Vers une approche systémique et multivues pour l'analyse de données et la recherche d'information : un nouveau paradigme

Jean-Charles Lamirel

UNIVERSITE DE NANCY 2

MEMOIRE D'HABILITATION A DIRIGER DES RECHERCHES :

VERS UNE APPROCHE SYSTEMIQUE ET MULTIVUES POUR L'ANALYSE DE DONNEES ET LA RECHERCHE D'INFORMATION : UN NOUVEAU PARADIGME

JEAN-CHARLES LAMIREL

Président du jury (rapporteur) : Pr. Bernard DOUSSET (IRIT, Toulouse)

Rapporteurs :

Dr. HDR. Michel ZITT (INRA, Nantes)
Pr. Wolfgang GLANZEL (KU, Leuwen)

Examinateurs :

Pr. Ronald ROUSSEAU (KHBO, Leuwen)
Pr. Mike THELWALL (Univ. Wolverhampton)
Pr. Odile THIERY (LORIA, Nancy)
Pr. Hildrun KRETSCHMER (Univ. Humbold, Berlin)
Dr. HDR. François CHARPILLET (LORIA, Nancy)
Dr. HDR. Claire GARDENT (LORIA, Nancy)
Pr. Amos DAVID (LORIA, Nancy)

Invités :

Dr ING. Jacques DUCLOY (LORIA, Nancy)

Table des matières

A. INTRODUCTION

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 compétences 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 globaux obtenus par les méthodes de fouille de données. Ces dernières, 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 d'analyse symboliques est le temps de calcul lié aux stratégies combinatoires utilisées [Simo99]. 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 [Koho01]. Un autre problème important des techniques de traitement 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 en combinant les techniques de classification avec des techniques complémentaires de projection.

Par ailleurs, 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 techniques 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 pour 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, voire locales. Cette dernière limitation nuit fortement à la qualité, 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 avoir 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 représentées dans un espace fortement multidimensionnel [Samm69]. De plus, la prise en compte de relations entre ces données pose des problèmes supplémentaires de surcharge cognitive propres à la représentation des graphes.

Les techniques d'apprentissage supervisé s'avèrent complémentaires aux techniques d'apprentissage non supervisé pour appréhender les résultats des analyses de données ou pour valider ces résultats. Utilisées dans un contexte plus large, elles doivent également permettre de prédire le comportement du système d'analyse lors de l'arrivée de données nouvelles, tout comme d'adapter celui-ci à partir de données propres au déroulement de l'analyse et aux actions de l'analyste. Les méthodes de catégorisation supervisées sont cependant souvent lourdes et peu flexibles. Elles ne tiennent pas compte de paramètres liés à la dynamique de l'apprentissage et se comportent de manière approximative sur des données complexes.

Le problème de l'analyse de données incrémentale, basée sur des données susceptibles de varier au cours du temps, est un problème complexe pour lequel peu de solutions fiables ont été proposées jusqu'à présent, bien que ce type d'analyse présente un très vaste champ d'applications. Il s'agit en effet de pouvoir détecter de nouveaux sujets au fil de l'arrivée des nouvelles données, ceci avec une réactivité suffisamment bonne, ce qui revient à définir des méthodes qui offrent un compromis optimal entre la plasticité et la stabilité lors de l'analyse, ainsi qu'une indépendance vis-à-vis de l'ordre d'arrivée des données [Prud04].

Le but de notre travail est donc de proposer un cadre général qui permette de résoudre de manière globale les problèmes liés à l'analyse et à la fouille de données complexes et volumineuses 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, certaines des tâches courantes de l'analyste utilisant une méthode de classification numérique seront de :
  - déterminer subjectivement le nombre de groupes, ou de classes, à utiliser,
  - estimer empiriquement la qualité de la méthode proprement dite,
  - supprimer les 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, voire 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 institutionnelles 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, sur la répartition géographique des thèmes de recherche, 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 termes 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, lesquelles 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 les comportements de référencement entre les universités par l'intermédiaire de la distribution des liens, entrants ou sortants, entre les sites associés, de figurer la répartition géographique des domaines de recherche, etc.
- le traitement des données variant au cours du temps.
- la validation de l'information produite par l'analyse de données à partir d'informations externes au processus d'analyse et la catégorisation d'informations à partir d'ensembles de données-exemples.
- l'adaptation au comportement de l'analyste et aux paramètres dynamiques de l'analyse.

B. TRAVAUX PRELIMINAIRES

Notre travail de recherche sur les Systèmes de Recherche et d'Analyse de l'Information (SRAI) a débuté dans le cadre de notre fonction d'ingénieur d'étude au Département Recherche et Produits Nouveaux de l'INIST (CNRS). Notre travail de base à l'INIST a consisté à mettre au point des outils de conception de réseaux de connaissances pour permettre l'analyse, aussi bien que la consultation associative, du contenu des bases documentaires. Il s'agissait notamment d'appliquer ces outils sur les bases de références bibliographiques PASCAL et FRANCIS. Les résultats de ce travail, et notamment la plate-forme hypertexte d'infométrie SDOC-HENOCH, ont été présentés dans le cadre du projet européen KWICK, et ont fait l'objet de plusieurs publications internationales [IC91a][IC91b][IC91c][IC93a]. La plate-forme SDOC-HENOCH est encore opérationnelle à ce jour dans le cadre du portail STANALYST de l'INIST.

Lorsque nous avons débuté notre travail de recherche doctorale dans le cadre de l'équipe EXPRIM du LORIA (Laboratoire LOrrain de Recherche en Informatique et ses Applications), nos travaux antérieurs sont rentrés directement en synergie avec certains des travaux de cette équipe, notamment ceux prenant en compte le comportement de l'utilisateur dans le processus de recherche et d'analyse de l'information, à l'image du système d'interrogation de bases d'images RIVAGE [Hali90] exploitant un concept de thesaurus dynamique dont le marquage évolue en fonction des actions de l'utilisateur. Cela nous a amenés à faire plusieurs constatations importantes :

- Le mode de communication usuel entre un utilisateur et un SRAI dans lequel l'utilisateur joue le rôle d'émetteur actif et le système celui de récepteur passif est sans doute à l'origine d'un grand nombre des problèmes posés par les systèmes traditionnels. Ce mode de

communication, naturellement peu riche, ne permet pas de tenir correctement compte des diverses imperfections qui touchent aussi bien la définition du besoin de l'utilisateur que la connaissance du SRAI concernant le contenu du fonds documentaire incriminé. Ceci amène à penser que les processus de recherche et d'analyse de l'information, pour être améliorés, doivent être envisagés comme des **actes de communication à double sens** entre l'utilisateur et le SRAI, permettant l'enrichissement mutuel des connaissances de l'un comme de l'autre. Cette dernière approche, qui s'inspire fortement de la théorie systémique de la communication [Watz67], impose non seulement d'améliorer l'interactivité entre l'utilisateur et le SRAI, mais également d'équilibrer leurs rôles respectifs lors des opérations de recherche et d'analyse de l'information. Elle suggère aussi d'accorder au bruit, apparaissant inévitablement dans la communication entre l'utilisateur et le SRAI, un **rôle constructif,** aussi bien lors de la phase opérationnelle que lors de la phase d'apprentissage pour le SRAI.

- Le développement des réseaux mondiaux d'information a déjà commencé à entraîner un glissement de la recherche et de l'analyse de l'information traditionnelles, opérant sur une base unique, vers des formes fortement distribuées impliquant des fonds multiples. L'avenir de ces derniers types de recherche et d'analyse repose cependant sur la mise en place de nouvelles méthodes d'accès aux informations contenues dans les différents fonds qui, pour être efficaces, devront faire usage de structures fortement parallélisées. Mener à bien de tels types de recherche et d'analyse nécessitera également de pouvoir **acquérir dynamiquement des connaissances** concernant les différents fonds interrogés, ainsi que de pouvoir **harmoniser l'interaction entre ces connaissances**. Ceci implique bien évidemment que la connaissance sur un fonds, au cas où elle n'est pas directement disponible, puisse être **construite** de manière non supervisée et visualisée de manière adéquate. L'avènement de la recherche et de l'analyse distribuées encourage également à penser que la notion d'évolutivité d'un SRAI devra elle-même être remise en cause puisqu'elle devra entre autres prendre en compte la capacité de ce SRAI à intégrer de nouveaux fonds à une échelle de temps proche de celle du temps réel.

Ces constatations nous ont incités à revoir en profondeur l'architecture des systèmes de recherche et d'analyse de l'information, ainsi que la forme des connaissances qu'ils sont amenés à manipuler. En nous inspirant fortement de la théorie systémique de la communication, nous avons donc pu généraliser l'approche SDOC-HENOCH afin de concevoir un modèle de système d'interrogation de bases documentaires, qui, tout en intégrant des fonctions spécifiques de consultation associative et d'analyse de contenu, permette également la formulation de requêtes, ainsi que la gestion de l'interaction avec l'utilisateur. Le modèle intègre finalement des mécanismes de correction de ses connaissances et de celles de l'utilisateur afin d'améliorer ses performances à court et à long terme.

Le modèle NOMAD ainsi obtenu est le premier modèle de système de recherche documentaire à introduire la représentation des données documentaires selon des vues multiples et la communication entre les vues. Cette approche symbolico-connexionniste s'inspire à la fois des travaux de Kohonen sur les structures neuronales auto-organisatrices [Koho01] et de ceux de Grossberg sur la théorie de la résonnance adaptative [Gross87]. Elle définit un nouveau modèle de représentation de connaissances fondé sur la notion de cartes topographiques SOM multiples communiquant entre elles. Chaque carte est constituée d'une structure neuronale de faible dimensionnalité. Elle fournit un point de vue synthétique particulier sur le contenu de la base

documentaire ainsi qu'une interprétation de la distance entre les documents de la base selon ce point de vue. Elle sert aussi bien d'outil de raisonnement au système que d'outil de navigation à l'utilisateur. Le modèle gère l'interaction avec l'utilisateur grâce à un mécanisme original de mémoire de session fondé sur la détection de nouveauté [Koho01]. Ce mécanisme permet à la fois de synthétiser le besoin de l'utilisateur, de mettre en évidence de nouvelles alternatives de recherche, et d'évaluer la cohérence des décisions prises par l'utilisateur. Un apprentissage à posteriori permet la modification dynamique de la position des documents sur les cartes topographiques ainsi que la création ou la modification des liens entre ces documents. Il permet également de moduler l'effet des différents points de vue lors de l'interrogation. Cet apprentissage est conditionné par le niveau de compétence estimé de l'utilisateur. Une gestion générale des stratégies de recherche est également assurée par le modèle. La stratégie de recherche peut changer en cours de session en fonction des résultats obtenus et du comportement de l'utilisateur. En cela, le modèle proposé présente un certain nombre d'avantages par rapport aux modèles existants :

- En fournissant des mécanismes élaborés de complétion de requêtes, d'interrogation par l'exemple et d'interrogation dirigée par les thèmes, fonctionnant de manière unifiée, il permet à l'utilisateur de combiner de manière constructive et complète les phases de consultation interactive et les phases de formulation de requête lors d'une session d'interrogation.
- L'utilisation conjointe de la connaissance de base (cartes topographiques multiples) et de celle enregistrée dans la mémoire de session a permis de mettre en place des mécanismes originaux de gestion des contradictions intervenant lors des décisions de l'utilisateur. Ces mécanismes qui s'avèrent particulièrement cruciaux dans tout système documentaire utilisant l'interaction avec un utilisateur (bouclage de pertinence, interrogation par l'exemple) n'avaient pourtant pas été utilisés jusqu'ici.
- Le modèle permet à la fois de caractériser et de gérer différents types de recherche, comme la recherche précise, la recherche exploratoire, la recherche thématique, et la recherche connotative.
- Étant donné la forme des connaissances qu'il manipule, le modèle NOMAD offre des possibilités d'applications inédites, comme l'interrogation multi-bases intelligente et réactive avec mise en correspondance dynamique du contenu des différentes bases.
- Le modèle est exploitable sur des fonds existants sans modification directe de leur indexation initiale.

Une première implémentation prototype du modèle complet a permis de démontrer l'intérêt spécifique de cette approche pour la recherche d'information. Ce travail a fait l'objet de plusieurs publications internationales [IC94a][IC94b][IC94d]. Il représente également le thème central de notre thèse. La **figure 43** décrit l'architecture du modèle NOMAD original.

C. LE MODELE NOMAD-MULTISOM

Le fort potentiel de la composante de raisonnement thématique du modèle NOMAD nous a amené à en étudier une exploitation plus ciblée pour l'analyse de données. Nous avons développé pour cela une spécialisation du modèle NOMAD, nommée NOMAD-MULTISOM, qui en reprend uniquement son principe de cartes topographiques multiples, ou modèles de classification multiples, et celui de ses mécanismes noyaux associés, à savoir le mécanisme de généralisation

des modèles et le mécanisme de communication entre les modèles et entre les vues. En proposant une interface d'interrogation et de visualisation interactive élaborée, cette spécialisation avait plus directement pour but d'adresser le problème de l'analyse multi-vues des données documentaires et multimédia.

La **Figure 1** illustre le principe de partition par vues exploité dans l'analyse multi-vues proposée par le modèle NOMAD-MULTISOM. La **Figure 2** et la **Figure 3** illustrent respectivement le principe de la communication entre les modèles, et celui de la généralisation des modèles. La **Figure 4** synthétise le fonctionnement du noyau du modèle NOMAD-MULTISOM.

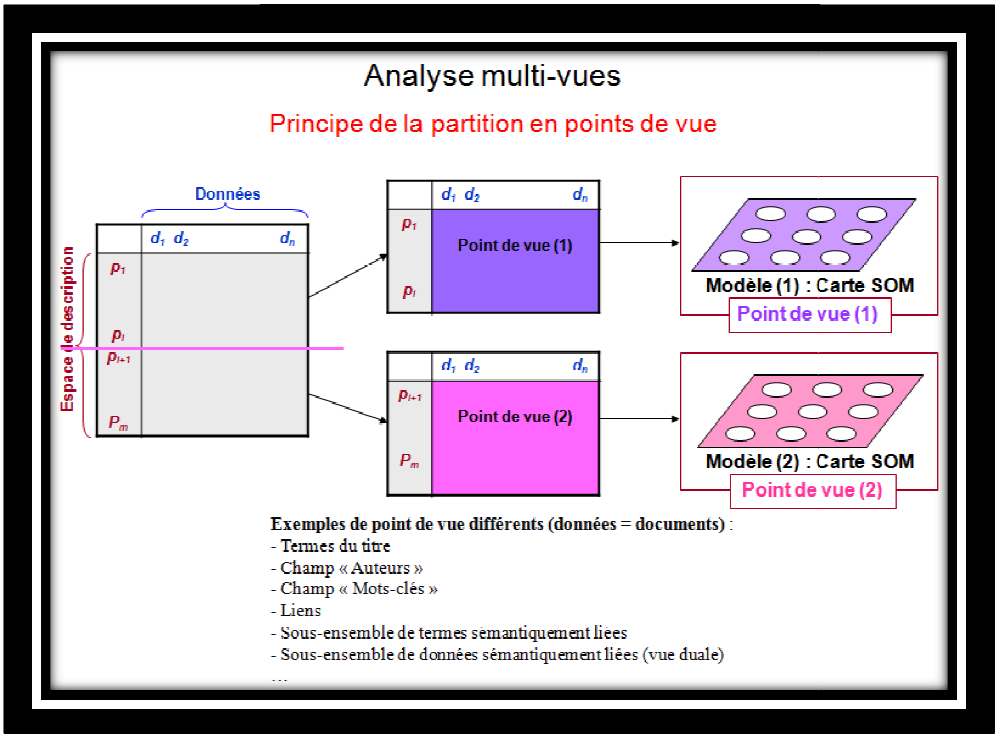

***Figure 1** : Le principe de l'analyse par vues, établi pour la première fois par l'approche NOMAD (cf. **section B**), revient à considérer l'espace de description des données comme un espace « partitionnable » en sous-espaces, éventuellement recouvrants, attachés à des domaines sémantiques différents. La partition en vues peut être simplement basée sur une structuration logique explicite inhérente aux données, comme la structuration en champs souvent présente dans les données textuelles. Elle peut également s'opérer de manière duale en constituant des sous-groupes de données associées à des étiquettes ou à des critères prédéfinis. Dans tous les cas, cette opération est liée à un travail d'expertise préalable, opérée par l'analyste, et dépend des finalités de l'analyse.*

Dans l'approche NOMAD-MULTISOM, une carte SOM (modèle de classification) est spécifiquement construite (i.e. apprise) pour chaque point de vue.

C.1. Fouille de données textuelles

Les capacités de déduction automatique du modèle NOMAD-MULTISOM représentent un sérieux atout par rapport aux méthodes de fouille et d'analyse usuelles. Ces dernières ne permettent en effet pas d'exploiter dynamiquement plusieurs points de vue, qui peuvent être considérés comme plusieurs macro-dimensions, relatives aux mêmes données.

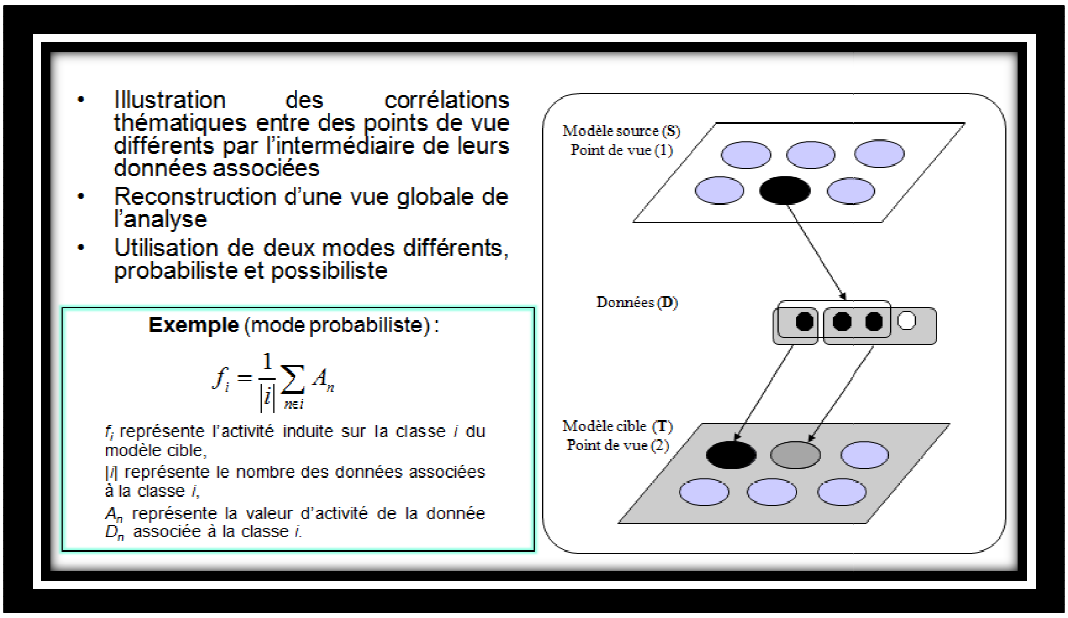

Figure 2 *: Dans l'approche NOMAD-MULTISOM, la communication entre les modèles s'opère par activation de classes sur un modèle source, puis par transmission de l'activité aux données associées à ces classes. Les classes destinations qui partagent les données activées depuis les classes sources avec ces dernières deviennent à leur tour active dans des proportions qui dépendent de celles de leurs propres données activées. Ce processus s'apparente à un processus de raisonnement bayésien à partir d'un réseau bayésien construit de manière non supervisée. Il est également dynamique étant donné que de nouveaux modèles peuvent être intégrés à tout moment dans le processus.*

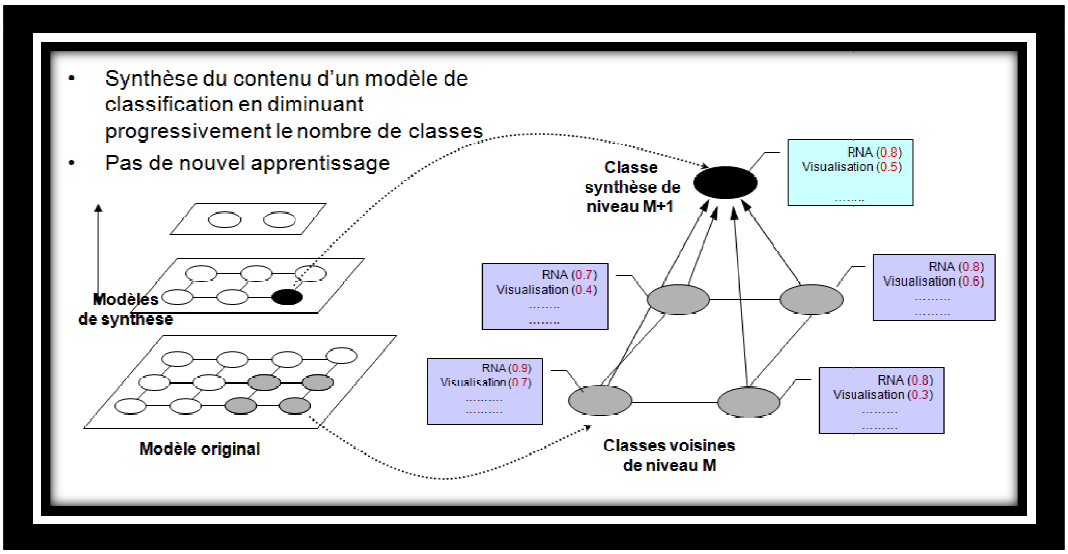

Figure 3 *: Dans l'approche NOMAD-MULTISOM, le processus de généralisation, propre à une vue, consiste à construire des modèles de classification « de synthèse » de plus en plus généraux en opérant par étapes successives à partir du modèle original de la vue, obtenu par apprentissage. Cette méthode, qui ne nécessite pas de nouvel apprentissage, se base sur la combinaison des profils de classes voisines. Elle permet aux modèles généraux d'une vue de conserver les propriétés topographiques du modèle original de la vue.*

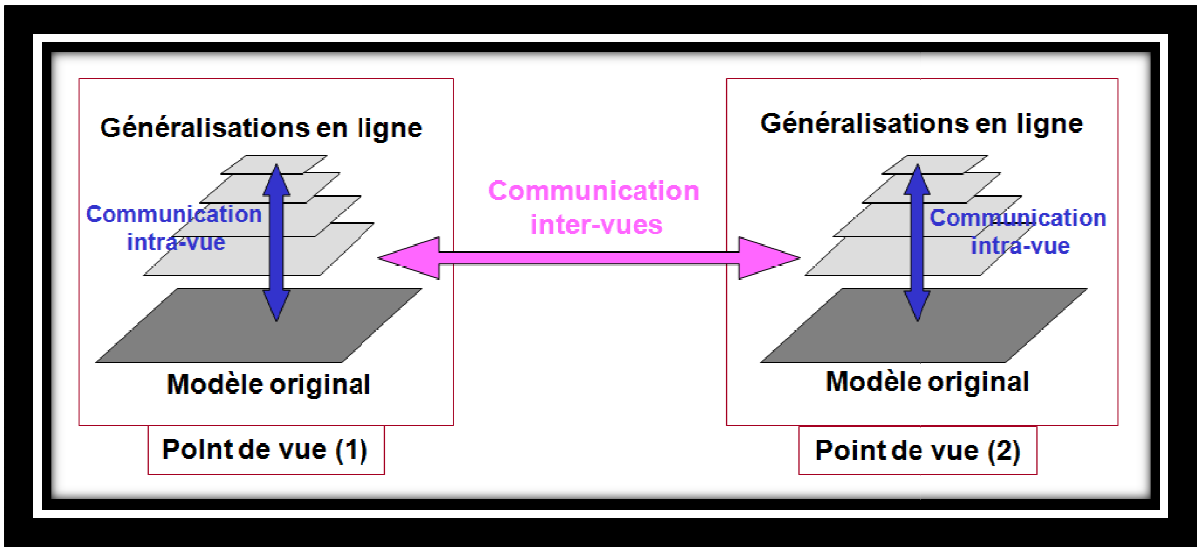

***Figure 4** : Dans l'approche NOMAD-MULTISOM, le principe de communication inter-modèles, décrit à la **Figure 2**, peut aussi bien s'appliquer entre des modèles appartenant à des points de vue différents qu'entre un modèle original et ses propres modèles de synthèse dans un point de vue donné.*

La première série d'expériences que nous avons menées sur l'analyse de brevets, avec l'aide d'experts du domaine, a révélé l'intérêt manifeste de l'approche NOMAD-MULTISOM relativement aux méthodes de type k-means [MacQ67], classification du simple lien [Mich88], ou perceptron neuronal élémentaire [Ould97]. Ces expériences mettent en jeu une indexation de l'information plein texte extraite des brevets, une répartition des index selon leurs contextes sémantiques, et une exploitation des contextes sous forme de classification multiples, non supervisées. Elles montrent que, contrairement aux autres méthodes, l'approche NOMAD-MULTISOM permet à un expert de répondre simplement, et si besoin interactivement, à des questions complexes comme : "Quelles sont les technologies utilisées par tel déposant et non par tel autre et quelles sont les propriétés émergentes de ces technologies et leur domaines d'utilisation potentiels?". Comme le montre la **Figure 5**, les questions et les réponses peuvent être appréhendées de manière interactive sur les cartes topographiques associées aux différentes vues. Il est également possible de faire usage de la négation lors de l'interrogation. Cette approche permet de plus d'opérer des analyses plus fines qu'une approche globale puisqu'elle permet d'éliminer le bruit inhérent à cette dernière, tout en offrant la possibilité de restituer un point de vue global par communication entre des points de vue locaux, ce que nous avons pu démontrer objectivement par la suite (cf. **section D.1**).

L'analyse intelligente des bases de références bibliographiques et des bases de brevets a donné lieu à plusieurs publications dans des conférences internationales [IC01c][IC01e][IC03c][IC06h] et à un article dans un revue internationale [IJ01a]. La référence [IC06h] est ajoutée en annexe. Ce travail de communication scientifique sur le modèle, de même que les résultats présentés, ont éveillé l'intérêt de la communauté d'évaluation des sciences vis à vis de l'approche multi-vues pour mettre en place des analyses à grande échelle des données Web liées à la recherche institutionnelle. Ceci qui a permis à l'approche NOMAD-MULTISOM d'être choisie comme l'une des deux approches de référence du projet européen IST-EISCTES [TR03a][TR03b], centré autour de la mise en place de méthodologies dédiées à l'analyse de l'impact de la recherche

européenne sur la nouvelle économie. Dans ce contexte, l'approche NOMAD-MULTISOM a principalement été exploitée pour croiser dynamiquement des analyses de liens avec des analyses de contenu, elles-mêmes menées selon des critères d'analyses multiples, des pages Web liées à la recherche institutionnelle européenne. Ce type d'expérience de webométrie, totalement inédite jusqu'alors, a bénéficié d'un impact important dans le domaine, de même qu'il a été très favorablement accueilli par les experts chargés de rapporter le projet auprès de la commission européenne. Il a mené à 2 publications de niveau international [IC03d][IC04a] à 2 publications dans des journaux internationaux [IJ04a][IJ04b], et à 2 rapports techniques de projet [TR03a] [TR03b].

