec précaution, étant donné que les qualités des cartes **InLinks** et **OutLinks** sont modérées.

Les discussions précédentes montrent que l'analyse de consistance dans le contexte de multi-points de vue de notre approche comprenant la méthode MultiSOM représente une base solide pour obtenir une vue globale du comportement de liens dans un contexte à grande échelle. Des stratégies plus focalisées basées sur la même approche sont présentées dans le chapitre suivant. Elles peuvent être utilisées pour illustrer des phénomènes locaux liés au comportement de référencement. En effet, ces derniers peuvent venir étayer ou au contraire infirmer, des hypothèses globales.

5.4 Mécanisme de généralisation

Nous rappelons que les rôles principaux de la méthode de généralisation sont à la fois, d'évaluer la cohérence des thématiques qui sont calculées dans la classification originale, et, de faire une synthèse de contenus de cette dernière en mettant en lumière des thématiques plus générales.

5.4.1 Mécanisme de généralisation de Neural Gas

Dans le cas du modèle NG, les connexions originales entre les neurones, qui sont créés par la méthode de Neural Gas plus Competitive Hebbian Learning (cf. section 2.3.4), ne déterminent pas de relation de voisinage du même type que celle du modèle SOM. Ainsi, ces connexions ne peuvent pas être utilisées dans le mécanisme de généralisation parce qu'elles ne reflètent aucune topologie fixe. De plus, elles sont spécifiquement instables du fait qu'elles dépendent d'un paramètre d'âge de connexion empirique. Notre mécanisme de généralisation de NG (cf. algorithme 9) résout ce problème en créant sa structure spécifique de lien dans laquelle chaque neurone d'un niveau donné est lié à ses 2 plus proches voisins (cf. figure 5.6) [Al Shehabi and Lamirel, 2005b].

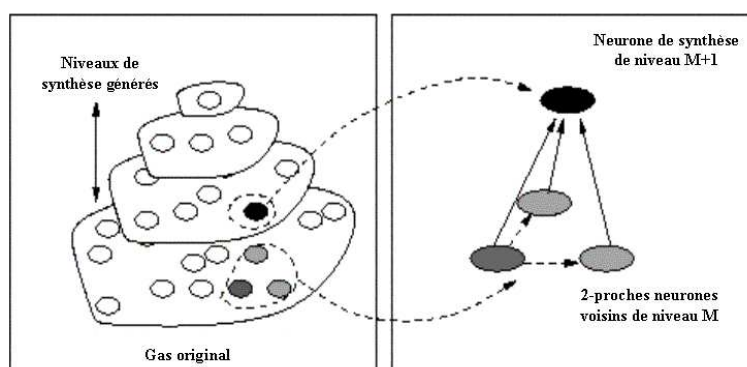


FIG. 5.6 – Le mécanisme de généralisation de NG (représentation 2D de gaz est utilisée seulement pour la clarté de la figure).

Algorithme 9 Algorithme de la généralisation de NG

Notations : $A^M = \{c_1^M, c_2^M, \dots, c_{N_M}^M\}$: représente l'ensemble de neurones au niveau M , c_i^M est le neurone i au niveau M , $N_M = |A^M|$.

A^0 : représente l'ensemble de neurones originaux générées par le modèle qui sont considérés au niveau original ($M = 0$)

$\mathcal{C}^M \subset A^M \times A^M$ représente un ensemble de connexions au niveau M , telles que (c_1^M, c_2^M) représente une connexion entre c_1^M et c_2^M ,

$\{c_1^M, c_2^M, c_3^M\}$ représente un triangle de neurones, $w_{c_i}^M$ est le vecteur de référence de neurone c_i au niveau M .

Entrée : $A^0 = \{c_1^0, c_2^0, \dots, c_{N_M}^0\}$

3: /* Construire le réseau des liens entre les neurones */

4: Déterminer les deux neurones de A^M , $c_{s_1}^M$ et $c_{s_2}^M$ les plus proche de c^M :

$$\|w_c^M - w_{s_1}^M\| \leq \|w_c^M - w_{s_2}^M\| \leq \|w_c^M - w_{c_i}^M\|, \quad \forall i \in A^M$$

5: $\mathcal{C}^M = \mathcal{C}^M \cup \{c^M, c_{s_1}^M\}, \{c^M, c_{s_2}^M\}$

6: Créer un nouveau neurone c^{M+1} au niveau $M+1$
 $A^{M+1} = A^{M+1} \cup \{c^{M+1}\}$

Son vecteur de référence w_c^{M+1} est donnée par :

$$w_c^{M+1} = \frac{1}{3}(w_c^M + w_{s_1}^M + w_{s_2}^M) \quad (5.15)$$

7: Tous les neurones du niveau $M+1$ qui partagent les mêmes neurones fils sont fusionnés.

8: **Si** $|A^{M+1}| = 1$ **alors** Fin de l'algorithme

9: **Si** $|A^{M+1}| = 1$ (i.e. $A^{M+1} = \{c_1^{M+1}, c_2^{M+1}\}$) **alors**

$$w_c^{M+2} = \frac{1}{2}(w_{c_1}^{M+1} + w_{c_2}^{M+1}) \quad (5.16)$$

Fin de l'algorithme

10: Retourner à l'étape 3.

5.4.2 Évaluation du mécanisme de généralisation

Afin d'évaluer la performance de mécanisme de généralisation que nous avons présenté dans le contexte des modèles de MultiSOM (cf. section 2.3.7) et de MultiGAS (cf. section 5.4.1), nous utiliserons deux types de mesures d'évaluation [Al Shehabi and Lamirel, 2005b] :

Évaluation de la qualité des topographies de généralisation

Le premier type de mesure d'évaluation sera basé sur nos propres critères de qualité, il s'agit des mesures de **Précision** (cf. équation 4.1) et de **Rappel** (cf. équation 4.2) qui rapportent aux propriétés des membres de classe (cf. section 4.2).

Évaluation de la cohérence de la généralisation

Le deuxième type de mesure d'évaluation sera la mesure de cohérence de la propagation (cf. équation 5.14) qui est basée sur notre modèle de propagation bayésien (cf. section 5.2.1).

Alors, deux mesures seront nécessaires pour comparer une classification originale à un de ses niveaux de généralisation :

- **Cohérence montante** : la classification originale est considérée comme une topographie source et son niveau de généralisation comme une topographie destination. Cette mesure évalue la qualité du mécanisme de généralisation. La plus élevée sera la mesure de cohérence montante, le meilleur sera le mécanisme de généralisation. En effet, même si le nombre de neurones du niveau de généralisation est moins que celui du niveau original, un bon mécanisme de généralisation ne dispersera pas les données déjà rassemblées sur la topographie originale dans le niveau généralisé.
- **Cohérence descendante** : la classification originale est considérée comme une topographie destination et son niveau de généralisation est considéré comme une topographie source. Cette mesure évalue la puissance de discrimination de la classification originale par rapport à sa généralisation. La plus forte sera la mesure de cohérence descendante, la moins discriminante sera la classification originale vis à vis de son niveau de généralisation. En effet, une classification originale ne sera pas discriminante par rapport à sa généralisation si elle utilise seulement le même nombre de neurones que cette dernière pour projeter les données originales.

5.4.3 Expérimentation

Données d'expérimentation

Nos expérimentations [Al Shehabi and Lamirel, 2005b] consistent à évaluer la performance des deux mécanismes de généralisation (généralisation de SOM et NG) que nous avons décrits précédemment. Notre corpus de test est le corpus des 1000 notices de brevets (cf. section A.2). Un seul point de vue, **Utilisations**, sera considéré dans notre expérimentation. Chacune de nos expérimentations est lancée avec une topographie optimale générée grâce à notre algorithme d'optimisation basé sur les critères de qualité (cf. section 4.2.3).

Notre première expérimentation consiste à générer un ensemble de cartes pour le modèle MultiSOM :

- Premièrement, des cartes originales de 16×16 (optimale), 15×15 , 14×14 et 13×13 neurones sont générées.
- Des cartes généralisées de 15×15 , de 14×14 et de 13×13 neurones sont générées en appliquant respectivement le mécanisme de généralisation aux cartes originales 16×16 (optimale), 15×15 , 14×14 neurones.
- Finalement, six niveaux de généralisation (de 15×15 neurones à 10×10 neurones) ont été appliqués à la carte originale et optimale de 16×16 neurones.

Notre deuxième expérimentation consiste à générer un ensemble de gaz pour le modèle MultiGAS :

- Premièrement, des gaz originaux de 121 neurones (optimal), et 81 neurones sont générés.
- Des gaz généralisées de 98 et de 64 neurones sont générés en appliquant respectivement le mécanisme de généralisation aux gaz originaux de 121 (optimal) et 81 neurones.
- Un nouveau gaz original de 64 neurones est généré dans une troisième étape (le même nombre de neurones qu'un gaz généralisé).
- Dans une dernière étape, un gaz généralisé de 53 neurones est généré dans une dernière

étape en appliquant le mécanisme de généralisation au gaz original de 64 neurones.

Pour chaque modèle, la cohérence de propagation entre chaque topographie originale et son voisin original de nombre de classes directement inférieur est calculée, ainsi que la cohérence de propagation entre chaque topographie originale et sa topographie généralisée directe (cf. figure 5.7 et figure 5.8). De plus, les mesures de qualité sont calculées pour chaque topographie qu'elle est originale ou généralisée (cf. figure 5.9 et figure 5.10). Pour le modèle MultiSOM, l'évaluation de différents niveaux de généralisation d'une même carte est également illustrée (cf. figure 5.11).

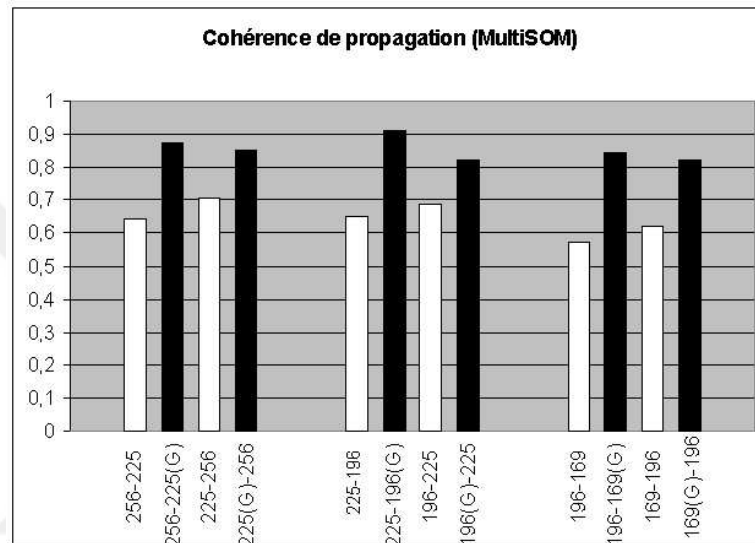


FIG. 5.7 – Modèle MultiSOM : Évaluation de la mesure de cohérence de propagation pour le point de vue **Utilisations** ($x - y$: représente la propagation de la carte source x à la carte destination y et $x(G)$ ou $y(G)$ représente un niveau de carte généralisée).

Discussion

Pour le premier résultat, la comparaison entre les mesures de qualité du gaz optimal de NG de 121 neurones et la carte optimale de SOM de 256 neurones démontre encore une fois clairement l'avantage du modèle NG par rapport au modèle SOM. En effet, en terme de qualité de classification, pour un certain nombre de neurones qui est presque deux fois inférieur, les valeurs de **Rappel** et de **Précision** de gaz de NG optimal sont plus élevées que celles de la carte optimale de SOM (cf. figure 5.9 et figure 5.10).

La meilleure homogénéité des résultats du modèle NG par rapport à ceux du modèle SOM est également mise en évidence en examinant les valeurs de cohérence montante entre topographies originales.

En effet, les valeurs de cohérence montante entre les gaz originaux qui ont des nombres différents de neurones sont plus élevées que les valeurs de cohérence montante entre les cartes originaux qui ont des nombres différents de neurones, bien que ces topographies aient été générées en utilisant les mêmes données (cf. figure 5.7 et figure 5.8).

Dans le cas des deux modèles, la figure 5.7 et la figure 5.8 montrent que les valeurs de

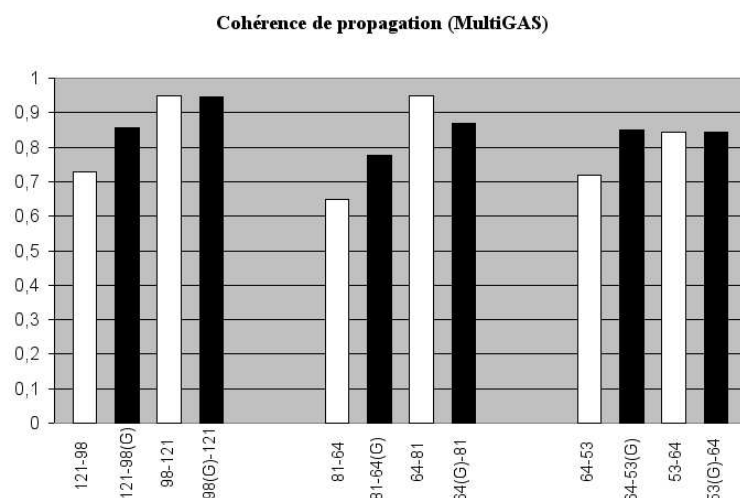


FIG. 5.8 – Modèle MultiGAS : Évaluation de la mesure de cohérence de propagation pour le point de vue **Utilisations**.

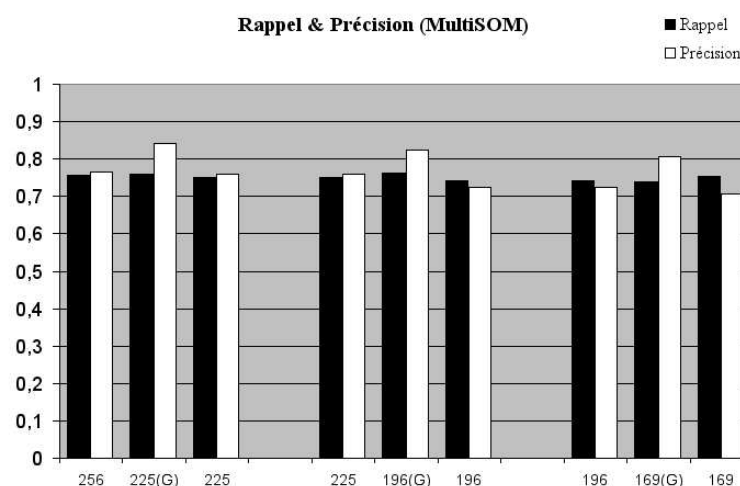


FIG. 5.9 – Modèle MultiSOM : Évaluation des mesures de qualité de cartes originales et leurs généralisations pour le point de vue **Utilisations** ($x(G)$: représente le premier niveau de la carte généralisée de x neurones).

cohérence montante entre les topographies originales et leurs topographies généralisées sont élevées. De plus, ces valeurs sont significativement plus élevées que les valeurs de cohérence montante entre topographies originales décrites précédemment. Les résultats montrent l'exactitude du processus de généralisation pour les deux modèles. Dans le cas du modèle MultiSOM, les valeurs systématiquement plus faibles de **Précision** (pour une valeur similaire du **Rappel**) des cartes originales par rapport à leurs cartes généralisées peuvent être considérées comme un résultat paradoxal (cf. figure 5.9). En effet, ce résultat indique que la qualité des cartes généralisées est plus élevée que celle de leurs cartes originales parentes. Il signifie également que le mécanisme de généralisation donne la possibilité de réduire de manière significative l'effet du

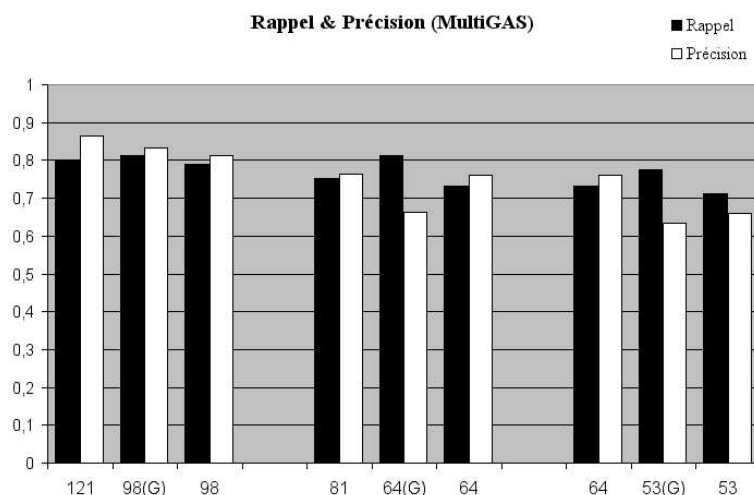


FIG. 5.10 – Modèle MultiGAS : Evaluation des mesures de qualité des gaz originaux et leurs généralisations pour le point de vue **Utilisations**.

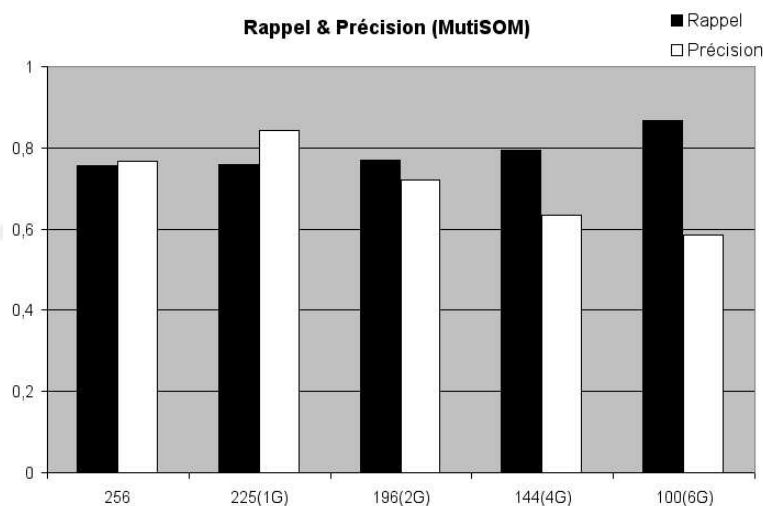


FIG. 5.11 – Modèle MultiSOM : illustration de l'évolution des mesures de qualité des niveaux de généralisation pour le point de vue **Utilisations** (xG : représente le $x^{\text{ième}}$ niveau de la carte de généralisation).

bord de la carte originale de SOM par le premier niveau de sa généralisation. Quand le nombre de données marginales est suffisamment élevé, comme dans les données documentaires, notre hypothèse est que tous les vecteurs de référence de la carte deviennent sensibles à l'effet du bord. Notre mécanisme de généralisation fournira alors, dans une première étape, un ajustement des vecteurs de référence au lieu de fournir une généralisation efficace. L'hypothèse de l'ajustement des vecteurs de référence dans les cartes généralisées peut être confirmée par le fait que les valeurs de cohérence descendante entre les cartes originaux et leurs cartes généralisées sont significativement plus élevées que les valeurs de cohérence descendante entre les cartes originales (cf. figure 5.7). En effet, ce dernier résultat indique que la carte et sa généralisation tendent à

devenir similaires (c'est-à-dire, même si le nombre de neurones d'une carte originale de SOM est plus élevé que celui de sa carte généralisée et associée, le même nombre sera utilisé dans les deux cartes pour la représentation de données effectives). Ce phénomène n'est par contre pas vérifié avec le modèle MultiGAS (cf. figure 5.8). En effet, le mécanisme de généralisation fourni par le modèle MultiGAS se comporte directement d'une manière logique : la valeur de **Précision** baisse et la valeur de **Rappel** augmente au premier niveau de la généralisation (cf. figure 5.10). L'examen de la figure 5.11 montre que le mécanisme de généralisation fourni par le modèle MultiSOM se comporte également d'une manière appropriée à partir du deuxième niveau de généralisation. Finalement, les résultats du mécanisme de généralisation sont légèrement meilleurs dans le modèle MultiSOM que dans le cas de modèle MultiGAS (meilleures valeurs en total de cohérence montante pour la généralisation avec le modèle MultiSOM) (cf. figure 5.7 et figure 5.8). Dans l'espace N-dimensionnel de NG, les liens de voisinage générés entre les neurones d'un gaz peuvent être partagés par beaucoup de triangles générées, ce qui amène une inexactitude potentielle dans la projection de données sur la généralisation de gaz. Le mécanisme du modèle MultiSOM qui généralise la carte originale à partir d'un ensemble voisinage carrée de neurones doit être logiquement moins sensible à ce phénomène.

A titre de conclusion, ce mécanisme est transposable entre les topographies générées par ce mécanisme de généralisation, il donne la meilleur résultats que la transposition qui est faites entre des topographies générées indépendamment.

5.5 Extraction des règles d'association par des modèles neuro-naux non supervisés

Les modèles les plus populaires utilisés dans ECBD sont les modèles symboliques (cf. section 1.4.3). Malheureusement, ces modèles souffrent des limitations très sérieuses. La génération des règles est un processus très coûteux et long, qui génère un nombre énorme de règles, incluant une large proportion de règles redondantes. Par conséquent, ceci interdit de calcul et de sélection de règles dès que les données sont nombreuses et qu'elles sont représentées par un espace de description de dimension très élevée. Cette dernière situation survient très souvent dans le cas des données documentaires.

Pour résoudre ces problèmes, quelques tentatives préliminaires d'ECBD utilisant des modèles numériques ont été faites. A titre d'exemple, un algorithme d'extraction de connaissances basé sur un réseau auto-organisateur est proposé par Hammer et al. [Hammer *et al.*, 2002]. Cette approche repose sur une quantification vectorielle généralisée et supervisée (GRLVQ) qui est utilisée pour extraire des arbres de décision. Les différents chemins des arbres produits servent ensuite à dénoter des règles. Le défaut principal de cette méthode est qu'elle nécessite d'utiliser des données préalablement étiquetées, ou classes de référence.

D'un autre côté, une méthode de classification hybride a été proposée par Lamirel et al. [Lamirel *et al.*, 2003c] pour projeter une structure explicative issue d'une classification symbolique de type treillis de Galois sur une carte SOM. La carte SOM et le treillis de Galois sont générés à partir des mêmes données. La projection cosinus est alors utilisée pour associer des concepts du treillis aux classes de carte SOM. Les propriétés des concepts agissent en tant qu'explication pour les classes de ladite carte. Des méthodes d'élagage peuvent être ensuite utilisées pour identifier une hiérarchie de classes de la carte à l'intérieur du treillis. Les règles d'association sont alors obtenues à partir du treillis élagué résultant. En fait, malgré qu'elle établisse des liens intéressants entre les approches numériques et symboliques, cette approche

nécessite le calcul coûteux d'un treillis de Galois complet.

Dans notre approche nous exploitons le mécanisme de généralisation et le mécanisme de communication fournis par le modèle. Le mécanisme de généralisation peut être spécifiquement utilisé pour contrôler le nombre de règles d'association extraites par une seule base de données (ou mono point de vue). Le mécanisme de communication est utile pour mettre de évidence des règles d'association en trouvant les relations entre les thèmes appartenant aux différents points de vue. Nous nous servons également de nos mesures originales de **Rappel** et de **Précision** qui sont dérivées de la théorie des treillis de Galois et le SRD (cf. section 4.2).

5.5.1 Critères principaux

Pour extraire des règles d'association, nous nous basons sur nos mesures originales, telles que le **Rappel** (ici appelé *Rapp*) (cf. équation 4.2) et la **Précision** (ici appelée *Prec*) (cf. équation 4.1). Nous rappelons ici que, le **Rappel** et la **Précision** d'une propriété (p) donnée d'une classe (c), sont données comme suit :

$$Rapp_c(p) = \frac{|c_p^*|}{|C_p^*|}, \quad Prec_c(p) = \frac{|c_p^*|}{|c|} \quad (5.17)$$

telles que, C est un ensemble de classes issues d'une méthode de classification appliquée sur un ensemble de données D , $c \in C$, et

$$c_p^* = \{d \in c, \zeta_d^p > 0\} \quad (5.18)$$

où ζ_d^p est le poids d'une propriété p pour une donnée d .

Nous avons démontré à la section 4.2.2 que si les deux valeurs de **Rappel** et de **Précision** atteignent la valeur d'unité pour une classe donnée, l'ensemble des classes propres (cf. équation 4.4) représente un sous-treillis de Galois. Une classe appartient à l'ensemble de classes propres d'une classification donnée si elle possède des propriétés propres (cf. équation 4.3). Finalement, une propriété est considérée comme propre à une classe donnée si elle est maximisée par les membres de la classe (cf. équation 4.5).

Comme nous l'avons montré précédemment, les moyennes de **Rappel** et de **Précision** peuvent être utilisées, pour la comparaison globale des méthodes de classification et pour l'optimisation des résultats d'une méthode relativement à une base de données. Dans cette section, nous nous concentrons plus spécifiquement sur les propriétés propres des classes et sur des mesures locales de **Précision** et **Rappel** singulièrement associées aux classes. Dès lors que cette information peut être exploitée pour générer des explications sur les contenus individuels des classes singulières, elle représente également une base solide pour extraire des règles desdites classes [Al Shehabi and Lamirel, 2005a; Lamirel and Al Shehabi, 2005].

5.5.2 Extraction par le mécanisme de généralisation

Nous agissons sur nos propres critères de qualité de classe pour extraire des règles à partir de classes issues d'un gaz original et de ses généralisations. Pour une classe donnée c , la forme générale de l'algorithme d'extraction est donnée dans l'algorithme 10.

où $\text{Extent}(p)$ représente l'extension d'une propriété p (c'est-à-dire, un ensemble de données aux quelles la propriété p est associée (cf. section 1.4.3)), et P_c^* représente l'ensemble de propriétés propres de la classe c . Les règles extraites selon ce principe peuvent être considérées

Algorithme 10 Algorithme d'extraction des règles d'association par le mécanisme de généralisation

Entrée : $\forall p_1, p_2 \in P_c^*$

- 2: **si** ($Rapp(p_1) = Rapp(p_2) = Prec(p_1) = Prec(p_2) = 1$) **alors** $p_1 \leftrightarrow p_2$ (règle d'équivalence)
- 3: **sinon si** ($Rapp(p_1) = Rapp(p_2) = Prec(p_2) = 1$) **alors** $p_1 \rightarrow p_2$
- 4: **sinon si** ($Rapp(p_1) = Rapp(p_2) = 1$) **alors**
 - si** ($Extent(p_1) \subset Extent(p_2)$) **alors** $p_1 \rightarrow p_2$
 - si** ($Extent(p_2) \subset Extent(p_1)$) **alors** $p_2 \rightarrow p_1$
 - si** ($Extent(p_1) \equiv Extent(p_2)$) **alors** $p_1 \leftrightarrow p_2$

Entrée : $\forall p_1 \in P_c^*, \forall p_2 \in P_c - P_c^*$

- 6: **si** $Rapp(p_1) = 1$ **alors**
 - si** ($Extent(p_1) \subset Extent(p_2)$) **alors** $p_1 \rightarrow p_2(*)$
-

comme des règles propres, ou règles de classe, étant donné qu'elles mettent en jeu des propriétés propres d'une classe à la fois dans leur prémisse et dans leur conclusion.

L'étape optionnelle (4) peut être utilisée pour extraire des règles étendues. Pour les règles étendues, la contrainte de propriétés propres n'est pas appliquée sur la propriété la plus générale. Ainsi, l'extension de cette dernière propriété peut référer des données qui se trouvent dehors de la classe courante c .

5.5.3 Extraction par le mécanisme de communication

Une stratégie d'extraction de règles qui s'avère intéressante, et complémentaire de la première, consiste à combiner l'algorithme d'extraction avec le principe de communication entre les points de vue pour extraire les règles. La forme générale de l'algorithme d'extraction entre les points de vue v_1 et v_2 est donnée ci-après.

Algorithme 11 Algorithme d'extraction des règles d'association par le mécanisme de communication

Entrée : $\forall p_1 \in P_c^*, \forall p_2 \in P_{c'}^*$ et $c \in v_1, c' \in v_2$

- 2: **si** ($Rapp(p_1) = Rapp(p_2) = Prec(p_1) = Prec(p_2) = 1$) **alors** *TestTypeRègle*
 - 3: **sinon si** ($Rapp(p_1) = Rapp(p_2) = Prec(p_2) = 1$) **alors** *TestTypeRègle*
 - 4: **sinon si** ($Rapp(p_1) = Rapp(p_2) = Prec(p_1) = 1$) **alors** *TestTypeRègle*
 - 5: **sinon si** ($Rapp(p_1) = Rapp(p_2) = 1$) **alors** *TestTypeRègle*
où la procédure de *TestTypeRègle* est exprimée comme :
 - si** ($Extent_{v_1}(p_1) \subset Extent_{v_2}(p_2)$) **alors** $p_1 \rightarrow p_2$
 - si** ($Extent_{v_2}(p_2) \subset Extent_{v_1}(p_1)$) **alors** $p_2 \rightarrow p_1$
 - si** ($Extent_{v_1}(p_1) \subset Extent_{v_2}(p_2)$) **alors** $p_1 \leftrightarrow p_2$
 - 6: /*Les règles étendues peuvent être obtenues à l'aide des étapes optionnelles suivantes :*/
 - a) $\forall p_1 \in P_c^*, \forall p_2 \in P_{c'}$: substituant respectivement $Rapp(p_2)$ et $Prec(P_2)$ par des mesures relatives à un point de vue source $Rapp_{v_1}(p_2)$ et $Prec_{v_1}(p_2)$, relié au point de vue source, dans l'algorithme précédent.
 - b) $\forall p_1 \in P_c, \forall p_2 \in P_{c'}^*$: substituant respectivement $Rapp(p_1)$ et $Prec(P_1)$ par des mesures relatives à un point de vue destination $Rapp_{v_2}(p_1)$ et $Prec_{v_2}(p_1)$, relié au point de vue destination, dans l'algorithme précédent.
-

	Niveau original	Niveaux généralisés												
		1	2	3	4	5	6	7	8	9	10	11	12	13
Utilisation	100	79	62	50	40	31	26	16	11	-	-	-	-	-
Avantages	121	100	83	75	64	53	44	34	28	23	18	13	-	-
Déposants	49	31	24	19	16	12	-	-	-	-	-	-	-	-
Titres	144	114	111	105	95	83	71	59	46	35	27	22	18	14

TAB. 5.3 – Sommaire du processus de génération de gaz. Pour tous les points de vue, la limite de généralisation a été fixée à des niveaux qui comprenant ou plus 10 neurones. Donc, pour un point de vue donné, le nombre de niveaux de généralisation dépend à la fois du nombre initial de neurones de son gaz optimal associé et de l'homogénéité de la distribution de données relativement à ce point de vue.

5.5.4 Expérimentation

Modèle MultiGAS

Cas de l'analyse de brevets : Notre corpus de test est le corpus de 1000 notices de brevets (cf. section A.2). Chaque expérimentation est basée sur la classification optimale du modèle NG. Cette classification est obtenue par l'algorithme d'optimisation basé sur nos critères de qualité (cf. section 4.2.3).

Dans la première étape, des gaz optimaux originaux sont générés pour tous les points de vue (**Utilisations**, **Avantages**, **Déposants** et **Titres**). Dans la deuxième étape, des gaz généralisés sont générés pour chaque point de vue en appliquant les étapes successives de généralisation sur les gaz optimaux originaux (cf. section 5.4). Les résultats de ces deux étapes sont résumés dans le tableau 5.3.

Notre première expérimentation consiste à extraire des règles de façon indépendante à chaque point de vue. Les gaz originaux et leurs généralisations sont utilisés pour extraire les règles. L'algorithme est premièrement utilisé sans son étape optionnelle. Ensuite, cette étape est appliquée (pour plus de détails, cf. algorithme 10). Quelques exemples des règles extraites reliées à chaque point de vue sont donnés ci-après :

Refrigerator oil $\rightarrow$ Gear oil (supp = 3, conf = 100%) (**Utilisations**)

Wide viscosity range $\rightarrow$ Thermal and oxidative stability (supp = 3, conf = 100%) (**Avantages**)

Surfactant system $\rightarrow$ Calcium (supp = 3, conf = 100%) (**Titres**)

Infineum $\rightarrow$ Hexxon (supp = 10, conf = 100%) (**Déposants**)

Pour évaluer la complexité de notre algorithme basé sur une approche numérique, par rapport à une approche symbolique, nous utilisons le calcul du facteur d'efficacité (EF) :

$$EF = \frac{RC}{LC} * \frac{MLC}{MRC} \quad (5.19)$$

où RC =nombre de règles, MRC =nombre de règles maximum (symbolique), LC =nombre de boucles et MLC =nombre de boucles maximum (symbolique).

Le facteur EF illustre un compromis entre le nombre de règles générées par le modèle numérique et le nombre de boucles générées par ce dernier, en considérant en même temps le nombre de règles et le nombre de boucles générées par le modèle symbolique. Cette mesure

peut être utilisée comme un indicateur pour déterminer le niveau optimal de la généralisation pour l'extraction des règles d'association.

Un résumé global des résultats est donné dans le tableau 5.4. Ledit tableau inclut une comparaison de notre algorithme d'extraction avec une méthode symbolique standard d'extraction de règles, en ce qui concerne à la quantité de règles extraites. Dans l'expérimentation par points de vue, quand notre algorithme d'extraction est utilisé avec son étape optionnelle, il peut extraire une proportion significative des règles qui peuvent être extraites par un modèle symbolique classique, alors que ce dernier utilise essentiellement une approche combinatoire. Dans certains cas, tel que celui du point de vue **Déposants**, toutes les règles de confiance 100% peuvent même être extraites à partir d'un seul niveau du gaz (cf. figure 5.15). Alternativement, comme dans le cas du point de vue **Utilisations**, la combinaison des niveaux de gaz du même point de vue peut être utilisée pour extraire toutes les règles de confiance 100% (cf. tableau 5.4). La performance d'extraction la plus mauvaise est obtenue avec le point de vue **Titres**. Cette performance relativement basse (58% de règles de confiance 100% extraites en utilisant tous les niveaux de gaz) peut être expliquée à la fois par la dispersion et par la diversité des données qui s'avèrent les plus élevées relativement à ce point de vue. Cependant, elle est compensée la meilleur efficacité d'extraction du modèle numérique par rapport au modèle symbolique (100 fois moins de boucles pour le modèle numérique). De plus, dans le cas de ce point de vue, les règles extraites ont un support moyen qui est plus élevé que le support moyen de l'ensemble des règles global (cf. tableau 5.4 et figure 5.16).

Même si aucune sélection de règle n'est effectuée quand l'algorithme d'extraction est utilisé avec son étape optionnelle, l'avantage principal de cette version de l'algorithme, par rapport à une méthode symbolique classique, est le temps de calcul. En effet, le temps de calcul est sensiblement réduit, puisque notre algorithme est basé sur les classes. De plus, d'une manière générale, plus le niveau de généralisation exploité pour l'extraction des règles sera bas, plus spécialisées seront les classes, et par conséquent, plus l'effet combinatoire pendant le calcul sera réduit (cf. figures 5.12, 5.14, 5.15 et 5.16). Un autre résultat intéressant est le comportement de notre algorithme d'extraction quand il est utilisé sans son étape optionnelle. La figure 5.13 montre que, dans ce cas, un processus de sélection de règles qui dépend du niveau de généralisation, est effectué : au fur et à mesure que nous allons vers le niveau élevé de la généralisation, les règles seront plus extraites.

L'utilité du facteur EF apparant clairement dans le cas du point de vue **Titres**. En effet, ce facteur permet de mettre en évidence un compromis pertinent entre le nombre des boucles et le pourcentage des règles.

Notre deuxième expérimentation consiste à extraire des règles en utilisant le mécanisme communication entre les points de vue **Utilisations** et **Avantages**. La communication est réalisée entre le gaz original de chaque point de vue, ensuite, entre les mêmes niveaux de la généralisation de chaque point de vue (cf. figure 5.17). Pour chaque étape de communication, l'algorithme d'extraction de cette communication est appliquée d'une manière bidirectionnelle. Quelques exemples des règles extraites sont donnés ci-après :

Natural oil (**Avantages**) $\rightarrow$ Catapult oil (**Utilisations**) (supp = 2, conf = 100%)

Natural oil (**Avantages**) $\rightarrow$ Drilling fluid (**Utilisations**) (supp = 2, conf = 100%)

Les résultats de notre expérimentation par points de vue multiples sont similaires à ceux de notre expérimentation par point de vue indépendant (cf. tableau 5.4). De plus, dans tous

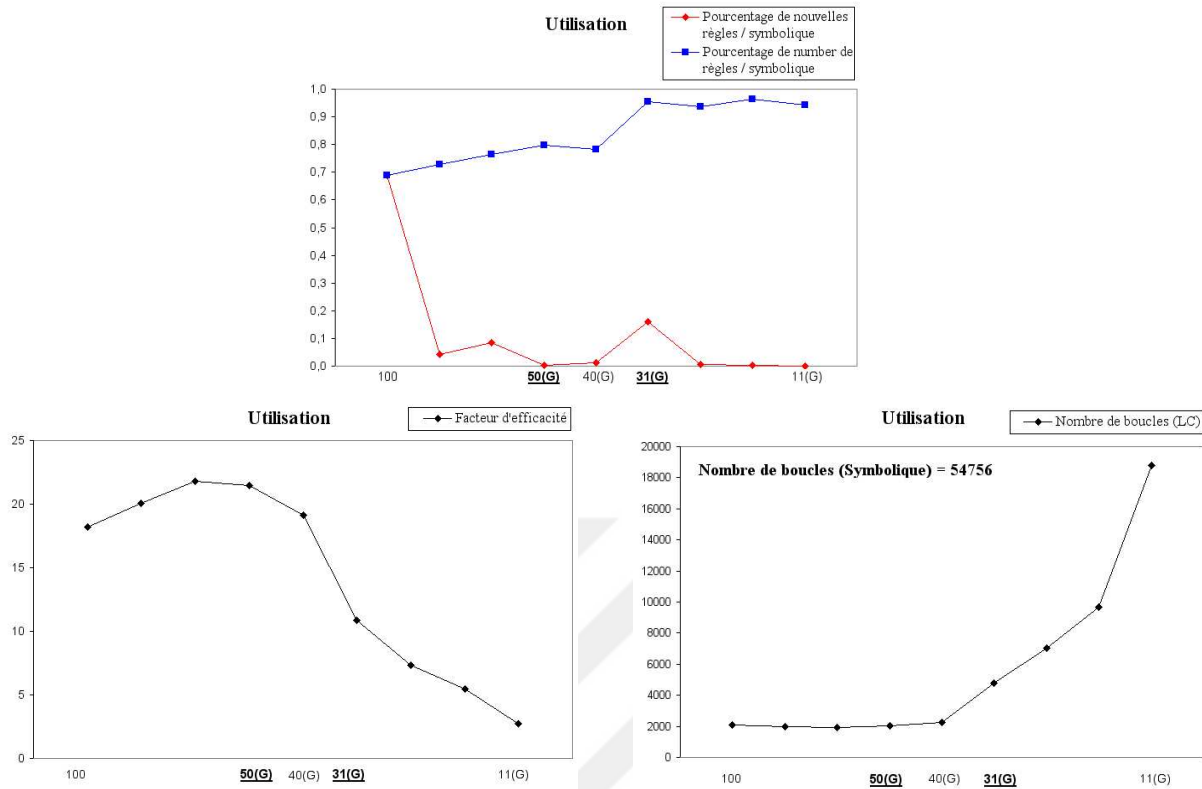


FIG. 5.12 – Les résultats d'extraction de règles pour le point de vue **Utilisations** (Algorithme d'extraction 10 avec l'étape optionnelle) : Le niveau 50(G) pourrait être considéré comme un niveau optimal, dès lors qu'il fournit le meilleur compromis entre le pourcentage de règles extraites (80% de toutes les règles) et le calcul de complexité (2033 boucles). Cependant, si le pourcentage de règles extraites est considéré comme prioritaire par rapport à la complexité de calcul, le niveau de 31(G) (95% de toutes les règles, 4794 boucles) pourrait à son tour être considéré comme le meilleur niveau. Nouvelles règles : règles qui sont trouvées à un niveau donné mais pas aux niveaux précédents. Nombre de boucles (Symbolique) : nombre de boucles utilisées par l'approche symbolique pour extraire des règles. $x(G)$: représente un niveau de généralisation de x neurones.

les cas, un processus de sélection de règle est effectué quand la version standard de notre algorithme est utilisée. La performance d'extraction maximale est obtenue quand le Rappel et la Précision relatifs à un point de vue sont utilisés (cf. algorithme 11). Cela prouve à la fois la pertinence de notre mécanisme d'extraction de règles, mais également l'efficacité de notre mécanisme de communication entre les vues.

Cas de l'analyse de données web : Cette expérimentation met en évidence l'intérêt de l'exploitation des règles en webométrie et qu'elle permettant d'isoler des connaissances précises qui ne peuvent pas être obtenues par un autre type d'analyse de type réseau de graphe présenté dans la section 3.4.2.

Dans cette expérimentation [Lamirel *et al.*, 2005], qui est basée sur le corpus des sites web

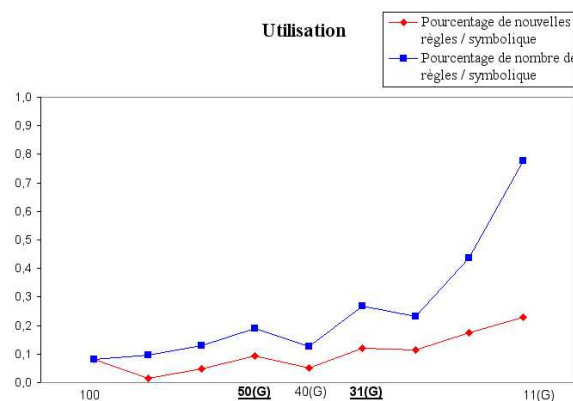


FIG. 5.13 – Les résultats d'extraction de règles pour le point de vue **Utilisations** (Algorithme d'extraction 10 sans l'étape optionnelle) : Seules des règles propres sont extraites par cette version de l'algorithme. Dans ce cas, un processus de sélection de règle qui dépend du niveau de la généralisation est exécuté : plus le niveau de la généralisation est élevé, la sélection va diminuer. Les bonnes performance du niveau de 50(G) et du niveau de 31(G) sont également accentuées avec cette version de l'algorithme (voir figure 5.12 pour une comparaison avec l'autre version de l'algorithme).

des laboratoires de recherche européens (cf. section A.3.2), des gaz optimaux originaux sont générés pour les points de vue (**UCodes**, **Cities**, **InLinks** et **In-Coupling**) (cf. tableau 5.5).

Dans une première expérimentation nous appliquons l'algorithme d'extraction 10, pour extraire des règles étendues des résultats de l'analyse de gaz propre au point de vue (**In-Coupling**). Le rôle de l'extraction des règles liées à un tel point de vue est de trouver des expressions de dépendance entre les laboratoires de recherche allemands par la vue externe d'autres laboratoires de recherche européens. En effet une règle du type $A \rightarrow B$ signifie dans un tel point de vue que A est vu (du point de vue européen) comme un laboratoire satellite du B .

Notre algorithme a produit des résultats très significatifs. En effet, 1909 règles étendues sont extraites à partir du gaz, parmi lesquelles, 604 règles étendues avec un *support* > 2 (plus le support et la confiance d'une règle sont élevés, la plus grande sera sa plausibilité). Les résultats globaux de l'extraction de règles sont présentés dans le tableau 5.6. Quelques exemples de règles extraites sont donnés ci-après.

www.fh-niederrhein.de $\rightarrow$ www.tu-chemnitz.de (*supp* = 5, *conf* = 100%) : propre
www.tu-chemnitz.de $\rightarrow$ www.sc.rwth-aachen.de (*supp* = 19, *conf* = 100%)

La combinaison de ces deux règles met en évidence une dépendance enchaînée entre les universités vus à partir du contexte de recherche européen.

Dans une deuxième expérimentation nous appliquons l'algorithme d'extraction 11, pour extraire des règles étendues à partir de résultats de l'analyse de gaz entre le point de vue **UCodes** et le point de vue **Cities** d'un côté et entre le point de vue **UCodes** et le point de vue **InLinks** d'un autre côté, respectivement. pour chaque cas, l'algorithme d'extraction est appliqué d'une manière bidirectionnelle.

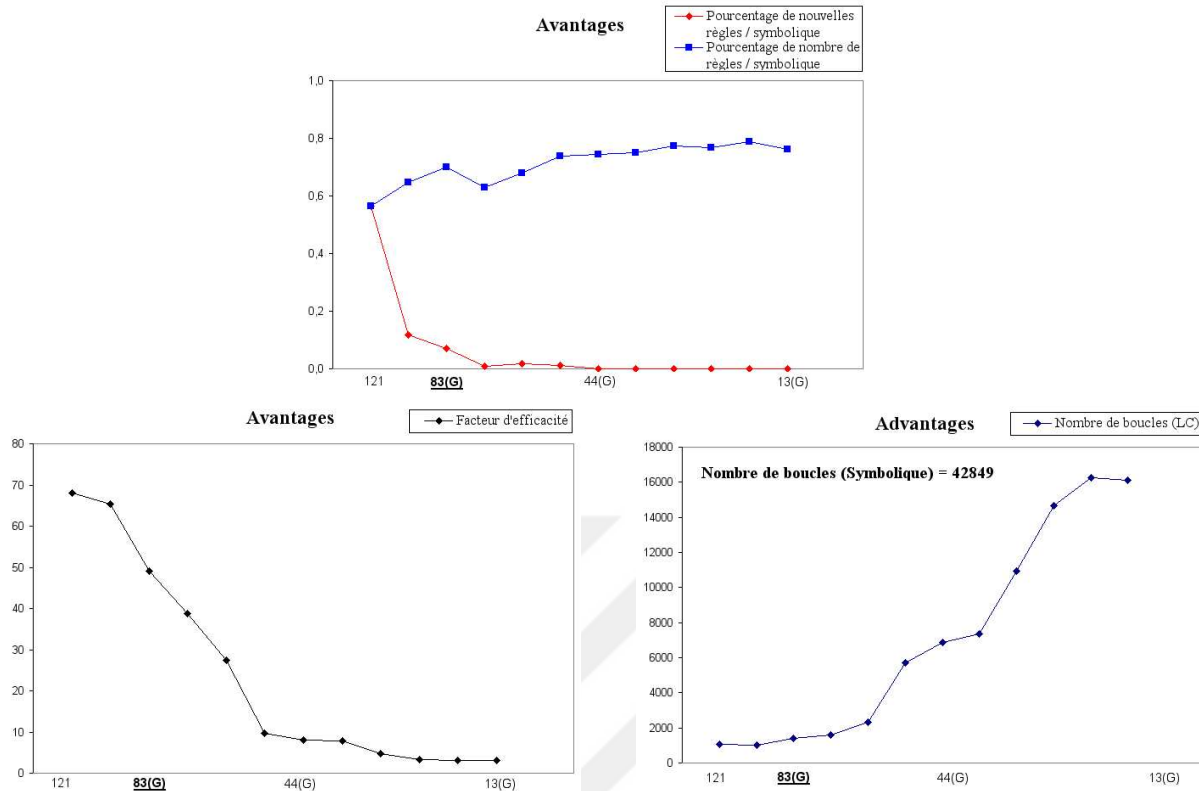


FIG. 5.14 – Les résultats d'extraction de règles pour le point de vue **Avantages**. Le niveau 83(G) peut être considéré autant qu'un niveau qui représente le compromis optimal : 70% de toutes les règles sont extraites en utilisant 1402 boucles.

Le rôle de l'extraction de règles liée à la communication entre le point de vue **UCodes** et le point de vue **Cities** est de trouver des expressions de la dépendance entre les villes et les thèmes de recherche. En effet, une règle du type $A \rightarrow B$, où A est un sous-domaine de recherche et B est une ville, signifie que le sous-domaine de recherche A est une spécialité de la ville B . Ainsi, 46 règles (toutes propres) ont été extraites à partir du gaz, y compris 2 règles avec un $support > 2$. Les résultats globaux de l'extraction de règle sont présentés dans le tableau 5.6. Un exemple de règle extraite est donné ci-après.

Data transmission. Videofax. Fax. Keys. Cryptography $\rightarrow$ Saarbruecken ($supp = 2$, $conf = 100\%$)

Le rôle de l'extraction de règles liée à la communication entre le point de vue **InLinks** et le point de vue **UCodes** est de trouver des expressions de la dépendance entre les liens et les sous-domaines de recherche. Ces règles peuvent être particulièrement utilisées pour expliquer des liens. En effet, une règle du type $A \rightarrow B$, où A est un lien et B est un UCode, signifie que le lien A est entièrement expliqué par le sous-domaine de recherche décrit par B . Ainsi, 1002 règles ont été extraites à partir du gaz. Les résultats globaux de l'extraction de règle sont présentés dans le tableau 5.6. Un exemple de règle extraite est donné ci-après.

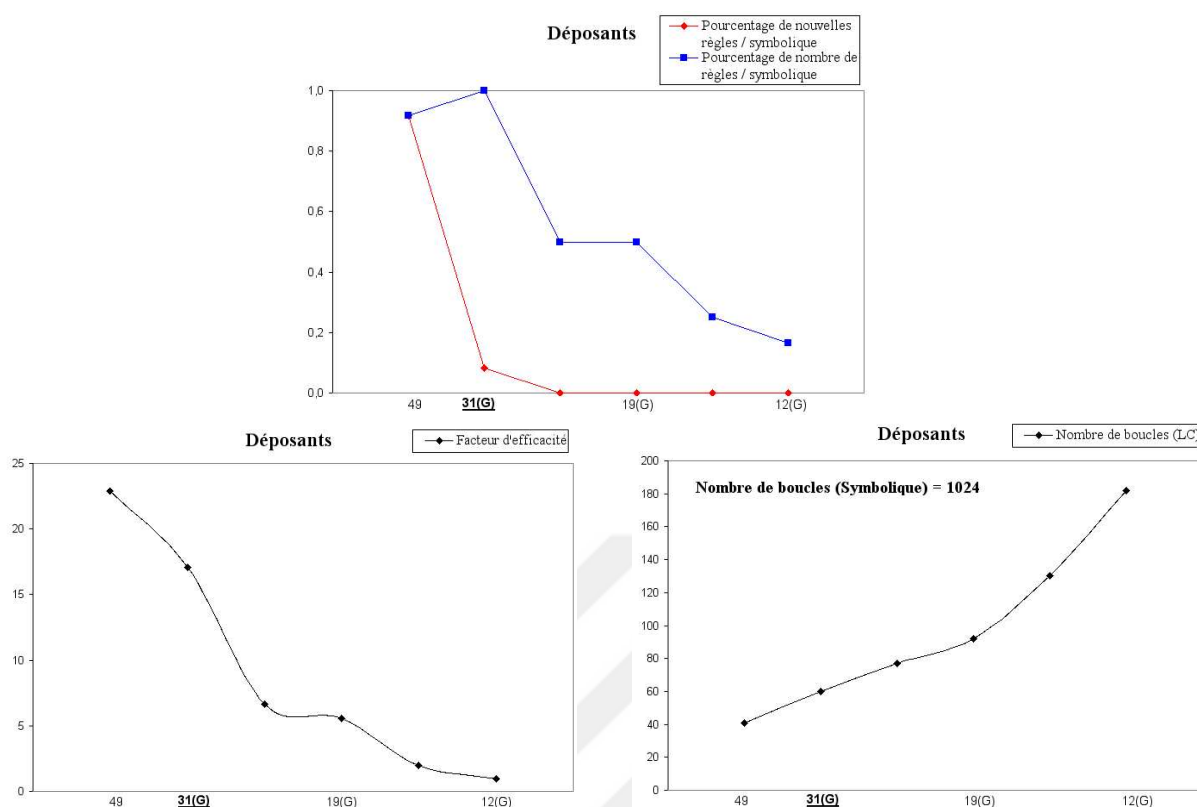


FIG. 5.15 – Les résultats d'extraction de règles pour le point de vue **Déposants**. Le niveau 49(G) peut être considéré autant qu'un niveau optimal puisqu'il fournit le meilleur compromis entre le pourcentage des règles extraites (92% de toutes les règles) et de la complexité de calcul (41 boucles). Néanmoins, si le pourcentage des règles extraites est considéré comme prioritaire vis à vis de la complexité de calcul, le niveau 31(G) devrait être considéré comme le niveau le plus optimal : toutes les règles sont extraites en utilisant 61 boucles. La baisse de la performance d'extraction de règles pour les niveaux élevés de généralisation est due à la fois aux caractéristiques spécifiques d'indexation de ce point de vue (le nombre moyen d'index par document est près de 1), à l'effet de regroupement initial de la stratégie de Winner-Takes-Most de NG, à l'effet supplémentaire du processus de généralisation et à la stratégie d'extraction de règles basée sur la distribution des propriétés singulières. Cependant, quand la généralisation est exécutée, les documents décrits par des combinaisons de termes d'index qui ont été regroupés dans les mêmes classes par la stratégie d'apprentissage seront redistribués dans des classes spécifiques. Ainsi, la distribution des propriétés singulières dans les classes deviendra plus hétérogène pour les niveaux intermédiaires de généralisation que pour le niveau initial.

www.medizin.uni-koeln.de → Medical science (*supp* = 3, *conf* = 100%)

La règle présentée donne la Science Médicale comme une explication thématique pour tous les liens issus du site web www.medizin.uni-koeln.de.