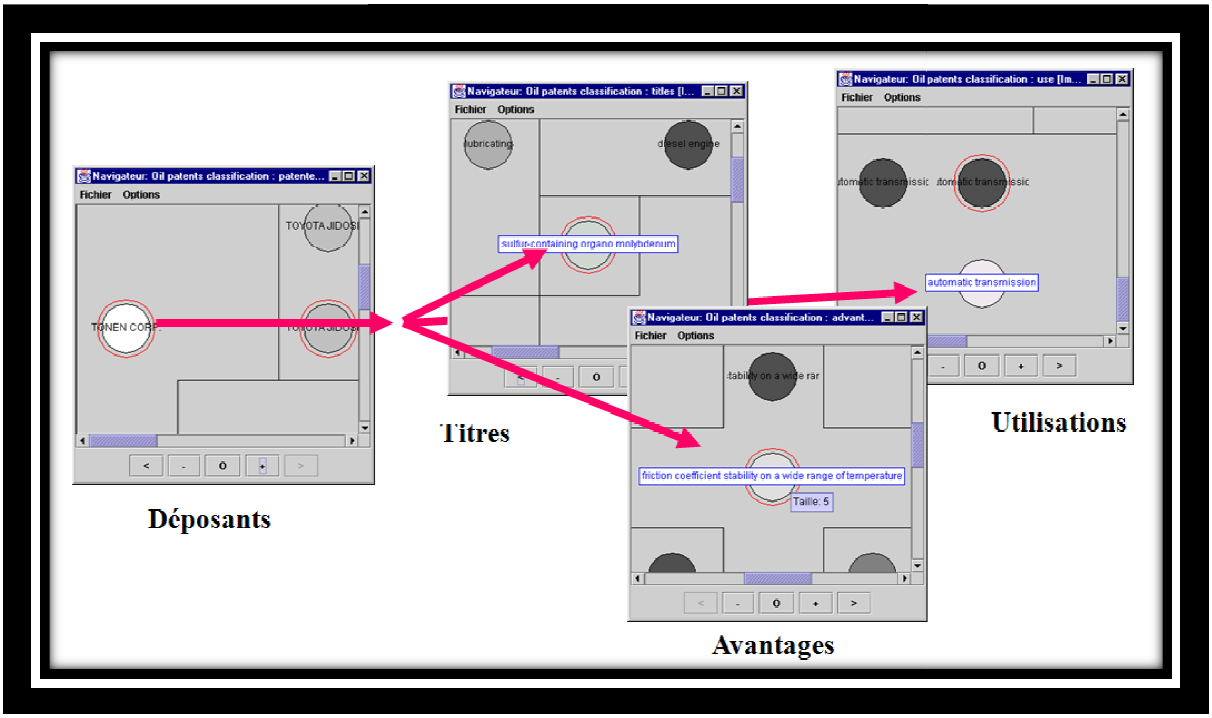

***Figure 5** : Exemple de déduction opérée de manière interactive par croisement non supervisé d'informations thématiques associées à 4 points de vue différents (**Déposants**, **Titres**, **Avantages**, **Utilisations**) sur les mêmes données (ici des brevets sur les **huiles moteur**) en utilisant l'approche NOMAD-MULTISOM.*

*Chaque point de vue est associé à un modèle de classification non supervisé représenté par une carte SOM spécifique. Les correspondances thématiques sont établies par transmission d'activité entre les classes par l'intermédiaire des données partagées par les modèles. Ici, à partir du point de vue **Déposants**, le système déduit dynamiquement que la société TONEN (classe activée par l'usager, en blanc sur la carte **Déposants**) produit spécifiquement des lubrifiants de synthèse à base de Molybdène exploitables dans les **transmissions automatiques** et dont la propriété est d'être **très stable en température** (classes déduites par inférence bayésienne, en grisé clair sur les cartes **Titres**, **Avantages** et **Utilisations**).*

Les travaux de développement d'une version logicielle opérationnelle de l'approche NOMAD-MULTISOM, incluant une interface interactive élaborée pour l'interrogation et la visualisation des résultats des analyses multi-vues, ont été assurés grâce à l'assistance et à

l'encadrement d'un ingénieur dont le financement de 18 mois a été pris en charge par le projet EICSTES. A l'issue de ce projet, une version finalisée du logiciel NOMAD-MULTISOM, complétée d'un manuel d'utilisation détaillé, a pu ainsi être mise à la disposition de l'ensemble des partenaires. Notre intervention dans le projet EICSTES nous a notamment amenés à modifier le mécanisme de communication inter-modèles et inter-vues, de manière à le généraliser à une transmission d'activité basée indifféremment sur l'espace des données (modèle original présenté à la **Figure 2**) ou sur l'espace dual de leurs propriétés (extension réalisée). Ce nouveau principe étend largement le champ de l'analyse multi-vues puisqu'il permet d'envisager, avec un seul et même modèle fédérateur, de nombreux types d'analyse dynamique et multi-vue supplémentaires (analyses basées sur le croisement de propriétés associées aux données, analyses basées sur le croisement de liens existants entre les données, analyses mixtes, etc.). Une nouvelle version du logiciel NOMAD-MULTISOM intégrant ces dernières modifications, a fait l'objet par l'INRIA, d'un dépôt à l'agence française pour la protection des programmes (APP) (Numéro APP : **IDNN.FR.001.460018.00.R.P.2000.000.1000**).

L'ensemble de ce travail nous a finalement ouvert de nombreuses portes pour des applications de notre approche dans le domaine de l'évaluation des sciences, domaine ou les chercheurs sont encore peu familiers avec les techniques de traitement issues de l'intelligence artificielle, alors que celles-ci s'avèrent extrêmement prometteuses, voir indispensables, pour y mener des analyses à grande échelle.

C.2. Fouille de données multimédia

La fouille d'information multimédia nécessite de pouvoir exploiter la synergie entre plusieurs représentations des mêmes objets, ou d'objets significativement proches, selon différents points de vue, susceptibles eux-mêmes être associés à différents média. L'avantage intrinsèque du modèle NOMAD-MULTISOM dans ce domaine est qu'il ne limite ni le nombre, ni le type des points de vue qui peuvent être utilisés, ni, enfin, les médias associés à ces points de vue. La nature intuitive du mécanisme de communication entre les cartes topographiques favorise des opérations de découverte interactive d'information et de déduction automatique guidées par les besoins propres d'un utilisateur non nécessairement expérimenté; ce qui fournit à ce modèle de nombreuses applications dans le domaine des librairies électroniques multimédia. Les expériences préliminaires que nous avons menées avec ce modèle confirment son utilité directe, notamment pour les tâches d'interprétation et de navigation qui mettent en jeu de fortes composantes iconographiques. Après de premiers tests particulièrement encourageants sur une base iconographico-textuelle traitant de l'art nouveau nous avons décidé, pour illustrer concrètement les capacités du modèle dans le domaine du multimédia, de l'expérimenter comme modèle-cœur dans deux applications-test significatives du domaine des librairies électroniques multimédia.

Un premier prototype d'interface de navigation multimédia a été mis en place à la Villette, au service des collections muséologiques. Cette interface a pour but d'assister au montage d'expositions à partir d'une collection de représentations 3D d'objets disponibles. Elle permet en outre d'établir dynamiquement des correspondances entre les objets qui dépendent du point de vue courant choisi par le conservateur, ou le monteur d'exposition, sur ces objets.

D'un autre côté, le travail sur les collections digitales de papillons de Taiwan a donné lieu à une autre approche, dans laquelle un modèle de SRAI utilisant en arrière-plan un moteur neuronal de

type NOMAD-MULTISOM a été proposé. Tout en permettant d'obtenir des résultats supérieurs à ceux d'un SRAI classique, cette solution décharge l'utilisateur novice de l'exploitation directe des résultats fournis par le modèle, dans le cas ou ceux-ci pourrait s'avérer trop complexes à interpréter : seuls les résultats finaux des interrogations et les cas ambigus nécessitant d'opérer des choix seront donc présentés à l'utilisateur (cf. **figure 6** et **Figure 7**). L'ensemble de ces travaux ont été présentés dans plusieurs publications de niveau international [IC00a][IC00e][IC01a][IC01b]. L'application « Papillons de Taiwan » nous a assuré, par la suite, une collaboration permanente avec le National Science Council (NSC) taïwanais.

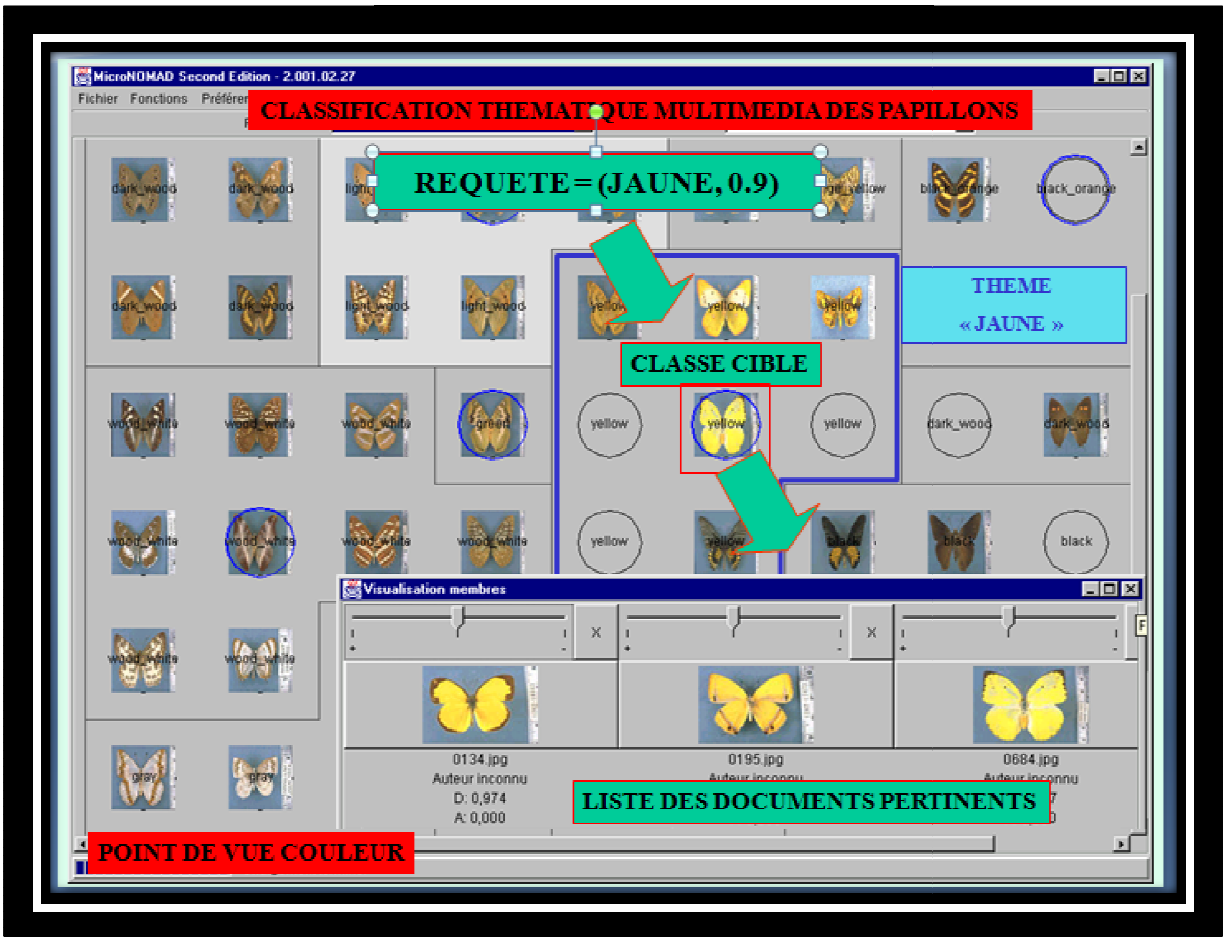

Figure 6 *: Dans le cas de la navigation multimédia, les cartes SOM du modèle NOMAD-MULTISOM peuvent faire office de mosaïque de navigation, aussi bien que servir d'interface d'interrogation thématique pour le système (en batch) ou pour l'utilisateur (en mode interactif). La carte SOM présentée ici correspond à une classification des papillons construite de manière non supervisée à partir d'une analyse des d'informations colorimétriques associées aux images de papillons. Dans le cas présent, la figure matérialise également le résultat du processus d'interrogation thématique généré par une requête de type* ***R=Jaune****.*

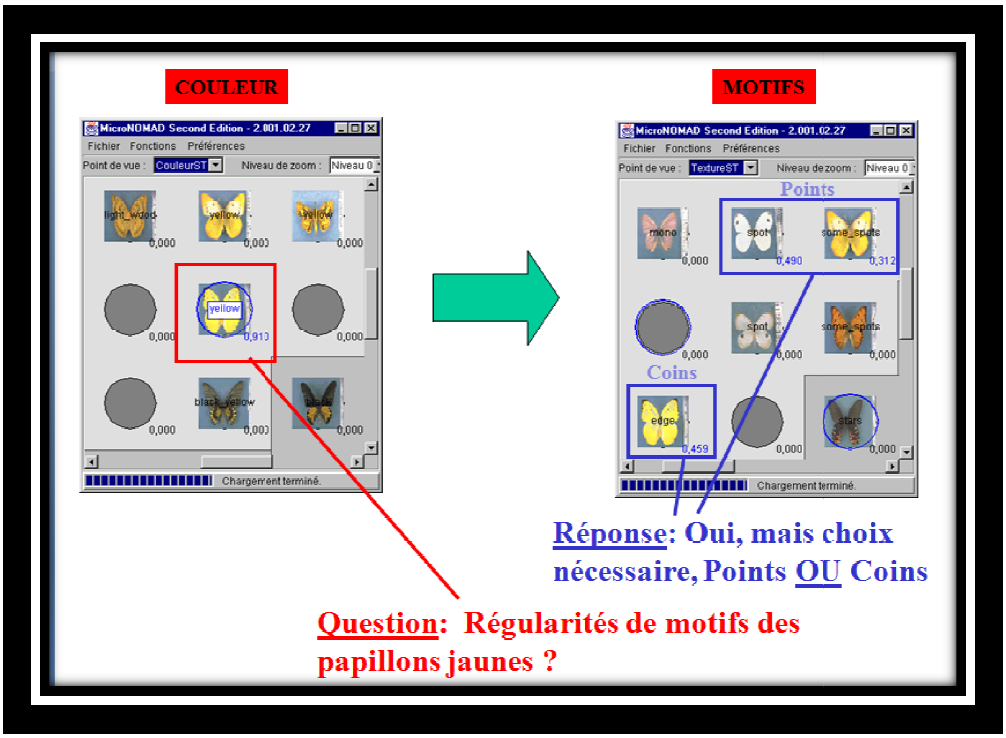

***Figure 7** : La figure matérialise le résultat du croisement d'informations entre un modèle de classification fondé sur la **Couleur** et un autre modèle fondé sur les **Motifs** (apparaissant sur les ailes des papillons). Le processus s'appuie sur l'activation de la classe des papillons **jaunes** sur le modèle de classification fondé sur la **Couleur**. L'activité résiduelle générée sur le modèle de classifications fondé sur les **Motifs** montre qu'il n'existe pas de réponse homogène concernant le type de motifs présents sur les ailes des papillons **jaunes** (soient des **Points**, soient des **Coins**). Cette situation matérialise également l'ambigüité susceptible d'apparaitre dans le cas où l'utilisateur formulerai une requête multi-vues de type **R=(Jaune ET Points ET Coins)**. En opérant à partir de cette requête, et en arrière-plan, un raisonnement thématique entre les vues du type de celui présenté ci-dessus, le système peut alors suggérer à l'utilisateur de faire un choix thématique entre les **Points** et les **Coins**, étant donné que ces deux critères sont thématiquement incompatibles au regard du contenu du fonds. Il peut même exposer à l'utilisateur visuellement l'ambigüité détectée. Ce mode opératoire est caractéristique du fonctionnement systémique de l'approche NOMAD-MULTISOM.*

D. VERS UN PARADIGME GENERAL D'ANALYSE DE DONNEES MULTIVUES

En suivant la même démarche que celle adoptée dans notre travail antécédent, et tout en privilégiant l'exploitation de modèles neuronaux qui présente l'avantage de procurer un parallélisme implicite au processus d'analyse de données, et celle des données documentaires qui s'avèrent être de nature complexe, nous avons continué à explorer les possibilités de coopération systémiques entre des modèles différents, spécialisés chacun dans des tâches précises, dans un processus global d'analyse et de fouille de données. Comme nous le montrons par la suite, cette démarche a finalement impliqué de développer la cohabitation entre le raisonnement neuronal et le raisonnement symbolique, ou entre des modèles de nature différente, de manière à couvrir l'ensemble des fonctions liées à l'analyse de données documentaires et à éliminer, sinon à réduire, les défauts inhérents à chacun des types d'approche. Elle nous a fourni par ailleurs une

base très solide pour définir un nouveau paradigme général d'analyse de données basée sur les points de vue multiples, à savoir le paradigme MVDA (**M**ulti **V**iewpoint **D**ata **A**nalysis).

La validation objective de ce paradigme nécessitait, dans un premier temps, de consolider les mécanismes d'évaluation des résultats des méthodes de classifications numériques. Pour cela, nous avons mis en place de nouveaux mécanismes d'évaluation originaux. Ces mécanismes nous ont également permis de mener à bien des comparaisons objectives entre différentes méthodes de classification non supervisée, de manière à sélectionner les méthodes les plus pertinentes, mais également de corriger certains des défauts, ainsi que d'optimiser le mode opératoire des méthodes les plus intéressantes.

Nous nous sommes ensuite attelés à proposer une approche générique pour la communication entre les modèles de classification et entre les vues pouvant être appliquée aux résultats de méthodes de classification différentes. Les mécanismes-noyaux inhérents à cette approche devaient permettre de gérer à la fois les interactions locales et les interactions globales entres les modèles et entre les vues, de manière à pouvoir mener parallèlement des analyses générales et des analyses ciblées. De manière connexe, nous nous sommes attachés à définir des mécanismes de généralisation des modèles de classification, à la fois compatibles avec l'ensemble des méthodes de classification et calculables en ligne, c'est-à-dire, sans faire intervenir une nouvelle phase d'adaptation aux données.

Ces premiers travaux nous ont ensuite permis de définir un nouveau cadre pour la fouille de données à partir des méthodes de classification numériques. Nous avons notamment mis en place une nouvelle méthodologie d'extraction de règles d'association, fonctionnant à la fois sur des modèles de classifications élémentaires, mais permettant également de tirer parti des mécanismes de généralisation des modèles et de communication inter-modèles et inter-vues.

Nous nous sommes ensuite attachés à corriger les défauts des méthodes de visualisation existantes. Nous avons cherché pour cela à mettre en place 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.

L'ensemble des points complémentaires découlent directement de ce travail, à savoir, la mise en place d'outils de validation fine pour la classification non supervisée, et le traitement supplémentaire des données en mode supervisé.

Par la suite, nos résultats seront donc présentés synthétiquement selon huit volets différents :

1. Comparaison et optimisation des résultats des classificateurs numériques sur les fonds documentaires.
2. Généralisation du paradigme d'analyse de données multi-vues MVDA.
3. Coopération numérique-symbolique.
4. Coopération de classificateurs.
5. Méthodes de visualisation des résultats de classification multidimensionnels.
6. Validation fine des résultats des classificateurs numériques.
7. Méthodes d'analyse diachroniques.
8. Nouveaux outils pour le filtrage personnalisé d'information et l'apprentissage supervisé.
9. Travaux annexes.

Dans le cadre général de ces recherches, nous nous sommes plus particulièrement intéressés à l'analyse des données du Web. Nous présenterons donc brièvement quelques résultats d'analyse obtenus. Nous terminerons notre présentation par une synthèse générale du fonctionnement du modèle MDVA.

Il est à noter que l'ensemble des travaux présentés ci-après ont mené à l'aboutissement de quatre thèses différentes [PH06a][PH06b][PH06c][PH09a]. Ils ont également généré une production scientifique très importante : 8 contributions dans des journaux internationaux, 9 conférences invitées, l'organisation d'une conférence internationale et 7 chaires de session dans des conférences internationales, 80 publications dans des conférences internationales, 3 chapitres de livres, 3 publications dans des conférences nationales, 5 rapports de projet européens. Il nous a également permis de valider 12 stages de fin d'étude d'ingénieur, de DEA, ou de Master, avec une politique de publication systématique en collaboration avec les étudiants concernés.

Ces travaux nous ont également valu dans leur ensemble la reconnaissance de nombreux partenaires étrangers institutionnels prestigieux comme NIEHS (USA), NSC (Taïwan), KU Leuwen (Belgique), NISTAD (Inde), et WISELAB (Chine). Dans le domaine spécifique de l'évaluation des sciences, notre activité nous a valu l'invitation au comité éditorial du journal international : "**Collnet Journal of Scientometrics and Information Management**".

D.1. Comparaison et optimisation des résultats des classificateurs numériques sur les fonds documentaires

Lorsqu'une classification numérique est menée sur de gros corpus, et en particulier sur des données textuelles, dont les caractéristiques sont souvent d'être peu discriminantes, il devient alors crucial de déterminer d'une manière objective la qualité des classifications obtenues, ainsi que d'optimiser le nombre de classes produites. Dans ce cadre, les méthodes usuelles basées sur l'inertie [Ould97] sont souvent inopérantes et sujettes à des biais qui dépendent fortement de la méthode de classification utilisée. Les critères de **Rappel** et de **Précision** que nous avons mis en place en vue de l'évaluation de la qualité des classifications sont inspirés à la fois des méthodes symboliques et des mesures de qualité issues des systèmes de recherche d'information. Ces critères mesurent de manière non supervisée l'homogénéité des classes en évaluant la cohérence et l'exhaustivité des propriétés des données qui ont été rattachées à la même classe. Avant la mesure, chaque propriété est associée à une classe support, qui représente la classe qui la maximise (cf. **Figure 8**). Ces critères présentent un avantage prépondérant par rapport aux critères usuels de qualité de classification, tels que les inerties intra-classe et inter-classes, ou leurs formes dérivées, qui est celui d'être indépendants de la méthode de classification utilisée.

Ils permettent donc de comparer les résultats de classification issus de différentes méthodes. De manière conjointe, ils peuvent être utilisés afin d'optimiser le nombre de classes obtenues par une méthode donnée, là où cette opération serait indécidable avec des critères usuels. Un exemple est donné à la **Figure 9**. Nous avons montré théoriquement que ces critères de qualité, dans le cas où ils atteignaient leur valeur limite, se comportaient comme des estimateurs symboliques. Ils permettent donc d'assimiler le comportement idéal d'une classification numérique à celui d'une classification symbolique opérant une sélection de classes dans un treillis de Galois virtuel.

Les résultats de ce travail, qui sont importants dans le domaine, ont fait l'objet d'une publication dans une conférence internationale [IC03d], qui nous a elle-même donné directement accès à une revue internationale [IJ03d]. Cette dernière référence est ajoutée en annexe.

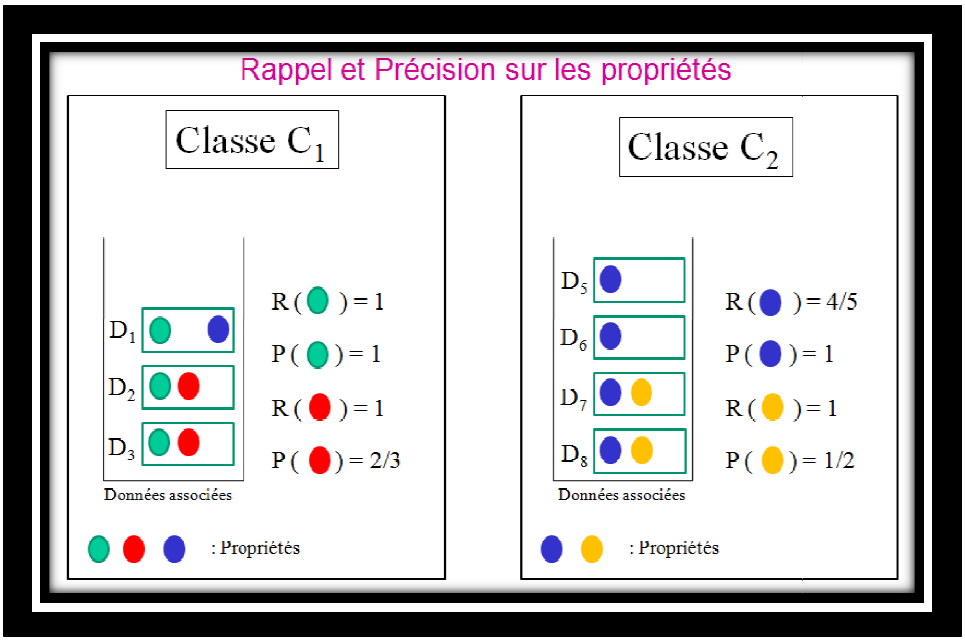

Figure 8 *: Illustration du principe des mesures de Rappel (R) et de Précision (P) sur les propriétés des données associées aux classes. Avant la mesure, chaque propriété est associée à la classe qui la maximise. Dans cet exemple la propriété « **bleue** » est maximisée par la classe C2. Ses valeurs de Rappel et de Précision seront donc calculées en se basant sur son comportement vis-à-vis des données de ladite classe.*

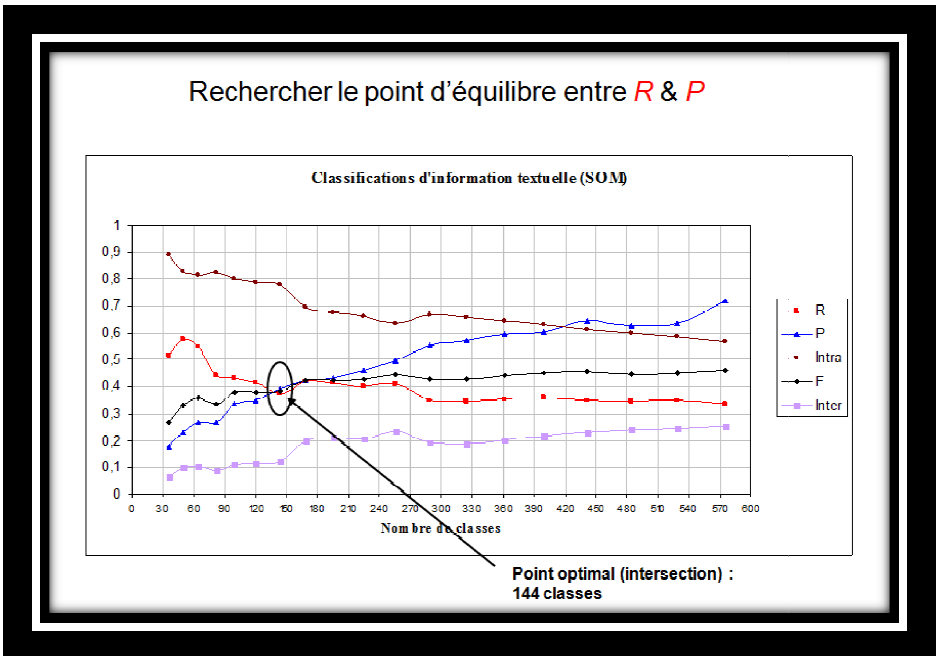

Figure 9 *: L'identification du modèle de classification optimal s'opère en recherchant le premier point de compromis (intersection, ou break-even) entre les valeurs de R et de P. Dans le cas où aucun point de compromis n'est trouvé, c'est le modèle qui maximise la moyenne harmonique (F-mesure) entre les valeurs R et P qui est considéré comme optimal. Comme le montre également cet exemple, l'exploitation des courbes d'inertie (Inter et Intra) dans le même but peut s'avérer très complexe, voire impossible.*

A l'aide de ces mêmes critères, nous avons également pu démontrer objectivement la supériorité de l'approche d'analyse de données basée sur les points de vues multiples vis-à-vis de l'approche classique d'analyse globale. Le principe de cette démonstration est présenté, sous forme résumée dans la référence [IC06h], donnée en annexe.