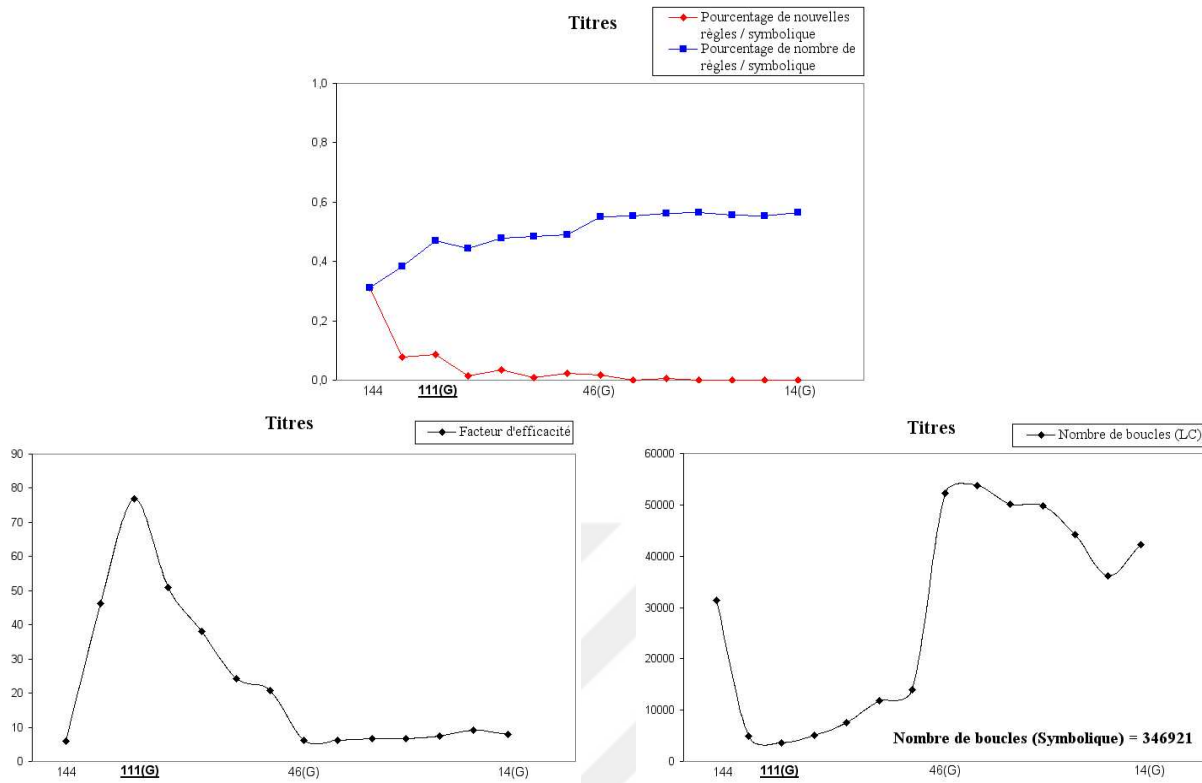


FIG. 5.16 – Les résultats d'extraction de règles pour le point de vue **Titres**. Le niveau 111(G) peut être considéré comme un niveau optimal. Cependant, même si le pourcentage des règles extraites est relativement faible (48% de toutes les règles), ce niveau fournit une efficacité d'extraction très bonne (facteur d'efficacité = 76 et 3619 boucles) par rapport au modèle symbolique. La complexité élevée de calcul de règles du premier niveau de gaz (niveau optimal), par rapport aux niveaux suivants de généralisation, est due à la fois à la dispersion des termes d'index du point de vue **Titres** et à l'effet de regroupement de type Winner-Takes-Most de la stratégie d'apprentissage de NG (voir également la figure 5.15 concernant à ce point). Cependant, quand la distribution de données est éparse, cette stratégie d'apprentissage tend à regrouper les données dans un petit nombre de classes, laissant quelques autres classes vides. Cet effet augmentera de manière significative la complexité du calcul de règle au niveau original puisque les classes non vides auront des descriptions relativement hétérogènes. Il faut noter que cette complexité sera réduite aux étapes intermédiaires de la généralisation puisque des données indépendantes seront distribuées dans différentes classes par le processus de généralisation (cf. figure 5.15).

Le potentiel d'une approche d'extraction de règles pour réaliser les analyses webométriques précises est mis en évidence par les exemples précédents, même si beaucoup de travail reste à effectuer pour trier les règles et pour combiner leurs résultats d'une manière plus synthétique.

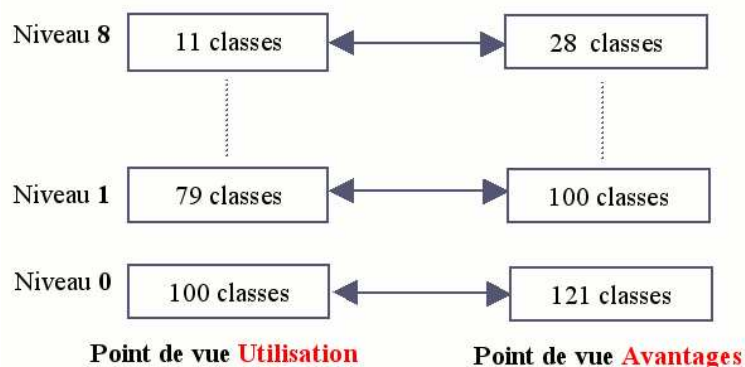


FIG. 5.17 – Vue générale de processus de communication : pour le processus de généralisation, seulement 9 niveaux ont été utilisés pour les deux points de vue, en commençant de niveau 0 qui représente la classification optimale de gaz.

Modèle MultiSOM

A titre de comparaison avec le modèle MultiGAS, nous avons également utilisé le modèle MultiSOM pour l'extraction des règles d'association [Lamirel and Al Shehabi, 2006] sur le même corpus de référence de notices de brevets. Les résultats obtenus avec ce modèle sont similaires à ceux qui sont obtenus avec le modèle MultiGAS, mais en nécessitant cependant d'utiliser un nombre de classes deux fois plus important qu'avec le modèle MultiGAS. Cette différence de capacité de synthèse entre les deux modèles est à rapprocher de celle que nous avons déjà illustré lors de nos précédentes expérimentations (cf. section 4.2.4).

5.6 Conclusion

Nous avons proposé dans ce chapitre plusieurs extensions du modèle MultiSOM, qui représente lui-même l'extension du modèle SOM à la gestion des points des vues multiples.

Nous avons premièrement proposé un mode de propagation optimisé en établissant une correspondance entre le modèle MultiSOM et un modèle de réseau d'inférence bayésien. Nous avons montré que les résultats de ce mode de propagation sont plus fiables que ceux du mode probabiliste initialement fourni par le modèle MultiSOM. De plus, il serait facile de démontrer que le modèle bayésien est le seul qui garantisse la cohérence des résultats quand des modalités différentes activent les données des mêmes classes des topographies destination. Finalement, la correspondance que nous avons établie entre le modèle multi-topographique et le modèle de réseau d'inférence bayésien nous permet également d'envisager d'implanter le mode de communication entre des points de vues multiples en prenant comme base une quelconque méthode de classification.

Nous avons ensuite proposé un nouveau mécanisme de généralisation pour le modèle NG. Etant donné qu'il s'affranchit des contraintes topologiques, en utilisant un principe proche de celui de la triangulation de Delaunay (cf. section 2.3.1), le mécanisme de généralisation de NG peut être considéré comme un mécanisme de synthèse en-ligne applicable à toute autre méthode de classification. Nous avons proposé une évaluation comparative de ce mécanisme avec le mécanisme de généralisation original proposé pour MultiSOM (cf. section 2.3.7). Cette évaluation a été opérée en utilisant de manière conjointe les mesures de qualité et le mécanisme

		Déposants	Titres	Utilisations	Avantages	Utilisations ↔ Avantages
Modèle Symbolique	Nb. de règles total	12	2326	536	404	649
	Taux de confiance	100%	100%	100%	100%	100%
	Taux de support	3.583	1.049	1.139	1.042	1.081
	Nb. de règle global	26	4912	2238	1436	2822
	Taux de confiance	53%	59%	59%	44%	45%
Modèle MultiGAS	Nb. de règles propres	12	422	251	287	250
	Taux de confiance	100%	100%	100%	100%	100%
	Nb. de règles étendues	12	1358	536	319	642
	Taux de confiance	100%	100%	100%	100%	100%
	% de total symbolique	100%	58%	100%	79%	99%
	Taux de support	3.583	1.081	1.139	1.050	1.073

TAB. 5.4 – Sommaire des résultats. Le tableau présente une comparaison synthétique entre la méthode symbolique standard d'extraction des règles et la méthode basée sur MultiGAS. Le nombre global de règles défini pour le modèle symbolique inclut le nombre de règles approximatives (confiance < 100%) et le nombre de règles totales (exactes) (confiance = 100%). Dans nos expérimentations, les règles générées par le modèle de MultiGAS sont seulement des règles totales (exactes). Le nombre de règles propres est le nombre de règles obtenues avec les versions standards des algorithmes d'extraction. Le nombre de règles étendues est le nombre de règles obtenues avec les versions étendues des algorithmes d'extraction comprenant leurs étapes optionnelles.

	UCodes	Cities	InLinks	In-Coupling
Nombre de classes de NG	50	50	50	15

TAB. 5.5 – Sommaire de processus de génération de gaz pour les quatre points de vue.

de comparaison des vues que nous avons également proposés. Nos résultats expérimentaux ont démontré l'efficacité des deux mécanismes. Un résultat intéressant de l'évaluation du mécanisme de généralisation fourni par le modèle MultiSOM est que, celui-ci permet, dans un premier temps, autrement dit avant d'opérer un synthèse, de réduire de manière significative les effets de bord produits sur une carte originale par les données marginales. Les expérimentations relatives à la généralisation ont de nouveau mis en évidence l'exactitude et l'homogénéité globales plus élevées des résultats du modèle MultiGAS.

Le mécanisme de comparaison des vues que nous avons mis en place constitue une approche

	In-Coupling (intra)	Cities ↔ UCodes	UCodes ↔ InLinks
Règles étendues (<i>supp</i> > 0)	1909	46	1002
Règles étendues (<i>supp</i> > 2)	604	2	-
Taux de confiance de règles étendues	100%	100%	100%
Règles propres	4	46	507
Taux de confiance de règles propres	100%	100%	100%

TAB. 5.6 – Sommaire des résultats d'extraction des règles d'association par le mécanisme de généralisation (pour le point de vue **In-Coupling**) et par le mécanisme de communication inter-cartes (entre **Cities** et **UCodes** d'un coté et entre **UCodes** et **InLinks** d'un autre coté) en utilisant le modèle MultiGAS.

originale qui permet d'instrumentaliser ce type de comparaison, de manière à opérer à des mises en correspondance dynamique de résultats d'analyse de données. Ce dernier principe peut lui-même être considéré comme un principe totalement nouveau. À titre d'exemple, nous l'avons exploité pour effectuer une analyse webométrique complète d'un ensemble des données web en nous basant sur des facteurs multiples. Il a donné des résultats impossibles à obtenir par une analyse manuelle.

Nous avons finalement proposé une nouvelle approche pour l'extraction de connaissances partir du modèle multi-topographique. Notre approche se base sur nos mesures originales de rappel et de précision pour extraire des règles à partir de classes issues de ce modèle. Nos expérimentations ont été menées à bien sur des classifications associées à un seul point de vue, aussi bien qu'entre des classifications associées à des points de vue multiples, ceci en utilisant notre nouveau modèle MultiGAS. Elles tirent également parti des mécanismes de généralisation et de communication inter-topographies étendus que nous avons proposés. Même si des expérimentations complémentaires doivent être faites, nos premiers résultats sont très prometteurs. Ils tendent à montrer qu'une méthode numérique, dès lors qu'elle est suffisamment élaborée, représente un candidat naturel pour s'affranchir des problèmes relatifs à l'inflation de règles, au choix des règles et au temps de calcul qui sont inhérents aux méthodes symboliques.

Finalement, l'ensemble des modifications que nous avons proposées, en partant de la base du modèle MultiSOM, sont suffisamment génériques pour nous permettre de considérer que nous avons défini un nouveau paradigme pour l'analyse et la fouille de données. Nous avons donc choisi d'appeler ce paradigme, le paradigme MVDA (pour Multi Viewpoint Data Analysis).

Dans le chapitre suivant, nous proposerons d'implanter de nouveaux mécanismes de visualisation dans le paradigme MVDA. Le rôle de ces mécanismes sera celui de permettre la visualisation des résultats de classification, dans le cas où la méthode de classification utilisée n'opère pas par construction de projection de ces résultats sur un plan de représentation ; ce qui est par exemple le cas de la méthode NG.

Chapitre 6

Optimisation des stratégies de visualisation des résultats de classification

Sommaire

6.1	Introduction	149
6.2	Validation de la communication inter-cartes	150
6.2.1	Description des cartes	150
6.2.2	Analyse locale de la communication inter-cartes	151
6.2.3	Validation croisée des approches : MultiSOM et BibTechMon	154
6.3	Classification hiérarchique multi-niveaux orientée par la densité (DBHC)	157
6.3.1	Principe de DBHC	157
6.3.2	Visualisation hypertree de DBHC	157
6.3.3	Application à la webométrie	160
6.4	Conclusion	161

6.1 Introduction

L'interprétation des résultats d'analyse d'une méthode de classification, ainsi que ceux d'une méthode de fouille de données, nécessite de disposer de méthodes de visualisation puissantes et interactives. Nous illustrons dans ce chapitre l'exploitation de deux méthodes de visualisation interactives différentes, à savoir l'interaction visuelle topographique et la visualisation hyperbolique.

L'interaction visuelle topographique peut être spécifiquement exploitée pour l'interprétation précise des résultats obtenus par le mécanisme de comparaison des vues (cf. section 5.3), mais également pour produire des analyses ciblées permettant d'isoler des phénomènes locaux spécifiques, non nécessairement compatibles avec les résultats de l'analyse globale (cf. section 6.2). Pour faciliter ces opérations, nous proposons dans la première partie de ce chapitre de combiner l'approche graphique BibTechMon (cf. section 3.4.2) et l'approche MultiSOM orientée par les points de vue (cf. section 2.3.7).

Un grand nombre de tâches d'analyse de donnée, telles que la webométrie, nécessitent une vue globale des résultats d'analyse où l'interaction entre les classes obtenues est préservée. Dès lors qu'elle devra être appliquée aux résultats de méthodes d'analyse de données produisant des descriptions de classes multidimensionnelles, une visualisation de ce type de données ou de classes représente un problème compliqué qui n'a pas été encore résolu. Ainsi, quand des méthodes de projection linéaire sont utilisées pour ce but (cf. section 3.3), le résultat est une perte significative d'information. Quand des méthodes de projection non linéaire sont utilisées (cf. section 3.4), il n'y a aucune garantie d'une visualisation appropriée de la structure de voisinage entre les classes dans le cas d'espace de données fortement multidimensionnelles. Cependant, la visualisation hyperbolique (cf. section 3.5.2) est connue de ses capacités à traiter le problème de la surcharge cognitive inhérente aux approches basées sur les graphes (cf. section 3.4.2). Ainsi, elle permet de visualiser des relations complexes entre les données par un mécanisme de focus et contexte. Jusqu'alors, la visualisation hyperbolique avait été plutôt utilisée pour résoudre des problèmes d'organisation de données, comme la gestion des répertoires. Dans le domaine des modèles neuronaux, pour résoudre les problèmes de visualisation des cartes, soit de manière structurelle (cf. section 2.3.6), soit de manière indirecte, en traitant les résultats, comme l'ont également proposé Vesanto et Alhoniemi [Vesanto and Alhoniemi, 2000]. Cependant, ces dernières méthodes ne résolvent pas les limitations du modèle SOM lui-même par rapport au modèle NG, plus performant car n'utilisant pas d'étape de projection (cf. sections 2.3.3). Nous proposons donc dans la deuxième partie de chapitre une approche de visualisation hyperbolique originale qui peut être appliquée sur les résultats d'un modèle de classification tel que le modèle NG. Son avantage principal est de produire une classification hiérarchique en utilisant un algorithme hiérarchique original afin de représenter les classes sous une forme d'un hypertree dans lequel l'information sur la densité de données issues de l'espace de description original de classes est préservé, à chaque niveau. Son second avantage est de mettre en évidence d'une manière indirecte la structure originale de voisinage entre les classes.

6.2 Validation de la communication inter-cartes

Nous proposons dans cette section une validation des résultats expérimentaux de l'analyse locale et des mécanismes de comparaisons des vues (cf. section 5.3.2). Cette validation se base sur la combinaison du modèle MultiSOM avec l'approche graphique BibTechMon (cf. section 3.4.2). Elle se réfère aux résultats obtenus pour l'analyse des données du web (cf. section A.3.2).

6.2.1 Description des cartes

Pour le point de vue **Cities**, les classes de la carte regroupent des sites web partageant la même location géographique. Pour le point de vue **UCodes**, les classes de la carte regroupent des sites web partageant le même profil global de recherche (c'est-à-dire, une combinaison de codes Unesco). Pour le point de vue **OultLinks**, les classes de la carte regroupent des sites web partageant leurs liens sortants : ils sont décrits par les cibles des liens. Le point de vue **InLinks** est l'équivalent de **OultLinks**, en utilisant cette fois les liens entrants : les classes de la carte sont donc décrites par des profils constitués de cibles de liens.

La facilité d'interprétation d'une carte dépend non seulement de la qualité proprement dite de la carte (cf. section 5.3.2), mais également des facteurs complémentaires, comme le granularité de la description. Deux cas typiques de cartes sont décrits ci-après.

1. La carte **UCodes** (cf. figure 6.1) : c'est une carte relativement petite (81 classes) avec des aires logiques relativement grandes. Les classes de la carte possèdent des titres génériques, et certaines d'entre elles contiennent un grand nombre de sites web. Ainsi, cinq classes (parmi un total de 81 classes) regroupent 227 sites web (parmi un total de 438 sites web). Une indexation plus précise aurait permis une meilleure distribution des sites web dans des classes d'une taille plus homogène, et par conséquent, une interprétation plus précise du contenu de carte.

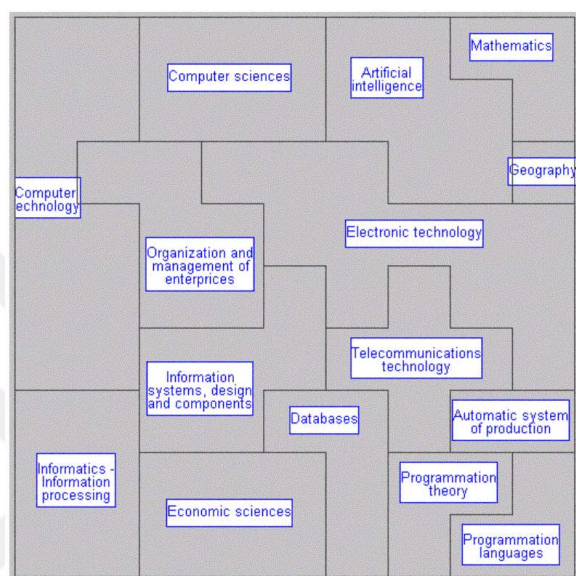


FIG. 6.1 – La carte SOM de UCodes ou Sous-domaines.

2. La carte **OutLinks** (cf. figure 6.2) : chaque site web a beaucoup de liens sortants différents, ce qui mène à une granularité élevée des profils de liens sortants des sites web, et ainsi, à une configuration potentiellement plus intéressante de carte. La carte **OutLinks** comprend des aires logiques de petites tailles, relativement homogènes. Une telle configuration de carte illustre un comportement homogène de référencement de la part des laboratoires de recherche. Cette interprétation sera confirmée par l'analyse de la structure du réseau de liens sortants produits par l'approche de BibTechMon (cf. figure 6.4).

6.2.2 Analyse locale de la communication inter-cartes

Le mécanisme de communication inter-cartes de MultiSOM peut être utilisé dans un mode interactif pour mettre en évidence les relations spécifiques entre des classes de cartes différentes (cf. section 2.3.7). Dans ce but, une activité est assignée à une classe, ou à une aire logique, d'une carte source. Puis, le mécanisme de la propagation de l'activité à toutes les autres cartes (cartes destinations) est appliqué (cf. figure 2.10). Finalement, le profil d'activité généré sur les cartes destinations est analysé (cf. figure 5.4). Ledit profil indique une cohérence faible ou forte entre les classes activées, ou aires, de la carte source, et, les classes activées des cartes destinations.

Nous présentons ci-après un exemple d'application d'un tel mécanisme pour opérer une analyse locale du comportement de référencement des laboratoires de recherche localisés dans

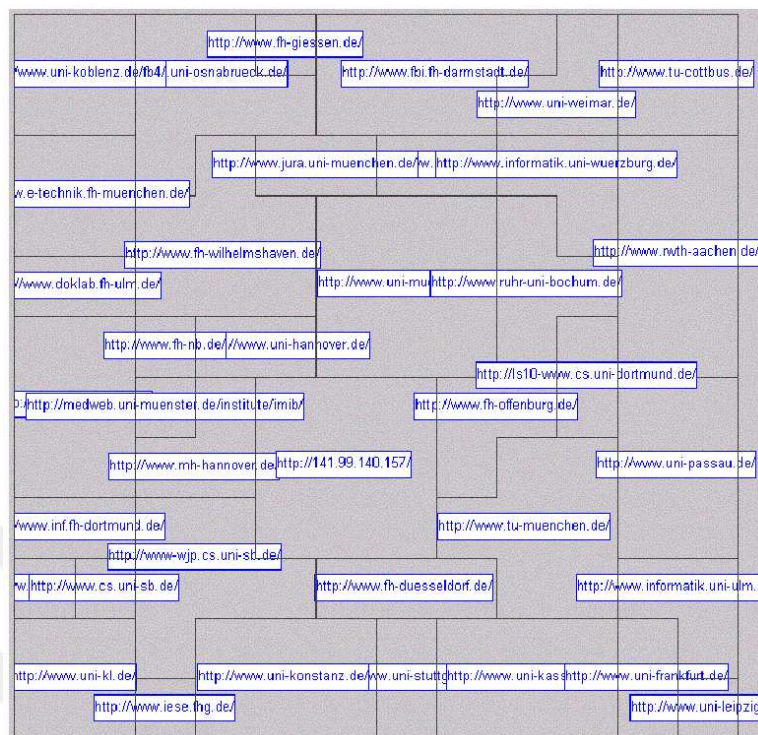


FIG. 6.2 – La carte SOM de OutLinks.

une ville. Pour cette analyse, nous avons choisi la plus grande classe de la carte **Cities**, en termes de nombre de sites web associés; celle-ci correspond à la ville de *Munich*.

La figure 6.3 illustre la stratégie globale qui a été utilisée pour réaliser l'étude sur la ville de *Munich*. Dans une première étape, une activité est faite sur l'aire logique *Munich* de la carte **Cities**, et, l'activation résultante distribuée sur la carte **OutLinks** est analysée. Dans une deuxième étape, la zone globale de la carte **OutLinks** sur laquelle l'activité s'est focalisée dans la première étape est réutilisée comme source d'activité, et, l'activité distribuée sur la carte **Cities** est analysée à son tour.

1. Étape 1 : La propagation de l'activité à partir de l'aire logique *Munich* vers la carte **OutLinks** se concentre sur une aire logique qui regroupe 3 classes dont le profil est dominé par l'URL <http://www.tu-muenchen.de/> (cf. figure 6.3). Les classes activées situées autour de cette aire logique possèdent les URLs dominantes suivantes dans leur profil :
 - <http://www.uni-passau.de/>
 - <http://www.informatik.uni-ulm.de/>
 - <http://www.fh-offenburg.de/>
 - <http://ls10-www.cs.uni-dortmund.de/>

Les URLs mentionnés ci-dessus correspondent aux sites web principaux cités par les sites web des laboratoires de *Munich*. Ils résument ainsi le comportement de référencement sortant de ces derniers laboratoires. Ceci nous amène à constater un comportement de référencement sortants à relativement faible échelle géographique, des laboratoires de *Munich*, c'est-à-dire que les villes référencées par ces laboratoires se trouvent, la plupart

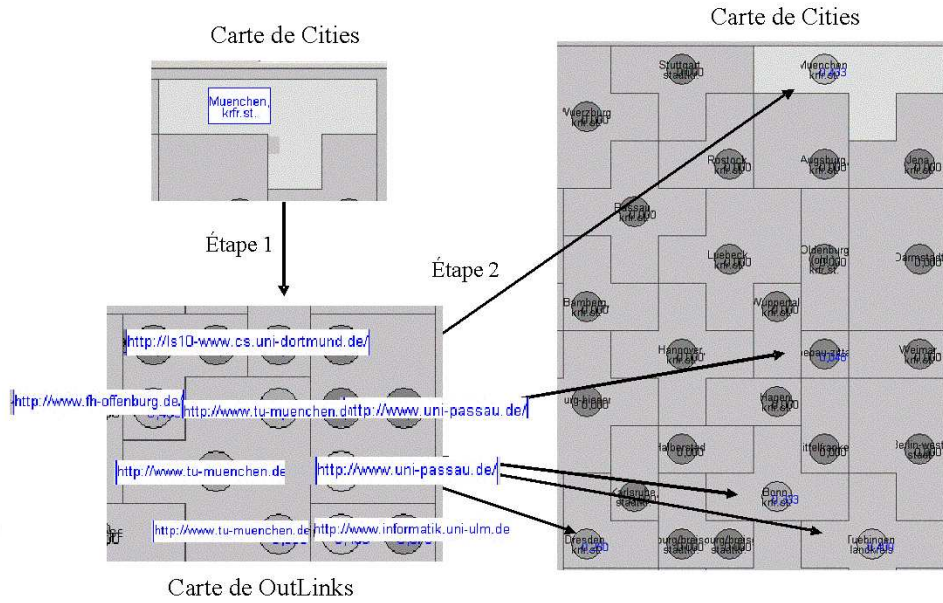


FIG. 6.3 – Mécanisme de communication inter-cartes à partir de la classe associée à la ville de *Munich*. Étape 1 : communication de la carte **Cities** vers la carte **OutLinks**, Étape 2 : communication de la carte **OutLinks** vers la carte **Cities**.