Selon une approche complémentaire à la précédente, nous avons adressé le problème de la classification des données associées des étiquettes multiples, indifféremment du fait que ces dernières soient de nature endogène ou exogène au processus de classification. Ce type de problème dépasse le cadre de l'évaluation des modèles de classifications, pour toucher également celui de la prédiction. Sa résolution donc peut donc être considérée comme un nouveau champ fertile de recherche. Dans ce contexte, nous avons proposé une alternative à notre approche précédente basée sur le **Rappel** et la **Précision**. Cette alternative exploite des mesures de similarité entre les données classées par une méthode de classification non supervisée et la distribution d'étiquettes projetées sur le modèle de classification. Contrairement à notre approche précédente, les nouvelles mesures de **C-Rappel** et de **C-Précision** résultantes, qui se basent la théorie des probabilités, sont calculées sur l'ensemble des classes et correspondent à des probabilités conditionnelles de type **étiquette|classe** et **classe|étiquette**. De plus, dans ce cas précis, ces mesures peuvent être complétées par des mesure d'entropie associées.

Comme les premières mesures de **Rappel** et de **Précision** que nous avons proposées, ces nouvelles mesures s'avèrent également utiles pour l'évaluation de la qualité de la classification et pour la sélection de modèle optimal. De plus, comme le montre la **Figure 10**, elles permettent de résoudre certains autres des problèmes centraux de la classification non supervisée, comme le choix du critère d'arrêt d'apprentissage [Mart91]. 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 (cf. **section D.6.1**).

Cette approche a fait l'objet de plusieurs publications internationales [IC06b][IC06c]. Elle a également été expérimentée avec succès pour la prédiction de gisements miniers. Ce dernier travail a fait l'objet de l'encadrement partiel d'une thèse [PH06c].

Le principe général des mesures de **C-Rappel** et de **C-Précision** probabilistes est repris dans la référence [IC08b], versée à l'annexe.

Un second volet du travail d'optimisation des résultats des classificateurs concerne les traitements "amont" qu'il est possible de mener sur les données documentaires. Ces traitements ont pour double but de faciliter l'analyse, ou même simplement de la rendre possible, en réduisant la dimension de l'espace de représentation des données, et d'optimiser les résultats de classification, en éliminant les informations marginales [Kask98]. Ils consistent principalement à extraire les directions intrinsèques de représentation des données. Les méthodes que nous proposons dans ce cadre combinent de manière originale des approches issues de l'analyse de données, telle que la décomposition en valeurs singulières des matrices (termes **x** documents), avec des approches issues de la théorie de l'information, telle que l'estimation de l'entropie ou du pouvoir d'information des termes. Nous avons pu mesurer leur efficacité en nous basant sur les critères de qualité que nous avons décrits précédemment. Ces critères nous ont également permis de mettre en évidence les principaux défauts des méthodes concurrentes, comme l'indexation par sémantique latente [Derw90], ou les méthodes de projection non linéaire, telle que l'analyse en

composantes curvilignes [Dema97]. De premiers résultats prometteurs de ce travail ont été obtenus sur des analyses webométriques de types documents-liens, réputées difficiles à mener sur des données brutes. Ils ont fait l'objet d'une publication internationale [IC04a].

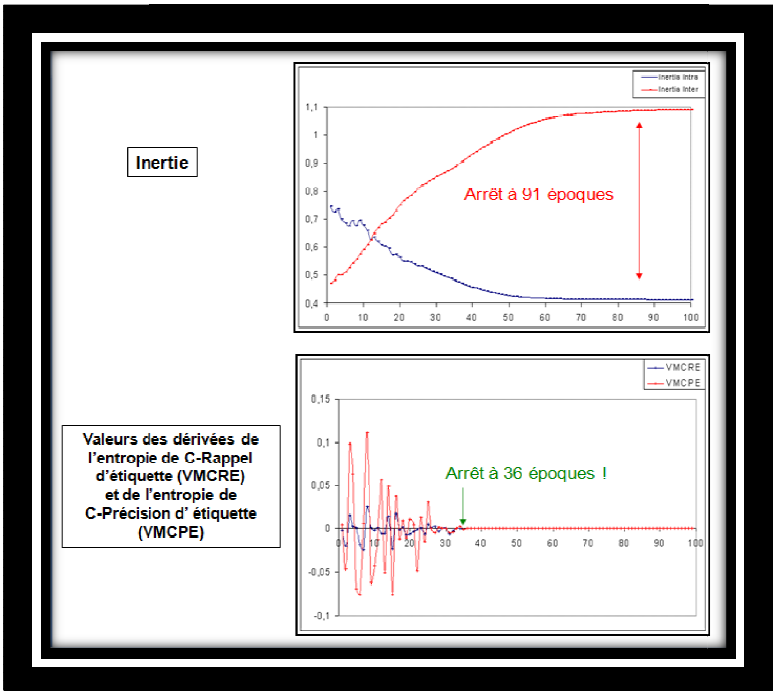

Figure 10 *: L'analyse de la migration des étiquettes entre les classes lors de l'apprentissage, qui peut s'opérer à partir des valeurs des dérivées d'Entropie de C-Rappel et d'Entropie de C-Précision, permet de définir un critère d'arrêt optimal pour l'apprentissage (ici cet arrêt, matérialisé par l'absence de migration d'étiquettes, est obtenu 3 fois plus tôt qu'en se basant sur les valeurs usuelles d'inertie).*

D.2. Généralisation du paradigme d'analyse de données multi-vues MVDA

Le paradigme d'analyse de données multi-vues MVDA représente un des thèmes centraux et originaux de notre de notre travail. Une partie importante de notre recherche concerne donc la validation théorique et expérimentale de ce paradigme. Elle concerne également l'application de ce paradigme à des approches de classification non supervisées alternatives à celles des cartes topographiques SOM de Kohonen.

Nous avons montré que l'on pouvait assimiler le mécanisme-noyau de communication inter-modèles de l'approche NOMAD-MULTISOM à un mécanisme d'inférence bayésien, ceci en définissant de nouvelles fonctions et un nouveau protocole de propagation [IC04b]. Cette démonstration représente une étape importante pour la validation théorique du paradigme MVDA. Elle a également permis d'améliorer la cohérence des résultats du mécanisme de communication inter-modèles propre à l'analyse multi-vues, et d'envisager ainsi une utilisation originale de l'approche NOMAD-MULTISOM pour la comparaison dynamique de modèles de classification, opération totalement inédite jusqu'alors, mais qui présente de très nombreuses applications. Nous avons notamment montré son utilité dans le cadre de la webométrie en proposant une méthode d'analyse du comportement de référencement des sites web des laboratoires européens qui s'appuyait sur ce principe.

Cette dernière expérimentation a fait l'objet d'une contribution dans une conférence internationale [IJ06a], elle-même sélectionnée, par la suite, pour être publiée dans un journal international [IJ06a].

Nous avons enfin continué à nous intéresser, d'un point de vue bibliographique, aux méthodes de classifications hiérarchiques, ainsi qu'aux méthodes d'apprentissage neuronales non supervisées à topologie libre, telles que les gaz de neurones. Les critères de qualités que nous avons présentés à la **section D.1** nous ont permis de confirmer expérimentalement l'avantage, pour le traitement des données de nature homogène, des gaz de neurones sur les modèles topographiques de type SOM, ainsi que de définir des conditions optimales d'utilisation des fonctions d'apprentissage de ces modèles [IC04c][IC06h][IC06j]. D'un point de vue pratique, ce travail nous a permis de proposer une extension du modèle MULTISOM à un modèle MULTI-X, plus générique car applicable aux résultats de tout type de méthode de classification, en implantant notamment des fonctions de généralisation hiérarchique génériques basées sur une méthode originale de regroupement par triangulation, de type 2-plus-proches-voisins.

Finalement, l'exploitation de notre nouveau protocole de communication inter-modèles et inter-vues nous a permis de montrer que cette méthode fournissait un apprentissage plus homogène des généralisations, tout en étant moins coûteuse en temps de calcul, qu'une méthode d'apprentissage direct appliquée plusieurs fois sur les mêmes données, en utilisant un nombre de classes décroissant [IC05c].

D.3. Coopération numérique-symbolique

Les limitations des méthodes de classification numériques, telles que NOMAD-MULTISOM, sont liées aux erreurs d'interprétation qu'elles peuvent générer si elles sont utilisées sans précaution préalable par des non-spécialistes pour des tâches d'analyse fine d'un domaine. De leur coté, les méthodes de classification symboliques qui peuvent être utilisées dans le même but, telles que les treillis de Galois, présentent l'inconvénient de fournir des résultats souvent inexploitables, à la fois trop touffus et trop détaillés, en particulier sur des données complexes [Simo99]. Nous avons montré qu'il était possible d'établir une synergie entre ces deux types de méthodes en définissant des équivalences entre les treillis de Galois et les cartes topographiques SOM. Ces équivalences permettent de générer des interprétations strictes et hiérarchisées sur les cartes SOM, et permettent également, à l'inverse, de définir des points focaux d'entrée dans les treillis de Galois.

Ce travail à fait l'objet de 4 publications internationales [IC00g][IC01d][IC02a][IC03b] et de l'encadrement d'un stage de DEA [MA00a].

Les méthodes de classification symboliques, telles que les treillis de Galois, permettent de générer des règles d'association concernant les propriétés des documents qui peuvent être interprétées comme de la connaissance extraite de ces documents [Agra96]. Au vu du nombre important de règles généralement obtenues et des faibles capacités de synthèses des méthodes de classification symboliques, le problème de la sélection des règles pertinentes reste un problème ouvert [Hamm02]. Il est notamment difficile de différencier les règles pertinentes des règles triviales. L'approche que nous proposons prolonge à la fois les travaux que nous avons menés sur les méthodes d'optimisation des classifications numériques et sur les extensions du modèle de

classificateur multi-vues. Elle consiste à utiliser le mécanisme d'inférence bayésien entre les vues en sélectionnant les classes origines et les classes destination à partir des critères de qualité de Rappel et de Précision non supervisés. Cette approche peut s'apparenter à un mécanisme de sélection non supervisé de règles, fondé sur une représentation synthétique préalable du fonds documentaire étudié. Elle s'avère être compatible avec le mécanisme de généralisation en ligne associé au paradigme MVDA.

Ce travail à déjà donné lieu à plusieurs publications internationales [IC05d][IC05e][IC06k]. Il nous a notamment permis de monter qu'il était possible d'extraire autant de règles avec un modèle numérique qu'avec un modèle symbolique, mais en bénéficiant cependant d'un coût de calcul fortement réduit. La **Figure 11** donne un aperçu du comportement de la méthode en fonction du nombre de classes utilisées. La référence [BC10a], donnée en annexe, donne un aperçu de l'algorithme.

Une expérimentation webométrique d'analyse multi-vues des domaines de recherche et du référencement des universités allemandes dans le cadre européen, basée sur les données Web du projet IST EICSTES, nous a permis de montrer que notre méthode permettait l'extraction de règles pertinentes dans le contexte de grands ensembles de données très fortement multidimensionnelles, contexte dans lequel les méthodes symboliques s'avèrent totalement inopérantes. Ce travail a fait l'objet d'une communication dans un journal international [IJ06a].

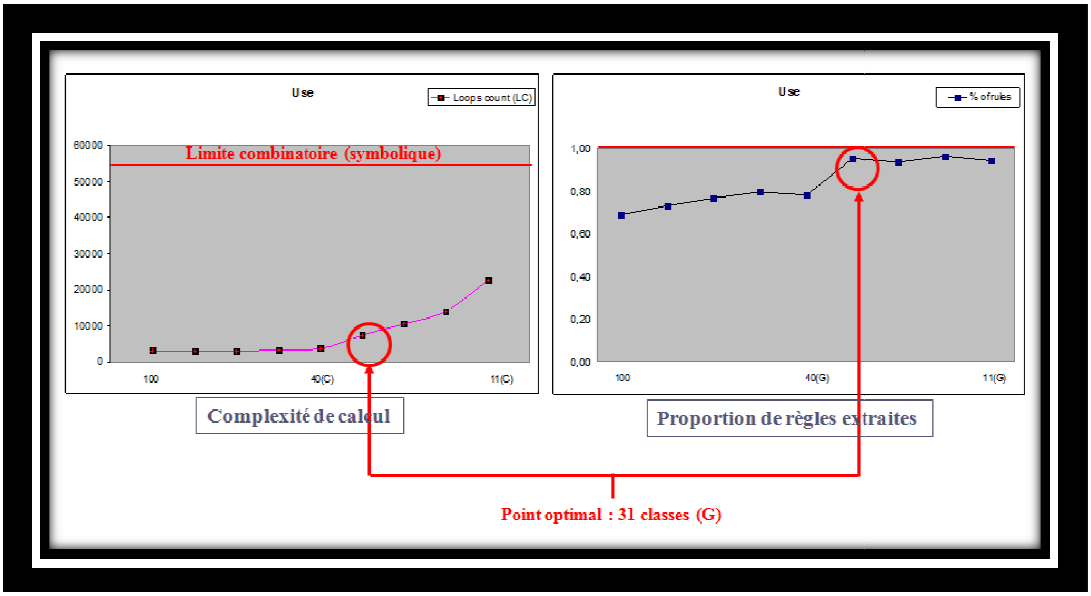

***Figure 11** : La stratégie numérique de recherche de règles d'association consiste à déterminer un modèle de classification optimal, puis à le généraliser pour tenter d'augmenter le nombre de règles extraites. L'extraction des règles proprement dites repose sur l'exploitation des mesures de Rappel et de Précision (cf. **section D.1**) localement à chacune des classes pour identifier des règles potentielles. La figure de droite illustre le fait qu'un nombre important de règles peuvent être extraites dès la prise en compte du modèle optimal, c.à.d. sans généraliser, puis que le nombre de règles extraites augmente progressivement de manière corrélée avec la complexité de calcul, quand le nombre de classes est réduit par généralisation de ce modèle. La complexité de calcul est présentée sur la figure de droite : la complexité symbolique, matérialisée ici par la notion de **Limite combinatoire**, correspond au nombre d'étapes de calcul nécessaires pour extraire toutes les règles en utilisant une classe unique contenant toutes des données.*

D.4. Coopération de classificateurs

La coopération de classificateurs a pour but de simuler des systèmes de recherche d'information qui permettent de répondre à différents niveaux de besoin au cours du processus de recherche d'information [Hali90]. Dans ce cadre, nous avons proposé un modèle spécifique de système de recherche d'information, nommé SARCI. Le modèle considéré est un modèle multi-agents, qui introduit de manière inédite la coopération entre différents types de classificateurs, de nature symbolique ou numérique. Il fonctionne en deux étapes, selon une stratégie proche du raisonnement à base de cas. La première étape consiste à chercher des situations similaires au besoin d'information de l'utilisateur. Pour ce faire, une phase de classification des requêtes antérieures est proposée, de manière à former des **profils requêtes**. Cette étape qui constitue la phase d'**analyse de surface** utilise un formalisme de classification symbolique de type treillis de Galois. Elle permet de répondre rapidement à des demandes de nature stéréotypique en recherchant le résultat de requêtes analogues déjà classifiées. En cas d'échec, la seconde étape de recherche est appliquée. Elle consiste à faire une **analyse en profondeur** en considérant les documents de la collection. Cette étape introduit une coopération entre plusieurs méthodes de classification (neuronale, SVM, génétique et symbolique) pour déterminer des **profils documents** et des **profils utilisateurs** en tenant compte du feedback de l'utilisateur et de son type de besoin. Finalement, l'ensemble des profils du modèle peut être géré selon des vues multiples.

Le modèle SARCI a fait l'objet d'une série d'expérimentations sur la collection-test CRAN qui ont permis de démontrer la pertinence de la démarche proposée.

Ce travail, a donne lieu à l'encadrement d'une thèse [PH06a]. Il a également mené à 1 publication de niveau national [NC00b] et à 6 publications de niveau international [IC00d][IC00e][IC01c] [IC02c][IC05f](IC05g].

D.5. Méthodes de visualisation des résultats de classification multidimensionnels

La plupart des méthodes de classification non supervisées, à l'image des gaz de neurones, fournissent explicitement leurs résultats dans l'espace de description des données. L'exploitation de ces résultats dépend donc directement de leur visualisation sous une forme synthétique. Ce problème s'avère complexe lorsque l'espace de description originel des données est fortement multidimensionnel, comme dans le cas du traitement de données textuelles. Après avoir testé les avantages et les inconvénients des différentes méthodes classiques de visualisation basées sur des projections linéaires ou non linéaires [Dema97], nous nous sommes orientés vers une approche de visualisation hyperbolique basée sur le cercle de Poincaré. Cette approche, habituellement utilisée pour la gestion de fichiers [Lamp96], présente l'avantage déterminant de maintenir une vision globale sur l'organisation de l'information, et sur sa structure, sans produire d'effet de surreprésentation. Elle exploite de plus un mécanisme type focus-contexte, ce qui la rend, contrairement aux approches classiques, adaptable à des espaces originaux fortement multidimensionnels présentant localement de fortes densités d'information. Dans le cas de son utilisation sur un nuage de classes, ou même plus directement sur un nuage de données, elle nécessite cependant d'être couplée avec une méthode de classification hiérarchique.

Après avoir testé les méthodes hiérarchiques classiques, telles que la classification hiérarchique ascendante standard et la méthode des voisins réciproques, nous avons proposé notre propre algorithme de classification hiérarchique qui préserve structurellement la densité des données. Le

principe de cet algorithme, que nous avons nommé DBHC (**D**ensity **B**ased **H**ierarchical **C**lustering) est celui de construire une hiérarchie de groupes de classes en utilisant des hypershères de rayon croissant, et en partant d'un rayon initial égal à la distance minimale entre les données à regrouper. Dans la version actuelle de l'algorithme, le nombre de niveaux de regroupement est fixé à l'avance. A la **Figure 12**, le comportement de cette méthode est comparé à celui des méthodes classiques, vis-à-vis de la construction d'hyper-arbres de classes. Un exemple plus général de résultat qu'elle fournit est donné à la **Figure 13**.

Ce travail à fait l'objet de 2 publications internationales [IC06e][IC08d]. Il est également présenté et justifié de manière plus exhaustive dans [PH06b].

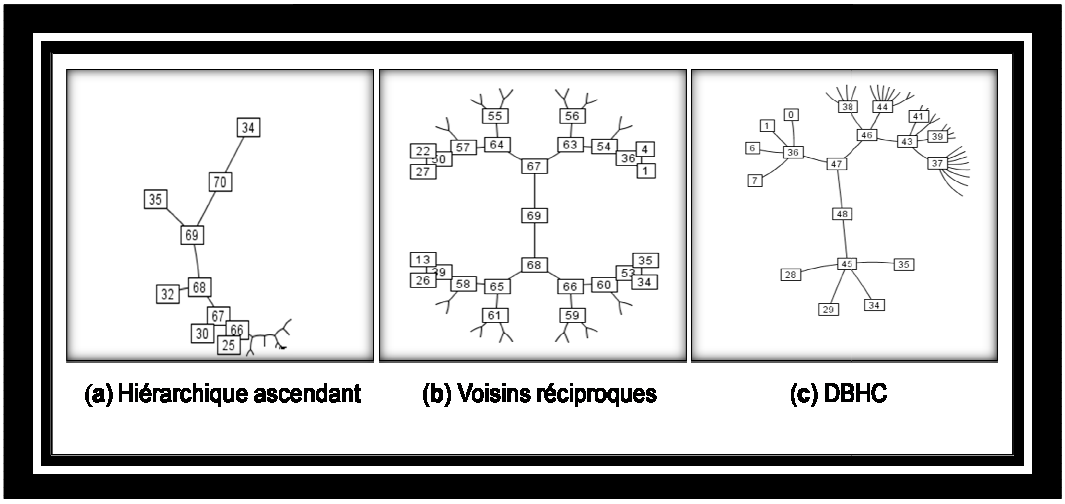

***Figure 12** : Comparaison entre les algorithmes de classification hiérarchiques pour la construction d'arbres hyperboliques à partir d'un même nuage de classes. Si les données-test, ici des classes, sont distribuées de manière non uniforme, seule la solution (c) fournit un arbre dont la structure préserve l'information sur les variations de densité dans le nuage original.*

Les numéros représentent les étiquettes des classes apparentes dans le contexte de visualisation courant.

La combinaison de telles approches globales avec des approches locales à base de modèles d'équilibre de forces, tels que les modèles de Spring [Kops98], a été également considérée dans notre travail. Cette dernière approche est notamment utile pour générer des graphes locaux permettant eux-mêmes de matérialiser des liens existants entre les classes associées aux nœuds d'un hyper-arbre de classes. La combinaison arbre-graphe ainsi obtenue peut se substituer aux méthodes à base de graphe employées dans ce contexte, telles que BibtechMon [Heim06], étant donné qu'elle fournit les mêmes informations tout en s'affranchissant du problème de surcharge cognitive desdites méthodes. Ce dernier travail a fait l'objet d'une publication internationale [IC08d].

Finalement, nous nous intéressons à l'adaptation de méthodes de classification ascendantes, de type dynamiques, telle que la méthode CHAMELEON [Kary99], pour la construction d'hyper-arbres de classes. L'exploitation de ce type de méthodes, qui tiennent également compte de l'homogénéité et de la cohérence des classes résultantes lors des étapes de fusion, permettrait de préserver la cohérence explicative lors de la navigation. Ce dernier travail est encore en cours de finalisation.

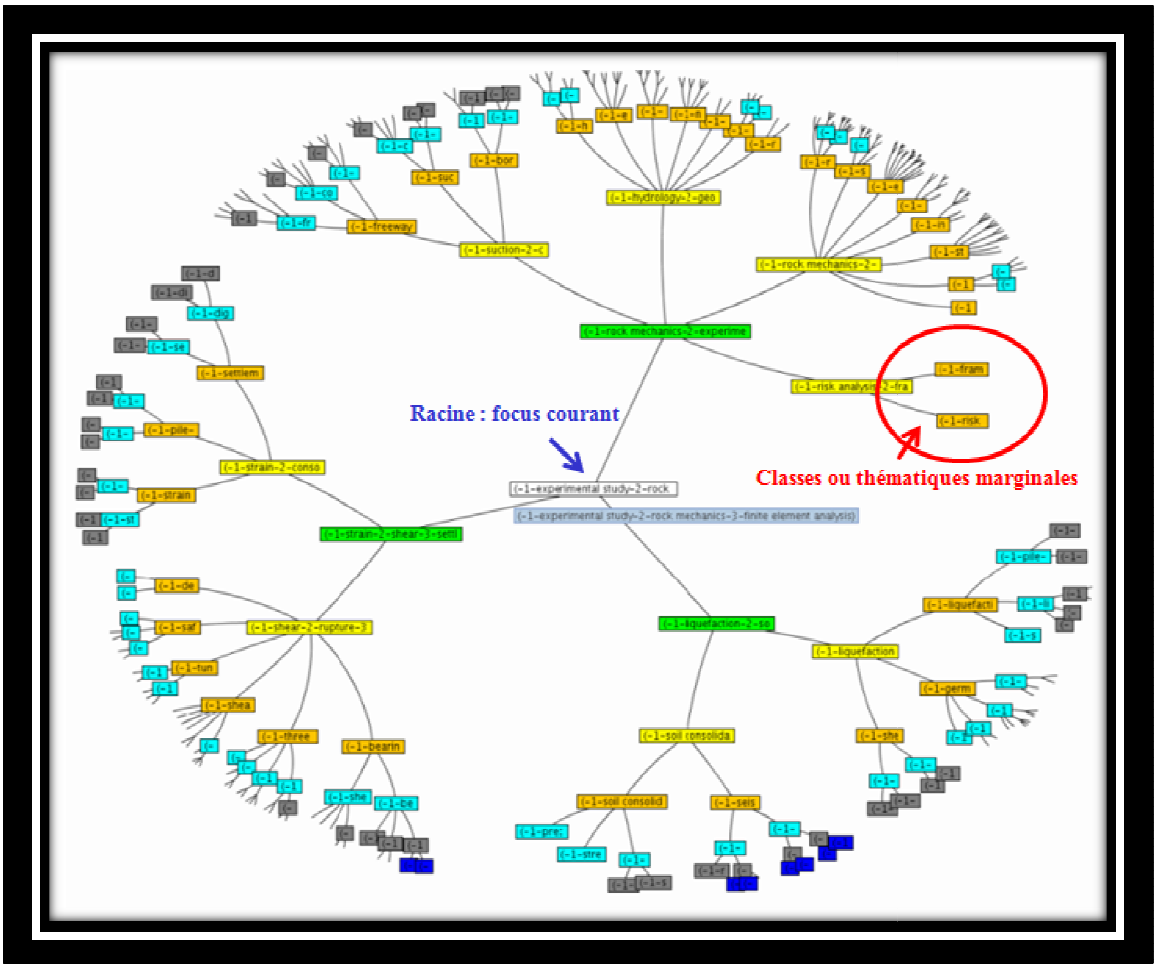

Figure 13 *: Représentation d'un nuage de classes multidimensionnel sous forme d'hyper-arbre après classification hiérarchique du nuage à l'aide d'un algorithme fondé sur la densité (ici, notre algorithme DBHC). Les classes initiales représentent les feuilles de l'hyper-arbre. Sa racine matérialise le nuage global de classes. Un nœud (c.-à-d. une classe) de l'hyper-arbre qui a beaucoup de liens (c.-à-d. beaucoup de fils), relativement aux autres nœuds du même niveau, représente une zone de forte densité du nuage original (zone plus riche de sujets). Chaque classe-feuille qui a la racine de l'hyper-arbre comme proche parent (ici, les classes inscrites dans le cercle rouge) peut être considérée comme supportant un sujet marginal. Le focus courant peut être changé à tout moment, ce qui revient à déplacer le point d'observation dans l'espace hyperbolique pour « découvrir » des branches cachées de l'hyper-arbre.*

D.6. Validation fine des résultats des classificateurs numériques

D.6.1. Etiquetage des classes issues d'une classification numérique

La mise au point de techniques d'étiquetage intelligent est nécessaire pour obtenir une vue d'ensemble des sujets principaux extraits par les méthodes de classification non supervisée, en particulier si celles-ci produisent originellement leurs résultats dans un espace fortement multidimensionnel. Ces techniques doivent permettre de fournir à la fois des informations synthétiques sur la nature des sujets couverts par chaque classe, mais également de discriminer de manière optimale les informations fournies par les différentes classes. Dans un cadre de synthèses plus générales, elles doivent également pouvoir s'appliquer à l'étiquetage de regroupements hiérarchiques de classes. Jusqu'ici, le développement de telles techniques demeurait un problème

ouvert. Nous avons récemment proposé des méthodes d'étiquetage originales basées sur l'utilisation des critères d'évaluation du contenu des classes, comme le **C-Rappel**, la **C-Précision** et la **C-F-mesure** probabiliste sur des propriétés des données associées aux classes (cf. **section D.1**). Nous avons montré que ces méthodes, qui se rapprochent des méthodes de recherche de maximum de vraisemblance (EM) [Demp77], surpassaient, en terme de pertinence, les techniques d'étiquetage élémentaires, comme celles basées sur la fréquence des propriétés des données associées aux classes [Chua04], ou la dominance des composantes de leur profil [Cutt93], aussi bien que les méthodes de discrimination statistiques usuelles, telle que la méthode du chi2 [Pope00]. Dans ce même cadre, nous avons également démontré l'efficacité de ces techniques pour l'étiquetage de structures complexes, telles que les hyper-arbres de classes. La **Figure 15** illustre les résultats théoriques obtenus avec l'ensemble des méthodes précitées dans le cas de l'étiquetage d'un arbre hyperbolique de classes de plusieurs centaines de nœuds. La **Figure 14** en illustre les résultats pratiques.

Ce travail a fait l'objet d'une communication internationale [IC08b], d'une soumission à un journal international [IJ09a] et de l'encadrement d'un stage de fin d'études d'ingénieur [Ma07a]. La référence [IC08b] est versée à l'annexe.

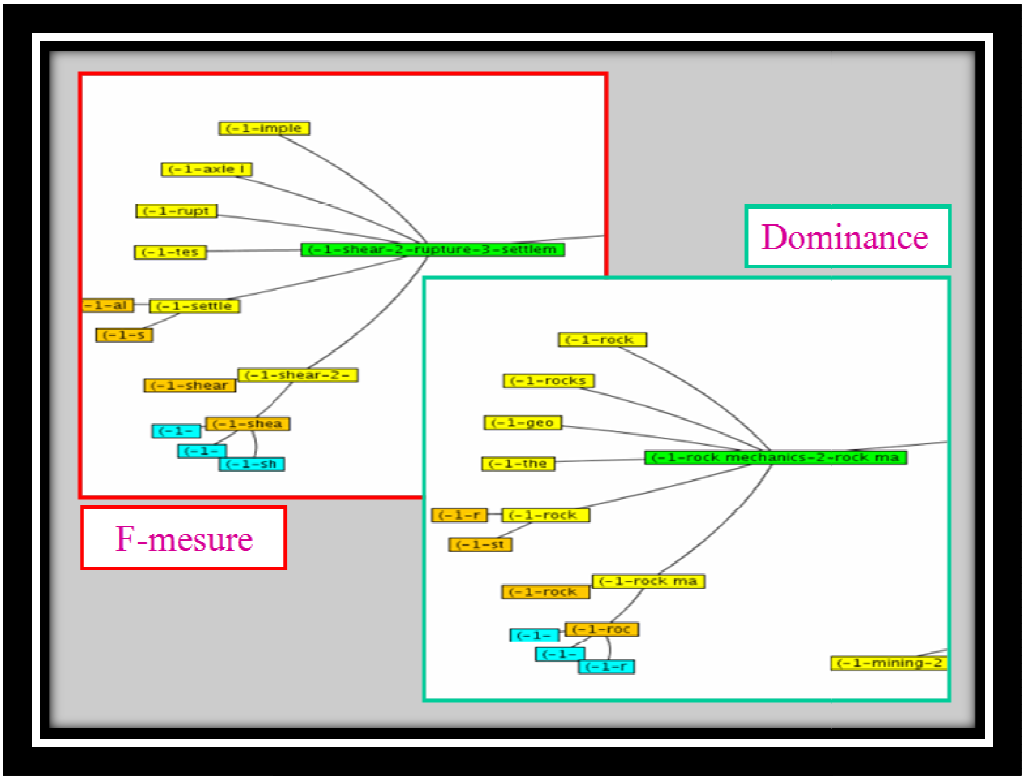

***Figure 14** : Dans la pratique, une mauvaise méthode d'étiquetage ne permet pas de discriminer entre les contenus de classes projetées sur un espace de représentation. Ces deux vues partielles du même hyper-arbre de classes étiquetées montrent la supériorité manifeste de la méthode d'étiquetage basée sur la **C-F-mesure** qui produit généralement des listes d'étiquettes de classes, à la fois différentes entre les classes et représentatives du contenu spécifique de chaque classe, alors que les étiquettes produites par la méthode **Dominance**, basées sur les profils de classes, sont la plupart du temps identiques et générales (sur l'exemple ci-dessus, c'est le cas de l'étiquette **rock**, très générale, récurrente dans l'étiquetage produit par cette dernière méthode).*

	LSP	ALP	SSP	LHP	UR	WR	δ
Dominance	1216	0.03	568	525	4	10.5	1.73
Fréquence	245	0.24	166	592	3.25	8	0.82
C-F-Mesure	155	0.26	112	760	2.25	4.75	0.82
X^2	121	0.21	89	1485	2.75	7.5	1.78

***Figure 15** : Résultats d'une expérience d'étiquetage sur un arbre hyperbolique de plusieurs centaines de classes. Les mesures présentées donnent une indication sur le pouvoir discriminant des étiquettes, mais également sur leur capacité à décrire de manière exhaustive le contenu des classes. (**LSP** : Pénalité de similarité des feuilles, **PLF** : Précision des étiquettes aux feuilles, **PSF** : Pénalité de similarité des fils dans l'hyper-arbre, **LHP** : Pénalité d'hétérogénéité de l'étiquetage, UR: Classement moyen de la méthode considérant les différents critères, **WR** : Classement moyen pondéré, δ : Ecart type de classement (Les valeurs les plus faibles sont les meilleures, sauf pour la Précision, où c'est l'inverse).*

*La méthode basée sur la **C-F-mesure** probabiliste donne bien les meilleurs résultats dans ce contexte. La méthode usuelle basée sur les profils de classes, ou **Dominance**, y fournit quant à elle les plus mauvais résultats. Des explications complémentaires sur le protocole expérimental associé à cette expérience peuvent être trouvées dans la référence [IC08b].*

D.6.2. Analyse locale des résultats de la classification non supervisée

Les méthodes de validation de la classification non supervisée sont généralement utilisées pour déterminer le nombre optimal de classes qui est possible d'associer à un ensemble de données. A une échelle plus fine, elles peuvent également permettre de distinguer les classes pertinentes des classes non pertinentes. En dépit de leur succès, l'efficacité des indices basés sur la distance ou sur les moindres carrés diminue très rapidement lorsque les données à traiter deviennent fortement multidimensionnelles. Pour compléter les méthodes de validation que nous avions déjà définies (cf. **section D.1**), nous avons donc mis en place de nouveaux indices basés sur l'identification de formes–cœur, qui représentent elles-mêmes des groupes de propriétés fortement corrélées dans les données rattachées à chacune des classes identifiées par une méthode de classification non supervisée. Cette technique se ramène à la caractérisation de sous-espaces de description « locaux » dans un espace de description global de grande dimension. Des valeurs d'inertie basées sur le pourcentage de formes-cœurs dans chaque classe (intra, ou CC) et sur le recouvrement de formes-cœurs entre les classes (inter, ou IC) permettent à la fois d'estimer précisément la qualité locale de chaque classe et d'évaluer la qualité globale d'un modèle de classification. La mesure finale, nommée OqC représente la F-Beta combinaison pondérée des mesures d'inertie. Des expériences étendues menées sur la partition temporelle de la collection-test Reuters-21578[1] nous ont permis de prouver l'efficacité de cette mesure sur des données complexes, données sur lesquelles les méthodes classiques, comme la méthode de Calinsky-Harabasz [Cali74] ou la méthode de Davies-Bouldin [Davi79], s'avèrent totalement inopérantes. Les résultats obtenus sont résumés sur la **Figure 16** et sur la **Figure 17**.

La mesure OqC est décrite plus précisément dans la publication internationale [IC08a]. Cette dernière référence est ajoutée en annexe.

[1] Une description plus précise de cette collection est donnée à la **section D.2**.

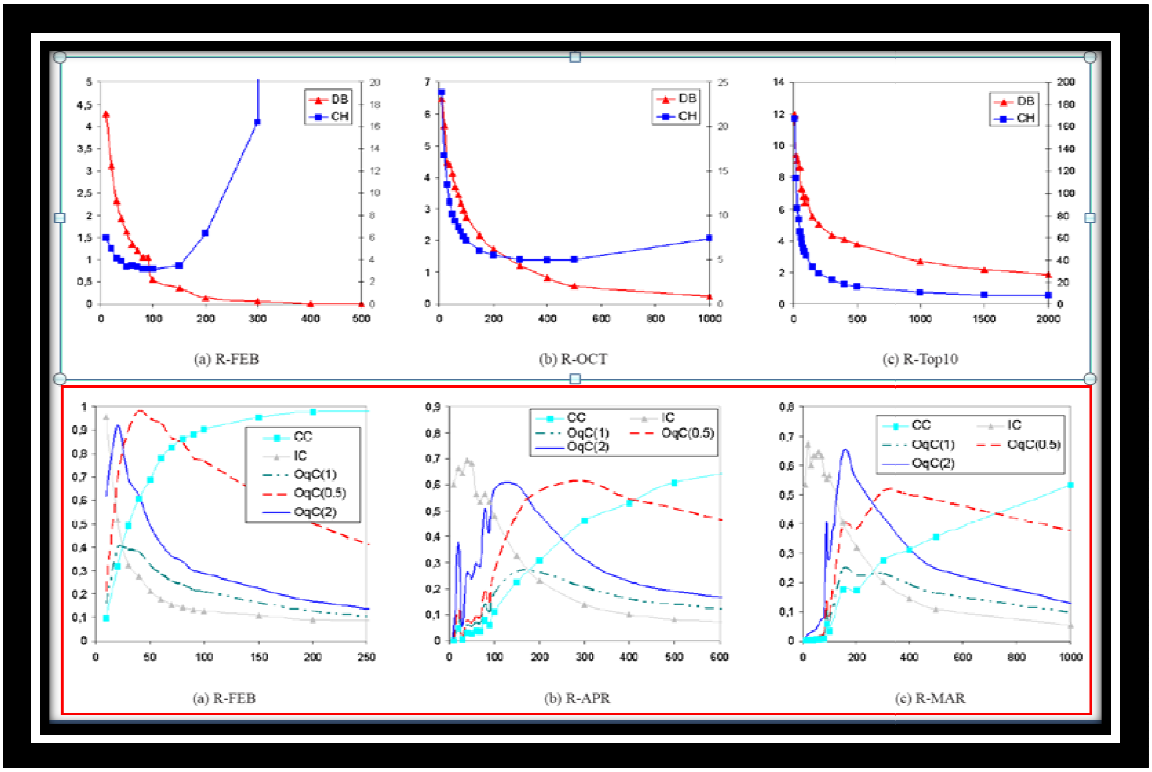

Figure 16 *: Dans le cas de la classification supervisée de données complexes, comme les données REUTERS, les indices classiques ne fournissent aucune information sur un nombre de classe optimal. Dans le cas présent l'indice de Calinsky-Harabasz n'est jamais minimisé et l'indice de Davis-Bouldin jamais maximisé (courbes du haut). Seule, la mesure OqC (courbes du bas) fournit des résultats stables en donnant des optima dans toutes les expériences (et ce, pour toutes les valeurs du paramètre de mélange inter-intra Beta). Le corpus REUTERS utilisé dans cette expérimentation est divisé en périodes de temps.*

Data Set	#Classes	OqC(2)		OqC(1)		OqC(0.5)		WPurity	
		#Clusters	WPurity	#Clusters	WPurity	#Clusters	WPurity	#Clusters	Value
R-FEB	34	20	0.6757	20	0.6757	40	0.7759	40	0.7759
R-MAR	81	150	0.6987	150	0.6987	300	0.7204	500 [200,500]	0.7210
R-APR	76	150	0.6834	150	0.6834	300	0.6687	150	0.6834
R-JUN	71	60	0.7396	90	0.7435	150	0.7595	300 [100,300]	0.7599
R-OCT	48	80	0.8325	100	0.8489	150	0.8474	100 [90-150]	0.8489
R-Top10	10	150	0.7997	300	0.7966	400	0.7910	40 [30,500]	0.8170

Figure 17 *: Dans le cas du corpus REUTERS, l'analyse de la pureté des classes obtenues de manière non supervisée peut être menée à partir les étiquettes de catégorie REUTERS associées aux données originales, et peut ainsi servir de base pour déterminer une configuration optimale en terme de nombre de classes. Les résultats présentés dans ce tableau prouvent que la mesure OqC fournit systématiquement un nombre de classes correspondant à une valeur se trouvant dans l'intervalle de pureté optimale (Cette valeur peut cependant être associée à des valeurs différentes du paramètre mélange inter-intra Beta).*

D.6.3. Détection de résultats incohérents de la classification non supervisée (2010)

Nous étudions également le comportement des méthodes d'analyse de la qualité de la classification non supervisée pour discriminer les résultats de classification cohérents des résultats incohérents. Dans le cas du traitement de données très complexes, comme les données textuelles fortement polythématiques, nos expérimentations récentes ont montré qu'aucun des indices de qualité existants, y compris ceux que nous avions définis jusqu'ici, ne permettait de discriminer des résultats de classification cohérents de résultats incohérents. Leur défaut principal est de n'être pas assez sensibles à la présence d'un petit nombre de classes incohérentes de grande taille, dans le cas de l'existence conjointe d'un grand nombre de classes de petite taille, généralement cohérentes, dans les résultats de classification [IC10a]. Nous avons même également montré que tous les indices étudiés pouvaient fournir des réponses contraires à la réalité en identifiant les meilleurs résultats de classification (i.e. les meilleures partitions) comme les plus mauvais, et inversement [IC10i].

Nous avons illustré ce phénomène en comparant le comportement des méthodes de classification non supervisées sur des données homogènes et sur des données hétérogènes, ou polythématiques, et en faisant appel, pour caractériser les résultats effectifs desdites méthodes, aux méthodes d'étiquetage que nous avons précédemment développées (cf. **section D.6.1**). Les données homogènes que nous avons utilisées sont des notices de brevets relatives au développement des huiles moteurs que nous avions exploitées dans plusieurs de nos expériences précédentes [IC03c][IC05e]. Cependant, seul le champ élémentaire des notices couvrant le domaine d'utilisation des brevets avec un vocabulaire normalisé, limité et contextuel, a été retenu dans ces nouvelles expériences. Nous avons parallèlement construit un corpus-test de données hétérogènes, ou polythématiques, en extrayant de la base PASCAL de l'INIST un ensemble de notices bibliographiques englobant une année complète de recherche, tous domaines confondus, et impliquant au moins un laboratoire de recherche lorrain. Seuls les mots d'index documentalistes, qui couvraient malgré tout un grand ensemble de sujets différents (aussi éloignés les uns des autres que la médecine, la physique structurelle ou la sylviculture,…), avec un fort taux de polysémie (système, âge, structure, …), ont été considérés dans ces expériences.

Un exemple des indications de qualité des résultats de classification qu'il est possible d'obtenir avec les indices d'inertie, ou avec des Macro indices de Précision et de Rappel, comme ceux que nous avions proposés précédemment (cf. **section D.1**), suite à l'analyse du corpus hétérogène par les méthodes K-Means et SOM, est donné à la **figure 18**. L'analyse des étiquettes des classes obtenues et celle de la répartition des données dans ces dernières (**figure 19** et **figure 20**), montre cependant que les indications données par ces indices sont contraires à la réalité.

Pour corriger cela, nous avons récemment proposé d'adapter notre approche basée sur les indices de Rappel/Précision non supervisés en définissant de nouveaux indices de Micro-Précision et de Micro-Rappel [IC10d][IC10g], ainsi que des indices de Micro-F-mesure cumulée [IC10c], moyennés directement sur les propriétés des données associées aux classes, sans donc finalement considérer pour cela le niveau de structuration intermédiaire des propriétés défini par la partition en classes. Nos dernières expériences menées sur différents corpus-test de données hétérogènes nous ont permis de montrer que ces nouvelles mesures pouvaient effectivement être exploitées pour identifier de manière sûre les résultats incohérents de classification non supervisée [IC10e][IC10i][IC10j]. La **figure 21** présente les résultats de qualité obtenus par les nouvelles mesures sur le même corpus et avec les mêmes méthodes de classification que celles présentées à la **figure 18**.

Le principe de ces nouvelles mesures est plus précisément décrit dans la publication [IC10c], versée à l'annexe. Cette approche, directement dérivée de nos travaux antécédents sur les indices de qualité basés sur la Précision et le Rappel non supervisés, ouvre de nouvelles voies pour l'évaluation précise du comportement des méthodes de classification non supervisées. Comme nous le montrons par la suite (**section D.6.5**), elle nous a également permis de mieux comprendre le comportement des méthodes de classification non supervisée usuelles sur des données complexes, comme les données fortement polythématiques, ainsi que de proposer de nouvelles méthodes plus adaptées au traitement de ce type de données.

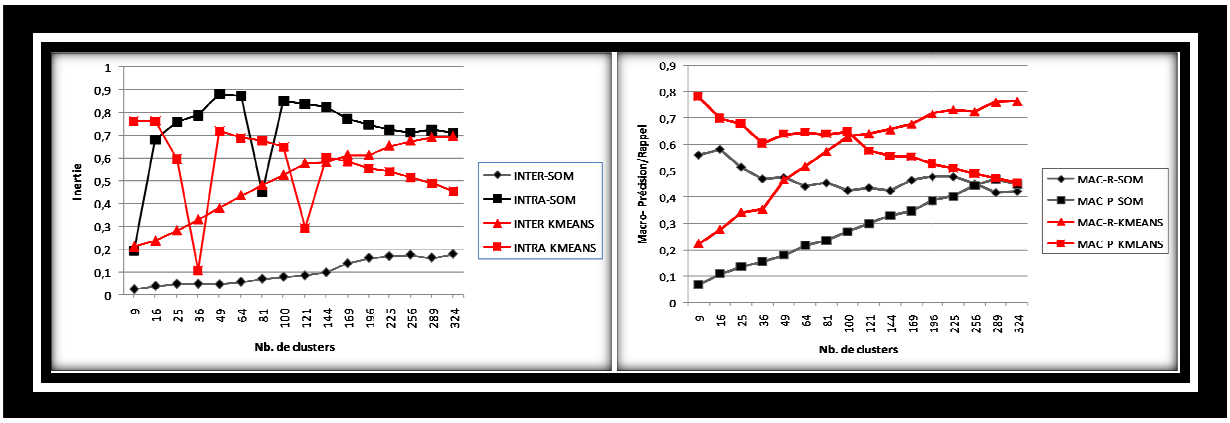

Figure 18 *: Evolution des valeurs des indices d'inertie (courbes de gauche) et de Macro-Précision et de Macro-Rappel (courbes de droite) en fonction du nombre de clusters non vides pour les méthodes SOM et K-Means pour des résultats de classification sur des données hétérogènes. Sur la figure 1B les valeurs d'inertie apparaissent meilleures pour la méthode K-means (plus faible inertie intra et plus forte inertie inter que pour la méthode SOM). Les courbes de Rappel et de Précision, bien que bénéficiant d'un comportement plus stable que les courbes d'inertie, indiquent également la méthode K-Means comme étant la meilleure (point de croisement entre les courbes apparaissant à une plus forte valeur avec un plus faible nombre de classes).*

[1189] 840-- Racine Densité Protéine Loi échelle Appareil respiratoire pathologie Condition aux limites Forêt Protéine lait Protection environnement Prévision Analyse donnée Pollution air Fibre In situ DNA Alcool Mécanique roche Sol Méthode analytique Cosmologie Traitement donnée Cartographie Végétation Rhizosphère Eau potable Application Méthode Monte Carlo Zone urbaine Grain Pédiatrie Cours eau Logiciel Partie aérienne végétal Inhibiteur enzyme Diagramme phase Photosynthèse Système nerveux pathologie Expression génique Variation saisonnière Nucléosynthèse Biodisponibilité Matière organique Ingénierie Vérification programme Cytométrie flux Lymphocyte T Diffusion(transport) Floculation Analyse composante principale Déprédateur Détection Produit contraste Arbre forestier feuillu Tolérance Industrie alimentaire Système à retard Tolérance faute Système temps réel Problème NP difficile Appareil respiratoire Polymère Synchronisation .....

Figure 19 *: Vue générale des étiquettes extraites de la partition générée par la méthode K-Means. Les mauvais résultats de la méthode sont confirmés par une observation globale de la répartition des étiquettes de classes. Ici une classe-poubelle de très grande taille (1189 données sur 1340 au total) "attire" un très grand nombre d'étiquettes (840) de différentes natures, figurant un contenu très hétérogène. Les étiquettes de classe sont extraites au moyen la méthode présentée à la* ***section D.6.1****.*

[28] 27-- Etude théorique Méthode dynamique moléculaire Simulation numérique Interface gaz liquide Continuum Modèle thermodynamique Fluide non newtonien Diffusion Nanoparticule Laser semiconducteur Propriété mécanique Analyse statistique Système binaire Compressibilité Point critique Propriété rhéologique Ecoulement conduite Perte charge Macromolécule Effet dimensionnel Mouillage

[21] 38-- Zone tempérée Foresterie Arbre forestier feuillu Arbre forestier Peuplement forestier Plante ligneuse Forêt décidue Montagne Sylviculture Futaie Arbre forestier résineux Formation végétale Gestion forestière Dendrométrie Peuplement forestier artificiel Filière bois Ecophysiologie Stade juvénile plante Entomologie Phytopathologie Caractéristique peuplement Industrie bois Qualité production Tolérance Economie forestière Dépérissement forêt Modèle simulation Peuplement forestier mélangé Symptomatologie Recommandation Bois feuillu Histoire Réglementation Bois résineux Contrainte Photosynthèse Climat Canopée

[16] 28-- Cardiopathie Insuffisance cardiaque Aigu Conduite à tenir Myocarde pathologie Pronostic Mortalité Chimiothérapie Epidémiologie Appareil circulatoire Chronique Evolution Cœur Stratégie Electrocardiographie Santé publique Electrodiagnostic Facteur prédictif Maladie Prévalence Soin Complication Coût Maladie héréditaire Tomodensitométrie Résultat Risque

Figure 20 *: Vue générale des étiquettes extraites de la partition générée par la méthode SOM. Les bons résultats de la méthode sont confirmés par une observation globale de la répartition de ces étiquettes. Il y a ici différentes classes de tailles assez semblables attirant des groupes d'étiquettes sémantiquement homogènes, qui représentent eux-mêmes les thèmes principaux traités par le corpus.*

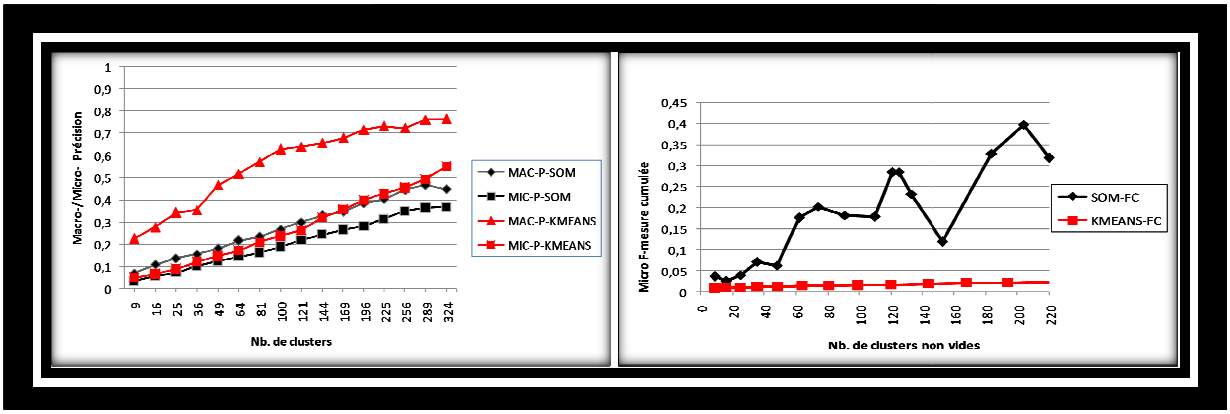

Figure 21 *: Les Micro-Rappel/Précision possèdent des caractéristiques générales analogues aux Macro-Rappel/Précision. Cependant, en les mixant avec ces derniers indices, ils permettent d'identifier des résultats de classification incohérents (courbes de gauche). En effet, dans ce dernier cas, les Précisions des classes de petite taille ne compenseront plus celles des classes de grande taille, et les propriétés imprécises présentes dans ces dernières, si elles s'avèrent hétérogènes, auront un effet considérable sur la Micro-Précision. Par conséquent, même si la Macro- et la Micro-Précision mesurent toutes deux le degré d'homogénéité des classes, l'écart entre ces deux mesures permet de confirmer la présence de clusters incohérents de taille importante (cas de K-means). La Micro F-mesure cumulée (courbes de droite) permet d'identifier directement, et à elle-seule, des résultats de classification incohérents, étant donné qu'elle pénalise plus fortement l'hétérogénéité des classes de taille plus importantes (cette mesure donne une valeur quasi-nulle dans le cas de K-means, quel que soit le nombre de classes considérées, matérialisant l'apparition systématique de classes-poubelles lors du traitement des données de l'expérience).*

D.6.4. Détection de structures implicites des données par la classification non supervisée

Les réseaux neuronaux à apprentissage compétitif préservant la topologie représentent de puissants outils pour la visualisation des interactions entre les données basée sur leurs relations implicites. Dans de tels modèles de réseaux, ce type d'approche exige cependant la prise en compte des interactions entre les neurones, et non plus uniquement la considération individuelle de ces derniers en tant que classes, ou groupes de données. Nous avons étudié comment les connexions entre des neurones qu'il est possible de produire dans ces réseaux, particulièrement celle obtenues par un apprentissage hebbien compétitif [Mart91], peuvent être exploitées pour matérialiser différents niveaux d'abstraction et de structuration implicites dans l'organisation des données d'analyse. Le principe de la méthode que nous proposons, nommée ACL-N, est celui de définir des pôles attracteurs en calculant des scores de centralité basés sur nombre et la stabilité des connexions originellement produites par l'apprentissage neuronal. Les pôles définissent eux-mêmes des groupes de classes, et seules les connexions entre les groupes sont conservées lors de la synthèse. Les expérimentations que nous avons menées sur l'analyse de distributions artificielles, telles que des mélanges de gaussiennes, ou d'autres distributions types, ont prouvé l'excellent pouvoir de synthèse et de discrimination de notre méthode, en comparaison avec les méthodes de référence, comme celles basées sur les voisinages de Voronoi [Rizz01]. La **figure 22** et la **figure 23** présentent les premiers résultats obtenus avec la méthode A-CLN. Cette méthode est décrite plus précisément dans la publication internationale [IC08a], ainsi que dans la thèse [PH09a].