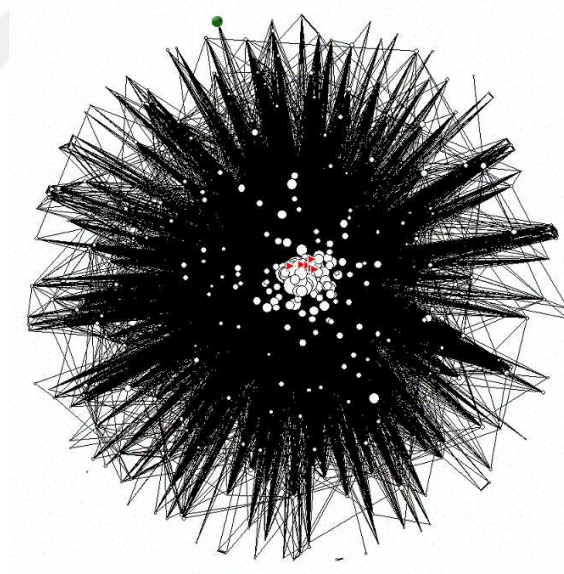


FIG. 6.4 – Réseau des sites web basé sur les liens sortants (OutLinks). On peut constater que la structure complexe de ce réseau, liée à l'affichage des liens entre les sites, représente une surcharge cognitive importante ce qui le rend de fait très difficile à interpréter et à utiliser de manière directe pour une analyse.

du temps, situées à proximité de *Munich*, dans le sud de l'Allemagne (Passau, Ulm, Offenburg).

Il est à noter que la forte valeur de consistance de la carte **Cities** vers la carte **OutLinks**

mesurée à la section 5.3.2 est en adéquation avec des comportements de référencement cohérents, sans toutefois permettre de les déterminer précisément, comme dans le cas de cette analyse locale.

2. Étape 2 : L'aire logique «<http://www.tu-muenchen.de/>» de la carte **OutLinks** regroupe des sites web ayant cette URL comme URL dominante dans leur profil. Une propagation d'activité à partir de cette aire logique est lancée vers la carte des villes (**Cities**). Son rôle est de mettre en évidence la propagation géographique des laboratoires partageant leur profil de liens sortants avec le profil de liens sortants dominant des laboratoires de *Munich*. Cette propagation d'activité a comme conséquence des classes activées de manière disparate sur la carte des villes (*Munich*, *Bonn*, *Tuebingen*, *Dresde*, et *Loebau-Zittau*). Ceci nous amène à conclure que : même si le comportement de liens de *Munich* est relativement local (voir l'étape 1), le type de liens sortants de *Munich* n'est pas spécifique à *Munich* et/ou à son entourage direct. Ainsi, il nous permet de rejeter l'hypothèse de «monde de référencement fermé» associé à *Munich*.

Il est à noter que la faible valeur de consistance de la carte **OutLinks** vers la carte **Cities** mesurée à la section 5.3.2 est également en adéquation avec l'hypothèse de «monde de référencement ouvert» obtenue, dans ce cas précis, par l'analyse locale.

6.2.3 Validation croisée des approches : MultiSOM et BibTechMon

La combinaison des approches MultiSOM et BibTechMon peut être analysée en termes d'utilité pour analyser le comportement de liens. D'une part, les fonctionnalités interactives, comme le processus de communication inter-cartes, permettent à l'analyste de faire une analyse locale focalisée sur une classe, mais elles ne lui fournissent pas de vue globale des interactions locales. D'autre part, les graphes sont utiles pour fournir l'analyste une telle vue, mais au prix d'une structure complexe, même au niveau local. Ainsi, la combinaison des deux méthodes semble être particulièrement pertinente pour valider les résultats des analyses locales.

La carte **OutLinks** de MultiSOM et le réseau **OutLinks** de BibTechMon peuvent être considérés comme deux manières différentes de décrire le même type de similarité : deux sites web seront regroupés dans la même classe de la carte **OutLinks** de MultiSOM s'ils partagent des liens sortants ; dans les mêmes conditions, ils seront également liés dans le réseau de **OutLinks** de BibTechMon. La combinaison des deux approches peut ainsi être exploitée pour la validation de l'analyse locale de comportement de liens sortants pour l'exemple de *Munich*.

La méthode BibTechMon (cf. section 3.4.2) est utilisée pour générer un graphe des sites correspondant à la carte **OutLinks** de MultiSOM. La construction de graphe sera basée sur l'analyse de co-occurrence de liens sortants. Les nœuds de graphe seront associés aux URLs de ces liens, et leur taille représentera le nombre de liens sortants, que nous appellerons également le degré de sortie.

La figure 6.5 montre une vue spécifique, obtenue par filtrage des sites web liés à *Munich* dans le réseau **OutLinks** global de BibTechMon (cf. figure 6.5). Dans cette vue, tous les nœuds du réseau sont dessinés, mais en ne conservant que les liens entre les 28 sites web des laboratoires relatifs à *Munich*. On peut constater que les sites web les plus forts degrés de sortie sont situés au centre de ce sous-réseau, alors que les sites web avec les plus faibles degrés de sortie sont situés à la marge de ce même sous-réseau. En effet, ces derniers entretiennent des relations fortes

avec d'autres sites web du réseau global. Leur position dans le réseau dépend donc directement de ces relations, qui matérialisent des relations potentielles avec d'autres villes.

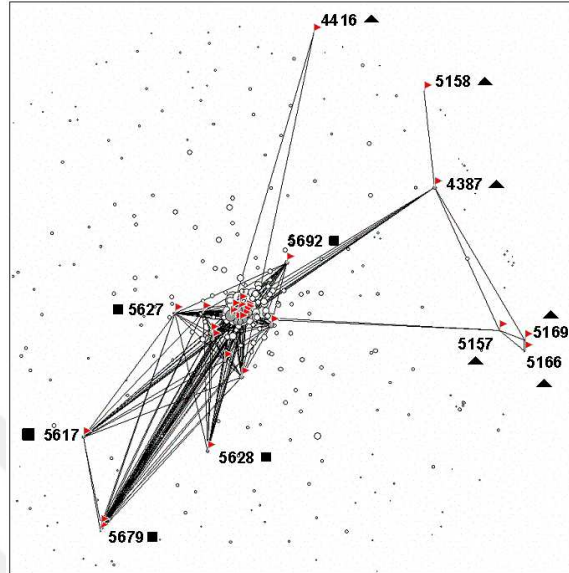


FIG. 6.5 – Réseau des sites web associés à *Munich* basé sur leurs liens sortants (OutLinks).

La figure 6.6 liste l'ensemble des sites web du sous-réseau associés à *Munich* sur le réseau **OutLinks** de BibTechMon. Cet ensemble de sites comprend les 5 sites web qui ont les degrés de sortie les plus élevés, au centre du sous-réseau, ainsi que les 11 sites web liés à *Munich* qui restent à la marge de ce sous-réseau. Différents pictogrammes sont utilisés afin d'établir la correspondance entre ce sous-réseau et l'information de la carte **OutLinks** de MultiSOM :

- Les cercles sont utilisés pour représenter les sites web qui ont les degrés de sortie les plus élevés, c'est-à-dire les sites web qui sont classés en tête de liste, situés au centre du sous-réseau de *Munich*. Lesdits sites web seront identifiés comme des sites-cercles.
- Les carrés sont utilisés pour représenter les sites web situés du côté gauche du sous-réseau de *Munich*. Ces sites web peuvent être caractérisés par leurs valeurs intermédiaires de degrés de sortie. Ils partagent également des liens sortants (outlinks) avec un nombre significatif de sites web classés en tête de liste. Lesdits sites web seront identifiés comme des sites-carrés.
- Les triangles sont utilisés pour représenter les sites web situés du côté droit du sous-réseau de *Munich*. Ces sites web peuvent être caractérisés par leurs faibles valeurs de degrés de sortie. Ils ne partagent des liens sortants qu'avec un petit nombre des sites web classés en tête de liste. Lesdits sites web seront identifiés comme des sites-triangles.

La figure 6.7 donne la position sur la carte **OutLinks** des sites web de la figure 6.6. Il met en évidence une analogie entre la répartition des sites web sur la carte et celle de ceux-ci sur le réseau. En effet, les 5 sites-cercles se concentrent sur la même zone focale de la carte **OutLink** (en bas à droite de la carte). Cette zone correspond également à la zone principale de focalisation d'activité de la classe *Munich*, décrite dans l'expérimentation de la section 5.3.2, portant sur la consistance. Les sites-carrés sont distribués dans l'entourage direct des sites-cercles. Les sites-triangles sont situés loin des autres sites web et distribués, de manière plus disparate, en haut à gauche de la carte.

			BibTechMon network	MultiSOM OutLink map	MultiSOM Sub-domains map
	Node number	Out-degree	Kernel		
●		161	Ludwig-Maximilians-Universität München, Institut für Informatik	11-9 http://www.informatik.uni-ulm.de/	0-0 Computer technology
●		160	Technische Universität München, Theoretische Informatik und Grundlagen der Künstlichen Intelligenz	11-9 http://www.informatik.uni-ulm.de/	6-1 Artificial intelligence
●		154	Technische Universität München, Teilbibliothek Mathematik/Informatik	8-11 http://www.uni-kassel.de/	8-1 Mathematics
●		113	Technische Universität München, Lehrstuhl für Effiziente Algorithmen	9-9 http://www.tu-muenchen.de/	0-0 Computer technology
●		113	Technische Universität München, Rechnerorganisation und Rechnerorganisation/Parallelrechnerarchitektur	9-9 http://www.tu-muenchen.de/	0-0 Computer technology
			Margin		
■	5692	22	München. Lehrstuhl für Numerische Mathematik und Wissenschaftliches Rechnen	8-8 http://www.tu-muenchen.de/	8-6 Automatic system of production
▲	4387	20	Fachhochschule München, Fachbereich Elektrotechnik und Informationstechnik	0-4 http://www.doklab.fh-ulm.de/	5-4 Elektronik technology
■	5627	12	Technische Universität München. Formale Sprachen, Übersetzerbau, Software-Konstruktion	11-7 http://www.uni-passau.de/	0-0 Computer technology
■	5617	10	Technische Universität München, Lehrstuhl für Reaktordynamik und Reaktorsicherheit	8-8 http://www.tu-muenchen.de/	5-6 Telecommunications technology
■	5628	7	Technische Universität München, Datenbanksysteme, Wissensbasen	4-9 http://www.uni-bremen.de/	0-8 Informatics - Information processing
▲	5157	3	Fachhochschule München, Labor Leistungselektronik	2-0 http://www.e-technik.fh-muenchen.de/	4-4 Information systems, design and components
■	5679	2	Technische Universität München. Algorithmische Graphentheorie	8-8 http://www.tu-muenchen.de/	5-8 Databases
▲	5169	2	Fachhochschule München Labor Betriebssysteme	0-2 http://www.e-technik.fh-muenchen.de/	0-0 Computer technology
▲	4416	1	Ludwig-Maximilians-Universität München, Institut für Rechtsphilosophie und Rechtsinformatik	5-2 http://www.jura.uni-muenchen.de/	8-5 Informatics - Information processing
▲	5158	1	Fachhochschule München, Labor Microcomputersysteme	0-4 http://www.doklab.fh-ulm.de/	4-4 Information systems, design and components
▲	5166	1	Fachhochschule München, Labor Mustererkennung und Künstliche Intelligenz	2-0 http://www.e-technik.fh-muenchen.de/	6-1 Artificial intelligence

FIG. 6.6 – Tableau des sites web associés à *Munich* et présentés dans la figure 6.7. La colonne **MultiSOM OutLink map** indique le nom de la classe où le site web est localisé et sa position (colonne, ligne) sur la carte **OutLinks**.

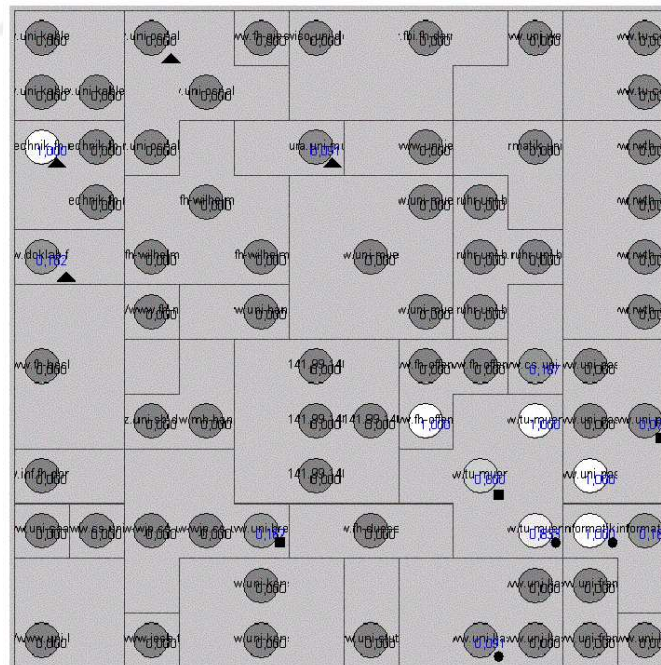


FIG. 6.7 – Répartition des sites web décrits dans le tableau 6.6 sur la carte **OutLinks**.

V_i	représente le $i^{\text{ème}}$ niveau de l'arbre hiérarchique ; $0 \leq i \leq n$
M_i	représente un ensemble de classes non agglomérées
$ M_i $	est le cardinal de M_i
M_0	représente l'ensemble initial de données (classes feuilles)

TAB. 6.1 – Notations de DBHC.

L'exemple précédent montre clairement que l'analogie observée entre les réseaux BibTechMon et les cartes MultiSOM peut servir de mécanisme de validation croisée des résultats des deux approches. D'une autre part, il montre également que le réseau de BibTechMon est utile pour obtenir des indications importantes pour comprendre l'organisation des cartes.

6.3 Classification hiérarchique multi-niveaux orientée par la densité (DBHC)

A notre sens, le choix d'un algorithme de classification hiérarchique s'avère déterminant, aussi bien dans le cadre de l'analyse de données que dans celui de la fouille de données. En effet, un bon algorithme doit permettre d'obtenir une synthèse incrémentale des relations entre les données à classer, tout en préservant des informations cohérentes sur la densité de ces données dans leur espace de description original. L'algorithme original que nous proposons ci-après permet de respecter à la fois ces deux contraintes.

6.3.1 Principe de DBHC

L'algorithme DBHC [Al Shehabi and Lamirel, 2006], que nous détaillons dans l'algorithme 12, se comporte à la fois comme un algorithme orienté par la densité (cf. section 2.2.4), et, comme un algorithme hiérarchique (cf. section 2.2.2). Son principe est celui de construire une hiérarchie de classes en utilisant un rayon de voisinage hiérarchiquement croissant. Le rayon de voisinage est défini automatiquement en fonction du nombre de niveaux désirés. Il est initialement calculé à partir des distances minimales et maximales entre les données originales. Ce rayon est ensuite adapté incrémentalement, de manière à préserver un degré de densité fixé pour chaque niveau hiérarchique. Selon ce principe, les propriétés suivantes seront vérifiées :

- Les données originales représenteront les feuilles de l'arbre hiérarchique.
- La racine de l'arbre représentera l'ensemble global des données.
- Une classe intermédiaire de l'arbre qui aura beaucoup de liens (c'est-à-dire, beaucoup de fils), par rapport aux autres classes du même niveau hiérarchique, représentera une zone plus dense par rapport à celles représentées par ces dernières.
- Toutes les feuilles qui auront la racine de l'arbre comme classe parente directe pourront être considérées comme des données marginales (outliers).

6.3.2 Visualisation hypertree de DBHC

La technique de la visualisation par arbre hyperbolique (hypertree) a été initialement créée pour faciliter la représentation des données hiérarchiques arborescentes à la structure complexe, telles que les systèmes de fichiers (cf. section 3.5.2). Un de ses points forts est celui de donner une idée intuitive de l'organisation globale des données, même si ces dernières sont nombreuses. Le principe directeur de l'approche originale que nous proposons ci-après est, dans

Algorithme 12 Algorithme de classification hiérarchique orienté par la densité (DBHC)

Entrée :

- Un ensemble de données
- n : le nombre maximal attendu de niveaux hiérarchiques

Construire la matrice des distances entre les classes de M_0

Extraire de cette matrice deux valeurs :

- $Dmin$: la distance minimale,
- $Dmax$: la distance maximale.
- Définir le seuil T comme : $T = \frac{Dmax-Dmin}{n}$

/* Construire les nouveaux parents de classes dans les différents niveaux */

pour chaque niveau V_i **faire**

$M_{i+1} = \phi$

$Dmin = Dmin + T$

si ($Dmin > Dmax$) **alors** Arrêt

pour chaque $C_j \in M_i$ **faire**

Construire une classe parente initiale $C_j^* : Parent(C_j)$

pour chaque $C_k \in M_i \setminus C_j$ **faire**

si $\|C_j - C_k\| \leq Dmin$ **alors**

Associer C_k à C_j^*

fin si

fin pour

le profil de C_j^* est le profil moyen des classes-filles associées.

$M_{i+1} = M_{i+1} \cup \{C_j^*\}$

fin pour

Procédure 1 (M_{i+1})

Procédure 2 (M_{i+1})

A cette étape, M_{i+1} contient toutes les classes parentes, ainsi que les données qui n'ont aucun parent.

Construire la matrice de distances entre les éléments de M_{i+1} .

fin pour

Sortie : un arbre hiérarchique.

Algorithme 13 Procédure 1 : élimination de classes parents répétées

Les classes parentes initiales répétées dans le nouveau niveau (i.e. les classes initiales du nouveau niveau qui partagent les mêmes classes-filles) sont résumées en une seule classe parente initiale.

un premier temps, d'utiliser l'algorithme hiérarchique basé sur la densité que nous avons décrit précédemment, afin d'organiser la visualisation de données initialement non structurées selon une structure arborescente, et puis dans un second temps, de visualiser cette structure sous la forme d'un arbre hyperbolique. Cette approche présente l'avantage déterminant de permettre de conserver, sur le plan de représentation, une bonne idée de l'organisation originale des données, sans pour autant introduire de surcharge cognitive. Elle peut aussi bien s'appliquer sur des données originales que sur les résultats préalables d'une méthode de classification. Dans le cas précis de notre expérimentation, nous l'appliquerons plus particulièrement à la visualisation des classes obtenues par la méthodes NG.

Algorithme 14 Procédure 2 : éviter les classes recouvrantes

Dans ce cas les classes fils qui ont plus d'un seul classe parente initiale vont être associées à un seul classe parent :

1. recalculer le profil de chaque classe parente initiale, sans ses classes fils associées qui sont partagées.
 2. associer chaque classe fils à la classe parente initiale dont le profil est le plus proche : ladite classe parente initiale devient alors la classe parente finale.
 3. si les classes parentes initiales d'une classe fils ont les mêmes distances à leur classe fils alors associer les classes fils partagées à l'une entre elles d'une manière arbitraire.
 4. si une classe parente initiale a seulement un classe fils, alors éliminer cette première.
-

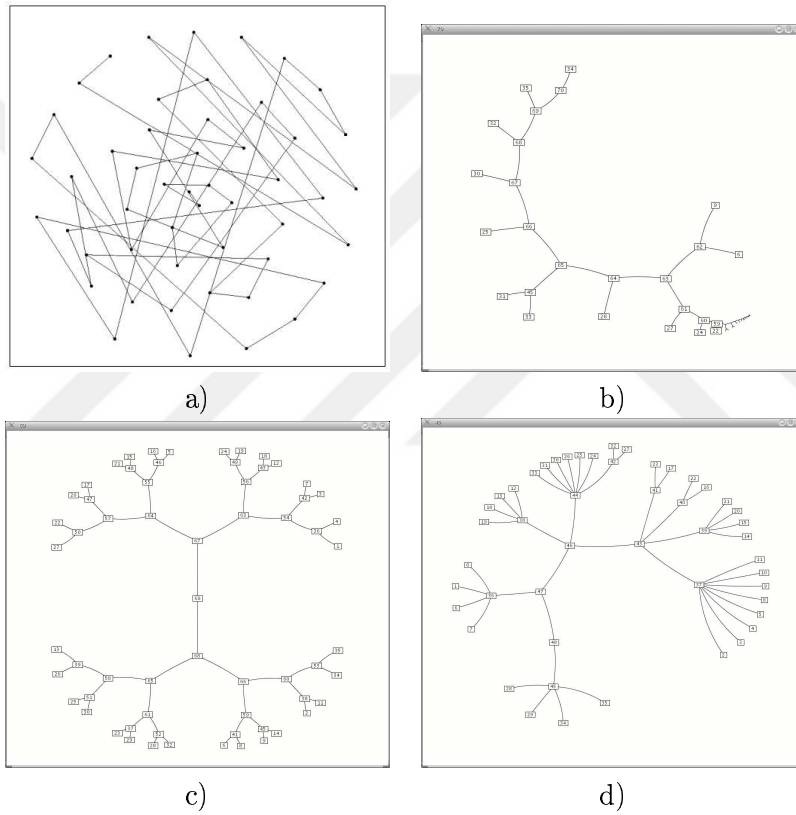


FIG. 6.8 – a) Visualisation des relations de voisinage dans 50 neurones (classes de Neural Gas). La complexité de la structure obtenue est trop élevée et ne permet aucune interprétation. Ce problème est typique d'une telle visualisation par projection non linéaire. b) Méthode hiérarchique ascendante/descendante. c) Méthode de voisins réciproques. d) Méthode de DBHC.

La figure 6.8 présente une comparaison entre notre approche (DBHC+Hypertree) (cf. figure 6.8 (d)), la projection non linéaire de Sammon (cf. figure 6.8 (a)), et, deux autres méthodes, à savoir, la méthode hiérarchique classique (cf. figure 6.8 (b)) et la méthode des voisins réciproques (cf. figure 6.8 (c)). Cette figure illustre le fait que les deux méthodes de classification hiérarchiques considérées, ainsi la méthode de Sammon, ne préservent aucunement les infor-

mations sur la densité des données originales et sur leur topologie, dans le cas où la dimension de l'espace de description est élevée. Ce problème survient par construction dans le cas des méthodes hiérarchiques précitées ; il apparaît quand la taille de l'espace de description des données devient grande, et plus particulièrement quand la distribution des données est creuse, comme dans le cas de la méthode de projection non linéaire. Il faut cependant noter que une méthode de classification hiérarchique de type lien simple (cf. section 2.2.2) ne présenterait pas ce défaut. Cette dernière serait cependant sujette à un autre défaut important qui est celui de produire des effets indésirables de chaînage lors de la construction des classes. La méthode que nous proposons représente donc naturellement le meilleur compromis.

Une autre remarque importante est liée à la gestion de liens matérialisant des distances entre les données originales. Cette technique est habituellement utilisée pour représenter des graphes. Nous avons illustré son application dans le cas de la méthode de projection non linéaire de Sammon, en matérialisant par des liens la relation de plus proche distance entre les données dans leur espace de description original. On peut cependant remarquer cet type de représentation introduit une forte surcharge cognitive, comme nous l'avons déjà illustré lors de l'utilisation de la méthode BibTechMon. La gestion de liens ne s'avère pas indispensable dans la cas de la méthode que nous proposons. Une telle méthode, basée sur la densité, permet en effet de représenter des graphes virtuels, c'est-à-dire, de matérialiser les graphes sans en représenter les liens, même si une certaine partie de l'information concernant ces derniers est évidemment perdue par la compression de la représentation.

6.3.3 Application à la webométrie

Notre expérimentation consiste à construire des arbres hyperboliques pour visualiser les résultats des analyses de gaz produites dans le cadre de l'analyse des données du web (cf. section A.3.2). Les deux points de vue considérés sont le point de vue **Cities** et le point de vue **UCodes**. Le but de notre expérimentation est de faire apparaître les groupes d'interaction de laboratoires, entre les villes, et les thèmes, respectivement. L'arbre résultant du point de vue **UCodes** est présenté aux figures 6.9 et 6.10. La figure 6.9 correspond à l'arbre dans sa position originale, où la racine est au centre, et la figure 6.10, à la position obtenue après le développement d'une branche, par l'intermédiaire du mécanisme focus+contexte. L'arbre résultant du point de vue **Cities** est présenté, en position centrée, à la figure 6.11.

L'examen de l'ensemble de ces figures montre qu'il est facile d'identifier, de manière hiérarchique, les sous-réseaux de coopération, géographiques ou thématiques, entre les laboratoires. En effet, chaque niveau hiérarchique décrit un niveau de coopération de portée différente, les réseaux les plus actifs se situant aux niveaux les plus bas.

Les deux exemples d'arbre générés montrent que l'approche que nous proposons représente un outil particulièrement utile pour l'interprétation des résultats d'une analyse de données. Son avantage principal est celui de donner une vision globale de l'interaction entre les données, tout en supprimant la surcharge cognitive de l'analyse d'interaction classique de l'analyse à base de graphes, telle que celle proposée par la méthode BibTechMon, décrite précédemment (cf. section 3.4.2). Au vu de la complexité des graphes obtenus par de telles méthodes, nous avons en effet montré qu'un filtrage sur les liens, permettant d'isoler des sous-graphes, s'avérerait souvent indispensable pour mener une analyse à bien. La méthode que nous proposons permet de s'affranchir de ces opérations en conservant naturellement l'information la plus importante pour l'analyse, tout en la décomposant en plusieurs niveaux de généralité.