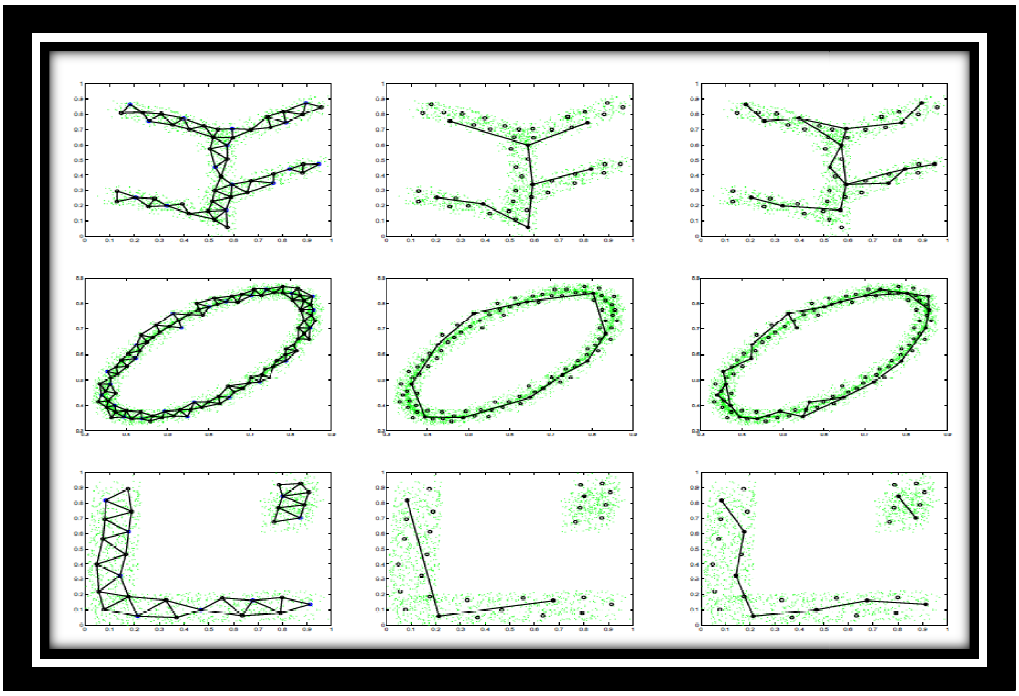

Figure 22 *: La méthode A-CLN permet également de synthétiser la structure de connexions entre des classes, comme celles produites par un apprentissage hebbien associé à une méthode de classification non supervisée. La structure latente des différentes distributions-type présentées dans cette expérience est clairement extraite par la méthode. La macro-abstraction (au centre) donne une vue générale de la structure. La micro-abstraction (à droite) permet d'obtenir un suivi plus précis de ses contours latents, mais au prix d'une perte de généralité.*

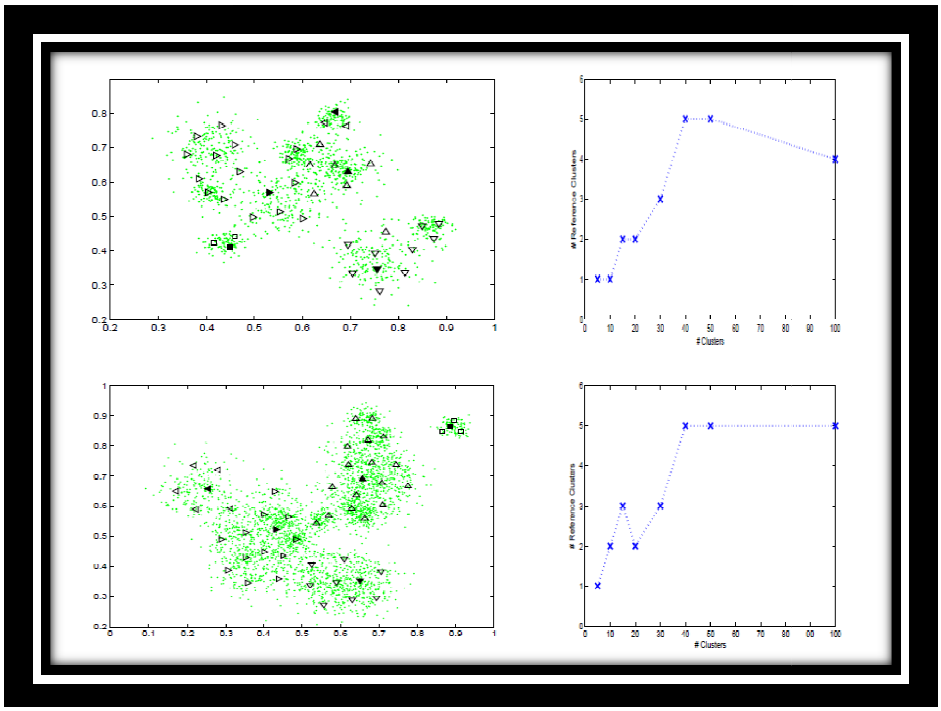

Figure 23 *: Si le nombre de classes initial est suffisant, La méthode A-CLN permet de regrouper les classes de même nature dans un même groupe cohérent. Les classes-pôles sont matérialisées par les formes pleines. Les formes de même type matérialisent des classes assimilées au même groupe par la méthode. Dans le cas présent, la méthode a clairement différencié les distributions souches d'un mélange de distributions gaussiennes.*

Les courbes de droite indiquent le nombre de groupes obtenus, en fonction du nombre de classes initales.

Les résultats d'expérimentations effectuées sur la partition temporelle du corpus Reuters-21578 nous ont permis de montrer la différence entre les deux modes d'abstraction de la méthode ACL-N, à savoir la macro- abstraction et de micro-abstraction, ainsi que de prouver son intérêt pour les tâches de généralisation (cf. **section C**). Ces résultats sont reportés à la **figure 24**.

	Niveau de base			Micro-abstraction			Macro-abstraction		
Corpus	#Cls	CC	IC	#Cls	CC	IC	#Cls	CC	IC
R-FEB	40	0.5950	0.2356	6	0.1841	0.6375	4	0.3214	0.4688
R-MAR	300	0.2917	0.2026	42	0.3163	0.6531	10	0.0053	0.7658
R-APR	300	0.4580	0.1324	49	0.3280	0.5671	11	0.1207	0.6114
R-JUN	150	0.4555	0.1776	25	0.1661	0.7442	8	0.0027	0.6250
R-OCT	150	0.5284	0.1712	24	0.1410	0.7923	9	0.0296	0.6667
R-Top10	400	0.2392	0.1181	55	0.1017	0.4045	10	0.0035	0.4254

Figure 18 *: Résultats de macro-abstraction et de micro-abstraction obtenus par la méthode ACL-N sur la partition temporelle du corpus Reuters-21578. La colonne #Cls représente le nombre de classes obtenues. Les indices CC (corrélation intra-classe) et IC (isolation inter-classes) sont ceux décrits à la section 0 et dans la référence [IC08b], versée à l'annexe. L'augmentation de la valeur de l'indice d'isolation IC lors des deux types de synthèse prouve la validité de la méthode ACL-N pour la généralisation.*

D.6.5. Analyse du fonctionnement des méthodes de classification non supervisées existantes sur les données hétérogènes et mise en place de nouvelles méthodes de classification incrémentale (2010)

Une analyse fine du comportement des méthodes de classification non supervisées sur les données hétérogènes a été rendue possible grâce au développement des nouveaux indices de qualité présentés à la **section D.6.3.** Le contexte expérimental des données hétérogènes, ou polythématiques, représente également un contexte-clé pour le test des méthodes de classification non supervisée incrémentales[2], étant donné qu'il matérialise la simulation statique des variations de sujets pouvant apparaître dynamiquement dans un corpus évolutif.

Nous avons donc poursuivi le double but de tester en parallèle les méthodes de classification non supervisées statiques et les méthodes dites incrémentales sur les données hétérogènes. Dans nos expérimentations, nous nous sommes plus particulièrement focalisés sur les méthodes de classification non supervisée neuronales. Ces méthodes partagent en effet le principe de prendre en considération des relations du voisinage entre les neurones (ou classes), qu'elles soient prédéfinies (topologie fixe), comme dans le cas des «cartes auto-organisatrices» (SOM) [Koho01], ou dynamique (topologie libre), comme dans celui des «gaz neuronaux» statiques (NG) [Mart91]. Cette stratégie les rend moins sensibles aux conditions initiales, ce qui représente un avantage déterminant dans le cadre de l'analyse des données textuelles, qui sont souvent représentées comme des données éparses associées à des espaces de description fortement multidimensionnels. Cet avantage reste de mise dans le cadre encore plus contraignant de l'analyse incrémentale de telles données.

Les cartes SOM ont été employées avec succès pour de nombreuses applications des domaines généraux de la recherche d'information [Kask98] [IC00a], et de l'analyse de données textuelles, comme pour la classification non supervisée de comptes-rendus de réunion [Orwi97] ou celle de données socio-économiques [Vars92]. Une version incrémentale de SOM intégrant des processus de croissance hiérarchique de la topologie neuronale permettant de suivre l'évolution des caractéristiques des données a été proposée par [Merk03]. Cependant, le défaut principal des méthodes de cette famille est celui de ne pas permettre de représenter très fidèlement les distributions de données complexes en raison de la structure topologique fixe exploitée. Dans l'algorithme «Neural Gas» ou NG [Mart91], les poids des neurones sont adaptés sans aucun arrangement topologique fixe dans le réseau. De fait, grâce à la perte de contraintes topographiques par rapport à SOM, NG tend à mieux représenter la structure des distributions des données, menant théoriquement à de meilleurs résultats de classification. L'algorithme «Growing Neural Gas» ou GNG [Frit95] résout le caractère statique de l'algorithme NG mettant en avant le concept du réseau évolutif. En effet, dans cette approche le nombre de neurones est adapté périodiquement pendant la phase d'apprentissage en fonction des caractéristiques de la distribution des données. GNG donne ainsi la possibilité de créer et de supprimer des neurones, ainsi que des connexions[3] entre ces derniers. L'algorithme «Incremental Growing Neural Gas» ou IGNG a été proposé par Prudent et Ennaji [Prud04]. Il représente une adaptation de l'algorithme

[2] Une présentation plus exhaustive du contexte d'application et des principes de la classification incrémentale est donnée à la **section E**.

[3] Ces connexions, dites « hebbiennes » [Hebb49] ou compétitives, sont crées au cours de l'apprentissage entre le neurone gagnant et son concurrent direct. Elles permettent de structurer dynamiquement le réseau créé, et peuvent être exploitées dans tous les algorithmes neuronaux à topologie libre.

GNG qui relaxe la contrainte d'évolution périodique du réseau. Par conséquent, dans cet algorithme, un nouveau neurone est créé chaque fois que la distance de la donnée d'entrée courante aux neurones existants est supérieure à un seuil qui correspond à la distance moyenne des données par rapport au centre de leur distribution. L'objectif principal de l'algorithme «Improved Incremental Growing Neural Gas» ou I^2GNG [Hamz08] est de résoudre la faiblesse de l'algorithme IGNG en termes de comportement incrémental. Dans ce but, l'algorithme I^2GNG exploite une valeur variable du seuil qui est calculée à chaque étape de l'apprentissage et qui dépend des données de chaque neurone.

A titre de comparaison, nous avons également inclus dans nos expérimentations une méthode non neuronale de référence classique, à savoir la méthode K-means [MacQ67], aussi bien qu'une méthode non neuronale de référence plus récente, à savoir la méthode Walktrap [Pons04]. En ce qui concerne les méthodes incrémentales non neuronales, nous avons pris en considération la très récente méthode Germen [Lelu06], basée sur la densité, et elle-même encore en cours de finalisation.

Nos expérimentations ont mis en évidence les difficultés énormes qu'on les méthodes de classification neuronales testées à traiter des données hétérogènes, celles-ci produisant par conséquent des résultats de qualité très médiocre, même pour leur nombre optimal de classes. Le principal phénomène observé est la création de classes-poubelles attirant la majeure partie des données parallèlement à celle de classes-miettes représentant les groupes marginaux ou encore non formés. Cette remarque s'est avérée également valable pour les méthodes non neuronales, qu'elles soient statiques, comme les méthodes K-means ou Walktrap, ou dites « incrémentales », comme la méthode Germen. Parmi l'ensemble des méthodes testées, seule la méthode SOM, pourtant non incrémentale, a eu un comportement à peu près conforme aux attentes dans notre contexte [IC10m]. La **figure 25** illustre le comportement comparatif de l'ensemble des méthodes testées pour un corpus homogène et pour un corpus hétérogène. Pour la description des corpus utilisés, l'on se reportera à la **section D.6.3.**

Une cause plausible du problème d'agglomération anarchique de données se produisant avec presque toutes les méthodes de classification examinées est la nature de la similarité exploitée par l'ensemble de ces dernières, à savoir la distance euclidienne. En effet, c'est un phénomène connu que cette distance devient faiblement discriminante dans les espaces fortement multidimensionnels contenant des données éparses [Verl04], comme c'est le cas de notre contexte expérimental. La méthode non incrémentale K-means souvent pourtant considérée comme la méthode de référence pour la classification non supervisée, est la méthode qui s'y avère la plus sensible (cf. **figure 25**). Les méthodes incrémentales testées y semblent également très sensibles, ce qui s'avère naturellement plus gênant. En particulier, la méthode IGNG définit la zone d'influence des classes en se basant sur la distance euclidienne. Le premier neurone établira donc son influence de manière non discriminante sur une part importante du corpus et provoquera par conséquent l'agglomération observée. L'apprentissage sous contrainte de grille proposé par la méthode SOM, qui exploite également la distance euclidienne, semble le plus robuste, étant donné qu'il assure l'homogénéité des résultats en répartissant à la fois les données et le bruit généré par l'utilisation de la distance euclidienne sur la carte topographique. Cette dernière méthode n'est cependant pas apte, par nature, à traiter les corpus évolutifs, et fournit en définitive des résultats de qualité assez moyenne.

	DONNEES HOMOGENES			DONNEES HETEROGENES		
METHODE DE CLUSTERING	NBR OPTIMAL DE CLUSTERS	F-MESURE MACRO	F-MESURE MICRO	NBR OPTIMAL DE CLUSTERS	F-MESURE MACRO	F-MESURE MICRO
WALLKTRAP	96	0.89	0.62	98	0.67	0.13
GERMEN	93	0.82	0.59	92	0.67	0.13
KMEANS	158	0.89	0.62	155	0.88	0.03
SOM	196	0.78	0.66	389	0.47	0.40
NG	160	0.85	0.86	160	0.59	0.33
GNG	170	0.80	0.80	–	–	--
IGNG	92	0.90	0.85	378	0.58	0.21
I²GNG	32	0.58	0.38	294	0.52	0.16

***Figure 25** : Résultats de la comparaison des méthodes de classification non supervisées sur des données homogènes et sur des données hétérogènes. Sur les données homogènes, les meilleurs résultats sont obtenus avec les méthodes basées sur une topologie libre, et particulièrement avec les méthodes NG ou GNG, et avec la méthode IGNG. Néanmoins, étant donné que la puissance de synthèse d'une méthode de clustering est aussi liée à sa capacité à produire de bons résultats avec un nombre de clusters aussi faible que possible, l'avantage tourne nettement en faveur d'IGNG qui obtient une Micro-F-mesure pratiquement équivalente à celle des méthodes NG ou GNG, tout en nécessitant un nombre de clusters deux fois moindre. Sur les données hétérogènes, les performances de l'ensemble des méthodes s'écroulent, à l'exception de celles de la méthode SOM qui continue à produire des partitions cohérentes (Voir également la **figure 20** de la **section D.6.3**). En termes de qualité, cette baisse de performance se manifeste par une forte différence entre les valeurs de Micro- et de Macro-Précision. Il est à noter que les méthodes Walktrap et Germen suppriment une partie importante des données « problématiques » de la partition finale. Malgré cette stratégie, elles affichent des performances inférieures aux autres méthodes sur les deux corpus testés, en produisant notamment des partitions incohérentes dans les deux cas. L'algorithme I2GNG souffre quant à lui, dans les deux cas, d'un problème d'initialisation de la zone d'influence des classes.*

Un deuxième problème que nous avons pu soulever à travers nos expérimentations est la dépendance, directe ou cachée, des méthodes dites « incrémentales » de paramètres globaux. Cette dépendance les rend en réalité inaptes à satisfaire aux conditions propres à l'incrémentalité.

Afin de pouvoir réaliser des analyses diachroniques, il s'avérait donc primordial de rendre les méthodes incrémentales plus robustes vis-à-vis des conditions hétérogènes. Nous avons donc, dans un premier temps, développé des améliorations des méthodes neuronales « incrémentales » existantes, puis dans un second temps, nous en avons opéré des adaptations plus profondes, pour en obtenir des résultats satisfaisants sur des données polythématiques, mais également pour obtenir des méthodes offrant un comportement réellement incrémental, et ne nécessitant donc aucun paramètre prédéfini dépendant des données initiales, tout en pouvant s'appliquer sur un espace de description des données d'entrée ouvert ou évolutif.

En particulier, nous inspirant du moule structurel de l'algorithme IGNG, nous avons développé un nouvel algorithme de gaz neuronal croissant, appelé **IGNG-F**, exploitant d'une manière incrémentale une mesure de cohésion des classes, comme alternative à la distance euclidienne utilisée dans l'algorithme IGNG original. La notion de cohésion est basée sur la **C-F-mesure** probabiliste, compromis entre la **C-Précision** et le **C-Rappel** non supervisés (cf. **section D.6.1**). Un nouveau document est affecté à une classe s'il permet d'augmenter sa valeur de **C-F-mesure**[4]. L'apprentissage du neurone associé à chaque classe correspond à la mise à jour de son étiquetage, sachant qu'un mot-clé est considéré comme étiquette d'une classe si sa valeur de **C-F-mesure** est maximale pour cette classe. Cet algorithme est présenté en détails dans la référence [IC10m], versée à l'annexe, où nous montrons également qu'il répond aux contraintes d'incrémentalité que nous avons mentionnées ci-avant.

La **figure 26** présente les résultats obtenus sur notre corpus hétérogène avec les différentes améliorations que nous avons apportées aux algorithmes IGNG et I^{2}GNG depuis l'initialisation et les stratégies d'affectation des données aux classes, jusqu'à l'exploitation globale de nouvelles mesures de similarité (**IGNG-F**). Ce tableau illustre notamment l'accroissement très significatif de performances qu'il est possible d'obtenir en utilisant l'algorithme IGNG-F. De plus amples détails sur les algorithmes originaux et sur les modifications effectuées sur ces derniers, ainsi que sur l'algorithme IGNG-F, sont donnés dans la référence [IC10m] versée à l'annexe.

METHODE DE CLUSTERING	NBR OPTIMAL DE CLUSTERS	F-MESURE MICRO
IGNG Original	378	0.21
IGNG-R (Hasard si égalité)	382	0.43
IGNG-L (Choix par F-maximisation d'étiquetage si égalité)	437	0.44
IGNG-F (Similarité basée sur F-maximisation d'étiquetage)	287	0.49
I²GNG Original	294	0.15
I²GNG-N (Initialisation par voisins réciproques)	221	0.38
Référence : SOM	389	0.40

***Figure 26** : Les résultats présentés dans ce tableau illustrent l'accroissement de qualité des partitions qu'il est possible d'obtenir après adaptation (IGNG-R, IGNG-L, I2GNG-N), ou modification en profondeur (IGNG-F), des algorithmes neuronaux « incrémentaux ». Les meilleures performances sont obtenues avec l'algorithme IGNG-F qui exploite de nouvelles mesures de similarité basées sur la maximisation de la qualité de l'étiquetage des classes.*

Les investigations que nous avons menées sur la comparaison du comportement des méthodes de classification non supervisées sur les données hétérogènes nous ont finalement permis de mettre

[4] En anglais, nous référençons également les C-mesures comme des **L-measures** [IC10l], ce qui représente l'abréviation de Labeling measures.

en place une nouvelle méthode de classification incrémentale. Cela nous ouvre donc des perspectives claires pour appliquer dans un futur proche notre nouvelle approche au domaine de l'analyse des données textuelles changeant au cours du temps. Dans un futur plus lointain, nous comptons également appliquer cette approche, en l'associant au paradigme d'analyse de données multi-vues MVDA que nous avons précédemment proposé (cf. **section D.2**), à d'autres domaines porteurs mettant en jeux des données numériques en parallèle avec des données textuelles, tout en intégrant une dimension temporelle, comme c'est le cas des données génomiques. Nous comptons également adapter le principe original de similarité par maximisation d'étiquetage des classes à plusieurs autres algorithmes de classification non supervisée. La latitude d'évolution de notre nouvelle approche pour l'analyse prospective est donc particulièrement grande.

D.7. Nouvelles méthodes d'analyse diachronique (2010)

Le principe de la comparaison des périodes de temps est établi depuis de nombreuses années. En 1986, à propos de ISI Atlas Of Science©, Garfield s'exprime ainsi : « En examinant des séquences chronologiques de cartes nous pouvons observer de quelle façon la connaissance scientifique évolue » [Garf86]. Il existe beaucoup de travaux pour afficher des graphes ou des cartes thématiques qui évoluent dans le temps, mais l'analyse de ces évolutions reste cependant jusqu'ici dévolue aux utilisateurs.

Ce type d'analyse se fonde usuellement sur l'application individualisée d'une méthode de classification non supervisée statique sur des données propres à plusieurs périodes de temps successives, et sur l'étude de l'évolution entre les différentes périodes du contenu des classes, ainsi que sur celle de leur projection sur un plan de visualisation. Cette démarche permet donc d'éviter d'avoir recours à une méthode de classification incrémentale, qui comme nous l'avons montré à la **section D.6.5**, reste difficile à mettre en œuvre. Shiebel et al [Shie10] ont proposé une première application de cette approche dans le contexte du projet européen PROMTECH, pour analyser l'évolution des domaines de recherche en optoélectronique sur deux large périodes de temps. Leur approche passe par la construction d'une matrice de comparaison de mots-clés, elle-même basée sur le pourcentage des mots-clés associés aux classes d'une période qui préexistent dans les classes de l'autre période. Grâce à cette matrice, il est alors possible à un expert du domaine de mettre en évidence différents comportements de classes : stabilité, mais également fusion ou éclatement. Une limitation importante de ce type approche est liée au fait que le processus final de comparaison doit être réalisé d'une manière supervisée, ce qui impose de travailler avec un faible nombre de classes, et donc à un faible degré de précision. Cuxac et al. [Cuxa09] ont par la suite proposé d'instrumentaliser ce processus de comparaison en exploitant des règles de correspondance floues entre les classes associées aux différentes périodes. Plus récemment, Thijs et al. [Thij10] ont proposé d'exploiter conjointement des classifications de mots-clés et des classifications de citations pour matérialiser les liens entre périodes, en tenant compte de changement éventuels de vocabulaire pouvant survenir entre ces dernières. Malheureusement, dans les deux cas mentionnés ci-avant, les méthodes conservent l'inconvénient de ne fournir que des informations de surface sur les correspondances et les divergences éventuelles entre les périodes. De plus, elles ne permettent pas de réellement de quantifier ces évolutions.

Les nouvelles expériences que nous décrivons synthétiquement dans cette section nous ont permis de montrer que le paradigme MVDA que nous avons mis en place (cf. **section D.2**) représente une excellente alternative aux approches existantes pour instrumentaliser efficacement

l'analyse diachronique en exploitant le principe de la séparation en périodes de temps. Sa combinaison avec de bonnes méthodes d'étiquetage des classes permet en effet de mettre en œuvre des stratégies d'analyse non supervisées fournissant des informations de correspondance et de divergence très précises et en contexte, entre périodes de temps, en partant du principe d'associer chaque période à un point de vue spécifique. L'identification des stabilités ou des variations entre les périodes peut alors s'appuyer sur la distribution des données ou sur la distribution de leurs propriétés dans les classes associées aux différents points de vue.

Cette dernière approche, que nous nommons « basée sur les étiquettes », s'avère être la plus précise [IC10a]. Selon celle-ci, après la construction des partitions pour chaque période de temps, une étape d'étiquetage des classes doit être réalisée. Cette étape consiste en la sélection des propriétés des donnés spécifiques à chaque classe, qui serviront à décrire cette dernière (cf. **section D.6.3**). L'identification des relations thématiques entre deux périodes de temps est ensuite réalisée par l'utilisation du raisonnement bayésien non supervisé (cf. **section D.2**). La probabilité de correspondance d'une classe source vis-à-vis d'une classe destination est directement liée à la somme des **C-F-mesure** d'étiquetage du groupe d'étiquettes qu'elle partage avec la classe destination, rapportée à la somme globale des **C-F-mesures** de l'ensemble des étiquettes de la classe destination (cf. **figure 28**, pour une vue générale sur le principe de la méthode). Cette mesure de probabilité n'est naturellement pas symétrique et permet donc d'identifier différents types d'évolution, tels que ceux illustrés à la **figure 27**.

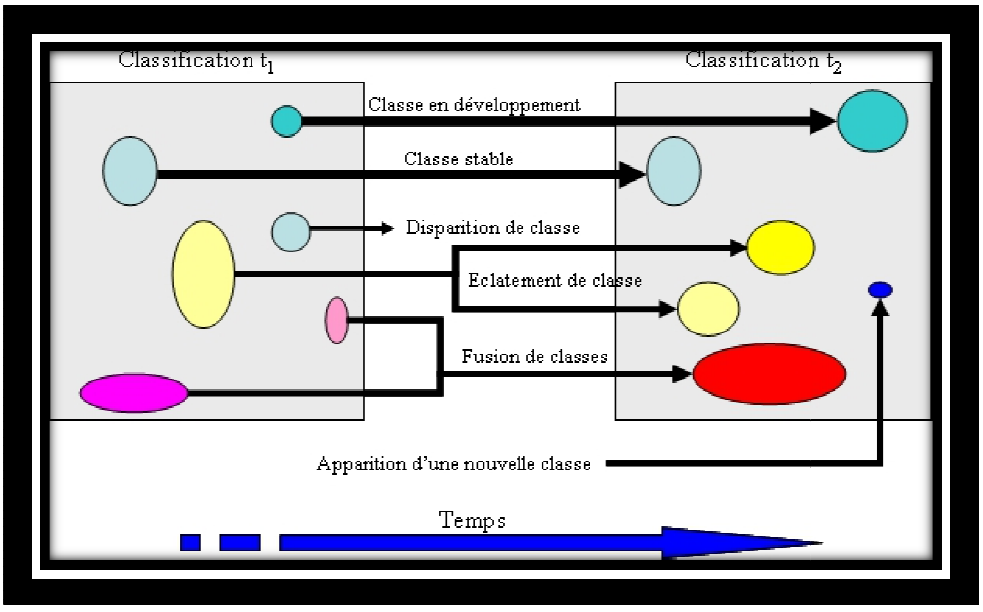

***Figure 27** : Les différents types d'évolution pouvant survenir dans une analyse diachronique basée sur la classification non supervisé et sur le découpage en périodes de temps.*

Les avantages principaux de cette approche sont sa précision intrinsèque due au procédé direct de comparaison entre partitions, ou modèles de classification, ainsi que sa conservation de l'indépendance entre les modèles comparés. Son fort pouvoir de synthèse lui évite également de produire des résultats de comparaison de trop faible granularité, lesquels pourraient rapidement s'avérer inexploitables.

Néanmoins, ces avantages ne peuvent être obtenus que sous la contrainte de l'utilisation de méthodes d'étiquetage des classes très efficaces, ainsi que par l'utilisation de méthodes de classification fournissant des partitions de qualité optimale pour chaque période. Nous avons justement abordés ces deux points dans nos travaux récents (cf. **sections D.6.1**, **D.6.3** et **D.6.5**).

Nous avons mené une expérimentation préliminaire pour tester le principe de cette nouvelle approche. Celle-ci est plus précisément décrite dans [IC10a]. Pour permettre de comparer ses résultats avec ceux des méthodes existantes, nous avons l'avons ensuite améliorée [IC10l], et plus récemment expérimentée sur les données d'étude du projet PROMTECH. Ces données sont des notices bibliographiques issues de la base PASCAL de l'INIST. Elles couvrent l'ensemble des publications de recherche indexées dans cette base sur le thème de l'optoélectronique durant la période 1996-2003. L'analyse menée par le projet PROMTECH est divisée en deux sous-périodes de temps. Nous avons donc respecté la même subdivision (1996-1999 : période 1) et (2000 2003 : période 2).

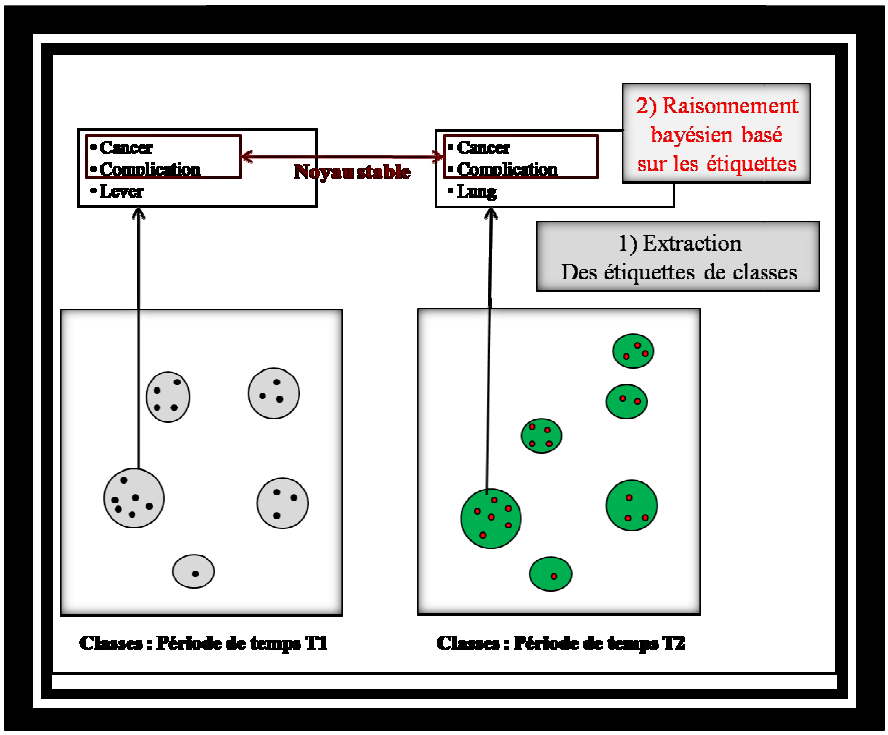

Figure 28 *: Principe de la comparaison par pas de temps exploitant le raisonnement bayésien fondé sur l'étiquetage des classes. Le raisonnement bayésien fondé sur les étiquettes est une variante du raisonnement bayésien non supervisé inhérent au paradigme MVDA (cf.* ***section D.2****). La mesure de probabilité de correspondance entre classes s'opère par l'intermédiaire de la somme des poids des étiquettes partagées entre les classes, relativement à la somme des poids de leurs étiquettes propres.*

Nous avons élaboré les classifications non supervisées à partir des mots-clés documentaliste des notices, comme cela a été réalisé dans le cadre du projet PROMTECH. Des détails plus précis sur le calcul des partitions associées à chaque période et sur les processus de caractérisation et de mise en correspondance du contenu des classes entre périodes peuvent être trouvés dans la référence [IC10l], versée à l'annexe.

La **figure 29** récapitule les résultats de notre expérience de comparaison de périodes de temps, en termes d'identification de correspondances et de différences. Comme cela est illustré à la **figure 30**, les similitudes entre les classes des différentes périodes sont identifiées par des groupes d'étiquettes partagées, que nous désignons également comme des **étiquettes noyaux**. Ces **étiquettes noyaux** illustrent de manière spécifique la nature des correspondances temporelles. D'une part, de petits changements temporels peuvent ainsi être identifiés dans le contexte environnant de ces étiquettes, et d'autre part, comme le montre la **figure 31**, les changements temporels plus importants peuvent être matérialisés par les classes isolées dont les étiquettes ne participent à aucun noyau. Dans notre approche, les résultats sont présentés sont forme de rapports de correspondance (**figure 30**), de différences (**figure 31**), de fusions ou d'éclatements.

PERIODE	NBR CLASSES	NBR SIMIL.	NBR DISPAR.	NBR APPAR.	NBR ECLAT.	NBR FUSIO.
1996-99	43	33	10		7	--
2000-03	50	38	--	12	--	3

***Figure 29** : Résultats du processus de comparaison entre périodes. Pour une période donnée, le nombre de classes impliquées dans la comparaison correspond à son nombre optimal de classes (cf. section D.1). Il est possible de noter que le nombre d'éclatements de classes de la première période dans la seconde, est plus important que leur nombre de fusions, signe de la diversification de la recherche dans le domaine de l'optoélectronique dans la seconde période.*

Les résultats produits par notre approche automatisée de comparaison de périodes de temps ont été finalement comparés à ceux de l'analyse effectuée par des experts du domaine sur les partitions produites sur des périodes de temps séparées dans l'expérience antérieure de Shiebel et al. [Shie10]. Cette dernière analyse avait principalement mis en évidence le fait que l'ensemble général de thèmes du corpus étudié correspondait aux dispositifs optoélectroniques contenant des semi-conducteurs minéraux ou organiques, et que les applications de l'optoélectronique dérivaient des applications du type « photodétecteur » (sondes, instruments de mesure,…), en période 1, vers la recherche et les applications liées aux « diodes électroluminescentes », en période 2.

Les conclusions susmentionnées présentent le défaut de rester très superficielles. L'examen des rapports de correspondances et de divergences fournis par notre nouvelle méthode d'analyse diachronique démontre qu'il est possible d'obtenir des conclusions à la fois synthétiques et précises, assorties d'indications claires de tendance (croissance ou décroissance), de façon non supervisée, tout en préservant la possibilité d'observer des orientations générales, telles que celles exprimées par les experts du projet PROMTECH. Un tel mode de fonctionnement était jusque là impossible à atteindre avec les méthodes existantes, qui par ailleurs restaient au mieux semi-supervisées. Il rend donc cette nouvelle approche particulièrement prometteuse.

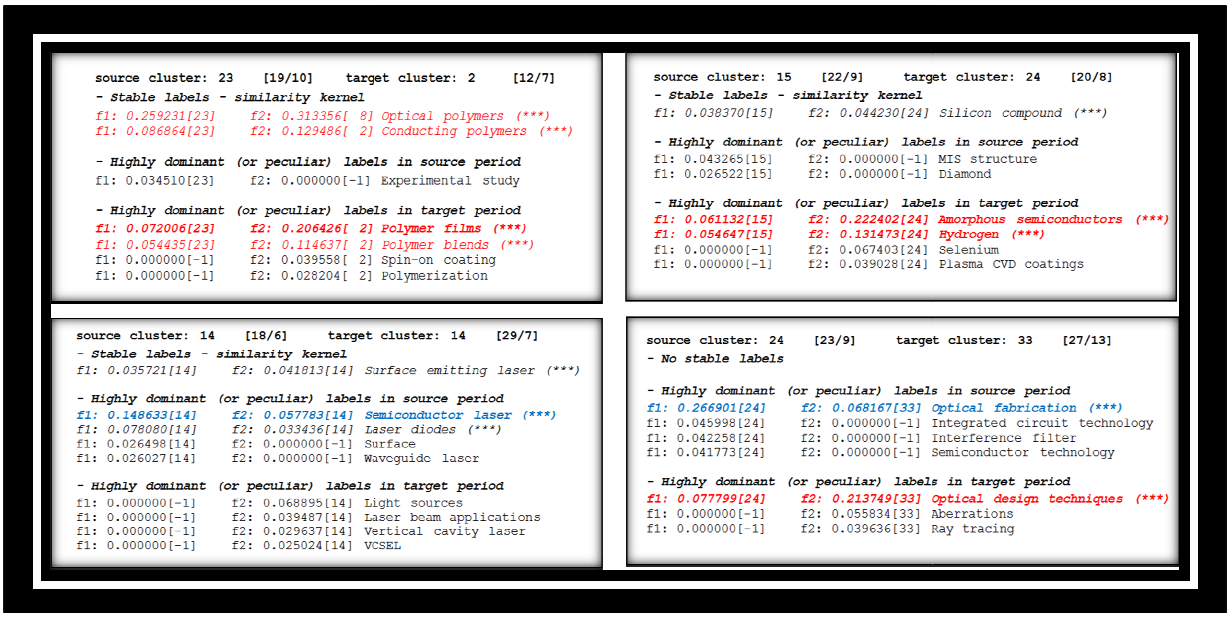

Figure 30 *: Exemples de rapports de correspondance entre périodes. Dans tous ces rapports, les étiquettes de classes sont associées à leur* ***C-F-mesure d'étiquetage****. Une valeur nulle de* ***C-F-mesure d'étiquetage*** *pour une étiquette signifie l'absence de l'étiquette concernée dans la classe considérée. La couleur bleue matérialise une décroissance très significative de la* C-F-mesure d'étiquetage *(i.e. de l'influence) d'une étiquette dans la seconde période; la couleur rouge matérialise à l'inverse une croissance très significative de cette même influence.*

Comme le montrent ces exemples, la méthode permet de faire apparaitre très clairement les développements des thématiques en contexte (ici les travaux sur les ***films polymères*** *et ceux sur les* ***semi-conducteurs amorphes****), le passage des études théoriques aux applications pratiques (****films polymères****,* ***lasers à cavité verticale****), mais également les changements de vocabulaire qui sont liés à l'évolution d'une thématique (passage de la notion de* ***fabrication optique*** *à la notion de* ***design optique****).*

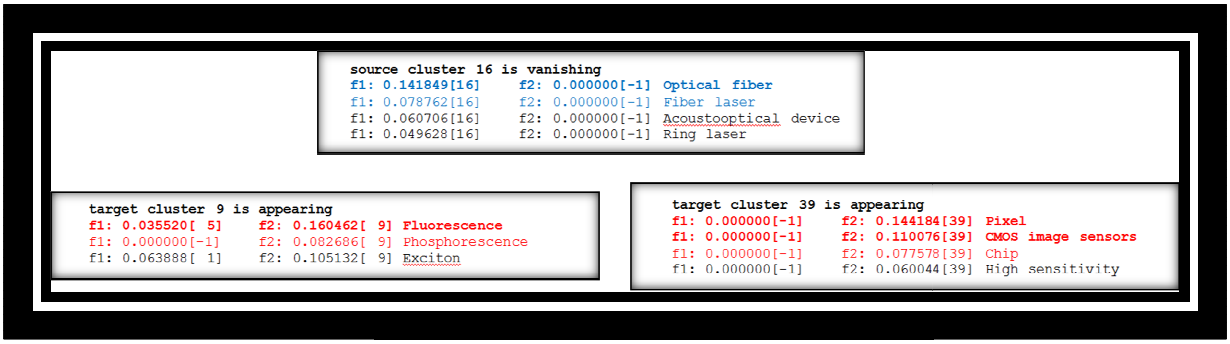

Figure 31 *: Exemples de rapports de différences entre périodes, matérialisant des disparations ou des émergences de sujets. Le premier rapport (en haut) met en évidence la disparition de la recherche sur les* ***fibres optiques*** *dans la seconde période. Le second rapport (en bas à gauche) met en évidence l'apparition de travaux sur la* ***phosphorescence****, conjointement au développement très significatif de ceux sur la* ***fluorescence****. Le troisième rapport (en bas à droite) met en lumière l'émergence de la recherche sur les* ***capteurs optiques à haute résolution*** *et sur leur* ***intégration****.*

D.8. Nouveaux outils pour le filtrage personnalisé d'information et l'apprentissage supervisé

Les modèles classiques d'analyse des décisions de l'utilisateur utilisés lors du processus de recherche d'information sont des mécanismes sans mémoire basés sur la reformulation du besoin en fonction des décisions prises par l'utilisateur (choix, rejets) en ce qui concerne les derniers résultats fournis par le système de recherche d'information (SRI). En dehors du fait d'être sans mémoire, ces mécanismes élémentaires présentent l'inconvénient de ne pas traiter de manière cohérente les décisions positives (choix) et les décisions négatives (rejets), ni de tenir compte du bruit inhérent aux décisions de l'utilisateur, pour faire le choix de nouvelles orientations de recherche. Le modèle systémique d'analyseur du comportement de l'utilisateur que nous avons développé est fondé sur la théorie de la détection de nouveauté [Koho91]. Ce modèle, nommé filtre détecteur de nouveauté, ou NDF, permet à la fois de synthétiser le besoin de l'utilisateur en fonction de l'ensemble de ses décisions, de mettre en évidence de nouvelles alternatives de recherche relativement au contenu du fonds disponible et d'évaluer la cohérence des décisions proprement dites (cf. **figure 32**). Il assure également la caractérisation du type de besoin de l'utilisateur : exploratoire, thématique, connotatif ou précis. Il bénéficie d'une trace variable qui permet de moduler l'effet des actions passées. Enfin, comme nous l'avons montré dans notre thèse, ce mécanisme présente l'avantage, relativement au bouclage de pertinence traditionnel [Rocc71], de traiter de manière homogène les décisions positives et les décisions négatives. Il a été exploité pour la première fois dans le modèle NOMAD pour gérer une composante de mémoire de session traitant des ensembles de données de taille réduite. L'approche a consisté à associer à chaque point de vue plusieurs filtres NDF traitant des types de décisions différents et partageant par la suite leurs résultats (cf. **figure 33**). Ce travail est décrit dans 3 publications internationales [IC94a][IC94b][IC94d]. Il est également détaillé dans notre propre thèse.

Nous avons envisagé par la suite l'adaptation du modèle NDF au traitement de grands ensembles de données et aux tâches de catégorisation. Dans ce cadre le modèle NDF doit être employé « à contre-sens », étant donné que se sont les directions apprises qui doivent être exploitées, et non les directions nouvelles. Ce dernier travail a permis de mettre en place une première série d'expérimentations de validation en vraie grandeur. Ces expérimentations ont notamment démontré les bonnes capacités d'apprentissage du modèle NDF dans le cas d'une catégorisation bi-classes. Ces expérimentations, qui ont fait l'objet de l'encadrement d'un DEA, ont conduit à deux premières publications, dont une dans une conférence nationale [NC05a] et une dans une conférence internationale [IC05b]. Le travail présenté par la suite a fait lui-même l'objet de l'encadrement d'une thèse.

L'adaptation du modèle NDF aux contraintes d'une catégorisation multi-classes a nécessité une modification en profondeur du modèle initial, de manière à concevoir un modèle étendu, plus général. En effet, nos expérimentations nous ont permis de constater que le modèle NDF original échouait à fournir des résultats de catégorisation satisfaisants dans le cas de corpus comprenant des données multi-étiquettes, tel que le corpus de référence Reuters. Dans de tels corpus, le nombre de caractéristiques discriminantes (i.e. propres à chaque catégorie) diminue alors que le nombre de non caractéristiques discriminantes augmente, relativement à des corpus bi-classes. Pour tenir compte de ces contraintes, nous avons défini une nouvelle approche, que nous avons baptisée ILoNDF, dont le principe consiste à réinjecter une composante de nouveauté à chaque étape d'apprentissage pour atténuer l'effet des caractéristiques non discriminantes sur ce dernier (cf. **figure 34**).

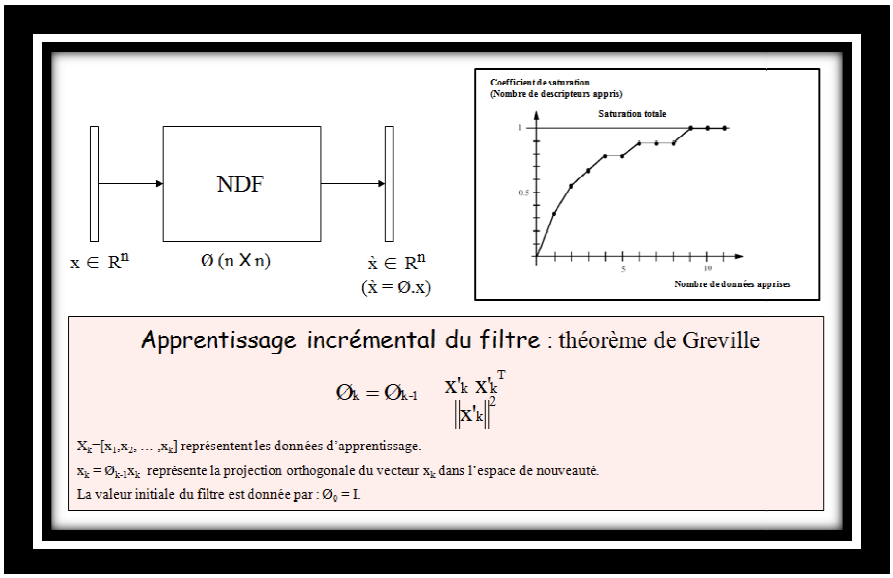

Figure 32 *: Le principe du filtre NDF est celui d'apprendre les directions propres inhérentes à un ensemble d'un ensemble de données de référence. Ces directions matérialisent un sous-espace spécifique « d'habituation » de l'espace de description global de l'ensemble des données disponibles. Le sous-espace de nouveauté est alors constitué des directions complémentaires à celles qui sont apprises à partir des données de référence. L'apprentissage du filtre peut s'opérer de façon incrémentale à partir des vecteurs descripteurs des données de référence. Une dynamique de l'apprentissage de directions « connues » peut alors être établie. Lorsque toutes les directions de l'espace de description ont été apprises, il est possible d'estimer que le filtre est saturé. Il ne produit alors plus de sortie. Ce fonctionnement est typique de celui d'une approche systémique, puisqu'il attribue au bruit un rôle constructif lors de l'apprentissage.*

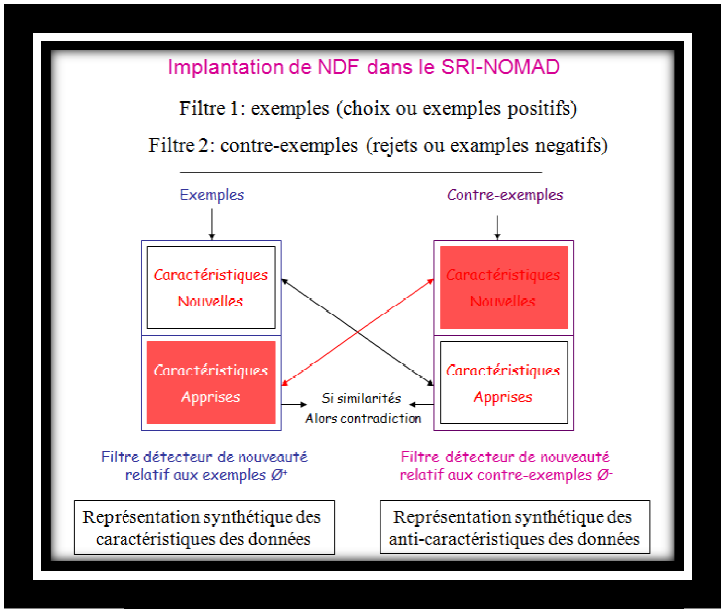

Figure 33 *: Dans le modèle NOMAD la mémoire de session est structurée autour du modèle NDF. Chaque point de vue comprend plusieurs modèles NDF attachés à chaque modalité de décision offerte à l'utilisateur.*

$$\emptyset_k = I + \emptyset_{k-1} - \frac{x\prime_k x\prime_k^T}{\|x\prime_k\|^2}$$

Figure 34 *: Dans le modèle ILoNDF, la nouveauté est réinjectée à chaque étape d'apprentissage, par l'intermédiaire de la matrice identité.*

Dans la suite de nos expériences, nous avons donc choisi d'utiliser la collection-test Reuters-21578 comme corpus de référence. La collection Reuters-21578 est constituée de 21578 documents représentant des extraits de dépêches de l'agence de presse Reuters. Les documents sont classés manuellement dans une ou plusieurs catégories, choisies parmi 135 catégories sémantiques. Les résultats de nos expérimentations se rapportent aux 10 catégories les plus fréquentes du corpus Reuters, qui représentent celles qui sont majoritairement exploitées lors des tests des méthodes de catégorisation et de filtrage. Le nombre de documents d'apprentissage associés à ces catégories varie de 181 à 2877, avec une moyenne de 719.3 documents par catégorie. La **figure 35** résume les informations sur ces catégories, en termes d'effectif d'apprentissage et d'effectif de test.

REUTERS		
Catégorie	Nb. doc apprentissage	Nb. Doc test
earn	2877	1087
acq	1650	719
money-fx	538	179
grain	433	149
crude	389	189
interest	347	131
trade	369	117
wheat	212	71
ship	197	89
corn	181	56

Figure 35 *: Les 10 catégories principales de la collection-test Reuters-21578 et leurs effectifs.*

Dans le cadre de l'étude sur ce corpus, de nouvelles méthodes d'initialisation, ainsi que des méthodes inédites de prétraitement des données ont donc été proposées [IC06a]. Les résultats obtenus pour la validation du modèle neuronal ILoNDF en multi-catégorisation, sont apparus d'emblée excellents. Pour démontrer les capacités véritables du modèle, nous avons cependant suivi plusieurs directions de recherche complémentaires. Celles-ci sont présentées ci-après.

Une de nos premières directions de recherche est celle de l'étude du comportement du nouveau modèle ILoNDF, en partant de l'exploitation des seuls exemples positifs (catégorisation à une seule classe). Cette approche fournit une solution pratique aux problèmes qui se posent dans les cas où les exemples négatifs sont difficiles à obtenir, très rares ou encore totalement indisponibles (inconnus), comme par exemple dans les applications de diagnostic ou de détection de défauts

(outliers), ou encore, dans les systèmes de recommandation, où les utilisateurs n'ont pas nécessairement la possibilité, ou la capacité, de fournir des exemples négatifs de leur besoin (votes négatifs). Plusieurs études, dont l'une des nôtres [PH09a], ont par ailleurs permis de prouver que l'apprentissage d'un modèle de classification uniquement à partir des exemples positifs permet d'obtenir de meilleurs résultats dans le cas d'un fort déséquilibre entre classes (le nombre d'exemples de la classe positive est faible et bien inférieur au nombre d'exemples de la classe négative) [Rask04][IC07a].

Après avoir expérimenté les méthodes statistiques multivariées, telles que la méthode des résidus de l'ACP [Jack79] ou la méthode T2 de Hotelling [Kim02], usuellement exploitées dans le cadre du problème de la classification à une seule classe, nous avons focalisé nos comparaisons sur nos deux modèles de référence, à savoir les modèles NDF et ILoNDF, ainsi que les modèles alternatifs à ceux-ci qui se sont avérés les plus performants, à savoir les réseaux de neurones auto-associatifs et les machines à vecteurs support (SVM) :

- Les réseaux de neurones auto-associatifs (AANNs), également connus sous le nom d'auto-encodeurs, sont un type spécial de réseaux multicouches de type feedforward. Ces réseaux, introduits à l'origine par Rumelhart et al. [Rume86], sont entraînés pour produire une approximation de la fonction identité entre les entrées et les sorties du réseau, dont l'effectif est identique, en terme de nombre de neurones. L'exploitation d'une couche cachée, de plus faible effectif, permet de capturer les variables significatives (similaires aux composantes principales de l'ACP) dans la représentation des données d'apprentissage en comprimant leurs redondances, sans chercher à mémoriser les données.
- La méthode 1-SVM est une version de machine à vecteurs supports adaptée à la classification à partir d'une seule classe. Elle a été proposée par Schölkopf et al. [Scho99]. Son idée est de séparer les exemples positifs de l'origine, le seul exemple négatif, avec une marge maximale dans l'espace de re-description associé à la fonction noyau choisie. En termes plus précis, 1-SVM recherche une hypersphère de volume minimal qui englobe la plupart des exemples positifs disponibles pour l'apprentissage.

Dans nos expérimentations, nous avons exploité les méthodes proposées, en optimisant séparément pour chacune d'entre elles les méthodes de pondération des données utilisées, ainsi que les paramètres de travail caractéristiques. Les résultats obtenus sont présentés globalement à la **figure 36** et à la **figure 37**, et par catégories à la **figure 38**.

À partir de ces résultats, nous avons pu soulever les constatations suivantes : le modèle NDF original ne s'avère pas particulièrement efficace sur le corpus de référence Reuters. Par contre, notre nouveau modèle ILoNDF dépasse sensiblement en performance le modèle NDF, produisant jusqu'à 30% d'amélioration sur ce même corpus. En outre, ILoNDF dépasse constamment les meilleures méthodes référencées jusqu'à aujourd'hui, apportant des améliorations d'environ 6-7 % sur les méthodes 1-SVM et d'AANN, respectivement. Comme l'illustrent les résultats présentés à la **figure 38**, ILoNDF tend à être particulièrement performant dans le cas des petites catégories ambiguës où la représentation des données peut être très bruitée. Pour sa part, la méthode 1-SVM semble être légèrement meilleure que la méthode AANN, avec une amélioration de l'ordre de +1%.

Pour conclure sur ces premiers résultats, il est important de noter que le modèle ILoNDF s'est avéré avoir un coût de calcul comparable à celui de la méthode 1-SVM, et notablement moins

important que celui de la méthode AANN. Par ailleurs, en raison des nombreux paramètres intervenant dans la méthode 1-SVM, une série d'expériences de validation doit généralement être effectuée pour déterminer leurs valeurs optimales. De telles optimisations sont coûteuses en temps de calcul, et augmentent ainsi de manière significative la complexité de calcul de 1-SVM. Pour sa part, ILoNDF ne requiert pas de telles optimisations et s'avère être moins sensible aux aspects expérimentaux tels que les schémas de pondération et la dimensionnalité de l'espace de représentation.

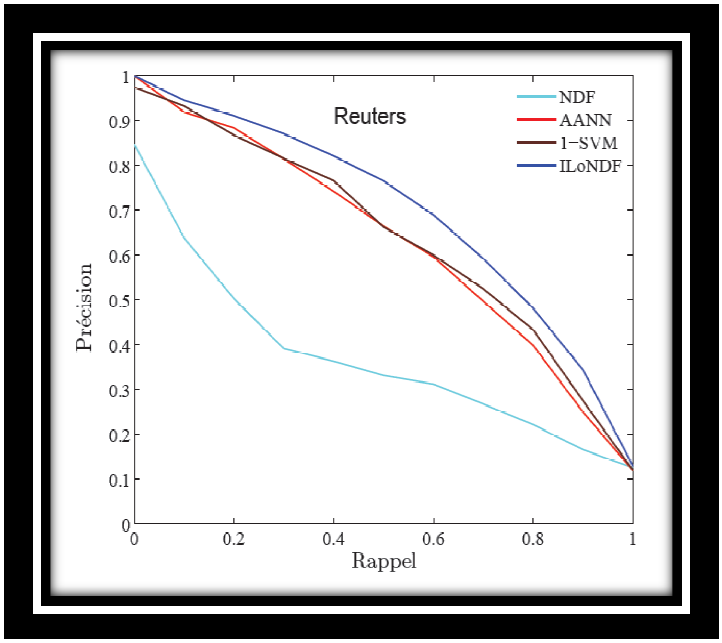

***Figure 36** : Courbes de Rappel-Précision moyennées sur les dix catégories principales de la collection-test Reuters-21578, pour la classification à une seule classe, et pour les différentes méthodes (NDF, ILoNDF, 1-SVM et AANN). Ces courbes illustrent bien la supériorité de la méthode ILoNDF sur l'ensemble des autres méthodes.*

	Reuters	
	MAP	BKP
NDF	0.3512	0.3584
AANN	0.6272	0.5935
1-SVM	0.6356	0.6086
ILoNDF	**0.6962**	**0.6552**

***Figure 37** : Résultats moyens des mesures usuelles BKP et MAP sur les dix catégories principales de la collection-test Reuters-21578, pour la classification à une seule classe, et pour les différentes méthodes (NDF, ILoNDF, 1-SVM et AANN). La mesure BKP est une F-mesure interpolée qui correspond à l'intégration des résultats fournis par la courbe Rappel-Précision (cf. Figure). La mesure MAP est une mesure moyenne de Précision, non interpolée, calculée de manière incrémentale en parcourant la liste des données de test classées par ordre décroissant de pertinence et en relevant les données correctement classées.*

Une seconde direction de recherche que nous avons suivie est celle de la mise en place de stratégies de filtrage en batch orientée-utilisateur à l'aide du modèle ILoNDF. Dans le contexte de cette nouvelle étude, toujours menée sur la collection-test Reuters, chaque catégorie principale de la collection a été utilisée pour simuler un besoin spécifique à l'utilisateur : les documents d'apprentissage associés à chaque catégorie ont donc été exploités en tant qu'exemples positifs du besoin de l'utilisateur, et les documents associés aux autres catégories, en tant qu'exemples négatifs de ce même besoin. En raison du nombre relativement plus important d'exemples négatifs, et à des fins d'optimisation de performance, nous avons cependant choisi d'opérer un choix plus sélectif de ces derniers exemples, en appliquant une technique de "Query Zoning" [Sing96], qui revient choisir les exemples négatifs dit « positifs proches », autrement dit ceux qui sont les plus difficiles à distinguer des exemples négatifs.

Les stratégies que nous avons proposées s'inspirent de la gestion des contradictions antérieurement proposée dans le modèle NOMAD [PH95a] (cf. **Figure**). Elles exploitent deux modèles ILoNDF associés à chaque type d'exemple (positifs, négatifs), et sont appliquées postérieurement à l'apprentissage.