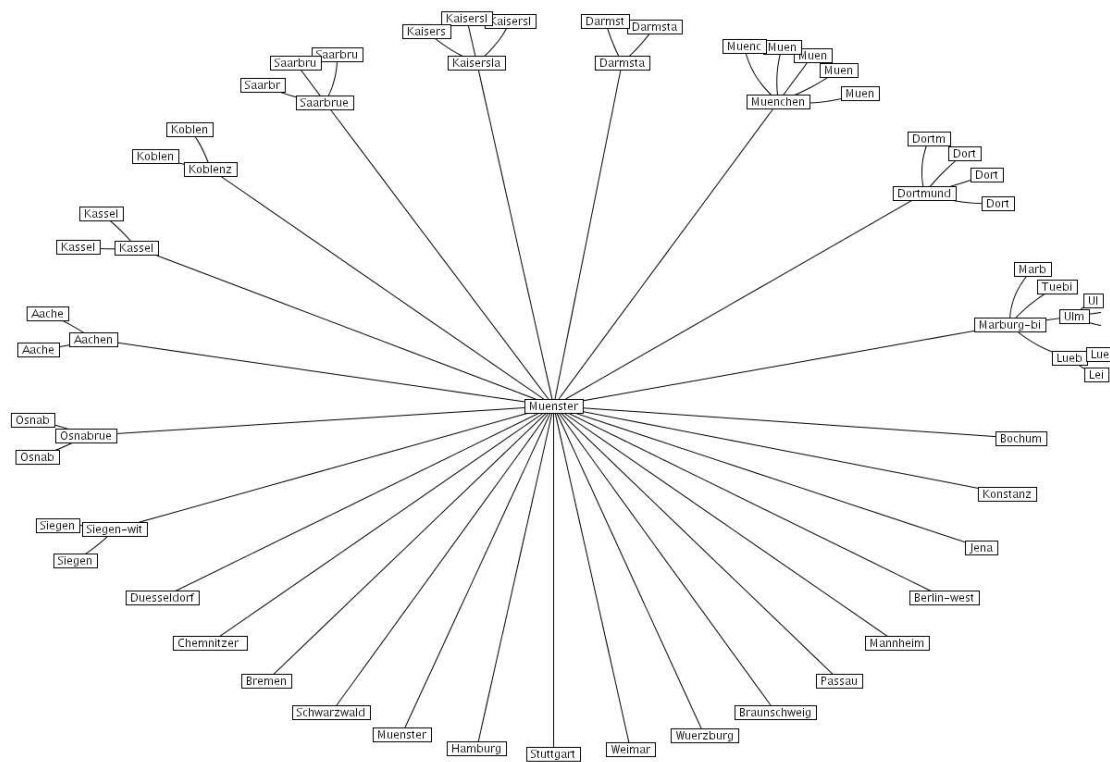


FIG. 6.11 – Arbre hyperbolique des villes allemandes possédant des laboratoires de recherche. La ville qui a été considérée comme centrale pour cette analyse est la ville de **Münster**. Cela signifie que cette recherche en computer science est particulièrement développée en Allemagne du nord.



Chapitre 7

Synthèse

Nous avons démontré objectivement la supériorité de l'analyse orientée par les points de vue par rapport à l'analyse globale pour les tâches de classification et d'extraction de connaissances. Dans ce chapitre nous présentons donc une synthèse de notre travail en donnant une vue générale de notre paradigme d'analyse de données orientée par les points de vue multiples (Multiple Viewpoint Data Analysis ou MVDA) (cf. figure 1). Ce paradigme couvre deux domaines complémentaires, à savoir celui de **l'analyse de données** proprement dite, et celui de **la fouille de données**. Il intègre également des méthodes de **visualisation** spécifiques.

Analyse de données

L'analyse de données représente la fonction centrale du paradigme MVDA. Cette analyse comprend les phases suivantes :

1. *Description des données par des points de vue multiples* : Le principe de partitionnement de la description des données en points de vue multiples permet à la fois d'éliminer le bruit inhérent à une analyse globale et de combiner différents types d'analyse pour en enrichir l'interprétation. Ce partitionnement est généralement piloté par un expert du domaine analysé, en fonction des besoins spécifiques de l'analyse (cf. section 1.2.3). Un point de vue supplémentaire restituant en description globale des données peut éventuellement être utilisé à titre de référence (cf. section 4.2.4).
2. *Prétraitement* : Si les données représentent des documents, chaque document sera décrit par un ensemble de vecteurs de termes (cf. section 1.2.2), dans lequel chaque vecteur sera associé à un point de vue différent. Chaque terme d'un vecteur pourra lui-même être pondéré en utilisant des stratégies de pondération telles que la stratégie $TF * IDF$ ou l'entropie (cf. section 1.2.1).
3. *Classification numérique neuronale non supervisée* : Différentes méthodes de classification neuronales non supervisées (SOM, NG ou GNG) (cf. section 2.3) pourront être choisies, en fonction des besoins spécifiques. Comme nous avons démontré la supériorité de la méthode NG relativement aux deux autres méthodes, à savoir les méthodes SOM et GNG (cf. section 4.2.4), nous avons décidé d'utiliser prioritairement cette méthode dans nos expérimentations, ce qui n'empêche nullement l'utilisation des autres méthodes dans des

cas spécifiques. La méthode SOM pourra par exemple être utilisée si une visualisation cartographique est nécessaire (cf. section 6.2).

4. *Génération de plusieurs classifications comprenant un nombre croissant de classes* : La méthode de classification sélectionnée est lancée plusieurs fois, ceci avec un nombre croissant de classes, et les différents modèles générés sont mémorisés.
5. *Évaluation globale des classifications* : Chacun des modèles générés est évalué par une des deux approches que nous avons proposées. La première approche utilise les mesures de **Rappel** et de **Précision** (cf. section 4.2). La deuxième approche utilise les mesures d'**Entropie Moyenne de Rappel** et d'**Entropie Moyenne de Précision** (cf. section 4.3).
6. *Détermination du modèle* : Dans le cas de l'application de la première méthode d'évaluation susmentionnée, le modèle optimal sera sélectionné par l'utilisation d'un algorithme spécifiquement basé sur le compromis (cf. section 4.2.3). Dans le cas de l'application de la deuxième approche, le modèle optimal sera sélectionné par le biais d'analyses plus générales, également basées sur le compromis (cf. section 4.3.2).
7. *Modèle multi-topographique étendu* : Si la topographie choisie est de type SOM, le modèle MultiSOM (cf. section 2.3.7), qui représente une extension du modèle SOM, sera appliqué ; chaque point de vue sera alors représenté par une carte SOM optimale spécifique. Si la topographie choisie est de type gaz, alors le modèle MultiGAS (cf. chapitre 5), qui représente une extension des modèles NG/GNG, sera appliqué ; chaque point de vue sera alors représenté par un gaz différent. Les trois mécanismes de base inhérents au modèle multi-topographique étendu pourront ensuite être utilisés :
 - (a) *Communication inter-topographies (analyse multi points de vue)* : Les modèles optimaux de type carte/gaz associés aux différents points de vue sont liés dynamiquement par un mécanisme de communication inter-topographies, lui-même basé sur un modèle de réseau d'inférence bayésien (cf. section 2.3.7 et section 5.2.1). Ce mécanisme permet d'opérer interactivement des déductions ciblées impliquant plusieurs analyses locales (cf. section 6.2.2). Il permet également de restituer une vue générale précise de l'analyse, dans le cas où un point de vue global, moins précis, n'est pas utilisé.
 - (b) *Comparaison des vues* : Elle est basée sur une méthode d'évaluation de la cohérence de propagation entre les topographies (cf. section 5.3.1). Ce mécanisme permet d'opérer dynamiquement des comparaisons globales entre des analyses différentes (cf. section 5.3.2).
 - (c) *Généralisation (analyse multi niveaux)* : Un mécanisme de généralisation en cascade, et en ligne, qui se différencie entre les cartes SOM (cf. section 2.3.7) et les gaz (cf. section 5.4.1), peut être appliqué sur le modèle optimal associé à chaque point de vue. Ce mécanisme permet d'opérer des choix de focus lors des analyses. Ces choix peuvent s'opérer dynamiquement car ce mécanisme de généralisation est compatible avec le mécanisme de communication inter-topographies. Il peut également être utilisé pour la sélection des règles extraites lors des opérations de fouille

de données (cf. section 5.5.2).

8. *Évaluation locale des classes* : Des évaluations locales de **Rappel** et de **Précision** peuvent être opérées indifféremment sur les classes d'un modèle optimal, ou sur celles de ses généralisations (cf. section 5.5.1). Ces évaluations permettront d'identifier les propriétés propres des classes concernées (pour plus de détail sur les propriétés propres, (cf. équation 4.3). Elles serviront ensuite de base aux opérations de fouille de données.

Fouille de données

Cette phase représente une phase d'exploitation des résultats produits par la phase d'analyse. C'est durant celle-ci que les règles d'association seront extraites, puis validées.

1. *Concept symbolique* : dans la section 4.2.2, nous avons démontré théoriquement que, si les deux valeurs de **Rappel** et de **Précision** atteignent conjointement la valeur unité pour l'ensemble de propriétés d'une classe donnée, cette classe représente alors un concept de treillis de Galois. Cette propriété peut être considérée comme une propriété de référence pour les tâches d'extraction de règles d'association qui sont présentées ci-après. De fait, un modèle numérique qui vérifie cette condition, peut être localement assimilé à un modèle symbolique.
2. *Combinaison de propriétés* : Toutes les propriétés propres dont la valeur de **Rappel** atteint la valeur unité ou dont les valeurs conjointes de **Rappel** et de **Précision** atteignent cette même valeur sont combinées deux à deux. Cette combinaison s'opère à l'intérieur d'une même classe, dans le cas d'une extraction mono-vue, ou entre des classes appartenant des points de vue différents, dans le cas d'une extraction multi-vues. Nous avons montré que ce type d'extraction orienté par les classes permet de diminuer sensiblement la complexité de calcul (cf. section 5.5).
3. *Extraction des règles d'association* : A partir de ces combinaisons, les règles d'association sont extraites en appliquant l'algorithme 10, dans le cas de mono-vue, et l'algorithme 11, dans le cas multi-vues, (cf. section 5.5).
4. *Validation des règles* : Toutes les règles extraites peuvent être validées par la mesure de confiance, ainsi que par tout autre type de mesure (cf. section 1.4.4). Le problème de la validation des règles extraites par les modèles numériques reste cependant un problème ouvert.

Visualisation

Toute tâche d'interprétation des résultats d'analyse d'une méthode de classification, ainsi que de ceux d'une méthode de fouille, nécessite de disposer de méthodes de visualisation puissantes et interactives. Deux méthodes de visualisation différentes correspondant à ces critères pourront être appliquées :

1. *Visualisation topographique* : Elle correspond, soit à la visualisation directe des résultats produits par une carte SOM associée à un point de vue donné, soit à la visualisation des résultats de l'interaction dynamique entre des cartes associées à plusieurs points de vue différents. Cette méthode permet également l'interprétation précise des résultats obtenus par le mécanisme de comparaison des vues (cf. section 6.2).
2. *Visualisation hyperbolique (hypertree)* : Cette méthode combine un algorithme original de classification hiérarchique, basé sur la densité, avec un mécanisme de visualisation par arbre hyperbolique (cf. section 6.3). Dans notre cas, elle est spécifiquement exploitée pour la visualisation des résultats des méthodes NG/GNG.

Appendice

Nous avons basé notre approche sur les méthodes neuronales non supervisées, à savoir, les méthodes SOM, NG et GNG. Cependant, il est à noter, comme nous l'avons déjà mentionné précédemment, que celle-ci pourrait parfaitement s'appliquer à d'autres méthodes de classification.

De plus, nous avons focalisé nos expérimentations sur deux types d'analyse, à savoir, l'analyse des brevets, et, l'analyse des données du web (cf. annexe A). Le paradigme que nous proposons est cependant suffisamment général pour pouvoir être appliqué à profit à tout autre type d'analyse, y compris à des analyses portant sur des données se trouvant hors du champ documentaire.

Conclusion générale

Conclusion

Le but principal de ce travail était de valider objectivement la supériorité de l'approche d'analyse numérique orientée par les points de vue par rapport à l'approche globale habituellement utilisée pour l'analyse et la fouille de données. Bien que cette supériorité, notamment en terme de flexibilité, ait déjà été mentionnée subjectivement par des experts, il fallait la démontrer objectivement. Pour mener à bien cette validation, il fallait prioritairement introduire de nouveaux concepts et de nouveaux outils pour l'analyse et l'interprétation des résultats des méthodes de classification. En partant d'une première approche déjà proposée pour l'analyse basée sur les points de vue, il s'agissait ensuite mieux comprendre ses principes de comportement pour pouvoir améliorer ses modes opérationnels et pour pouvoir les généraliser. Pour exploiter les résultats d'analyse de l'approche orientée par les points de vue, et même ceux de l'approche globale, dans le cadre plus spécifique de la fouille, il fallait également introduire une nouvelle méthodologie pour l'extraction de connaissances qui ne possède pas les inconvénients des techniques symboliques, en terme de complexité de calcul. Enfin, semblait nécessaire dans le cadre général de la visualisation des résultats d'analyse des données fortement multidimensionnelles, il semblait nécessaire d'introduire de nouvelles techniques de visualisation fournissant une vision globale sans surcharge cognitive sur les résultats tout en préservant la possibilité de comprendre les interactions spécifiques entre les données.

L'état de l'art nous a permis d'isoler les techniques les plus intéressantes pour l'analyse multi-vues des données documentaires en tenant compte de contraintes spécifiques liées à ce type de données et à leur représentation usuelles en machine, à savoir la représentation vectorielle. Il s'agissait donc de privilégier les techniques d'analyse et de visualisation pouvant s'appliquer efficacement sur des espaces fortement multidimensionnels ultra-creux. Ainsi, pour la combinaison des vues, nous avons retenus le principe très puissant de l'inférence bayésienne, qui permet de traiter de manière homogène des sources multiples. Pour la classification, nous avons celui des méthodes neuronales topographiques à fort degré de liberté, telles que les gaz de neurones, à la fois efficaces et stables dans le traitement des distributions complexes. Pour la visualisation, nous avons retenus le principe de la représentation hyperbolique qui permet très naturellement de s'affranchir de la surcharge cognitive. Nous avons mis l'accent sur l'absence de critères satisfaisants pour la comparaison des méthodes de classification numériques. En effet, les méthodes d'évaluation usuelles des classifications, telle que l'inertie inter-classes et intra-classe, sont souvent insuffisantes, car peu informatives et biaisées.

Nous avons mené à bien nos expérimentations en utilisant deux types de corpus différents dans nos expérimentations, à savoir un corpus de notice de brevets, et deux corpus différents des données web décrites à la fois par un contenu textuel et par des profils de liens hypertextes.

Pour mener à bien nos travaux, nous avons donc prioritairement mis en place des mé-

canismes d'évaluation originaux basés sur les rapprochements possibles entre les méthodes de classification numériques et les méthodes de classification symboliques, en démontrant théoriquement ces rapprochements. Les mesures de qualité que nous avons définies présentent la propriété très intéressante d'être indépendantes de la méthode de classification utilisée, ainsi des variations de son mode opératoire. Elles nous ont ainsi permis de démontrer objectivement la supériorité de l'approche orientée par les points de vue sur l'approche globale. Ils nous ont également permis de démontrer l'avantage en terme de précision de résultat d'une méthode de classification non recouvrante relativement à une méthode de classification recouvrante. En outre, elles ont permis de démontrer l'avantage des méthodes de gaz neuronal de type topologie libre par rapport à celles de type topologie fixe. Elles se sont montrées efficaces pour résoudre le problème de la recherche du modèle de classification optimal.

Nous avons ensuite abordé le problème de l'évaluation des classifications numériques de manière plus générale en nous intéressant à la distribution des étiquettes dans les classifications associées à des étiquettes multiples. Nous avons ainsi proposé des mesures de qualité de classification plus générales que celles proposées précédemment. Ces mesures, basées sur la théorie des probabilités, se sont avérées plus efficaces et plus stables que les précédentes pour la sélection du modèle optimal. Elles se sont montrées également efficaces pour contrôler le nombre des cycles d'apprentissage d'un algorithme de classification. Ces deux permettent de résoudre les deux problèmes principaux de la méthode de gaz neuronal Neural Gas, qui s'avère par ailleurs la plus efficace en terme de pouvoir de synthèse, à savoir le choix du nombre de classe et le critère d'arrêt de l'apprentissage. Nous avons ensuite proposé, en nous basant sur ces dernières mesures, une approche qui combine l'analyse de classes et l'analyse d'étiquette. Cette approche a permis d'identifier des catégories d'étiquettes, elles-mêmes utiles à la détermination des clusters pertinents. Nous avons montré, par une première expérimentation, que cette approche est susceptible de représenter une étape importante pour des tâches d'extraction de la connaissance basées sur une approche de classification. Un de ses avantages primordiaux est qu'elle ne nécessite pas l'intervention d'un expert.

Nous avons ensuite proposé un modèle générique d'approche pour l'analyse basée sur les points de vue multiples en nous attachement au développement de trois de ces composants essentiels, à savoir un mécanisme de communication inter-vues, un mécanisme de comparaison des vues et un mécanisme de généralisation de vues. Nous avons basé le mécanisme de communication entre les vues sur l'analogie avec un modèle de réseau d'inférence bayésien, en étendant ainsi le mécanisme de communication déjà proposé par le modèle MultiSOM. Une expérimentation a permis de montrer la valeur ajoutée du nouveau mécanisme, en termes de généralité et de performances. Nous avons ensuite proposé un mécanisme original de comparaison des vues qui évalue la cohérence entre les vues. Nous avons montré, par une nouvelle expérimentation, que ce mécanisme très puissant permet la comparaison dynamique d'analyses basées sur des critères différents. Nous avons ensuite étendu le mécanisme de généralisation de vues proposés par le modèle MultiSOM à un mécanisme à plus large champ, applicable aux méthodes de classification à topologie libre, qui représentent elles-mêmes la majorité des méthodes existantes. Nous avons démontré son efficacité en le couplant avec notre mécanisme de comparaison de vues.

Nous avons également proposé une approche originale pour l'extraction de connaissances, sous forme de règles d'association, à partir d'une méthode de classification numérique non supervisée. Cette approche est basée les mesures de qualité que nous avons précédemment définies. Les règles d'association sont donc extraites en utilisant les deux mécanismes du modèle multi-vues : le mécanisme de généralisation est utilisé pour une extraction relative un point

de vue élémentaire et le mécanisme de communication est utilisé pour une extraction faisant intervenir des points de vue multiples. Les résultats de nos expérimentations ont montré que cette technique permettait de s'affranchir des problèmes relatifs à l'inflation des règles, au choix des règles et au temps de calcul, inhérents aux modèles symboliques.

Nous avons finalement proposé une nouvelle approche de visualisation qui combine l'approche de clustering basée sur la densité avec une visualisation par arbre hyperbolique. Elle permet notamment de visualiser, sans surcharge cognitive, des résultats d'analyse à différentes échelles de précision. Nous avons illustré son fonctionnement par une première expérimentation.

Le résultat principal de notre travail est celui d'avoir proposé un nouveau paradigme d'analyse de données orientée par les points de vue multiples (Multiple Viewpoint Data Analysis ou MVDA). Ce paradigme couvre deux domaines complémentaires, à savoir celui de l'analyse de données proprement dite, et celui de la fouille de données.

Un des résultats secondaires important de notre travail est celui d'avoir ouvert de nouvelles voies dans l'évaluation des méthodes de classification numériques en introduisant de mécanismes d'évaluation stables et puissants, indépendants des méthodes de classification proprement dites. Nous avons ensuite défini, en collaboration avec un autre chercheur, le cadre général de l'étude de la distribution des étiquettes dans les classifications associées à des étiquettes multiples. Nous pensons que ce cadre représente un nouveau champ fertile de recherche avec de nombreuses applications la validation, l'évaluation des classifications numériques et de leurs résultats.

Nous pensons également que les techniques de visualisation que nous avons proposées représentent un excellent compromis pour mener à bien de manière parallèle des tâches de fouilles et d'analyse de données. En effet ces techniques répondent à de nombreux problèmes de techniques traditionnelles. En effet, elles traitent à la fois les problèmes des représentations à base de graphes et ceux liées à la représentation des données multidimensionnelles sur un plan d'interprétation. Elles permettent de plus d'exploiter les résultats des méthodes de classification très performantes, plutôt que d'utiliser des méthodes moins performantes qui intègrent leur propre fonction de projection.

Nous avons montré que notre nouveau paradigme MVDA permettaient de traiter, de manière non supervisée, des problèmes concrets très complexes, voire jusqu'ici impossibles à résoudre, notamment dans le cadre de la webométrie. Nous avons démontré son efficacité en en faisant plusieurs expérimentations inédites dans ce domaine en utilisant l'ensemble des mécanismes du paradigme MVDA et en faisant appel à la méthode Neural Gas. En utilisant la comparaison de vue nous avons démontré, pour la première fois, qu'il était possible d'effectuer une identification précise et automatique du comportement de référencement dans un domaine scientifique par l'analyse de la communication des institutions académiques, telles que les laboratoires de recherche sur le web. Nous avons également illustré l'intérêt général de la communication entre les vues pour croiser des analyses de liens avec des analyses de contenu. Nous avons également montré, que l'extraction de règles pouvait permettre, dans le même contexte, d'isoler des connaissances fines comme des explications thématiques précises sur les liens matérialisant des relations de dépendances entre des acteurs. Toujours dans le même contexte, nous avons montré que la visualisation hyperbolique permettait, par exemple, de revenir à une vision plus globale, en matérialisant les différents niveaux de coopération entre les laboratoires de recherche.

Perspectives

Nous comptons à l'avenir poursuivre nos études dans trois directions complémentaires, à savoir l'analyse de données, la fouille de données et la visualisation.

En ce qui concerne l'analyse de données, les mécanismes de base de l'approche MVDA peuvent encore être optimisés. Premièrement, le modèle bayésien que nous avons défini prend en considération les choix de modalités (i.e. ce modèle fonctionne dans l'espace discret du choix des modalités). Nous pensons donc le rendre de plus général en l'étendant à un modèle de décision contenu. Deuxièmement, une optimisation de notre mécanisme de comparaison des vues sera nécessaire pour amener l'interprétation de résultats. Il s'agit d'introduire la qualité locale de classes pour un modèle de classification donnée, mais également de croiser les résultats de l'analyse locale avec ceux de l'analyse globale. Enfin, le mécanisme de la généralisation peut être amélioré en tenant compte de la densité des classes.

En ce qui concerne la fouille de données, nous comptons par la suite mettre en place de nouveaux mécanismes d'évaluation et de valider les règles d'association extraites en utilisant les techniques classiques de calcul de pouvoir d'information. En effet, ce type d'évaluation nous permettrait de mieux illustrer et montrer l'avantage de la génération des règles à partir de l'approche numérique. De plus, Nous n'avons pas encore généré de règles complexes dont les prémisses et les conclusions représentent des ensembles des propriétés. Nous comptons notamment exploiter la notion de motifs générateurs d'un motif fermé (cf. section 1.4.3 pour plus de détail sur les types de motifs), pour extraire de telles règles, qui s'avèrent naturellement plus informatives.

Par ailleurs les étiquettes associées aux données peuvent être considérées comme des propriétés de ces données. Par conséquent, nous pensons à définir de nouveaux types des règles d'association prenant en compte les catégories d'étiquettes qu'il est possible d'associer à leurs composants, ceci grâce à nos méthodes d'analyse de la distribution des étiquettes. Selon leur catégorie résultante, ces règles pourraient par ailleurs servir de base à de nouvelles stratégies d'étiquetage des classes.

Enfin, pour la visualisation, nous comptons à optimiser notre méthode hiérarchique DBHC. Il s'agit d'une optimisation de nombre de niveaux générés dans l'arbre hyperbolique. De plus, nous pouvons appliquer notre algorithme de manière originale sur les résultats des cartes SOM, de manière à déterminer le degré de densités de ses classes, et, d'isoler ainsi les classes qui correspondent aux données marginales. L'introduction de la communication entre les arbres hyperboliques, à l'image de la communication entre les cartes SOM dans le modèle MultiSOM, s'avère également être une étape centrale pour la montée en puissance de notre approche.

Nous comptons poursuivre nos expérimentations en webométrie, mais nous pensons que notre paradigme est suffisamment général pour être applicable à d'autres domaines de recherche, et à des types de données autres que les données documentaires, ceci au prix d'adapter nos mesures de qualité à ces types de données. Nous nous intéressons en particulier aux domaines où le croisement de sources de connaissances multiples représente un problème central. Nous pensons en particulier au domaine de la bioinformatique, et plus spécifiquement à celui de l'interprétation des résultats expérimentaux des puces à ADN. Nous pensons finalement que notre paradigme MVDA représente également un outil puissant pour les analyses diachroniques en général. De nombreuses portes restent donc ouvertes

Annexe A

Description des données

Sommaire

A.1	Introduction	173
A.2	Analyse des brevets	173
A.2.1	Définition des points de vue	173
A.2.2	Multi-indexation des brevets	174
A.3	Analyse du Web	176
A.3.1	Analyse du site web du Laboratoire d'Informatique de l'Université de Cambridge	176
A.3.2	L'analyse de sites des laboratoires de recherche européens	177

A.1 Introduction

Nos expérimentations et nos évaluations ont été réalisées sur deux type de données. Le premier type est un ensemble de 1000 notices de brevets dans le domaine de l'ingénierie des huiles moteur. Le deuxième type correspond à des données extraites du web dans le cadre du projet européen EICSTES (European Indicators, Cyberspace and the Science - Technology - Economy System)¹³.

A.2 Analyse des brevets

Notre premier corpus regroupe un ensemble de 1000 notices de brevets issus du domaine de l'ingénierie des huiles moteur. Ces brevets ont été publiés durant l'année 1999.

A.2.1 Définition des points de vue

L'application d'une méthode de classification (par exemple le modèle multi-topographique (cf. chapitre 5)) sur ces notices de brevets dans le domaine des huiles moteur permet de répondre directement à des questions aussi variées que :

1. Quelles sont les relations de coopération/dépendance entre les déposants?,
2. Quels sont les avantages des différents types d'huile moteur?,

¹³<http://www.eicstes.org/>

3. Un déposant donné travaille-t-il sur une technologie précise, pour quels avantages et pour quelle utilisation ?,
4. Quelle technologie est utilisée par tel déposant et n'est pas utilisée par tel autre ?,
5. Quels sont les avantages majeurs de tel composant intervenant dans la fabrication des huiles ? Ces avantages ont-ils été mentionnés dans tous les brevets utilisant ce composant ?.

Ces questions correspondent à différents domaines sémantiques qui peuvent être associés aux différents champs des notices. La structure de ces notices nous permet donc de définir différents points de vue qui permettront de répondre aux questions ci-dessus.

Les différents points de vue qui ont été définis, sont les suivants :

1. **Déposants** : les déposants, correspondent à un champ spécifique,
2. **Titres** : les titres, décrivent précisément le contenu des brevets,
3. **Utilisations** : le champ «Utilisations», sous-ensemble du résumé, décrivant les utilisations potentielles du brevet,
4. **Avantages** : le champ «Avantages», sous champ du résumé décrivant les apports de ce brevet par rapport aux technologies existantes.