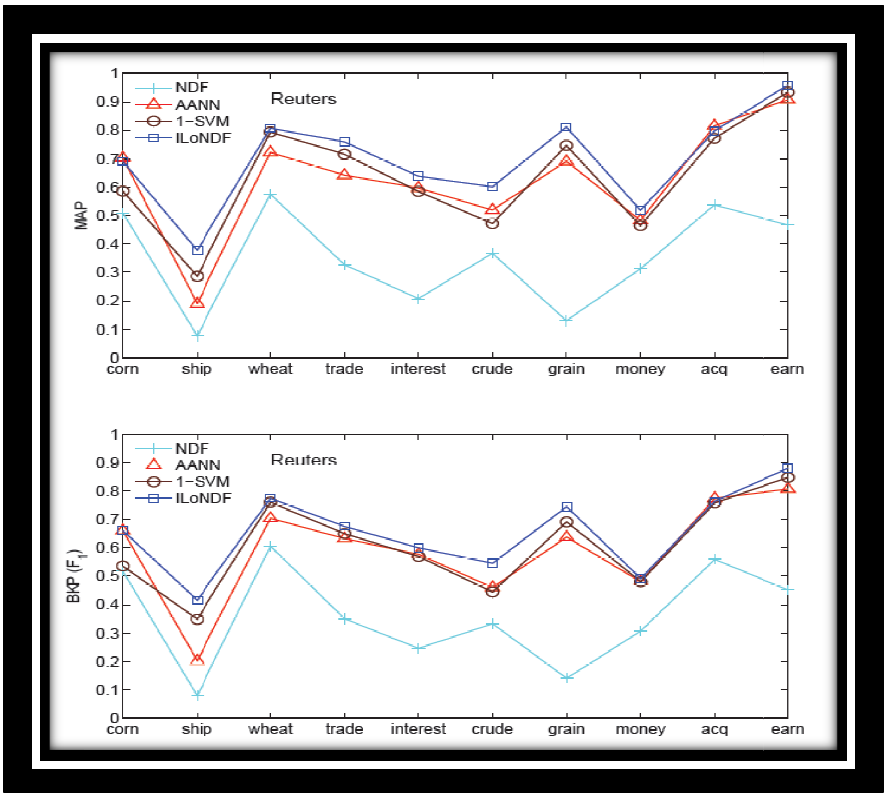

Figure 38 *: Résultats présentés par catégories, pour les dix catégories principales de la collection-test Reuters-21578, pour la classification à une seule classe, et pour les différentes méthodes (NDF, ILoNDF, 1-SVM et AANN). Ces courbes illustrent le fait que la méthode ILoNDF fournit bien les meilleurs résultats pour les catégories bruitées. Les catégories "**ship**" et "**crude**" sont connues pour être fortement recouvrantes, et partagent ainsi beaucoup de termes communs (le pourcentage de dispersion des termes sur ces deux catégories est de 70.5%). De même, la catégorie "**trade**" est aussi en lien avec la catégorie "**earn**" (le pourcentage de dispersion des termes est 55.5%), mais la catégorie "**earn**" est moins affectée par ce phénomène en raison de sa taille importante par rapport à celle de "**trade**".*

Une des premières stratégies que nous avons envisagée consiste à mener à bien une analyse non paramétrée du contenu du filtre ILoNDF. Cette approche nous a permis de définir une méthode de seuillage basée sur la précision attendue et qui dépend directement du comportement du flux d'entrée (i.e. de l'utilisateur) et de ses caractéristiques intrinsèques (i.e. de son type de besoin). Cette stratégie est tout à fait originale et pertinente dans le domaine, et, point intéressant à mentionner ici, seul les modèles neuronaux de type NDF ou ILoNDF permettent de la mettre en œuvre. Nous avons également pu démontrer que, dans certains cas, elle induisait indirectement une amélioration des résultats bruts du filtrage, ceci en exploitant une optimisation du seuil de filtrage basée sur la F-mesure étendue, ou F-Beta mesure. Les principes et les résultats de cette stratégie sont présentés plus explicitement dans la référence [IC07a], versée à l'annexe.

De manière connexe, nous avons cherché à optimiser l'exploitation des paramètres de mélange associés au traitement des informations acquises à partir des exemples positifs et de celles acquises à partir des exemples négatifs, lors de la phase d'évaluation des méthodes de classification et de filtrage. Ce travail avait pour but d'évaluer l'incidence du traitement des choix et des rejets dans un processus de bouclage de pertinence associé aux décisions de l'utilisateur dans le contexte du filtrage. Il devait également permettre de comparer de manière objective les différentes méthodes de classification et de filtrage susceptibles de traiter à la fois les exemples positifs et les exemples négatifs. Nous avons retenu dans ce contexte la méthode de Rocchio [Rocc71], la méthode SVM Light [Joac99] et le modèle ILoNDF. Nous avons examiné le cas d'une optimisation globale des paramètres de mélange et celui d'une optimisation locale, propre à chaque catégorie Reuters, par validation croisée [Mosc03]. Les résultats que nous avons obtenus se sont avérés proches de ceux obtenus pour la classification à une seule classe. Ils sont présentés de manière synthétique à la **figure 39**.

		Macro			Micro		
	MAP	Précision	Rappel	F_1	Précision	Rappel	F_1
Reuters10							
Rocchio (Opt.locale)	0.9298	0.8767	0.8463	0.8595	0.9208	0.8671	0.8931
(Opt.globale)	0.9215	0.8810	0.8290	0.8521	0.9160	0.8617	0.8880
ILoNDF(Opt.locale)	**0.9457**	0.9129	0.8564	**0.8821**	0.9419	0.9086	**0.9250**
(Opt.globale)	0.9415	0.9072	0.8423	0.8724	0.9392	0.8992	0.9187
SVM	0.8424	0.7665	0.5077	0.5839	0.9501	0.7903	0.8629
CS-SVM	0.9310	0.8079	0.9414	0.8669	0.8686	0.9556	0.9100
Reuters90							
Rocchio (Opt.locale)	0.6911	0.5978	0.5903	0.5096	0.4389	0.7975	0.5662
(Opt.globale)	0.6580	0.5836	0.5538	0.4962	0.5558	0.7598	0.6420
ILoNDF(Opt.locale)	0.7006	0.5906	0.4930	0.4701	0.6366	0.7665	0.6955
(Opt.globale)	0.6775	0.5940	0.4928	0.4712	0.6479	0.7649	0.7016
SVM	0.5220	0.3239	0.1715	0.2102	0.9214	0.6178	0.7397
CS-SVM	**0.7024**	0.5810	0.4847	**0.4982**	0.7888	0.8821	**0.8328**

***Figure 39** : Comparaison des moyennes des scores MAP, Précision, Rappel et F-mesure obtenus pour le modèle ILoNDF et les méthodes Rocchio et SVM. (Pour plus de détails, voir la référence [IC06a] versée à l'annexe, ou également, la référence [PH09a]).*

D'autres expérimentations menées sur collection-test de pages Web institutionnelles WebKB, constituée par une équipe de l'université de Carnegie Mellon (CMU), nous ont également permis de confirmer les résultats obtenus sur le corpus Reuters.

L'ensemble de ce travail ce travail a globalement fait l'objet de plusieurs publications internationales [IC06a][IC06d][IC07a] et d'un article en révision dans la revue Machine Learning [IJ07b]. La référence [IC06a] est rajoutée en annexe.

Un travail d'application exploitant notre approche pour améliorer le comportement des systèmes de distribution d'information à la demande à également été réalisé. Ce travail a permis de définir l'architecture finale du prototype de distribution personnalisé d'information par satellite CASABLANCA [TR05a][TR05b][TR05c]. Il a impliqué la collaboration avec une autre équipe de recherche travaillant sur le filtrage collaboratif.

D.9. Travaux annexes

D.9.1. Fouille de données symbolique

Une méthode classique d'extraction de règles d'association utilisée dans le domaine de la fouille de données symbolique consiste à identifier, dans un premier temps, les groupes de propriétés, ou motifs, les plus fréquents présents dans les données de la base à analyser [Agra96]. Certains sous-ensembles de motifs, comme les fermés, ou les générateurs, sont cependant plus intéressants à considérer pour isoler des règles pertinentes. Les motifs fermés sont les groupes de propriétés qui maximisent l'information sur les données et les générateurs sont les groupes de propriétés qui la minimisent. [Pasq00] a montré comment ces deux types de motifs pouvaient être employés pour extraire des règles d'association informatives, autrement dit, des règles avec les prémisses, ou causes, minimales (basées sur les générateurs) et avec des conséquences maximales (basées sur les motifs fermés). D'une part, [Pasq99] ont proposé l'algorithme Close pour extraire de manière incrémentale les motifs fermés. D'autre part, [Bast06] ont proposé l'algorithme Pascal pour extraire efficacement les générateurs. L'algorithme Zart, proposé par [Szat05], représente quant à lui la première tentative pour combiner l'algorithme Pascal avec l'extraction de motifs fermés. Même si cette dernière approche est particulièrement intéressante pour assister dans le choix de règles informatives, elle ne fournit aucune solution satisfaisante concernant le temps de calcul de ces règles. En effet, son mode opératoire rend nécessaire de nombreux accès coûteux à l'information de la base des données source.

Nous avons donc proposé un nouvel algorithme original permettant d'extraire de manière optimale les motifs fermés à partir des générateurs. Tout comme l'algorithme Zart, notre algorithme fait appel à l'algorithme Pascal pour extraire efficacement les générateurs. Il ne rend cependant plus nécessaire l'accès à l'information de la base des données source pour l'extraction des motifs fermés. Le principe de notre algorithme est celui d'extraire les motifs fermés associés aux générateurs de longueur n en se servant des générateurs de longueur n + 1. Le motif fermé d'un générateur donné de longueur n est obtenu en éliminant de ce dernier tous les motifs élémentaires (i.e. les propriétés) obtenues en différenciant ce générateur des générateurs de longueur n + 1 ayant un plus petit support.

Cet algorithme a fait l'objet d'une communication internationale [IC06k] et de l'encadrement d'un stage de DEA [MA06a]. Il a nécessité la démonstration d'un théorème important sur les générateurs. Cette démonstration est explicitée dans la référence [IC06k]. Elle est susceptible d'être publiée par la suite dans un journal international.

D.9.2. Apprentissage semi-supervisé

L'apprentissage semi-supervisé s'avère nécessaire dans le cas de corpus importants pour lesquels l'étiquetage n'a pu être réalisé que de manière partielle. Dans ce cadre, les capacités de généralisation de la méthode d'apprentissage utilisée s'avèrent primordiales. Les machines à vecteurs supports, ou SVM [Vapn98], sont connues pour posséder de telle capacités. Elles représentent donc une alternative intéressante, d'autant plus que Demiriz et al. [Demi98] en ont proposé une formulation transductive, TSVM, qui se ramène à la résolution d'un problème de programmation linéaire en variables mixtes. La résolution d'un tel problème, NP-complet, peut cependant s'avérer très couteuse en temps de calcul, notamment lorsque les données concernées sont fortement multidimensionnelles. Dans ce cadre, nous avons proposé une méta-heuristique de recherche de solution optimale, que nous avons nommé recherche taboue réactive (RTR), et qui représente une elle-même extension de la méta-heuristique de recherche taboue classique (RT) [Glov89]. Celle-ci présente l'avantage de permettre d'ajuster automatiquement les paramètres génériques de recherche par apprentissage au cours du processus de recherche. Nos résultats préliminaires, sur les corpus de données-test issus de la librairie de référence LIBSVM ont montré que notre méthode RTR pouvait effectivement trouver les solutions globales optimales pour la formulation TSVM sur des problèmes de dimension relativement grande et que l'approche s'avérait concurrentielle, en terme de pouvoir de généralisation, avec les approches de référence, telle que l'approche Transductive SVMlight [Joac99].

Cette proposition a fait l'objet de 2 communications internationales [IC06i][IC07e], et d'un stage de recherche [Re06a]. Elle a également fait l'objet de la participation à l'encadrement d'une thèse étrangère.

D.10. Wikis sémantiques évolutifs (2010)

Nous participons au projet WICRI dont le but est de développer le concept novateur de réseau auto-organisateur de wikis sémantiques. Notre rôle spécifique dans ce projet est de proposer des stratégies automatisées d'enrichissement du réseau de wikis par l'intermédiaire de l'exploitation de données du Web. En effet, la fouille du Web représente un challenge important pour améliorer à la fois la réactivité, la flexibilité et la portée d'un tel type de réseau. D'un coté, ce processus est nécessaire pour assister les contributeurs potentiels lors de la phase de construction du réseau en leur fournissant des règles rédactionnelles fiabilisées. D'un autre coté, il s'avère également décisif pour fournir aux utilisateurs finals du réseau de l'information externe à celui-ci, avec comme valeur ajoutée de pouvoir la rapporter au contexte sémantique du réseau. Bien que le projet WICRI soit encore dans sa phase de démarrage, notre prototype de réseau de wikis est déjà opérationnel, et peut ainsi être exploité comme une plate-forme de recherche et d'investigation collaborative disponible en ligne [IC10h].

D.11. Applications de recherche en webométrie et en scientométrie

L'étude de l'utilisation des liens entre les sites Web académiques pour créer un mode informel de communication savante est un champ de recherche prometteur. Des analyses qualitatives récentes [Wilk03] ont montré que les métriques s'appuyant sur le dénombrement des liens permettaient d'identifier d'une large variété de manières les agglomérations de connections liées aux activités savantes. Récemment, il y a eu une croissance forte des études sur les liens hypertextes dans le domaine de la recherche sur l'Internet [Park03]. L'idée fondamentale est que la collaboration et l'échange d'information entre les organismes académiques peuvent être reflétés par les réseaux de liens hypertextes [Heim06].

Selon Chu [Chu05], les analyses de référencement devraient non seulement considérer le dénombrement des liens, mais également les causes du référencement pour assurer leur propre validité. D'ailleurs, le comportement de référencement a été démontré comme dépendant du domaine étudié. A titre d'exemple, Thelwall et al. [Thel03] ont constaté que les domaines des mathématiques et de l''informatique avaient tendance à plus « s'inter-relier » que d'autres domaines scientifiques. Des travaux complémentaires des mêmes auteurs ont permis de franchir une nouvelle étape, en démontrant qu'une analyse précise du comportement de référencement dépendait de la capacité de la méthode d'analyse à prendre en considération plusieurs facteurs complémentaires, comme la discipline étudiée et les facteurs géographiques [Thel02]. Néanmoins, la réalisation d'un tel type d'analyse à large échelle condamne l'utilisation des approches « manuelles », comme celle réalisées jusqu'à aujourd'hui. L'analyse des bases des liens extraits des sites Web doit de plus se heurter à l'absence d'une véritable typologie des liens [Ingv03].

L'emploi d'une stratégie d'analyse composite et automatisée représente donc une manière prometteuse de résoudre ces problèmes: celle-ci revient à interpréter de manière non supervisée les regroupements de liens en incluant les caractéristiques des sites Web, selon des points de vue multiples dans l'analyse.

L'approche que nous avons proposée dans ce cadre consiste à combiner des classifications thématiques associées au contenu et à la description des pages Web avec des classifications de liens, d'abord dans le contexte du modèle d'analyse multi-vues NONAD-MultiSOM, et puis ultérieurement, dans celui plus général du modèle MVDA. Cette approche exploite le raisonnement bayésien inhérent à ces modèles pour caractériser les regroupements de liens, de manière locale ou globale, à partir des informations fournies par les classifications thématiques, Elle exploite également les principes d'extraction automatique de règles d'association exploitables à grande échelle que nous avons définis pour identifier les dépendances dans le référencement. Nous avons réalisé plusieurs applications-test pour démontrer les capacités de cette approche. Pour ces applications, nous nous sommes basés sur un corpus de référence constitué dans le cadre du projet européen IST-EICSTES [Arro03]. Les données de ce corpus couvrent les sites Web des universités et des instituts de recherche de la plupart des pays de l'Union européenne. Elles concernent 1736 sites différents et mentionnent, entre autres, les URL des sites, les URL de leurs liens entrants et sortants, la situation géographique des institutions concernées, et leurs thèmes de recherche. Elles représentent donc une base assez riche pour expérimenter de nouvelles méthodes d'analyse.

Une première expérimentation du modèle NOMAD-MultiSOM dans ce cadre nous a permis de montrer qu'il était possible d'identifier la nature de la politique de recherche et des relations industrielles d'un laboratoire et de ses équipes en croisant dynamiquement des analyses de contenu avec des analyses de liens sortants opérées sur le même ensemble de pages Web, extraites du site du laboratoire. Cette expérimentation préliminaire a fait l'objet d'un chapitre dans un des rapports du projet européen EISCTES [TR03b].

Nous avons étendu le principe de cette expérimentation au contexte général des laboratoires européens en nous basant sur les données globales du corpus EISCTES et en prenant en compte des facteurs multiples tels que la situation géographiques des laboratoires, leurs thèmes de recherche privilégiés, leurs liens sortants, et leur profil de co-référencement. Cette dernière

expérimentation nous a permis de montrer qui était possible de mettre en évidence de manière non supervisée, et à grande échelle, des relations privilégiées entre les pays, les thèmes de recherche et les types de communication savante, en associant chaque facteur d'analyse à un point de vue différent et en croisant dynamiquement les analyses obtenues pour chacun des points de vue. Ce travail fait l'objet d'une communication dans une conférence internationale [IC04a], et d'une publication dans un journal international [IJ04b].

Nous avons ensuite mené une nouvelle série d'expérimentations à plus large portée concernant l'analyse du référencement. Ce type inédit d'expérimentation avait pour but de démontrer l'efficacité particulière, dans ce contexte, des mécanismes de comparaison dynamique de classifications que nous avons mis en place dans le cadre du modèle MVDA. Nous avons pour cela réalisé l'étude du comportement de référencement des laboratoires de recherche allemands à partir des données du corpus EISCTES en appliquant ces mécanismes. Notre ensemble de données couvrait 438 sites Web de laboratoires de recherche en informatique allemands. Nous avons considéré 4 points de vue différents représentés par les villes d'implantation des laboratoires, leurs domaines de recherche, leurs liens entrants et leurs liens sortants, puis exploité des classifications associées aux différents points de vue, construites à partir de gaz de neurones. Cette approche s'est avérée très fructueuse, puisqu'elle nous a permis aussi bien d'avancer des hypothèses précises sur le référencement, comme la consistance géographique, que de caractériser des épiphénomènes intéressants liés aux réseaux de liens, comme la différence générale de comportement entre le référencement entrant et le référencement sortant. Ce travail a fait l'objet d'une communication dans une conférence internationale [IC06f], et d'une publication dans un journal international [IC06b].

La **figure 41** illustre le type de résultats d'analyse de comportement de référencement qu'il est possible d'obtenir en exploitant un mécanisme de comparaison dynamique de classifications tel que celui que nous avons proposé dans le modèle MVDA. Ce mécanisme est également présenté synthétiquement à la **figure 40**. De plus amples détails sur l'ensemble de cette expérience sont donnés dans la référence [IJ06b], versée à l'annexe.

Dans le domaine plus large de la scientométrie, nous nous sommes par la suite intéressés à l'analyse et à la validation du contenu des réseaux sociaux construits à partir des méthodes d'analyse des co-publications [Kret85]. Nous avons proposé pour cela une méthodologie basée sur la distribution d'étiquettes exploitant les mesures de qualité d'étiquettage que nous avions préalablement définies (cf. **section D.1**). Le principe consiste à projeter sur un réseau d'auteurs préalablement établi des étiquettes construites à partir de l'identification des thèmes de recherches extraits des publications des auteurs. Les mesures de qualité, telles que l'entropie de C-Rappel et l'entropie de C-Précision permettent ensuite d'analyser la répartition des étiquettes sur les groupes d'auteurs pour identifier les groupes ou les thèmes de recherche pertinents ou privilégiés. Ce travail fait l'objet d'une communication dans une conférence internationale [IC07b]. Il reste cependant à finaliser pour faire très prochainement l'objet d'une soumission dans un journal international.

Nous avons complété notre contribution au domaine de la scientométrie en démontrant les avantages des méthodes de visualisation hyperbolique relativement aux méthodes de visualisation par graphes pour l'identification des réseaux sociaux [IJ06b], et en proposant également de nouvelles méthodes d'étiquetage adaptées à ce contexte [IC08b].

Nous travaillons actuellement sur l'exploitation d'une méthode d'analyse diachronique automatisée applicable au domaine très porteur de l'analyse de l'évolution des thèmes de recherche et des groupes de collaboration au cours du temps. La méthode que nous proposons est décrite à la **section D.7** et dans la référence [IC10l]. Elle intègre une grande partie de nos travaux et de notre expérience antérieure. Nous avons montré qu'elle permettait de sursoir aux problèmes posés par les méthodes existantes. Les résultats qu'elle a produits sont en cours de validation pour une soumission très prochaine dans une conférence internationale en scientométrie, et dans un journal international dans le même domaine.

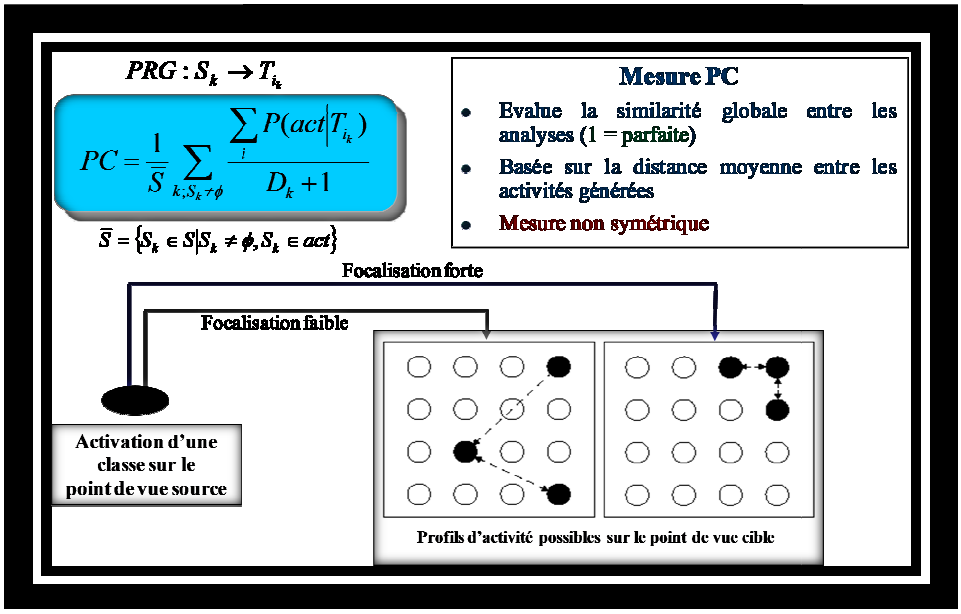

Figure 40 *: Le mécanisme de comparaison dynamique de classifications du modèle MDVA est une mesure de similarité globale non symétrique et adaptative qui permet d'analyser la cohérence entre des partitions construites à partir des mêmes données. Il est fondé sur une mesure de cohérence moyenne (PC) des classes d'un modèle source vis-à-vis des classes d'un modèle destination. La mesure de cohérence individuelle de chaque classe source tient compte de la distance moyenne entre les classes cibles qu'elle active sur le modèle destination, par l'intermédiaire de ses données ou de ses étiquettes associées.*

	Towns	Sub-domains	Outlinks	Inlinks
Towns	1	0.096	0.250	0.270
Sub-domains	0.216	1	0.554	0.416
Outlinks	0.117	0.095	1	0.270
Inlinks	0.080	0.055	0.130	1

Figure 41 *: Le mécanisme de comparaison de classifications du modèle MDVA est ici appliqué à la comparaison entres différents types de classifications (villes, thèmes de recherche, lien entrants et liens sortants) menées à partir des mêmes données web institutionnelles relatives aux laboratoires de recherche allemands. Cette analyse permet par exemple d'observer une forte cohérence de comportement des domaines de recherche, à la fois vis-à-vis du référencement entrant et vis-à-vis du référencement sortant, avec cependant une valeur sensiblement plus importante vis-à-vis du référencement sortant.*

E. PERSPECTIVES DE RECHERCHE

E.1. Cadre général

Nos perspectives de recherche s'avèrent assez larges, étant donné les nombreux domaines d'application potentiels de notre travail, et en particulier ceux du paradigme d'analyse de données multi-vues MVDA.

D'une manière générale, une réflexion théorique approfondie est nécessaire pour aboutir à une meilleure appréhension du potentiel et à une meilleure coordination des méthodes que nous avons proposées, qu'il s'agisse des méthodes de synthèse des résultats de l'apprentissage, des méthodes d'évaluation de la qualité de classification, ou des méthodes de détection de nouveauté, de filtrage d'information et d'apprentissage supervisé, tant celles-ci sont nombreuses, et somme toutes complémentaires. Cette approche intégratrice nous permettra sans doute de mettre en place, à terme, une plate-forme intelligente et réactive traitant la fois les tâches complexes liées à l'analyse de données, mais également celles liées à la veille et au traitement des flux d'information dans un environnement évolutif et distribué, et selon des principes systémiques.

De même, de nombreux ponts restent à explorer en ce qui concerne la coopération entre les approches numériques et les approches symboliques dans les processus d'analyse et de fouille de données, qui comme nous l'avons déjà montré à plusieurs reprises, pouvait s'avérer très fructueuse.

Nous avons cependant de bonnes de penser qu'un contexte d'enrichissement privilégié du travail de recherche que nous avons mené jusqu'alors est celui de la classification incrémentale.

Le problème de la classification incrémentale est un problème complexe pour lequel peu de solutions on été proposées jusqu'à aujourd'hui. Il s'agit en effet de pouvoir détecter de nouveaux sujets au fil de l'arrivée des nouvelles données, ceci avec une réactivité suffisamment bonne, ce qui revient à trouver un compromis optimal entre la plasticité et la stabilité du modèle de classification lors de l'apprentissage. Le terme « Incrémental » est souvent associé aux termes « dynamique, adaptatif, interactif, on-line, batch…». Dans un rapport sur la classification incrémentale, Memmi [Memm01] utilise l'expression Algorithme Incrémental (ou Dynamique) « quand la nouvelle donnée peut être traitée sans avoir à reconsidérer l'ensemble des données antécédentes. L'algorithme sera complètement incrémental s'il met à jour les résultats antérieurs si nécessaire ». Et il ajoute une remarque importante pour la suite : « Les données doivent être sous forme de vecteurs dans un même espace vectoriel. Si les dimensions de l'espace changent, le problème est plus compliqué ». De façon plus générale on peut définir un algorithme incrémental par les quatre points suivants [Mfou98][Lelu06][IC10f] :

1. Possibilité de l'appliquer sans connaître au préalable tous les objets à classifier.
2. La classification d'un nouvel objet doit être exécutée sans faire usage intensif des objets déjà classifiés.
3. Résultat disponible après insertion de tout nouvel objet.
4. Adaptation aux changements potentiels de l'espace de description des données.

La littérature prenant en compte l'aspect chronologique dans les flux d'information se focalise généralement sur le "DataStream" dont l'idée principale est le traitement à la volée et en-ligne de

données non préalablement stockées. Les applications aux textes (bases de données bibliographiques, journaux en ligne…) sont encore balbutiantes. Les recherches sur le "DataStream" on été initiées, entre autres, en 1996 par le DARPA[5] à travers le projet TDT[6] [Alla98][Wayn98]. Mais les algorithmes issus de ces travaux sont avant tout destinés à traiter de manière globale de très gros volumes de données (dépêches d'agences…) et ne sont pas optimaux pour les tâches nécessitant une analyse incrémentale plus précise, comme la détection de thèmes émergents et le suivi de l'évolution d'un domaine.

Parmi les méthodes de classification non supervisée les plus utilisées, les méthodes procédant par agrégation autour de centres mobiles, comme les K-means [MacQ67] et leurs nombreuses variantes, font partie d'une famille basée sur l'optimisation d'un indicateur de qualité global de la partition. Ce problème d'optimisation étant NP-difficile, on ne sait que les faire converger vers un optimum local qui dépend de leur initialisation (par ex. positions initiales des centres choisies arbitrairement, ou en fonction des données), voire de l'ordre des données. Un bon nombre de variantes incrémentales de ces méthodes ont été proposées [Bins02][Chen03][Lin04]. Beaucoup sont issues de l'action DARPA TDT, comme [Gaud05][Gabe05].