Un cinquième point de vue «le point de vue global», représente la combinaison de tous les points de vue, il est également considéré pour effectuer notre comparaison entre une approche globale et une approche orientée par les points de vue.

Durant la phase de l'analyse, une des préoccupations principales de l'expert a été celle de pouvoir utiliser chaque point de vue séparément pour répondre à des questions monodomaines (telles que les questions 1,2), tout en maintenant la possibilité d'une communication multi points de vue dans le but de répondre à des questions multidomaines (telles que les questions 3,4,5).

A.2.2 Multi-indexation des brevets

Pour chaque point de vue, un index est construit en analysant le contenu des champs des notices correspondant. Cette construction peut être découpée en trois phases :

1. Phase 1 : la structure des résumés des brevets est analysée de manière à en extraire les sous-champs correspondant aux points de vue **Utilisations** et **Avantages** .
2. Phase 2 : un index brut est constitué pour chaque point de vue en utilisant un outil d'analyse lexicologique. L'outil choisi, développé en interne, nous permet d'extraire des termes simples ou des expressions composées en se basant sur une terminologie normalisée et sur les variations syntaxiques possibles, tout en éliminant les tournures langagières usuelles.
3. Phase 3 : la normalisation des index associés aux points de vue **Titres**, **Utilisations** et **Avantages** consiste en la sélection d'un meilleur représentant dans chaque groupe de termes/expressions composées décrivant le même concept (par exemple les expressions «fabrication des huiles» et «ingénierie des huiles» peuvent être toutes les deux assimilées à la seule expression «ingénierie des huiles»). La normalisation de l'index associé au point de vue **Déposants** est réalisée de manière similaire en considérant qu'un déposant était susceptible d'apparaître sous plusieurs noms différents dans les brevets. Après la normalisation des index, les brevets ont été réindexés selon chaque point de vue, en

	Déposants	Titres	Utilisations	Avantages
Nombre de documents indexés	1000	1000	745	624
Index brut : nombre de termes	73	605	252	231
Index final : nombre de termes	32	589	234	207

TAB. A.1 – Tableau résumé des caractéristiques résultantes de brevets

s'appuyant sur ces index. La figure A.1 montre un exemple de notice de brevet avec l'indexation obtenue.

Title:	Lubricating oil composition - includes a base oil, sulphur-containing organic molybdenum compound, one organic acid salt compound,
Patentee(s):	TONEN CORP
Abstract:	Lubricating oil composition includes a base oil, (a) a sulphur-containing organic molybdenum compound, and (b) at least one organic acid salt compound selected from (b-1) and (b-2), wherein (b-1) is an organic acid metal salt compound excluding copper carboxylate and being a metal salt of an organic acid selected from a carboxylic acid, an aliphatic sulphonic acid, an aromatic sulphonic acid, an alkyl salicylic acid and an alkyl phenolic compound in which the metal is a metal selected from 1A, 3A-7A, 8 and 1B-6B groups of the Periodic Table, and (b-2) is an ammonium salt and an amine salt compound the organic acid selected from the carboxylic acid, the aliphatic sulphonic acid, the aromatic sulphonic acid, the alkyl salicylic acid and the alkyl phenolic compound
Use:	The lubricating oil composition is used for internal combustion engines, automatic gear changers, gears,
Advantages:	The lubricating oil composition exhibits excellent friction-reducing effect over a wide range from low temperatures to high temperatures
Final indexation:	adv friction coefficient stability on a wide range of temperature ; adv friction reduction ; adv low temperature ; soc.TONEN CORP ; titre base oil ; titre lubricant composition ; titre lubricating ; titre organic acid salt ; titre sulfur-containing organo molybdenum ; use automatic transmission ; use engine oil ;

FIG. A.1 – Exemple de notice de brevet. L'indexation qui a été générée pour ce brevet est matérialisé par le contenu du champ «Final indexation». Ces termes d'indexation sont préfixés par le nom du point de vue auquel ils sont associés : «adv.» pour le point de vue **Avantages**, «titre» pour le point de vue **Titres**, «use» pour le point de vue **Utilisations**, «soc.» pour le point de vue **Déposants**.

Le tableau A.1 résume les résultats de l'indexation des brevets. Ces résultats appellent quelques remarques :

1. La taille de l'index associé au point de vue **Titres** est significativement plus importante que celle des index associés aux autres points de vue. Une analyse des index, réalisée par l'expert du domaine, a mis en évidence le fait que l'information contenue dans les titres des brevets était à la fois plus diversifiée, moins dense et plus précise que celle contenue dans les autres champs descriptifs de contenu, à savoir les sous-champs **Utilisations** et **Avantages**. Selon l'expert, le niveau de généralité plus important des sous-champs **Utilisations** et **Avantages**, qui conduit par conséquent à un index plus pauvre, peut être expliqué comme étant la conséquence d'une stratégie délibérée des déposants visant à protéger leur technologie.
2. Le nombre final de déposants (32) a été réduit de manière significative par l'expert, en comparaison de celui généré initialement par l'analyseur lexical. La majeure partie de

cette réduction n'est pas due à des variations apparaissant dans les noms de déposants, mais au fait que l'expert a décidé de centrer l'étude sur les déposants les plus importants et sur leur interrelations. Ainsi les déposants les moins importants ont été regroupés sous le même index : «Divers».

3. Seuls 62% des brevets possèdent un sous-champ **Avantages**, et 75% un sous-champ **Utilisations**. En conséquence, certains brevets ne seront pas indexés selon l'ensemble des points de vue.

Le mécanisme de pondération par $TF * IDF$ (cf. équation 1.2) est ensuite appliqué aux index des différents points de vue afin de réduire l'influence des termes les plus répandus avant l'application des méthodes de classification.

A.3 Analyse du Web

Dans cette section, nous présentons deux ensembles de données web issues du projet européen EICSTES¹⁴. Le projet EICSTES a pour objectif d'offrir de statistiques et d'obtenir des indicateurs sur le Système Européen de Science-Technologie-Economie dans l'Internet permettant de mettre en lumière des relations entre le secteur de la Recherche et Développement et les acteurs de la Nouvelle Economie. Ces indicateurs seront diffusés avec un environnement graphique convivial mettant en œuvre des nouvelles techniques de visualisation du web.

A.3.1 Analyse du site web du Laboratoire d'Informatique de l'Université de Cambridge

Ce corpus regroupe l'ensemble des pages HTML extraites du site du Laboratoire d'Informatique de l'Université de Cambridge «The Computer Laboratory, University of Cambridge» en mars 2003. Il contient 1353 pages HTML.

Définition des points de vue

D'une façon similaire aux brevets, l'application d'un modèle de classification sur les pages du site web étudié permet de répondre aux questions suivantes :

1. Quels sont les domaines de recherche de ce laboratoire ?
2. Vers quels autres laboratoires sont dirigés les liens sortant de ce site ?
3. Vers quels sites sont orientés les liens d'une équipe de ce laboratoire ?

Pour répondre à ces questions, deux points de vue ont été définis. A partir de chaque page HTML sont extraits l'information textuelle ainsi que les liens sortant. L'information textuelle est indexée à l'aide d'un analyseur lexicographique. Après la validation humaine, cet index contient 1230 termes différents. Les liens sortant sont normalisés en gardant seulement la racine de l'URL désignant le site cible du lien. Par exemple :

le lien «<http://raw.cs.berkeley.edu/texpoint/index.htm>» est transformé en :

«<http://raw.cs.berkeley.edu>». Après la normalisation, cet index a fourni 1912 racines différentes de site web. Pour chaque point de vue spécifique l'index résultant est pondéré par la fonction $TF * IDF$ (cf. équation 1.2).

¹⁴Site miroir en français : <http://eicstes.inist.fr/miroir>

	Information textuelle	Lien sortant
Nombre de pages HTML indexés	1066	1127
Index final : nombre de termes	1230	1912

TAB. A.2 – Tableau résumé des caractéristiques résultantes de donnée issues de site du Laboratoire d'Informatique d'Université de Cambridge

L'analyse de l'information disponible et des questions de veille définies ci-dessus, nous permet de définir les deux points de vue suivants :

1. **L'information textuelle** : représentée par les termes d'indexation des pages HTML.
2. **Les liens sortants** : représentés par les sites cibles des liens de chaque page HTML.

Le tableau A.2 résume les résultats de l'indexation des pages du site web ainsi leurs liens sortants.

A.3.2 L'analyse de sites des laboratoires de recherche européens

Notre troisième corpus de données est extrait de sites web de laboratoire de recherche européens. Dans le cadre du projet EICSTES, l'ensemble de données est généré automatiquement par un outil développé par Isidro F. Aguillo de CINDOC-CSIC Madrid, Espagne, il décrit les sites web des laboratoires et instituts de recherche dans les 15 pays de l'union européenne avant mars 2004. Ces données sont décrites par Arroyo et al. [Arroyo *et al.*, 2003].

Définition des points de vue

Les données sont présentées initialement sous le forme de tableaux multiples décrivant la structure des sites web. La première étape de la préparation des données correspond à la fusion des différents tableaux de description en une structure XML décrivant chaque site web. Les éléments de cette structure sont :

- URL,
- Nom de l'organisation et du laboratoire,
- Nombre de pages,
- Nombre de liens,
- Nombre de liens répétés,
- Nombre de liens externes récursifs,
- Situation géographique : le nom du pays et de la ville ainsi que le code du pays et de la ville,
- Domaine : code, libellé et autres codes associés,
- InLinks : liste de liens entrants avec l'URL et le nombre de liens venant de cet URL,
- OutLinks : liste de liens sortants avec l'URL et le nombre de liens allant à cet URL.

La catégorisation par domaine des sites web est basée sur la classification de UNESCO. Cette classification est utilisée pour assigner un domaine scientifique à un site web à partir de son contenu. Nous avons décidé dans une première étape de concentrer l'étude sur un domaine thématique et spécifique. Le code UNESCO 1203, qui correspond d'une manière globale à l'informatique est utilisé pour la sélection des sites Web. Dans une deuxième étape, nous

	UCodes	Cities	InLinks	OutLinks	In-Coupl	Out-Coupl
Nombre de documents indexés	438	438	388	386	377	385
Index final : nombre de termes	93	96	2839	2839	378	378

TAB. A.3 – Tableau résumé des caractéristiques résultantes de donnée issues de sites web de laboratoires de recherche allemands.

choisissons de concentrer l'étude sur les sites web de laboratoires de recherche allemands. Un ensemble de 438 sites web est donc choisi de cette façon.

Parmi tous les liens sortant des sites web, nous avons sélectionné ceux dont la destination est également un site de laboratoire de recherche européen (c'est-à-dire qu'il appartient à l'ensemble des sites défini par nos collègues du CINDOC-CSIC Madrid). Pour chaque lien sélectionné, nous avons commencé conservé la racine de l'URL désignant le site web de la page cible. Ceci nous permet d'étudier les relations entre sites de laboratoires.

Les liens entrants sont également calculés à partir de l'ensemble des sites de laboratoires de recherche européens (défini par le CINDOC-CSIC Madrid).

Les informations associées aux sites web choisis nous permettent de définir six points de vue différents :

1. Domaine de recherche : Codes UNESCO (**UCodes**),
2. Situation géographique : villes des laboratoires (**Cities**)
3. Liens provenant des laboratoires de recherche européens (**InLinks**),
4. Liens sortants vers les laboratoires de recherche européens (**OutLinks**),
5. Profils de couplage de laboratoires allemands basés sur la similarité de leurs InLinks (**In-Coupling**),
6. Profils de couplage de laboratoires allemands basés sur la similarité de leurs OutLinks (**Out-Coupling**).

Chaque point de vue est représenté sous forme d'une matrice comprenant les codes de sites web en lignes et les critères de point de vue spécifique en colonnes. Pour chaque point de vue des tableaux et des index complémentaires sont également construits (tableau A.3).

D'une façon similaire aux autres corpus susmentionnés dans cette annexe, l'application d'un modèle de classification sur les différents points de vue permet aux questions suivantes :

1. Quelles sont les domaines scientifiques d'une ville ou d'un laboratoire de recherche allemand ?
2. Quelles sont les villes qui partagent un domaine scientifique donné ?
3. Quel est le comportement de référencement (OutLinks) des sites d'une ville ou d'une thématique donnée ?
4. Peut-on définir une stratégie de référencement (OutLinks) des sites en fonction de leur situation géographique ou de leur thématique ?

Liste des figures

1	Paradigme de traitement de l'information orienté par les points de vue (MVDA).	5
1.1	Exemple de réseau bayésien élémentaire représentant les dépendances conditionnelles entre trois variables	16
1.2	Réseau Bayésien à deux niveaux	18
1.3	Evaluation de SRD	22
1.4	Exemple de taux de rappel et de précision	23
1.5	Les étapes du processus d'extraction de connaissances	25
1.6	Représentation du treillis de Galois $\mathcal{L} = (\mathcal{O}, \mathcal{P})$	28
1.7	Représentation du treillis de Galois $\mathcal{L} = (\mathcal{O}, \mathcal{P})$ avec ses classes d'équivalence	32
2.1	Méthode de K-means.	37
2.2	KMA : exemple d'un corpus de documents indexés par les 3 mots-clés t_1 , t_2 , et t_3 . Ces mots-clés définissent l'espace $\mathbb{R}^3$	38
2.3	Exemple de classification hiérarchique	39
2.4	a) Un ensemble de points dans $\mathbb{R}^2$, b) la tessellation de Voronoi, c) la triangulation de Delaunay.	45
2.5	Modification de la topologie initiale dans le modèle SOM suite à la présentation d'une donnée ($\bullet$).	46
2.6	Options de voisinage du modèle SOM : ficelle (1D), rectangulaire (2D), hexagonale (2D) et tridimensionnelle (3D)	46
2.7	Classification de Neural Gas pour une forme de bague.	48
2.8	Classification de Growing Neural Gas d'une distribution en «bague».	51
2.9	Exemple de structure générée par GHSOM.	51
2.10	Mécanisme de communication intercartes	53
2.11	Mécanisme de généralisation	54
3.1	Distribution du «fer à cheval».	63
3.2	Évolution du volume de sphère en fonction de nombre de dimensions.	63
3.3	Distribution en «fer à cheval» : (a) Distribution et plan principal trouvé par l'ACP	64
3.4	Projection faite par CCA de $\mathbb{R}^3$ à $\mathbb{R}^2$ de la distribution du «fer à cheval».	67
3.5	(a) deux points d'une spirale, (b) la distance euclidienne entre ces deux points et (c) la distance curviligne ou géodésique	67
3.6	Approximation de la distance curviligne à l'aide du chemin le plus court par l'intermédiaire des liens entre les centroïdes (ici la distance entre les deux centroïdes noircis)	67
3.7	CDA : Projection non-linéaire d'un «nœud de tresse» (de dimension 3 à 1)	68

3.8	Isomap : Exemple du «rouleau suisse» (à droite) et de la projection de 20000 échantillons tirés du rouleau par Isomap.	68
3.9	La base de construction du système WEBSOM. Le document est encodé en un vecteur numérique. La dimension de représentation peut être réduite pour des raisons de calcul.	69
3.10	Exemple d'accès à un document à partir d'une carte WEBSOM. L'accès à un document fait intervenir 3 niveaux : (a) La carte WEBSOM de la collection. (b) Le contenu d'un nœud de la carte. (c) Le document proprement dit (ici un article de newsgroup).	70
3.11	Vue partielle d'une carte de 12×12 neurones (classes). La carte est organisée comme une grille carrée $2D$ de neurones. Les profils de classe sont ensuite générés par l'intermédiaire d'un apprentissage compétitif non supervisé. Le point de vue choisi ici est celui des Termes d'index . Les noms des classes illustrent les thématiques qui ont été isolées par le processus d'apprentissage. Après l'apprentissage, les neurones globalement rattachés aux mêmes thématiques ont été regroupés en aires logiques en utilisant les propriétés topographiques de la carte.	72
3.12	Exemple d'exploitation interactive du mécanisme de communication inter-carte pour la veille technologique. La décision de l'analyste d'activer l'aire correspondant au déposant TONEN CORP. sur la carte des Déposants et de propager cette activité vers les cartes thématiques associées aux points de vue Utilisations , Avantages et Titres correspond à une requête transthématique virtuelle dont la formulation explicite pourrait être la suivante : «Je voudrais connaître les domaines de compétence spécifiques du déposant TONEN CORP. (concernant l'utilisation des huiles, leur composition et les avantages attendus), si ceux-ci existent». L'application MULTISOM lui permet de trouver de manière interactive que le déposant TONEN CORP. est un spécialiste de la lubrification des transmissions automatiques [flèche n°2 sur la carte] et que ce déposant a adopté pour ce type de lubrification des composés à base d'organomolybdène [flèche n°1] dont les avantages principaux sont de fournir des huiles dont le coefficient de friction est stable sur une large gamme de températures [flèche n°3]. Dans ce cas de figure, une propagation inverse depuis les thématiques cibles aurait également dû être utilisée pour vérifier que ces thématiques sont parties exclusives des domaines de compétence du déposant (ce qui est le cas). Plus la couleur des nœuds des cartes se rapproche du blanc, plus forte est leur activité résultante.	73
3.13	BibTechMon : réseau de mots basé sur les relations entre eux. Ce réseau contient 28 nœuds et 131 connexions	74
3.14	Deux types de géodésique : un diamètre passant par O et P et un arc de cercle AB orthogonal au cercle unité	75
3.15	La visualisation de l'arbre hyperbolique (Hypertree)	76
3.16	Tesselations régulières à base de triangles d'une carte hyperbolique, projetées sur le disque unité de Poincaré. La figure de gauche illustre le cas où le nombre n de triangles équilatéraux pour chaque sommet est de 7 (en dehors du bord) ; la figure de droite est construite avec $n = 10$	77

3.17	(a) Les nœuds aux sommets - dans ce cas $n_b = 8$ - des triangles équilatéraux forment le premier niveau dans la hiérarchie du H^2SOM . (b) Quand un nœud répond à un critère de croissance, un ensemble de $n_b - 3$ nœuds fils va être généré. Dans (c) la hiérarchie présentée est une hiérarchie complète à 3 niveaux, et un chemin virtuel correspondant à l'ensemble de nœuds visités par un utilisateur est accentué.	78
4.1	Évolution des mesures de qualité en considérant le nombre de classes pour le point de vue liens sortants (R = Rappel, P = Précision, Intra = inertie intra-classe, F = F-mesure, Inter = inertie inter-classes).	92
4.2	Évolution des mesures de qualité en considérant le nombre de classes pour le point de vue information textuelle (R = Rappel, P = Précision, Intra = inertie intra-classe, F = F-mesure, Inter = inertie inter-classes).	92
4.3	Distribution des termes et des liens dans les pages web : a) Nombre de descripteurs inclus dans des motifs de descripteur des documents, b) Nombre de différents motifs de descripteur inclus dans le motif de descripteur de documents, c) Nombre de couples des documents/nombre de documents indexés par motif.	93
4.4	Comparaison entre la carte 12×12 à gauche et la carte 16×16 à droite qui représentent la classification d' information textuelle . Les aires Opera Group et Rainbow Group , qui représentent le focus de la comparaison, sont spécifiquement entourées par des lignes noires épaisses.	94
4.5	Évolution des critères de qualité selon le nombre de classes pour le point de vue Utilisations en utilisant la méthode de NG-normalisé.	95
4.6	Evolution des critères de qualité selon le nombre de classes pour le point de vue Utilisations en utilisant la méthode de NG-standard.	96
4.7	Evolution des critères de qualité selon le nombre de classes pour le point de vue information textuelle en utilisant la méthode de NG.	97
4.8	Évolution des critères de qualité selon le nombre de classes pour le point de vue information textuelle en utilisant la méthode GNG.	97
4.9	Mesures d'inertie intra-classe et inter-classes.	108
4.10	Mesures de MCRE et de MCPE.	108
4.11	Mesures de MCR, de P^* et de MCP.	109
4.12	Mesures de MCRE et de MCPE : Approche basée sur la normalisation de min-max.	109
4.13	Mesures de MCRE et de MCPE : Approche basée sur la normalisation de z-score.	109
4.14	Mesures de MCRE et de MCPE : Approche basée sur la dérivation.	109
4.15	Mesures de MCRE et de MCPE.	111
4.16	Mesures de MCRE et de MCPE : Approche basée sur la dérivation.	111
4.17	Mesures de MCRE et de MCPE : Approche basée sur la normalisation de min-max.	111
4.18	Mesures d'inertie intra-classe et inter-classes.	111
4.19	Évaluation de P_{l_k} pour chaque étiquette l_k dans le modèle de classification NG.	112
4.20	Évaluation de H_{l_k} pour chaque étiquette l_k dans le modèle de classification NG.	112
4.21	Caractérisation de la catégorie de chaque étiquette l_k dans le modèle de classification NG.	112
4.22	Évaluation de R_{c_l} pour chaque classe c_l dans le modèle de classification NG.	114
4.23	Évaluation de H_{c_l} pour chaque classe c_l dans le modèle de classification NG.	114

4.24	a) Distribution des étiquettes nouvelles. b) Distribution des étiquettes implicites générales (à l'exclusion des étiquettes implicites spécifiques). c) Distribution des étiquettes implicites spécifiques.	115
5.1	La structure de réseau bayésien pour la communication inter-topographies. . . .	121
5.2	La figure illustre le comportement de mode probabiliste basé sur la relation entre le pourcentage de données activées dans les classes destination et la mesure de probabilité a posteriori générée par le processus de propagation.	124
5.3	La figure illustre le comportement du modèle de réseau bayésien d'inférence basé sur la relation entre le pourcentage de données activées dans les classes destination et la mesure de probabilité a posteriori générée par le processus de propagation.	124
5.4	Types de profil d'activités sur des topographies destination.	125
5.5	Qualité de la carte pour chaque point de vue.	128
5.6	Le mécanisme de généralisation de NG (représentation 2D de gaz est utilisée seulement pour la clarté de la figure).	130
5.7	Modèle MultiSOM : Évaluation de la mesure de cohérence de propagation pour le point de vue Utilisations ($x - y$: représente la propagation de la carte source x à la carte destination y et $x(G)$ ou $y(G)$ représente un niveau de carte généralisée).	133
5.8	Modèle MultiGAS : Évaluation de la mesure de cohérence de propagation pour le point de vue Utilisations	134
5.9	Modèle MultiSOM : Évaluation des mesures de qualité de cartes originales et leurs généralisations pour le point de vue Utilisations ($x(G)$: représente le premier niveau de la carte généralisée de x neurones).	134
5.10	Modèle MultiGAS : Evaluation des mesures de qualité des gaz originaux et leurs généralisations pour le point de vue Utilisations	135
5.11	Modèle MultiSOM : illustration de l'évolution des mesures de qualité des niveaux de généralisation pour le point de vue Utilisations (xG : représente le $x^{\text{ième}}$ niveau de la carte de généralisation).	135
5.12	Les résultats d'extraction de règles pour le point de vue Utilisations (Algorithme d'extraction 10 avec l'étape optionnelle) : Le niveau 50(G) pourrait être considéré comme un niveau optimal, dès lors qu'il fournit le meilleur compromis entre le pourcentage de règles extraites (80% de toutes les règles) et le calcul de complexité (2033 boucles). Cependant, si le pourcentage de règles extraites est considéré comme prioritaire par rapport à la complexité de calcul, le niveau de 31(G) (95% de toutes les règles, 4794 boucles) pourrait à son tour être considéré comme le meilleur niveau. Nouvelles règles : règles qui sont trouvées à un niveau donné mais pas aux niveaux précédents. Nombre de boucles (Symbolique) : nombre de boucles utilisées par l'approche symbolique pour extraire des règles. $x(G)$: représente un niveau de généralisation de x neurones.	141

-
- 5.13 Les résultats d'extraction de règles pour le point de vue **Utilisations** (Algorithme d'extraction 10 sans l'étape optionnelle) : Seules des règles propres sont extraites par cette version de l'algorithme. Dans ce cas, un processus de sélection de règle qui dépend du niveau de la généralisation est exécuté : plus le niveau de la généralisation est élevé, la sélection va diminuer. Les bonnes performance du niveau de 50(G) et du niveau de 31(G) sont également accentuées avec cette version de l'algorithme (voir figure 5.12 pour une comparaison avec l'autre version de l'algorithme). 142
- 5.14 Les résultats d'extraction de règles pour le point de vue **Avantages**. Le niveau 83(G) peut être considéré autant qu'un niveau qui représente le compromis optimal : 70% de toutes les règles sont extraites en utilisant 1402 boucles. 143
- 5.15 Les résultats d'extraction de règles pour le point de vue **Déposants**. Le niveau 49(G) peut être considéré autant qu'un niveau optimal puisqu'il fournit le meilleur compromis entre le pourcentage des règles extraites (92% de toutes les règles) et de la complexité de calcul (41 boucles). Néanmoins, si le pourcentage des règles extraites est considéré comme prioritaire vis à vis de la complexité de calcul, le niveau 31(G) devrait être considéré comme le niveau le plus optimal : toutes les règles sont extraites en utilisant 61 boucles. La baisse de la performance d'extraction de règles pour les niveaux élevés de généralisation est due à la fois aux caractéristiques spécifiques d'indexation de ce point de vue (le nombre moyen d'index par document est près de 1), à l'effet de regroupement initial de la stratégie de Winner-Takes-Most de NG, à l'effet supplémentaire du processus de généralisation et à la stratégie d'extraction de règles basée sur la distribution des propriétés singulières. Cependant, quand la généralisation est exécutée, les documents décrits par des combinaisons de termes d'index qui ont été regroupés dans les mêmes classes par la stratégie d'apprentissage seront redistribués dans des classes spécifiques. Ainsi, la distribution des propriétés singulières dans les classes deviendra plus hétérogène pour les niveaux intermédiaires de généralisation que pour le niveau initial. 144
- 5.16 Les résultats d'extraction de règles pour le point de vue **Titres**. Le niveau 111(G) peut être considéré comme un niveau optimal. Cependant, même si le pourcentage des règles extraites est relativement faible (48% de toutes les règles), ce niveau fournit une efficacité d'extraction très bonne (facteur d'efficacité =76 et 3619 boucles) par rapport au modèle symbolique. La complexité élevée de calcul de règles du premier niveau de gaz (niveau optimal), par rapport aux niveaux suivants de généralisation, est due à la fois à la dispersion des termes d'index du point de vue **Titres** et à l'effet de regroupement de type Winner-Takes-Most de la stratégie d'apprentissage de NG (voir également la figure 5.15 concernant à ce point). Cependant, quand la distribution de données est éparse, cette stratégie d'apprentissage tend à regrouper les données dans un petit nombre de classes, laissant quelques autres classes vides. Cet effet augmentera de manière significative la complexité du calcul de règle au niveau original puisque les classes non vides auront des descriptions relativement hétérogènes. Il faut noter que cette complexité sera réduite aux étapes intermédiaires de la généralisation puisque des données indépendantes seront distribuées dans différentes classes par le processus de généralisation (cf. figure 5.15). 145

5.17	Vue générale de processus de communication : pour le processus de généralisation, seulement 9 niveaux ont été utilisés pour les deux points de vue, en commençant de niveau 0 qui représente la classification optimale de gaz.	146
6.1	La carte SOM de UCodes ou Sous-domaines.	151
6.2	La carte SOM de OutLinks.	152
6.3	Mécanisme de communication inter-cartes à partir de la classe associées à la ville de <i>Munich</i> . Étape 1 : communication de la carte Cities vers la carte OutLinks , Étape 1 : communication de la carte OutLinks vers la carte Cities	153
6.4	Réseau des sites web basé sur les liens sortants (OutLinks). On peut constater que la structure complexe de ce réseau, liée à l’affichage des liens entre les sites, représente une surcharge cognitive importante ce qui le rend de fait très difficile à interpréter et à utiliser de manière directe pour une analyse.	153
6.5	Réseau des sites web associés à <i>Munich</i> basé sur leurs liens sortants (OutLinks).	155
6.6	Tableau des sites web associés à <i>Munich</i> et présentés dans la figure 6.7. La colonne MultiSOM OutLink map indique le nom de la classe où le site web est localisé et sa position (colonne, ligne) sur la carte OutLinks	156
6.7	Répartition des sites web décrits dans le tableau 6.6 sur la carte OutLinks	156
6.8	a) Visualisation des relations de voisinage dans 50 neurones (classes de Neural Gas). La complexité de la structure obtenue est trop élevée et ne permet aucune interprétation. Ce problème est typique d’une telle visualisation par projection non linéaire. b) Méthode hiérarchique ascendante/descendante. c) Méthode de voisins réciproques. d) Méthode de DBHC.	159
6.9	Arbre hyperbolique des thèmes de recherche couverts par les laboratoires de recherche allemands (1). Les données de l’arbre sont étiquetées selon les stratégies définies dans le cas d’étiquetage de modèle MultiSOM à la section 3.4.2, chaque étiquette pouvant être assimilée un thème relatif au point de vue considéré. Un thème qui reste stable en montant vers la racine de l’arbre correspond à un thèmes influent. De plus, les thèmes qui apparaissent dans les sous branches voisines de l’arbre sont fortement liés un à l’autre. Par exemple, les thèmes Simulation et Medical Science sont fortement liés dans la recherche allemande. Le thème Information systems (à la droite de l’arbre) est un thème très riche car il inclut beaucoup de sous-branches. Le thème central de la recherche informatique allemande est le thème Programmation languages . De fait, il représente la racine de l’arbre hyperbolique.	161
6.10	Arbre hyperbolique des thèmes de recherche couverts par les laboratoires de recherche allemands (2). L’analyse qui peut être exécutée avec un tel arbre est interactive (focus+contexte). Le focus peut être placé par l’analyste sur un nœud spécifique de l’arbre. Dans ce cas, le nœud choisi migrera au centre de l’arbre (focus), et, les sous-arbres contenant l’ensemble des fils de ce nœud (contexte) sera automatiquement déplié. Dans cet exemple, l’analyste a choisi de placer le focus sur le nœud associé au thème Information System Design and Components	162
6.11	Arbre hyperbolique des villes allemandes possédant des laboratoires de recherche. La ville qui a été considérée comme centrale pour cette analyse est la ville de Munster . Cela signifie que cette recherche en computer science est particulièrement développée en Allemagne du nord.	163

-
- A.1 Exemple de notice de brevet. L'indexation qui a été générée pour ce brevet est matérialisé par le contenu du champ «Final indexation». Ces termes d'indexation sont préfixés par le nom du point de vue auquel ils sont associés : «adv.» pour le point de vue **Avantages**, «titre» pour le point de vue **Titres**, «use» pour le point de vue **Utilisations**, «soc.» pour le point de vue **Déposants**. 175





Liste des tableaux

1.1	La table d'éventualité de la distribution de document pour t	20
1.2	Indexation des documents.	28
3.1	Le nombre de nœuds N_R de grille hyperbolique pour des valeurs différentes du rayon R , et pour $n = 7$	77
4.1	Résultats de l'évaluation de l'inertie, I-intra : inertie intra-classe, I-moyenne : inertie moyenne, I-inter : inertie inter-classes.	89
4.2	Résultats de l'évaluation du Rappel , de la Précision et de la F-mesure.	89
4.3	Résultats de l'évaluation de l'Inertie, du Rappel et de la Précision pour la méthode GNG.	90
4.4	Résultats de l'évaluation de l'Inertie, du Rappel et de la Précision pour la méthode de NG.	90
4.5	Résultats globaux de processus de catégorisation des étiquettes.	114
4.6	Distribution des types d'étiquettes selon les classes	116
5.1	Nombre optimal de classes pour chaque point de vue.	128
5.2	Valeurs de consistance entre les cartes issues de MultiSOM pour différents points de vue.	129
5.3	Sommaire du processus de génération de gaz. Pour tous les points de vue, la limite de généralisation a été fixée à des niveaux qui comprenant ou plus 10 neurones. Donc, pour un point de vue donné, le nombre de niveaux de généralisation dépend à la fois du nombre initial de neurones de son gaz optimal associé et de l'homogénéité de la distribution de données relativement à ce point de vue.	139
5.4	Sommaire des résultats. Le tableau présente une comparaison synthétique entre la méthode symbolique standard d'extraction des règles et la méthode basée sur MultiGAS. Le nombre global de règles défini pour le modèle symbolique inclut le nombre de règles approximatives (confiance < 100%) et le nombre de règles totales (exactes) (confiance = 100%). Dans nos expérimentations, les règles générées par le modèle de MultiGAS sont seulement des règles totales (exactes). Le nombre de règles propres est le nombre de règles obtenues avec les versions standards des algorithmes d'extraction. Le nombre de règles étendues est le nombre de règles obtenues avec les versions étendues des algorithmes d'extraction comprenant leurs étapes optionnelles.	147
5.5	Sommaire de processus de génération de gaz pour les quatre points de vue.	147

5.6	Sommaire des résultats d'extraction des règles d'association par le mécanisme de généralisation (pour le point de vue In-Coupling) et par le mécanisme de communication inter-cartes (entre Cities et UCodes d'un coté et entre UCodes et InLinks d'un autre coté) en utilisant le modèle MultiGAS.	148
6.1	Notations de DBHC.	157
A.1	Tableau résumé des caractéristiques résultantes de brevets	175
A.2	Tableau résumé des caractéristiques résultantes de donnée issues de site du Laboratoire d'Informatique d'Université de Cambridge	177
A.3	Tableau résumé des caractéristiques résultantes de donnée issues de sites web de laboratoires de recherche allemands.	178

Liste des algorithmes

1	Algorithme de K-means	37
2	Algorithme d'apprentissage des cartes auto-organisatrices de Kohonen (SOM) .	47
3	Algorithme d'apprentissage du modèle Neural Gas (NG)	48
4	Algorithme d'apprentissage du modèle Growing Neural Gas	50
5	Algorithme de SOM incrémental hiérarchique (GHSOM)	52
6	Algorithme de recherche du modèle de classification optimal	118
7	Algorithme de l'évaluation de la cohérence de propagation	126
8	Fonction(act, T_{i_k}, S_k)	127
9	Algorithme de la généralisation de NG	131
10	Algorithme d'extraction des règles d'association par le mécanisme de généralisation	138
11	Algorithme d'extraction des règles d'association par le mécanisme de communi- cation	138
12	Algorithme de classification hiérarchique orienté par la densité (DBHC)	158
13	Procédure 1 : élimination de classes parents répétées	158
14	Procédure 2 : éviter les classes recouvrantes	159



Liste des publications

- [Al Shehabi and Lamirel, 2004a] Shadi Al Shehabi and Jean-Charles Lamirel. Inference bayesian network for multi-topographic neural network communication : a case study in documentary data. In *IEEE International Conference on Information and Communication Technologies : from Theory to Applications - ICTTA 2004, Damascus, Syria*, Apr 2004.
- [Al Shehabi and Lamirel, 2004b] Shadi Al Shehabi and Jean-Charles Lamirel. Unsupervised neural networks of topographic and gas families for documentary data classification. In *The 8th World Multi-Conference on Systemics, Cybernetics and Informatics - SCI'04, Orlando, Florida, USA*, Jul 2004.
- [Al Shehabi and Lamirel, 2005a] Shadi Al Shehabi and Jean-Charles Lamirel. Knowledge extraction from unsupervised multi-topographic neural network models. In *International Conference on Artificial Neural Networks, ICANN (2)*, pages 479–484, 2005.
- [Al Shehabi and Lamirel, 2005b] Shadi Al Shehabi and Jean-Charles Lamirel. Multi-topographic neural network communication and generalization for multi-viewpoint analysis. In *Proceeding of the IEEE International Joint Conference on Neural Networks, IJCNN 2005, Montreal, Canada, IEEE Computer Societ*, pages 0–0, August 2005.
- [Al Shehabi and Lamirel, 2006] Shadi Al Shehabi and Jean-Charles Lamirel. A new hyperbolic visualization method for displaying the results of a neural gas model : application to webometrics. In *Proceedings of the 13th European Symposium On Artificial Neural Networks (ESANN 2005)*, pages 351–356, Bruges, Belgium, April 2006.
- [Attik et al., 2006a] Mohammed Attik, Shadi Al Shehabi, and Jean-Charles Lamirel. Clustering analysis for data samples with multiple labels. In *IASTED International Conference on Databases and Applications, part of the 24th IASTED International Multi-Conference on Applied Informatics, DBA 2006, Innsbruck, Austria*, pages 0–0, february 2006.
- [Attik et al., 2006b] Mohammed Attik, Shadi Al Shehabi, and Jean-Charles Lamirel. Clustering quality measures for data samples with multiple labels. In *IASTED International Conference on Databases and Applications, part of the 24th IASTED International Multi-Conference on Applied Informatics, DBA 2006, Innsbruck, Austria*, pages 0–0, february 2006.
- [Francois et al., 2004] Claire Francois, Marianne Hoerlesberger, Xavier Polanco, Jean-Charles Lamirel, and Shadi Al Shehabi. Visualization of web data : Complementarities between a graph and a multi-maps approach. In *8th International Conference on Science and Technology Indicators - ST 2004, Leiden, Nederlands*, pages 94–95, Sep 2004.
- [Lamirel and Al Shehabi, 2005] Jean-Charles Lamirel and Shadi Al Shehabi. Efficient knowledge extraction from unsupervised multi-topographic neural network models. In *5th Workshop on Self-Organizing Maps, WSOM 2005, Paris 1 Panthéon-Sorbonne University*, pages 291–298, 2005.

- [Lamirel and Al Shehabi, 2006] Jean-Charles Lamirel and Shadi Al Shehabi. Multisom : a multiview neural model for accurately analyzing and mining complex data. In *4th International Conference on Coordinated and Multiple Views in Exploratory Visualization, CMV*, 2006.
- [Lamirel *et al.*, 2003a] Jean-Charles Lamirel, Shadi Al Shehabi, Martial Hoffmann, and Claire Francois. Intelligent patent analysis through the use of a neural network : experiment of multi-viewpoint analysis with the multisom model. In *The ACL 2003 Workshop on Patent Corpus Processing, Sapporo, Japan*, pages 7–23, Jul 2003.
- [Lamirel *et al.*, 2003b] Jean-Charles Lamirel, Claire Francois, Shadi Al Shehabi, and Martial Hoffmann. New classification quality estimators for analysis of documentary information : application to patent analysis and web mapping. In *9th International Conference on Scientometrics and Informetrics - ISSI 2003, Beijing, chine*, Aug 2003.
- [Lamirel *et al.*, 2003c] Jean-Charles Lamirel, Yannick Toussaint, and Shadi Al Shehabi. A hybrid classification method for database contents analysis. In *The 16th International FLAIRS Conference - FLAIRS 2003, St. Augustine, Florida*, May 2003.
- [Lamirel *et al.*, 2004a] Jean-Charles Lamirel, Shadi Al Shehabi, Claire Francois, and Martial Hoffmann. New classification quality estimators for analysis of documentary information : application to patent analysis and web mapping. *Scientometrics*, 60(3) :445–462, Feb 2004.
- [Lamirel *et al.*, 2004b] Jean-Charles Lamirel, Shadi Al Shehabi, Claire Francois, and Xavier Polanco. Using a compound approach based on elaborated neural network for webometrics : An example issued from the eicstes project. *Scientometrics*, 61(3) :427–441, Jan 2004.
- [Lamirel *et al.*, 2004c] Jean-Charles Lamirel, Shadi Al Shehabi, Claire Francois, Xavier Polanco, and Martial Hoffmann. Using a compound approach based on elaborated neural network for webometrics : an example issued from the eicstes project. In *5th COLNET Meeting - COLLNET 2004, Roorkee, India*, Mar 2004.
- [Lamirel *et al.*, 2005] Jean-Charles Lamirel, Shadi Al Shehabi, and Claire François. Evaluation of collaboration between european universities using dynamic interaction between multiple sources. In *10th International Conference of the International Society for Scientometrics and Informetrics - ISSI-COLLNET 2005, Stockholm, Suède*, Jul 2005.

Bibliographie

- [Agrawal and Srikant, 1994] R. Agrawal and R. Srikant. Fast algorithms for mining association rules. In *Proc. of the Int. Conf. on Very Large Databases, VLDB*, pages 487–499, SanTiago, Chile, June 1994. VLDB.
- [Ahn, 1997] C. Ahn. Statistical methods for the estimation of sensitivity and specificity of site-specific diagnostic tests. *Journal of Periodontal Research*, 32 :351–354, 1997.
- [Al Shehabi and Lamirel, 2004a] Shadi Al Shehabi and Jean-Charles Lamirel. Inference bayesian network for multi-topographic neural network communication : a case study in documentary data. In *IEEE International Conference on Information and Communication Technologies : from Theory to Applications - ICTTA 2004, Damascus, Syria*, Apr 2004.
- [Al Shehabi and Lamirel, 2004b] Shadi Al Shehabi and Jean-Charles Lamirel. Unsupervised neural networks of topographic and gas families for documentary data classification. In *The 8th World Multi-Conference on Systemics, Cybernetics and Informatics - SCI'04, Orlando, Florida, USA*, Jul 2004.
- [Al Shehabi and Lamirel, 2005a] Shadi Al Shehabi and Jean-Charles Lamirel. Knowledge extraction from unsupervised multi-topographic neural network models. In *International Conference on Artificial Neural Networks, ICANN (2)*, pages 479–484, 2005.
- [Al Shehabi and Lamirel, 2005b] Shadi Al Shehabi and Jean-Charles Lamirel. Multi-topographic neural network communication and generalization for multi-viewpoint analysis. In *Proceeding of the IEEE International Joint Conference on Neural Networks, IJCNN 2005, Montreal, Canada, IEEE Computer Societ*, pages 0–0, August 2005.
- [Al Shehabi and Lamirel, 2006] Shadi Al Shehabi and Jean-Charles Lamirel. A new hyperbolic visualization method for displaying the results of a neural gas model : application to webometrics. In *Proceedings of the 13th European Symposium On Artificial Neural Networks (ESANN 2005)*, pages 351–356, Bruges, Belgium, April 2006.
- [André Pantel, 2003] Patrick André Pantel. *Clustering by Committee*. Ph.d. thesis, university of alberta, Edmonton, Alberta, Canada, 2003.
- [Arroyo et al., 2003] N. Arroyo, V.M. Pareja, and I. Arguillo. *Description of Web data in D3.1. Eiscetes European Project IST-1999-20350*, 2003.
- [Attik et al., 2006a] Mohammed Attik, Shadi Al Shehabi, and Jean-Charles Lamirel. Clustering analysis for data samples with multiple labels. In *IASTED International Conference on Databases and Applications, part of the 24th IASTED International Multi-Conference on Applied Informatics, DBA 2006, Innsbruck, Austria*, pages 0–0, february 2006.
- [Attik et al., 2006b] Mohammed Attik, Shadi Al Shehabi, and Jean-Charles Lamirel. Clustering quality measures for data samples with multiple labels. In *IASTED International*

- Conference on Databases and Applications, part of the 24th IASTED International Multi-Conference on Applied Informatics, DBA 2006, Innsbruck, Austria*, pages 0–0, february 2006.
- [Barbut and Monjardet, 1970] M. Barbut and B. Monjardet. *Ordre et classification algèbre et combinatoire*. Hachette université, 1970.
- [Bergé and Bouthier, 2003] Benjamin Bergé and Christophe Bouthier. Mathematics and algorithms for the hyperbolic tree visualization. Rapport technique, Sep 2003.
- [Birkoff, 1967] G. Birkoff. *Lattice Theory*. Providence, Amer. Math. Soc., New York, first edition 1940 edition, 1967.
- [Björneborn and Ingwersen, 2004] Lennart Björneborn and Peter Ingwersen. Toward a basic framework for webometrics. *Journal of the American Society for Information Science and Technology JASIST*, 55(14) :1216–1227, 2004.
- [Bottou and Bengio, 1995] L. Bottou and Y. Bengio. Convergence properties of the K-means algorithm. pages 585–592. MIT Press, Cambridge, MA, 1995.
- [Bouthier, 2004] Christophe Bouthier. *Mise en contexte de la conscience de groupe : adaptation et visualisation*. Thèse d’université, INPL, Sep 2004.
- [Caid *et al.*, 1995] William R. Caid, Susan T. Dumais, and Stephen I. Gallant. Learned vector-space models for document retrieval. *Inf. Process. Manage.*, 31(3) :419–429, 1995.
- [Calinski and Harabasz, 1974] R. B. Calinski and J. Harabasz. A dendrite method for cluster analysis. *Communications in Statistics*, 3 :1–27, 1974.
- [Chartron, 1988] G. Chartron. *Analyse des corpus de données textuelles, sondage de flux d’informations*. PhD thesis, Université de Paris 7, Paris, France, 1988.
- [Cherfi, 2004] Hacène Cherfi. *Étude et réalisation d’un système d’extraction de connaissances à partir de textes*. Thèse d’université, UHP - Nancy 1, Nov 2004.
- [Chu, 2005] Heting Chu. Taxonomy of inlinked web entities : What does it imply for webometric research ? *Library and Information Science Research*, 27(1) :8–27, 2005.
- [Davey and Priestley, 1994] B.A. Davey and H.A. Priestley. *Introduction to Lattices and Order*. Cambridge University Press, 1994.
- [de Campos *et al.*, 2000] L. M. de Campos, J. M. Fernandez, and J. F. Huete. Building bayesian network-based information retrieval systems. In *DEXA ’00 : Proceedings of the 11th International Workshop on Database and Expert Systems Applications*, page 543, Washington, DC, USA, 2000. IEEE Computer Society.
- [de Campos *et al.*, 2002] Luis M. de Campos, Juan M. Fernández-Luna, and Juan F. Huete. A layered bayesian network model for document retrieval. In *Retrieval, Proceedings of the 24th BCS-IRSG European Colloquium on IR Research : Advances in Information Retrieval, March 25-27, 2002*, Lecture Notes in Computer Science, vol. 2291, pages 169–182. Springer, London, UK, 2002.
- [Deerwester *et al.*, 1990] Scott C. Deerwester, Susan T. Dumais, Thomas K. Landauer, George W. Furnas, and Richard A. Harshman. Indexing by latent semantic analysis. *Journal of the American Society of Information Science*, 41(6) :391–407, 1990.
- [Defays, 1977] D. Defays. An efficient algorithm for a complete link method. volume 20, pages 364–366, 1977.

-
- [Dekker, 2001] Anthony Dekker. Visualisation of social networks using cavalier. In *CRPITS '01 : Australian symposium on Information visualisation*, pages 49–55, Darlinghurst, Australia, Australia, 2001. Australian Computer Society, Inc.
- [Demartines and Hérault, 1997] P. Demartines and J. Hérault. Curvilinear component analysis : A self-organizing neural network for nonlinear mapping of data sets. 8(1) :148–154, January 1997.
- [Demartines, 1994] P. Demartines. *Analyse de données par réseaux de neurones auto-organisés*. PhD thesis, Institut National Polytechnique de Grenoble, Grenoble, France, 1994.
- [Dijkstra, 1959] E. W. Dijkstra. A note on two problems in connexion with graphs. *Numer. Math.*, 1 :269–271, 1959.
- [Duda and Hart, 1973] R. O. Duda and P. E. Hart. *Pattern Classification and Scene Analysis*. New York, 1973.
- [Dudoit and Fridlyand, 2002] S. Dudoit and J. Fridlyand. A prediction-based resampling method for estimating the number of clusters in a dataset. *Genome Biology*, 3(7) :1–21, 2002.
- [Dumais, 1991] Susan T. Dumais. Improving the retrieval of information from external sources. *Behavior Research Methods, Instruments, & Computers*, 23(2) :229, 1991.
- [Duquenne, 1996] V. Duquenne. On lattice approximations : Syntactic aspects. *Social Networks*, 18 :189–199, 1996.
- [England and Beynon, 1981] R. England and D. Beynon. Statistical algorithms : Remark AS R39 : A remark on AS 136 : A K -means clustering algorithm. 30(3) :355–356, 1981.
- [Ester *et al.*, 1996] Martin Ester, Hans-Peter Kriegel, Jorg Sander, and Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In Evangelos Simoudis, Jiawei Han, and Usama Fayyad, editors, *Second International Conference on Knowledge Discovery and Data Mining*, pages 226–231, Portland, Oregon, 1996. AAAI Press.
- [Fayyad *et al.*, 1996] Usama M. Fayyad, Gregory Piatetsky-Shapiro, Padhraic Smyth, and Ramasamy Uthrusamy, editors. *Advances in Knowledge Discovery and Data Mining*. AAAI/MIT Press, 1996.
- [Francois *et al.*, 2004] Claire Francois, Marianne Hoerlesberger, Xavier Polanco, Jean-Charles Lamirel, and Shadi Al Shehabi. Visualization of web data : Complementarities between a graph and a multi-maps approach. In *8th International Conference on Science and Technology Indicators - ST 2004, Leiden, Netherlands*, pages 94–95, Sep 2004.
- [Fritzke, 1995] B. Fritzke. A growing neural gas network learns topologies. In G. Tesauero, D. S. Touretzky, and T. K. Leen, editors, *Advances in Neural Information Processing Systems 7*, pages 625–632. MIT Press, Cambridge MA, 1995.
- [Fritzke, 1997] Bernd Fritzke. Some competitive learning methods. Technical report, Ruhr-Universität Bochum, 1997.
- [Gallant *et al.*, 1992] Stephen I. Gallant, Robert Hecht-Nielsen, William R. Caid, Kent Pu Qing, Joel Carleton, and David Sudbeck. Hnc’s matchplus system. *SIGIR Forum*, 26(2) :34–38, 1992.
- [Gallant *et al.*, 1993] Stephen I. Gallant, William R. Caid, Joel Carleton, T. Gutschow, Robert Hecht-Nielsen, Kent Pu Qing, and David Sudbeck. Feedback and mixing experiments with matchplus. In *TREC*, pages 101–104, 1993.

- [Gallant, 1991] Stephen I. Gallant. Context vector representations for document retrieval. In *AAA-91 Natural Language Text Retrieval Workshop : Anaheim, CA*, 1991.
- [Godin and Missaoui, 1994] Robert Godin and Rokia Missaoui. An incremental concept formation approach for learning from databases. *Theor. Comput. Sci.*, 133(2) :387–419, 1994.
- [Godin *et al.*, 1995] R. Godin, G. Mineau, R. Missaoui, and H. Mili. Méthodes de classification conceptuelle basées sur les treillis de galois et applications. *Revue d'intelligence artificielle*, 9(2) :105–137, 1995.
- [Golub and Loan, 1991] G. H. Golub and C. F. V. Loan. *Matrix Computation*. John Hopkins University Press, Second Edition, 1991.
- [Grossberg, 1976] S. Grossberg. Adaptative pattern classification and universal recording, i : Parallel development and coding of neural feature detectors. *Biological Cybernetics*, 23 :121–134, 1976.
- [Guha *et al.*, 1998] S. Guha, R. Rastogi, and K. Shim. Cure : An efficient clustering algorithm for large databases. In *ACM SIGMOD'98*, 1998.
- [Guha *et al.*, 1999] S. Guha, R. Rastogi, and K. Shim. Rock : a robust clustering algorithm for categorical attributes. In *In Proc. of the 15th Int'l Conf. on Data Eng.*, 1999.
- [Guigues and Duquenne, 1986] J.L. Guigues and V. Duquenne. Famille minimales d'implications informatives résultant d'un tableau de données binaires. *Mathématiques, Informatiques et Sciences Humaines*, 95 :5–18, 1986.
- [Hammer *et al.*, 2002] Barbara Hammer, Andreas Rechten, Marc Strickert, and Thomas Villmann. Rule extraction from self-organizing networks. In *ICANN*, pages 877–883, 2002.
- [Han *et al.*, 2001] J. Han, M. Kamber, and A. K. H. Tung. Spatial clustering methods in data mining : A survey. In *Miller, H. and Han, J. (Eds.) Geographic Data Mining and Knowledge Discovery, Taylor and Francis.*, pages 188–217, 2001.
- [Hartigan and Wong, 1979] J. A. Hartigan and M. A. Wong. Statistical algorithms : Algorithm AS 136 : A K -means clustering algorithm. 28(1) :100–108, March 1979.
- [Hartigan, 1975] J. A. Hartigan. *Clustering Algorithms*. Wiley, New York, 1975.
- [Hebb, 1949] D. O. Hebb. *The organization of behaviour*. Wiley, New York, 1949.
- [Henle, 2001] M. Henle. *Modern Geometries*. Prentice Hall, 2001.
- [Honkela *et al.*, 1996a] Timo Honkela, Samuel Kaski, Krista Lagus, and Teuvo Kohonen. Exploration of full-text databases with self-organizing maps. In *Proceedings of the ICNN96, International Conference on Neural Networks*, volume I, pages 56–61. IEEE Service Center, Piscataway, NJ, 1996.
- [Honkela *et al.*, 1996b] Timo Honkela, Samuel Kaski, Krista Lagus, and Teuvo Kohonen. News-group exploration with WEBSOM method and browsing interface. Technical Report A32, Helsinki University of Technology, Laboratory of Computer and Information Science, Espoo, Finland, 1996.
- [Honkela, 1997] T. Honkela. *Self-Organizing Maps in Natural Language Processing*. PhD thesis, Helsinki University of Technology, Espoo, Finland, 1997.
- [Hopfield, 1982] J.J. Hopfield. Neural Network and physical systems with emergent collective computational abilities. volume 79 of *Biophysics*, pages 2554–2558, 1982.
- [Imola K. Fodor, 2002] Imola K. Fodor. A survey of dimension reduction techniques, June 2002.