Parmi les méthodes hiérarchiques, certaines ont été adaptées à l'incrémentalité, comme JERARTOP [Pons04], COBWEB [Fish87] ou CURE [Guha98]. Ces méthodes, divisives ou agglomératives, le plus souvent conviviales et efficaces, sont indépendantes de l'ordre de présentation des données et des conditions initiales. Cependant, elles manquent de robustesse quant aux perturbations mineures des valeurs de similarité entre vecteurs-données, et au regard de la qualité des partitions obtenues à un niveau donné de l'arbre de classe généré. Un consensus existe ainsi pour leur préférer les méthodes à centres mobiles [Leba82].

Certains auteurs ont exploré la piste des algorithmes biomimétiques. Nous mettons dans cette catégorie les algorithmes évolutionnaires (algorithmes génétiques, stratégies d'évolution…), les essaims intelligents (Swarm intelligence : nuages d'insectes volants, fourmis artificielles…) et les systèmes immunitaires [Lee00][Monm01]. Azzag souligne qu'«il existe très peu de méthodes biomimétiques incrémentales, bien que l'incrémentalité soit possible dans de nombreux algorithmes. Un grand avantage de certaines de ces méthodes (fourmis, essaims, SI) est leur capacité à donner une classification et également une visualisation de cette classification » [Azza04]. Jusqu'à présent, les automates cellulaires ont été peu utilisés pour la classification. [Azza04][Azza05] proposent un algorithme les exploitant dans un tel contexte. Le but est d'obtenir un résultat" visuel" permettant une exploration directe de la classification et éventuellement de visualiser d'autres informations complémentaires (images…). Les fourmis artificielles, s'inspirent du comportement collectif des fourmis vivantes pour résoudre des problèmes de classification (rassembler des « objets » en tas, reconnaissance « olfactive »). Cette méthode permet de déterminer automatiquement des classes sans connaissance a priori sur le nombre de classes, sans partition initiale et sans paramétrages complexes. Les fourmis se déplacent sur une grille à deux dimensions et peuvent ainsi ramasser ou déposer des objets afin de partitionner les données [Azza05][Lave07].

---

5 Defense Advanced research Projects Agency ; www.darpa.mil

6 (Topic Detection & Tracking ; http://projects.ldc.upenn.edu/TDT)

Nous ne pouvons pas terminer ce rapide panorama sans parler des méthodes par densité. Celles-ci définissent une classe en s'appuyant sur la notion de densité d'un nuage de points. Cette notion caractérise le voisinage d'un point, voisinage caractérisé par un seuil de distance ou un nombre de plus proches voisins ; le paysage de densité qui en découle est unique et parfaitement défini.

Trémolières [Trém79] a proposé un algorithme général, dit de percolation, indépendant de la définition de la densité et du type de données, pour délimiter rigoureusement les noyaux, les points-frontière ambivalents et les points atypiques. Il procède par baisse progressive du niveau de densité depuis le point le plus dense, et diffusion autour des noyaux qui apparaissent successivement. D'autres travaux retrouvent le même principe de repérage des noyaux denses, le plus souvent avec une définition spécifique de la densité, et d'extensions de diverses sortes à partir des noyaux [Mood01][Bata02][Hade03][Guen04]. A noter que ces méthodes peuvent se traduire en termes de partitionnement de graphe, car définir une densité implique d'avoir fixé des relations de voisinage, donc un graphe. DBSCAN (Density Based Spatial Classification of Applications with Noise), décrit dans [Este96], utilise une définition propre de la densité au moyen de deux paramètres, dont l'un fixe le seuil à partir duquel les noyaux sont constitués et étendus. Pour mémoire, on citera également CHAMELEON [Kary99], OPTICS [Anke99] et DENCLUE [Hine98]. Ces algorithmes n'ont cependant pas été envisagés, à notre connaissance, dans une perspective dynamique d'incrémentalité.

Dans [Cuxa09], un nouvel algorithme incrémental fondé sur la densité est également proposé et testé. Cependant, celui-ci travaillant à partir d'un graphe de documents, la méthode même d'obtention d'un tel graphe reste un problème majeur. Il en va de même pour la pondération des arêtes du graphe : quelle valeur adopter ? Différents pistes on été étudiées (graphe des K-plus proches voisins, graphe de cooccurrences de mots, graphe des liens statistiquement valides…) sans pourtant obtenir de résultats probants à ce jour (cf. **section D.6.5**).

Les méthodes de classification non supervisées neuronales présentent quant à elles la particularité de tenir compte de relations de voisinages prédéfinies (topologie fixe), comme les cartes SOM [Koho01], ou de relations de voisinage dynamique (topologie libre), comme les réseaux de gaz de neurones, fixes, comme « Neural Gas » (NG) [Mart91], ou évolutifs, comme « Growing Neural Gas » (GNG) [Friz95]. Cette stratégie les rend moins sensibles aux conditions initiales, ce qui représente un atout important dans le cadre d'une utilisation incrémentale. Les réseaux à topologie libre permettent, de plus, de mieux s'adapter aux caractéristiques de chaque distribution de données, notamment s'ils sont évolutifs (cf. **section D.6.5**).

Notre expérience d'exploitation de cette dernière famille de méthodes, combinée à celle de l'ensemble des techniques spécifiques que nous avons proposées jusqu'alors pour l'optimisation de la classification non supervisée, pour l'analyse de données multi-vues, pour la détection de nouveauté, et pour l'analyse locale et la synthèse des résultats de la classification non supervisée [IJ07b] [IC08a] [IC08b], nous laissent à penser que nous disposons d'élément-clés pour apporter des solutions efficaces et stables au problème de la classification incrémentale. Nous choisirons donc d'orienter principalement nos travaux de recherche pour les 2 années à venir sur ce thème de recherche, associé par ailleurs à des domaines d'application extrêmement larges et extrêmement variés.

Dans ce cadre, nous venons récemment de monter un projet de collaboration CPER, avec plusieurs équipes de recherche régionales, appartenant au LORIA et à l'INIST. Cette

collaboration qui se met actuellement en place a, de notre côté, pour but d'explorer plusieurs voies complémentaires :

1. Une analyse à grande échelle des critères d'évaluation de la qualité des classifications numériques. Cette analyse inclue les méthodes basées sur l'exploitation des liens entre les classes que nous n'avions pas traitées jusqu'alors. Elle a principalement pour rôle de déterminer l'ensemble des critères qui peuvent être exploités en ligne lors du processus de classification incrémentale.

2. L'étude de l'adaptation d'algorithmes neuronaux existants, de la famille des gaz de neurones [Mart91][Fritz95][Prud04] à un fonctionnement incrémental. Nous chercherons, pour cela, à définir de nouvelles règles d'apprentissage locales qui remplacent les règles d'apprentissage globales usuellement employées dans ces méthodes. Ce travail rentrera directement en synergie avec celui présenté précédemment et devrait sans doute nous amener à proposer de nouvelles méthodes de classification incrémentale. Une première initiative en ce sens est présentée à la **section D.6.5**.

3. Une dernière voie originale que nous explorerons est celle qui consiste à ramener la comparaison entre données temporelles présentes à différents pas de temps à un problème de comparaison entre des vues constituées par des analyses de données associées à des macro-étapes de temps prédéfinies. Cette approche se base sur une adaptation du paradigme MVDA au raisonnement sur des groupes de données différenciées [IC09a]. Elle est en cours de mise au point (cf. section **D.7**).

L'environnement de test prioritaire pour ce travail sera celui du CPER, à savoir les bases bibliographiques PASCAL et FRANCIS. Nous comptons cependant privilégier d'autres expérimentations à large échelle de nos approches dans les domaines spécifiques de la Webométrie/Scientometrie et de la bio-informatique.

La visualisation des résultats de la classification incrémentale représente également une étape déterminante pour la compréhension des analyses correspondantes. Sans cette étape, des tableaux de nombres et de mots sont les seules sorties que l'expert peut exploiter, avec toutes les difficultés que l'on imagine. Ces dernières années, les progrès techniques ont permis l'émergence de nouvelles méthodes de représentation.

L'approche ThemeRiver que proposent Havre et al. [Havr02], permet de visualiser des variations d'effectifs. Les thématiques sont construites à partir d'occurrences de termes. Si cette méthode permet de bien mettre en évidence les relations d'effectif en fonction du temps, cette représentation a l'inconvénient de ne révéler aucune structure : rien ne permet de voir des relations entre les données. Erten et al. [Erte03] proposent de visualiser l'évolution des thématiques grâce au système TGRIP. Celui-ci illustre l'évolution des effectifs des thématiques sous forme de graphe. La taille de chaque sommet évolue en fonction du nombre de documents que contient la thématique représentée par ce sommet. Cette méthode permet de mettre en évidence l'existence d'une structure entre les thématiques, l'évolution étant suggérée par la superposition de niveaux. L'approche proposée par le système CiteSpace [Chen06] permet la représentation de l'évolution de réseaux de citations. Chen emploie deux dimensions temporelles

différentes : la date de publication et la date de citation. La date de publication détermine la place des articles (les nœuds du graphe) le long d'un axe temporel. La seconde dimension est celle qui correspond à l'année de citation : chaque nœud est caractérisé par différentes strates de couleurs qui représentent l'année ou l'article correspondant a été cité. L'ensemble, permet de distinguer les domaines de recherche actuels et les travaux plus anciens. Cette approche fondée sur les citations révèle une dynamique de construction de réseaux d'articles sur des thématiques proches. Nous voyons ainsi que différents auteurs ont proposé ou proposent des modes de visualisation qui sont peut-être adaptables à notre problématique, mais certainement pas de façon simple. A propos de l'évolution d'un résultat de classification incrémentale, Humbert se pose la question : « Quelles sont les informations utiles à l'utilisateur pour pouvoir analyser cette évolution ? Comment les représenter graphiquement à cet utilisateur de manière efficace et efficiente ? » [Hum06]. A partir de la définition de différents types d'utilisateurs ayant chacun des besoins bien spécifiques il propose ainsi de « découper la représentation en quatre approches : globale, quantitative, qualitative, et par la requête, ainsi qu'en une fonction d'alerte. ».

Comme nous le voyons, la visualisation des résultats de classification incrémentale reste un champ d'investigation important, et il est probable qu'après avoir exploré diverses pistes, la solution idéale ne réside pas dans un seul type de visualisation, mais plutôt dans une combinaison d'approches, comme nous l'avons déjà proposé auparavant pour la visualisation d'information sujettes à des relations complexes, comme les réseaux sociaux (cf. section D.5 et section D.11).

E.2. Webométrie et scientométrie

Les expériences que nous avons menées sur les données issues du Web, sur les publications scientifiques et sur les fichiers de citations, nous ont permis de montrer qu'il était possible de pallier aux carences des méthodes webométriques et scientométriques classiques. Dans ces expériences, nous avons proposé plusieurs approches originales et novatrices. Nous avons notamment expérimenté la combinaison d'analyses de contenu et d'analyses de liens dans un contexte multi-vues, l'exploitation de méthodes de visualisation et de synthèses avancées, et celle de techniques d'évaluation et de validation des résultats indépendantes des méthodes d'analyse utilisées. Ces expériences nous ont globalement permis de démontrer les bénéfices de l'utilisation des techniques avancées d'analyse de données et d'intelligence artificielle dans le cadre de la webométrie et de la scientométrie (cf. **section D.11**). Nous pensons cependant qu'une des valeurs ajoutées les plus importantes de la démarche que nous avons suivie est qu'elle nous a permis de faire des propositions concrètes et efficaces pour instrumentaliser, à grande échelle, les analyses webométriques et scientométriques.

Nos expériences nous ont également permis d'acquérir une visibilité internationale dans ce domaine sous des formes variées : invitation au comité éditorial d'un journal international, publications régulières dans des journaux internationaux spécialisés du domaine, réalisation sur demande de formations spécialisées sur les techniques avancées d'analyse de données dans les départements d'évaluation des sciences des universités étrangères, nombreuses propositions de séminaires invités. Elles nous ont notamment permis de développer des partenariats privilégiés avec des instituts importants dans le domaine de l'évaluation des sciences, comme l'Université Catholique de Louvain (KU Leuven), le NISTAD (**N**ational **I**nstitue of **S**cience **T**echnology **A**nd **D**evelopment **S**tudies) en Inde, ou des partenaires émergents, comme le WISELAB (**W**ebometrics, **I**nformetrics, **S**cientometrics and **E**conometrics **Lab**) en Chine. Ces partenariats

devraient nous aider continuer à défendre à long terme le bien-fondé de nos approches dans ce domaine, mais également à développer des collaborations plus approfondies dans le cadre de projets internationaux.

Dans ce dernier cadre, nous sommes en cours de montage du projet COLDLib. Ce projet vise à construire un espace numérique fédéré pour le réseau interdisciplinaire de recherches "Collaboration en Science et en Technologie" (COLLNET), afin de rassembler la connaissance existante relative au domaine des aspects quantitatifs de la science, et de favoriser ainsi la collaboration et la communication des chercheurs des pays émergents en analyse de la science et de la technologie et en politique de la science, par la combinaison et l'intégration d'approches qualitatives et quantitatives. Le but principal du projet est de fournir à ces chercheurs un accès global, efficace et fiable aux données, aux résultats, ainsi qu'aux sources de connaissances à valeur ajoutée de ce domaine de recherche, plutôt que de leur laisser conserver une vue isolée et locale sur les ces informations. La proposition de projet est actuellement en cours d'achèvement et de validation au niveau européen.

Nous nous sommes plus récemment intéressés aux modèles concurrents du modèle NOMAD-MULTISOM utilisables dans le cadre de la webométrie, tel que le modèle de représentation de réseaux BibtechMon, ce qui nous a permis de participer à une analyse comparative entre ces deux modèles menée à bien par l'INIST. Les résultats de cette analyse ont été présentés dans journal international [IJ06b]. Elle a de nouveau fait ressortir l'intérêt de l'exploitation des points de vue multiples pour l'analyse de données complexes, comme des graphes d'interaction, ou des réseaux sociaux, celle-ci permettant de générer des explications précises sur les interactions locales apparaissant dans un réseau construit à partir de telles données.

Grâce aux résultats obtenus dans le cadre de l'étude que nous menons sur la classification incrémentale, nous comptons également aborder toute une classe de problèmes liés à l'analyse des données temporelles dans le cadre spécifique de l'évaluation des sciences. Cette démarche aura certainement des retombées très larges étant donné qu'elle est susceptible de couvrir l'ensemble des études liées à la dynamique des laboratoires et des thèmes de recherche.

E.3. Analyse des données bioinformatiques

L'analyse des résultats d'expérimentation produits par les puces ADN en vue de la détermination des relations entre les gènes et les pathologies implique le croisement d'information provenant de nombreuses sources différentes (textuelles, ontologies, numériques, …). Les sources textuelles doivent être analysées pour en extraire l'information permettant d'établir des liens de causalité gènes-pathologies rapportées par les expériences des biologistes. De leur côté, les ontologies spécialisées, comme la **G**ene **O**ntology, permettent d'envisager de classifier les gènes selon différentes modalités rapportées à différents types d'annotations (fonctionnelles, processus chimiques...). Dans ce cas, les distances entre les gènes peuvent être matérialisées par des distances entre leurs annotations [Chia04]. Enfin, des classifications de gènes peuvent être établies en analysant la similarité de leur expression rapportée par les expériences basées sur les puces ADN [Chia06]. Le projet de recherche AREX que nous avons défini en partenariat avec le laboratoire **IIR** (**I**ntelligent **I**nformation **R**etrieval) de Taiwan (attaché au **N**ational **S**cience **C**ouncil de Taiwan) et le NIEHS (**N**ational **I**nstitute for **E**nvironmental **H**ealth) américain,

consiste à mettre en place un cadre fédérateur permettant de croiser les différents types d'analyse obtenus, de manière à produire des annotations intelligentes des gènes, contextuelles aux situations pathologiques et aux besoins d'information des analystes. Il fait plus particulièrement intervenir le modèle MVDA que nous avons développé pour le traitement des sources multiples, les analyseurs syntaxico-sémantiques mis au point par le laboratoire **IIR** pour celui des données textuelles impliquées dans l'analyse, et les données issues du traitement des puces ADN, normalisées par le **NIEHS**.

Dans ce cadre, nous veillerons naturellement à ré-exploiter les résultats obtenus dans celui de notre étude sur la classification incrémentale. Ces travaux pourraient en effet s'avérer extrêmement précieux pour analyser l'évolution du comportement des gènes au cours de temps, en situation pathologique.

E.4. Classification supervisée

Nous allons assurer la direction scientifique, dans le cadre du projet QUAERO, de la tâche de gestion intelligente des données brevets. Le but de cette tâche est de mettre en place une plate-forme permettant d'assister de manière efficace les ingénieurs experts en brevets dans le processus de contrôle de validation des dépôts. Elle repose donc principalement sur l'identification des relations pertinentes entre les classifications hiérarchiques de brevets et les références bibliographiques matérialisant les travaux de recherche se rapportant, directement ou indirectement, à ces différentes classes.

Les challenges associés à ce travail sont particulièrement nombreux. Les méthodes de classification numériques qu'il sera nécessaire d'exploiter devront en effet prendre en compte les contraintes inhérentes au traitement des données textuelles de gros volume comme la multidimensionnalité, ou la gestion de données déséquilibrées. Elles devront également permettre de traiter efficacement les interactions avec des approches symboliques, liée à l'exploitation nécessaire de hiérarchies de concepts, ou d'ontologies.

Pour relever ces challenges, nous comptons sur notre expérience antérieure des méthodes de classification supervisées, et surtout sur l'efficacité des nouvelles méthodes que nous avons proposées dans ce cadre, comme la méthode ILoNDF (cf. **section D.8**). Nous comptons également sur notre expérience approfondie des interactions numérique-symbolique. Nous explorerons en parallèle les méthodes se situant aux frontières de l'état de l'art dans le domaine de la sélection de variables et de l'équilibrage des résultats de classification, telles que celles basées sur les forêts aléatoires [Brei01], ou sur les forêts de rotations [Rodr06], de manière à en étudier la combinaison avec nos propres méthodes. Finalement, dans ce même contexte, nous envisageons également d'expérimenter des techniques de mesures vectorielles prenant en compte les relations hiérarchiques entre concepts, telles que celles que nous avons proposées dans notre propre thèse [PH95a].

E.5. Autres perspectives

Après tout cela, il restera encore de nombreux thèmes à aborder. Pour n'en citer qu'un, nous mentionnerons celui de la conception de mécanismes d'extraction des connaissances basés sur la structure latente des distributions de données. Ce thème représenterait une prolongation logique de travaux que nous avons démarrés autour de la synthèse de résultats de clustering (cf. **section D.6.4**).

Le traitement automatique des langues représente également un des cadres de travail que nous commençons à aborder dans le contexte de notre nouvelle équipe de recherche. Dans ce contexte, nous avons pu noter que les approches que nous proposons présentent de nombreuses possibilités d'application. Nous prenons ci-après deux exemples-types qui corroborent cette intuition. L'identification des classes verbales est une opération qui vise à isoler les caractéristiques (traits sémantiques) associées à différentes classes de verbes [Falk09]. Elle fait souvent usage de méthodes de classification non supervisées basées sur la prise en compte de plusieurs ressources différentes et nécessite donc de pouvoir exploiter des mécanismes de combinaison des ressources, analogues aux mécanismes de combinaison entre les vues que nous avons mis en place dans le paradigme MVDA. Le traitement automatique des langues implique également souvent la gestion de structures complexes, comme les graphes de synonymie (également appelés grand graphes « petits monde »), ou les espaces sémantiques de mots. Ces structures présentent la particularité de devoir être appréhendées à la fois de manière locale et de manière globale par les linguistes. Nous pensons que des modes de représentation spécifiques, tels que celui des hyper-arbres combiné aux modèles d'équilibre de forces que nous avons également mis en place dans notre approche, devraient parfaitement répondre à ces contraintes. Nous pensons de toutes les façons que le champ reste très ouvert pour l'exploitation de notre nouveau paradigme à plus large échelle dans ce domaine.

F. SYNTHESE

Notre travail de recherche nous a globalement permis de mettre en place un nouveau paradigme d'analyse de données orientée par les points de vue multiples. A la Figure, nous présentons donc une vue plus générale sur ce paradigme. Nous rappelons que celui-ci couvre plusieurs 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 catégorisation, de filtrage et de détection de nouveauté.

L'ensemble des opérations proposées peut se dérouler de manière interactive, selon un mode d'interaction systémique qui exploite des méthodes de visualisation spécifiques. Toutes ces opérations peuvent également se dérouler en arrière-plan.

F.1. Analyse de données

L'analyse de données représente la fonction centrale du paradigme MVDA. Cette analyse comprend les phases suivantes:

- **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.

- **Prétraitement** : Les descriptions des données attachées aux différents points de vue étant de nature vectorielle, cela permet d'envisager de nombreux mécanismes de pondération différents avant la phase d'analyse.

- **Classification numérique non supervisée** : Nous avons privilégié dans notre approche l'exploitation de méthodes de classification neuronales non supervisées (SOM, NG ou GNG), ce qui n'exclue nullement l'utilisation de méthodes non neuronales ad-hoc.

- **Évaluation globale des classifications** : Ce type d'évaluation pourrait s'opérer par de nombreux critères différents. Le paradigme MVDA exploite ses propres critères qui présentent tous la caractéristique d'être indépendants de la méthode de classification utilisée.

- **Modèle multi-topographique étendu** : Pour représenter les analyses associées aux différents points de vue, le modèle multi-topographique permet de mélanger des modèles à topographie fixe, comme les cartes SOM, avec des modèles à topologie libre, comme les gaz de neurones. Les trois mécanismes de base inhérents au modèle multi-topographique étendu peuvent ensuite être exploités sur l'ensemble des groupes de modèles associés aux vues :

  1. Communication inter-topographies (analyse multi-vues) : Les modèles associés aux différents points de vue sont liés dynamiquement par un mécanisme de communication inter-vues, lui même fondé sur un modèle de réseau d'inférence bayésien. Ce mécanisme permet d'opérer dynamiquement des déductions ciblées impliquant plusieurs analyses locales. Il permet également de restituer, si nécessaire, une vue générale précise de l'analyse.

2. **Comparaison des vues** : Elle est basée sur une méthode d'évaluation de la cohérence de propagation entre les vues. Ce mécanisme permet d'opérer dynamiquement des comparaisons globales entre des analyses différentes.

3. **Généralisation (analyse multi-niveaux)** : Un mécanisme de généralisation en cascade, et en ligne, 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.

- **Évaluation locale des classes** : Des évaluations locales peuvent être opérées indifféremment sur les classes d'un modèle optimal, ou sur celles de ses généralisations. Ces évaluations permettront d'identifier les propriétés propres aux classes concernées. Elles serviront ensuite de base aux opérations de fouille de données.

F.2. Fouille de données

Cette phase représente une phase d'exploitation précise des résultats produits par la phase d'analyse. Elle inclut les opérations d'étiquetage, ainsi que les opérations d'extraction de règles d'association.

F.3. 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:

- **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 SOM associées à plusieurs points de vue différents. Cette méthode permet également l'interprétation précise des résultats obtenus par le mécanisme de comparaison des vues.

- **Visualisation hyperbolique (hyper-arbre de classes)** : Cette méthode combine un algorithme original de classification hiérarchique fondé sur la densité, nommé DBHC, avec un mécanisme de visualisation par arbre hyperbolique. Dans notre cas, elle est spécifiquement exploitée pour la visualisation des résultats des méthodes produisant des nuages de classes représentées dans un espace multidimensionnel.

F.4. Détection de nouveauté, filtrage et apprentissage supervisé

Ces opérations sont susceptibles de s'appliquer à de nouvelles données entrantes dans le système. Elle permettent, après apprentissage à partir des données ou des classes existantes, de positionner ces nouvelles données relativement aux données déjà analysées, de valider les résultats obtenus, mais également de répondre à des besoins spécifiques en termes d'analyse du comportement de l'analyste et de son besoin d'information.

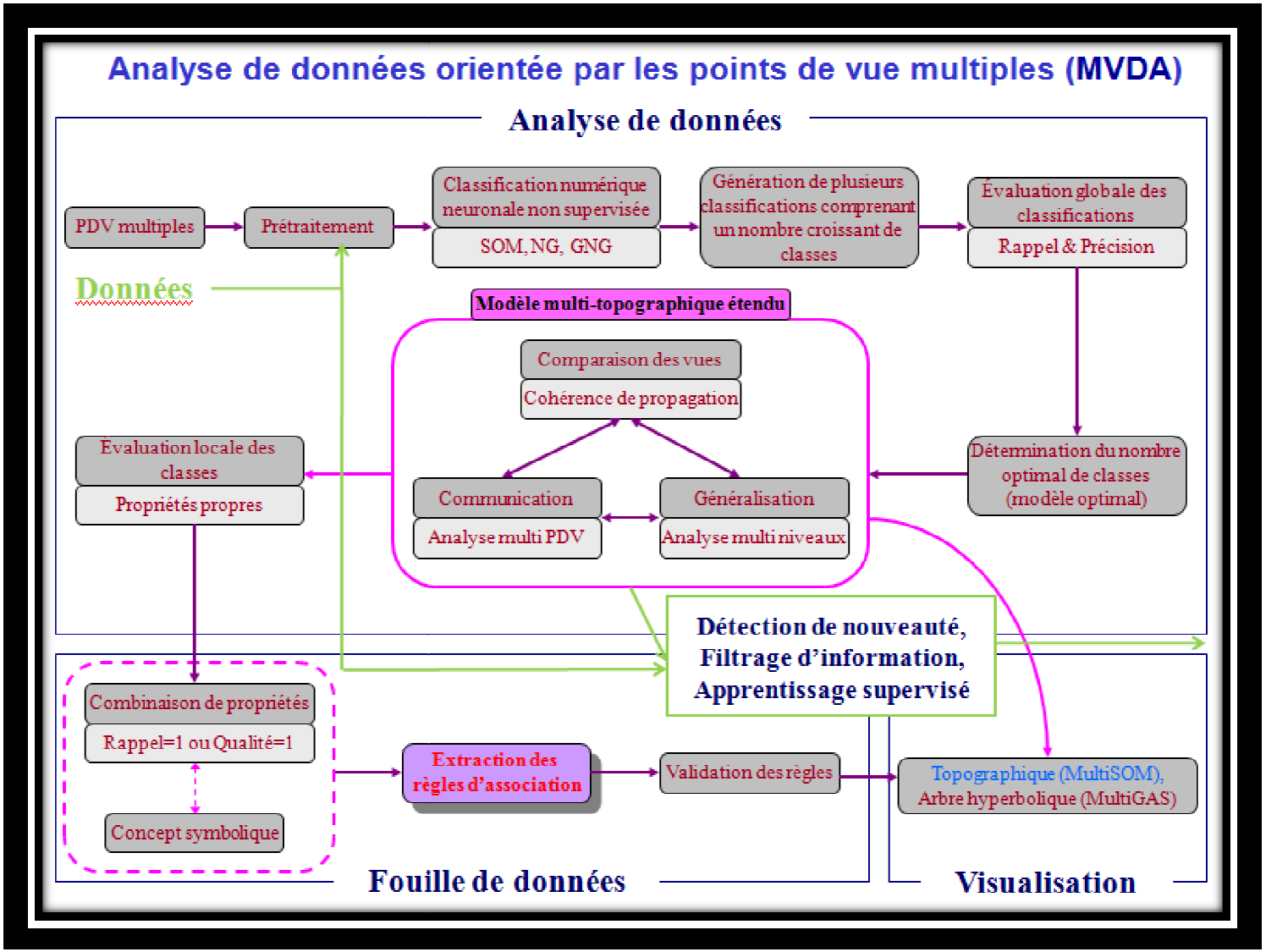

Figure 42 *: Le paradigme MVDA (vue générale).*

G. CV ETENDU
ET REFERENCES BIBLIOGRAPHIQUES PERSONNELLES