-
- [Jamoussi, 2004] Salma Jamoussi. *Méthodes statistiques pour la compréhension automatique de la parole*. Thèse d'université, UHP - Nancy 1, Dec 2004.
- [Jensen *et al.*, 1990] F.V. Jensen, S.L. Lauritzen, and K.G. Olesen. Bayesian updating in causal probabilistic networks by local computations. *Computational Statistics Quarterly*, 4 :269–282, 1990.
- [Jensen, 1996] Finn V. Jensen. *An Introduction to Bayesian Networks*. UCL Press, 1st edition, 1996.
- [Jiang *et al.*, 2002] D. Jiang, C. Tang, and A. Zhang. Cluster analysis for gene expression data : A survey. *IEEE Transactions on Knowledge and Data Engineering*, 2002.
- [Jolliffe, 1972] I. T. Jolliffe. Discarding variables in principal component analysis i : Artificial data. *Applied Statistics*, 21 :160–173, 1972.
- [Jolliffe, 1986] I. T. Jolliffe. *Principal Component Analysis*. Springer, 1986.
- [Jones, 1979] K. Sparck Jones. Search term relevance weighting given little relevance information. *Journal of Documentation*, 35 :30–48, 1979.
- [Karypis and Kumar, 1998] George Karypis and Vipin Kumar. Multilevel k-way partitioning scheme for irregular graphs. *Journal of Parallel and Distributed Computing*, 48(1) :96–129, 1998.
- [Karypis *et al.*, 1999] G. Karypis, E.-H. Han, and V. Kumar. Chameleon : Hierarchical clustering using dynamic modeling. *IEEE Computer*, 32(8) :68–75, 1999.
- [Kaski *et al.*, 1998] Samuel Kaski, Timo Honkela, Krista Lagus, and Teuvo Kohonen. WEBSOM—self-organizing maps of document collections. *Neurocomputing*, 21 :101–117, 1998.
- [Kaski, 1998] Samuel Kaski. Dimensionality reduction by random mapping : Fast similarity computation for clustering. In *Proceedings of IJCNN'98, International Joint Conference on Neural Networks*, volume 1, pages 413–418. IEEE Service Center, Piscataway, NJ, 1998.
- [Kaufman and Rousseeuw, 1990a] Leonard Kaufman and Peter J. Rousseeuw. *Finding groups in data : an introduction to cluster analysis*. 1990.
- [Kaufman and Rousseeuw, 1990b] Leonard Kaufman and Peter J. Rousseeuw. *Finding groups in data : an introduction to cluster analysis*. John Wiley and Sons, 1990.
- [Kohonen *et al.*, 1996] Teuvo Kohonen, Samuel Kaski, Krista Lagus, and Timo Honkela. Very large two-level SOM for the browsing of newsgroups. In C. von der Malsburg, W. von Seelen, J. C. Vorbrüggen, and B. Sendhoff, editors, *Proceedings of ICANN96, International Conference on Artificial Neural Networks, Bochum, Germany, July 16-19, 1996*, Lecture Notes in Computer Science, vol. 1112, pages 269–274. Springer, Berlin, 1996.
- [Kohonen *et al.*, 2000] T. Kohonen, S. Kaski, K. Lagus, J. Salojärvi, J. Honkela, V. Paatero, and A. Saarela. Self organization of a massive document collection. *IEEE Transactions on Neural Networks, Special Issue on Neural Networks for Data Mining and Knowledge Discovery*, 11, 2000.
- [Kohonen, 1984] T. Kohonen. *Self-Organization and Associative Memory*. Springer Verlag, 2 edition, 1984.
- [Kohonen, 1997] T. Kohonen. *Self-Organising Maps*. Springer-Verlag, Berlin, 2nd edition, 1997.

- [Kohonen, 2001] T. Kohonen. *Self-Organising Maps*. Springer-Verlag, Berlin, 3rd edition, 2001.
- [Koikkalainen and Oja, 1990] P. Koikkalainen and E. Oja. Self-organizing hierarchical feature maps. In *Proc. IJCNN-90, International Joint Conference on Neural Networks, Washington, DC*, volume II, pages 279–285, Piscataway, NJ, 1990. IEEE Service Center.
- [Kopcsa and Schiebel, 1998] Alexander Kopcsa and Edgar Schiebel. Science and technology mapping : a new iteration model for representing multidimensional relationships. *Journal of American Society for Information Sciences (JASIS)*, 49(1) :7–17, 1998.
- [Krzanowski and Lai, 1985] W. J. Krzanowski and Y. T. Lai. A criterion for determining the number of clusters in a data set. *Biometrics*, 44 :23–34, 1985.
- [Lagus and Kaski, 1999] Krista Lagus and Samuel Kaski. Keyword selection method for characterizing text document maps. In *Proceedings of ICANN99, Ninth International Conference on Artificial Neural Networks*, volume 1, pages 371–376. IEE, London, 1999.
- [Lagus *et al.*, 1996] Krista Lagus, Timo Honkela, Samuel Kaski, and Teuvo Kohonen. Self-organizing maps of document collections : A new approach to interactive exploration. In Evangelios Simoudis, Jiawei Han, and Usama Fayyad, editors, *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, pages 238–243. AAAI Press, Menlo Park, California, 1996.
- [Lagus *et al.*, 1999] Krista Lagus, Timo Honkela, Samuel Kaski, and Teuvo Kohonen. WEB-SOM for textual data mining. *Artificial Intelligence Review*, 13 :345–364, 1999.
- [Lagus *et al.*, 2004] Krista Lagus, Samuel Kaski, and Teuvo Kohonen. Mining massive document collections by the WEBSOM method. *Information Sciences*, 163 :135–156, 2004.
- [Lamirel and Al Shehabi, 2005] Jean-Charles Lamirel and Shadi Al Shehabi. Efficient knowledge extraction from unsupervised multi-topographic neural network models. In *5th Workshop on Self-Organizing Maps, WSOM 2005, Paris 1 Panthéon-Sorbonne University*, pages 291–298, 2005.
- [Lamirel and Al Shehabi, 2006] Jean-Charles Lamirel and Shadi Al Shehabi. Multisom : a multiview neural model for accurately analyzing and mining complex data. In *4th International Conference on Coordinated and Multiple Views in Exploratory Visualization, CMV*, 2006.
- [Lamirel *et al.*, 2000] Jean-Charles Lamirel, Jacques Ducloy, and G  rald Oster. Adaptative browsing for information discovery in an iconographic context. In CID, editor, *RIAO'2000, Coll  ge de France, Paris, France*. CID, Apr 2000.
- [Lamirel *et al.*, 2001] Jean-Charles Lamirel, Yannick Toussaint, Claire Francois, and Xavier Polanco. Using a multisom approach for mapping of science and technology. In *ISSI 2001*, Aug 2001.
- [Lamirel *et al.*, 2003a] Jean-Charles Lamirel, Shadi Al Shehabi, Martial Hoffmann, and Claire Francois. Intelligent patent analysis through the use of a neural network : experiment of multi-viewpoint analysis with the multisom model. In *The ACL 2003 Workshop on Patent Corpus Processing, Sapporo, Japan*, pages 7–23, Jul 2003.
- [Lamirel *et al.*, 2003b] Jean-Charles Lamirel, Claire Francois, Shadi Al Shehabi, and Martial Hoffmann. New classification quality estimators for analysis of documentary information : application to patent analysis and web mapping. In *9th International Conference on Scientometrics and Informetrics - ISSI 2003, Beijing, chine*, Aug 2003.

-
- [Lamirel *et al.*, 2003c] Jean-Charles Lamirel, Yannick Toussaint, and Shadi Al Shehabi. A hybrid classification method for database contents analysis. In *The 16th International FLAIRS Conference - FLAIRS 2003, St. Augustine, Florida*, May 2003.
- [Lamirel *et al.*, 2004a] Jean-Charles Lamirel, Shadi Al Shehabi, Claire Francois, and Martial Hoffmann. New classification quality estimators for analysis of documentary information : application to patent analysis and web mapping. *Scientometrics*, 60(3) :445–462, Feb 2004.
- [Lamirel *et al.*, 2004b] Jean-Charles Lamirel, Shadi Al Shehabi, Claire Francois, and Xavier Polanco. Using a compound approach based on elaborated neural network for webometrics : An example issued from the eicstes project. *Scientometrics*, 61(3) :427–441, Jan 2004.
- [Lamirel *et al.*, 2004c] Jean-Charles Lamirel, Shadi Al Shehabi, Claire Francois, Xavier Polanco, and Martial Hoffmann. Using a compound approach based on elaborated neural network for webometrics : an example issued from the eicstes project. In *5th COLNET Meeting - COLLNET 2004, Roorkee, India*, Mar 2004.
- [Lamirel *et al.*, 2005] Jean-Charles Lamirel, Shadi Al Shehabi, and Claire François. Evaluation of collaboration between european universities using dynamic interaction between multiple sources. In *10th International Conference of the International Society for Scientometrics and Informetrics - ISSI-COLLNET 2005, Stockholm, Suède*, Jul 2005.
- [Lamirel, 1995] J-C. Lamirel. *Application d'une approche symbolico-connexioniste pour la conception d'un système documentaire hautement interactif*. PhD thesis, Université de Nancy 1 Henri Poincaré, Nancy, France, 1995.
- [Lamirel, 2001] Jean-Charles Lamirel. Using images for enhancing discovering task in a dl context. In *SPIE 2001*, Jan 2001.
- [Lamping and Rao, 1994] J. Lamping and R. Rao. *Laying out and visualizing large trees using a hyperbolic space*. UIST '94 : Proceedings of the 7th annual ACM symposium on User interface software and technology, 1994.
- [Lamping and Rao, 1996] J. Lamping and R. Rao. The hyperbolic browser : A focus + context technique for visualizing large hierarchies. *Journal of Visual Languages & Computing*, 7, 1996.
- [Lamping J. and P., 1995] Rao R. Lamping J. and Pirolli P. A *focus+context* technique based on hyperbolic geometry for visualizing large hierarchies. In *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI-95)*, pages 401–408, Denver, 1995. ACM Press.
- [Lebart *et al.*, 1982] L. Lebart, A. Morineau, and J. P. Fénelon. *Traitement des données statistiques*. Dunod, Paris, 1982.
- [Lebart *et al.*, 1998] L. Lebart, A. Morineau, and M. Piron. *Statistique Exploratoire Multidimensionnelle*. Dunod, 2 edition, 1998.
- [Lee *et al.*, 2000] John Aldo Lee, Amaury Lendasse, Nicolas Donckers, and Michel Verleysen. A robust nonlinear projection method. In *European Symposium on Artificial Neural Networks*, pages 13–20, 2000.
- [Lee *et al.*, 2002] J.A. Lee, A. Lendasse, and M. Verleysen. Curvilinear distance analysis versus isomap. In *ESANN 2002, European Symposium on Artificial Neural Networks*, pages 185–192, Bruges (Belgium), 2002.

- [Lelu and Georgel, 1990] A. Lelu and A. Georgel. Neural models for orthogonal and oblique factor analyses : Towards dynamic data analysis of large sets of highly multidimensional objects. In *Proceeding of International Joint Conference on Neural Networks, IJCNN, Paris, France*, pages 829–832, 1990.
- [Lelu, 1994] A Lelu. *Clusters and factors : neural algorithms for a novel representation of huge and highly multidimensional data sets* *Statistique Exploratoire Multidimensionnelle*, pages 241–248. Springer Verlag, Berlin, 1994.
- [Lendasse *et al.*, 2001] A. Lendasse, J. Lee, E. De Bodt, V. Wertz, and M. Verleysen. Dimension reduction of technical indicators for the prediction of financial time series - application to the bel 20 market index. *European Journal of Economic and Social Systems*, 15(2) :31–48, 2001.
- [Lin *et al.*, 1991] X. Lin, D. Soergel, and G. Marchionini. A self organizing semantic map for information retrieval. In *Proceedings of 4th International SIGFIR Conference on R&D in Information retrieval*, pages 262–269, 1991.
- [Linde *et al.*, 1980] Y. Linde, A. Buzo, and R. M. Gray. An algorithm for vector quantizer design. *IEEE Transaction on Communications*, COM-28 :84–95, January 1980.
- [Magnus, 1974] W. Magnus. *Noneuclidean Tesselations and Their Groups*. Academic Press, 1974.
- [Manning and Schütze, 2000] Christopher D. Manning and Hinrich Schütze. *Foundation of statistical natural language processing*. The MIT Press, Cambridge, Massachusetts, USA, 2000.
- [Martinetz and Schulten, 1994] Thomas Martinetz and Klaus Schulten. Topology representing networks. *Neural Netw.*, 7(3) :507–522, 1994.
- [Martinetz, Thomas and Schulten, Klaus, 1991] Martinetz, Thomas and Schulten, Klaus. A “Neural-Gas” Network Learns Topologies. In Kohonen, T., M. akisara, K., Simula, O., and Kangas, J., editors, *Artificial Neural Networks – Proceedings of the 1991 International Conference on Artificial Neural Networks (ICANN’91)*, volume 1, pages 397–402. Elsevier Science Publishing Company B. V., 1991.
- [Mather, 2000] Laura A. Mather. A linear algebra measure of cluster quality. *Journal of the American Society for Information Science (JASIS)*, 51(7) :602–613, 2000.
- [Morrison and Chalmers, 2003] Alistair Morrison and Matthew Chalmers. Improving hybrid mds with pivot-based searching. In *INFOVIS*, 2003.
- [Murtagh, 1983] F. Murtagh. A survey of recent advances in hierarchical clustering algorithms. volume 26, pages 354–359, 1983.
- [Murtagh, 1985] F. Murtagh. *Multidimensional Clustering Algorithms*. Physica-Verlag, 1985.
- [Ontrup and Ritter, 2001] J. Ontrup and H. Ritter. Hyperbolic self-organizing maps for semantic navigation. In *Advances in Neural Information Processing Systems 14*, 2001.
- [Ontrup and Ritter, 2005] Jörg Ontrup and Helge Ritter. A hierarchically growing hyperbolic self-organizing map for rapid structuring of large data sets. In *Proceedings of the 5th Workshop on Self-Organizing Maps (WSOM 05)*, Paris, France, Sep 2005.
- [Pearl, 1988] Judea Pearl. *Probabilistic Reasoning in Intelligent Systems : Networks of Plausible Inference*. Morgan Kaufman, San Mateo, CA, 1988.

-
- [Polanco and François, 2000] X. Polanco and C. François. Data classing and class mapping or visualization in text processing and mining. In N. J. Williamson. Advances in Knowledge Organization C. Beghtol, C. L. Howarth, editor, *Dynamism and Stability in Knowledge Organization. Proceedings of the Sixth international ISKO Conference*, volume 7, pages 359–365, Toronto, Canada, July 2000.
- [Ratcliffe, 1994] J. G. Ratcliffe. *Foundations of hyperbolic manifolds*, volume 149 of *Graduate texts in mathematics*. Springer New York/Berlin/Heidelberg, 1994.
- [Rauber *et al.*, 2002] A. Rauber, D. Merkl, and M. Dittenbach. The growing hierarchical self-organizing map : Exploratory analysis of high-dimensional data. *IEEE Transactions on Neural Networks*, IEEE Press, 13(6) :1331–1341, November 2002.
- [Rham, 1980] C. De Rham. La classification hiérarchique ascendante selon la méthode des voisins réciproques. *Les Cahiers de l'Analyse des Données*, 5(2) :135–144, 1980.
- [Ribeiro and Muntz, 1996] Berthier A. N. Ribeiro and Richard Muntz. A belief network model for IR. In Hans-Peter Frei, Donna Harman, Peter Schäuble, and Ross Wilkinson, editors, *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 253–260, August 18–22 1996.
- [Ritter, 1999] H. Ritter. Self-organizing maps in non-euclidean spaces. In E. Oja and S. Kaski, editors, *Kohonen Maps*, pages 97–108. 1999.
- [Roberto J. Bayardo and Agrawal, 1999] Jr. Roberto J. Bayardo and Rakesh Agrawal. Mining the most interesting rules. In *KDD '99 : Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 145–154, New York, NY, USA, 1999. ACM Press.
- [Robertson and Jones, 1976] S.E. Robertson and K. Sparck Jones. Relevance weighting of search terms. *Journal of the American Society for Information Science*, 27(3) :129–146, 1976.
- [Rocchio, 1971] J. J. Rocchio. *Evaluation Viewpoints in Document Retrieval. In The SMART Retrieval System - Experiments in Automatic Document Processing*. Prentice Hall Inc., Englewood Cliffs, New Jersey, 1971.
- [Rosenblatt, 1958] F. Rosenblatt. The perceptron : A probabilistic model for information storage and organization in the brain. *Psychological Review*, 65(6) :386–408, 1958.
- [Salton, 1971] G. Salton. *The SMART Retrieval System - Experiments in Automatic Document Processing*. Prentice Hall Inc., Englewood Cliffs, New Jersey, 1971.
- [Salton, 1983] G. Salton. *Introduction to Modern Information Retrieval*. McGraw-Hill, 1 edition, 1983.
- [Salton, 1989] G. Salton. *Automatic text Processing*. Addison Wesley, 1989.
- [Sammon, 1969] J. W. Sammon. A nonlinear mapping for data structure analysis. *IEEE Transactions on Computers*, C-18(5) :401–409, 1969.
- [Shepard, 1962] R. N. Shepard. The analysis of proximities : Multidimensional Scaling with an unknown distance function. *Psychometrika*, 27 :125–139, 219–246, 1962.
- [Simon and Napoli, 1999] A. Simon and A. Napoli. In , *Atlanta, USA*, pages 104–108, Oct 1999.
- [Simon, 2000] Arnaud Simon. *Outils classificatoires par objets pour l'extraction de connaissances dans des bases de données*. Thèse d'université, Loria, Sep 2000.

- [Stephenson, 2000] Todd A. Stephenson. An introduction to Bayesian network theory and usage. IDIAP-RR 03, IDIAP, 2000.
- [Szathmary *et al.*, 2005] Laszlo Szathmary, Amedeo Napoli, and Sergei O. Kuznetsov. Zart : A multifunctional itemset miner algorithm. Rapport de recherche, Feb 2005.
- [Tenenbaum *et al.*, 2000] J. B. Tenenbaum, V. de Silva, and J. C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500) :2319–2323, December 2000.
- [Thelwall *et al.*, 2003] M. Thelwall, G. Harries, and D. Wilkinson. Why do web sites from different academic subjects interlink? *Journal of Information Science*, 29(6) :453–471, 2003.
- [Thelwall *et al.*, 2005] M Thelwall, L Vaughan, and L Björneborn. *The Annual Review of Information Science and Technology (ARIST)*, volume 39. 2005.
- [Thelwall, 2002] M. Thelwall. Evidence for the existence of geographic trends in university website interlinking. *Journal of Documentation*, 58(5) :563–574, 2002.
- [Toussaint *et al.*, 2000] Yannick Toussaint, Arnaud Simon, and Hacene Cherfi. Apport de la fouille de données textuelles pour l’analyse de l’information. In *Ingénierie des connaissances - IC’2000, Toulouse, France*, May 2000.
- [Tsikrika and Lalmas, 2002] Theodora Tsikrika and Mounia Lalmas. Combining web document representations in a bayesian inference network model using link and content-based evidence. In *Proceedings of the 24th BCS-IRSG European Colloquium on IR Research*, pages 53–72, London, UK, 2002. Springer-Verlag.
- [Turtle and Croft, 1991] Howard Turtle and W. Bruce Croft. Evaluation of an inference network-based retrieval model. *ACM Trans. Inf. Syst.*, 9(3) :187–222, 1991.
- [Van Rijsbergen, 1975] C. J. Van Rijsbergen. *Information Retrieval*. Butterworths, London, 1975.
- [Van Rijsbergen, 1979] C. J. Van Rijsbergen. *Information Retrieval*. Butterworths, London, 1979.
- [Verleysen, 2003] M. Verleysen. *Limitations and future trends in neural computation*, chapter Learning high-dimensional data, pages 141–162. IOS Press, 2003.
- [Vesanto and Alhoniemi, 2000] Juha Vesanto and Esa Alhoniemi. Clustering of the self-organizing map. *IEEE Transactions on Neural Networks*, 11(3) :586–600, May 2000.
- [Widrow and Hoff, 1960] B. Widrow and M. E. Hoff. Adaptive switching circuits. *Institute of Radio Engineers, Western Electronic Show and Convention, Convention Record*, pages 96–104, 1960.
- [Wilkinson *et al.*, 2003] David Wilkinson, Gareth Harries, Mike Thelwall, and Liz Price. Motivations for academic web site interlinking : evidence for the web as a novel source of information on informal scholarly communication. *Journal of Information Science*, 29(1) :49–56, 2003.
- [Wille, 1982] Rudolf Wille. Restructuring lattice theory : An approach based on hierarchies of concepts. In Ivan Rival, editor, *Ordered Sets*, volume 83 of *NATO Advanced Study Institute Series C*, pages 445–470. Reidel, Dordrecht, 1982.
- [Witten and Frank, 2000] Ian H. Witten and Eibe Frank. *Data mining : practical machine learning tools and techniques with Java implementations*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2000.

-
- [Yahya, 1997] M.A. Ould Mahamed Yahya. Comparaison de méthodes neuronales avec des méthodes d'analyse des données dans le cadre d'ingénierie de l'information. Master's thesis, Mémoire de stage de D.E.S.S. en Ingénierie mathématique et outils informatiques, Centre Elie Cartan, Université de Nancy I, France, 1997.
- [Young and Calvert, 1974] T. Y. Young and T. W. Calvert. *Classification, Estimation and Pattern Recognition*. American Elsevier, New York, 1974.
- [Zhang *et al.*, 1996] Tian Zhang, Raghu Ramakrishnan, and Miron Livny. Birch : An efficient clustering method for very large databases. In *ACM SIGMOD Workshop on Research Issues on Data Mining and Knowledge Discovery*, pages 103–114, Montreal, Canada, 1996.
- [Zhao and Karypis, 2001] Ying Zhao and George Karypis. Criterion functions for document clustering : Experiments and analysis. Technical report, 2001.
- [Zhao and Karypis, 2002] Ying Zhao and George Karypis. Evaluation of hierarchical clustering algorithms for document datasets. In *Proceedings of the International Conference on Information and Knowledge Management (CIKM)*, pages 515–524, New York, 2002.
- [Zhao and Karypis, 2003] Ying Zhao and George Karypis. Hierarchical clustering algorithms for document datasets. Technical report, Department of Computer Science, University of Minnesota, Minneapolis, MN 55455, 2003.
- [Zweig and Russell, 1998] Geoffrey Zweig and Stuart Russell. Speech recognition with dynamic bayesian networks. In *AAAI '98/IAAI '98 : Proceedings of the fifteenth national/tenth conference on Artificial intelligence/Innovative applications of artificial intelligence*, pages 173–180, Menlo Park, CA, USA, 1998. American Association for Artificial Intelligence.

Résumé

Nous proposons dans ce travail de thèse un nouveau paradigme basé sur les points de vue applicable à l'analyse et à la fouille de données. La mise en place d'estimateurs de qualité des classifications numériques indépendants des méthodes utilisées, tout comme de leur mode opératoire, nous permet de démontrer objectivement la supériorité de ce paradigme par rapport à une approche globale classique. Elle nous permet également d'intégrer à celui-ci des méthodes de classification non supervisées adaptées au traitement des données fortement multidimensionnelles s'organisant selon des distributions ultra-crauses, comme les données documentaires, tout en optimisant le fonctionnement de telles méthodes. Enfin, elle nous sert de base pour le développement de méthodes d'extraction numériques de règles d'association qui ne sont pas sujettes aux problèmes usuels des méthodes d'extraction symboliques. Ce travail se consacre également au développement et à la généralisation des mécanismes-noyaux du paradigme orienté par les points de vues, à savoir, le mécanisme de communication entre les vues, le mécanisme de comparaison dynamique de vues et le mécanisme de synthèse en ligne de vue, de manière à gérer à la fois les interactions locales et les interactions globales entre les vues. Ce travail s'attache finalement à corriger les défauts des méthodes de visualisation usuelles en proposant une approche originale basée sur les arbres hyperboliques qui permet de visualiser sans surcharge cognitive et à différentes échelles l'interaction entre les données multidimensionnelles analysées. Plusieurs expérimentations en vraie grandeur sont proposées sur différents types de données documentaires.

Mots-clés : analyse de données multi-vues, modèles neuronaux non supervisés, classification numérique, modèles multi-topographiques, méthodes symboliques, réseaux bayésiens, critères de qualité et optimisation des classifications numériques, extraction de connaissances, règles d'association, visualisation hyperbolique, analyse d'étiquettes, webométrie

Abstract

In this thesis, we propose a new paradigm of data analysis based on multiple viewpoints. This paradigm covers both the data analysis and the data mining fields. It also embeds specific visualisation methods. Setting up new classification quality estimators that are both independent of the classification methods and of their operational mode permits us to objectively demonstrate the superiority of the viewpoint-oriented paradigm as compared to the classical global approach. It also permits us to integrate in our paradigm clustering methods that are especially adapted to the management of highly multidimensional ultra sparse data, like documentary data, as well as to optimise the operational mode of such methods. Finally, it serves as a basis for developing association rule extraction methods that cope with the computation complexity and rule selection problems of the classical symbolic extraction rule methods. This work also focuses on the development and on the generalisation of the kernel mechanisms of the viewpoint-oriented data analysis paradigm, which are, the communication mechanism between viewpoints, the dynamic view comparison mechanism and the online view synthesis mechanism. It highlights how these mechanisms can be used for managing local, as well as global, view interactions. The final goal of this work is to cope with the defect of the classical visualisation methods by proposing an original approach based on hyperbolic tree allowing to visualise the interaction between the multidimensional data at different scales and without cognitive overload. Several realistic test experiments of our proposals are achieved on different kinds of documentary data.

Keywords : multiview data analysis, unsupervised neural models, clustering, multi-topographic models, clustering quality evaluation and clustering optimisation, symbolic methods, bayesian networks, data mining, association rules, hyperbolic visualization, label analysis, webometrics