Analyse et extraction de connaissances des bases de données spatio-temporelles

Karine Zeitouni

Université de Versailles Saint-Quentin-en-Yvelines

Habilitation à Diriger des Recherches
Spécialité Informatique

Analyse et extraction de connaissances des bases de données spatiotemporelles

Mémoire présenté par

Karine Zeitouni

Le 1er Décembre 2006

Devant le jury

Rapporteurs	Jiawei Han	Professeur - Université de l'Illinois
	Robert Laurini	Professeur - INSA de Lyon
	Philippe Rigaux	Professeur - Université de Paris Dauphine
Examinateurs	Georges Gardarin	Professeur - Université de Versailles-Saint-Quentin
	Philippe Pucheral	Professeur - Université de Versailles-Saint-Quentin
	Michel Scholl	Professeur - CNAM Paris
	Stefano Spaccapietra	Professeur - Ecole Polytechnique Fédérale de Lausanne

A Paul et Maria

A la mémoire de Lara

Remerciements

Je tiens en premier lieu à exprimer toute ma gratitude au Professeur Georges Gardarin pour sa confiance et pour l'autonomie qu'il m'a accordée durant toutes ces années passées au PRISM. Je le remercie aussi pour sa compétence exemplaire dans son métier, qui nous trace le chemin. J'ai toujours été fière d'appartenir à l'équipe qu'il a monté au sein du laboratoire qu'il a crée.

J'ai eu la chance de réunir dans mon jury des spécialistes de mon domaine parmi les plus reconnus sur le plan national et international. Aussi, je tiens à les remercier tout particulièrement pour m'avoir fait l'honneur de leur participation. Merci au Professeur Michel Scholl pour avoir présidé mon Jury, aux Professeurs Jiawei Han, Robert Laurini et Philippe Rigaux pour avoir rapporté sur ce mémoire et aux Professeurs Georges Gardarin, Philippe Pucheral et Stefano Spaccapietra pour leur enthousiasme à participer à ce jury.

Ce mémoire est le fruit de longues années de travail d'équipe et de collaborations scientifiques et industrielles. A ce titre, je témoigne du mérite des doctorants que j'ai encadrés dans ce résultat, nommément Thomas Delpy, Huaizhong Kou, Nadjim Chelghoum, Lionel Savary et Tao Wan. Je leur souhaite un avenir plein de succès. J'ai eu la chance d'avoir de multiples collaborations scientifiques et industrielles. Je remercie vivement les Professeurs et chercheurs français et étrangers, ainsi que les partenaires industriels avec qui j'ai eu le plaisir de travailler.

Au sein du laboratoire PRISM, je n'oublie pas de citer les collègues qui m'ont beaucoup apportée, que ce soit par leur conseil, leur aide ou tout simplement par leur amitié. Je tiens à remercier en premier lieu Mokrane Bouzeghoub pour ses précieux conseils ; Philippe Pucheral pour la lecture avisée de ce mémoire et pour son aide dans tout ce projet ; tous les collègues de l'équipe Systèmes de Bases de données ainsi que ceux de l'équipe SIAL pour leur volontariat. Je ne puis oublier nos secrétaires Annick Baffert et Chantal Ducoin pour leur amitié d'abord et pour leur dévouement. Je remercie collectivement les collègues qui m'ont soutenue, que ce soit au PRISM ou à l'IUT de Vélizy.

Enfin, sans le soutien et l'aide de mon mari, il aurait été difficile de mener à bien ce travail. Merci à Paul pour avoir assuré sur le plan familial lorsqu'il fallait que je prolonge une journée de travail ou que j'assure une mission loin du foyer. Merci à Maria d'être simplement ce qu'elle est, notre source de bonheur.

Résumé

Ces dernières années ont vu une croissance phénoménale dans la production et la diffusion des données spatiales de sources aussi variées qu'hétérogènes. Cela a généré des besoins d'intégration dans des entrepôts de données et des perspectives d'analyse exploratoire et de fouille de données spatiales et spatiotemporelles. Nos travaux se placent dans ce contexte visant l'analyse et l'extraction des connaissances depuis les bases de données spatiotemporelles. Ils traitent différents aspects allant de la modélisation avancée des données spatiales, à la fouille de ces données en passant par leur intégration dans un entrepôt, l'optimisation des requêtes et l'analyse en ligne. Ainsi, nous décrivons nos approches pour la modélisation 3D, puis pour la modélisation spatiotemporelle d'objets mobiles. Ensuite, l'intégration de données spatiales est traitées selon deux aspects : l'intégration de formats et l'intégration de données par l'appariement géométrique. Une architecture d'entrepôt de données spatiales basée sur les standards XML et GML est proposée, puis dotée d'une technique d'optimisation de requêtes spatiales basée sur un cache sémantique. L'exploration des données spatiotemporelles a donné lieu à des solutions originales extension de l'OLAP. Enfin, différentes approches sont proposées pour la fouille de données spatiales. Nous avons ouvert le spectre de nos recherches à la fouille d'autres données complexes, telles que les données séquentielles et textuelles. Ces travaux ainsi que les développements futurs sont exposés dans ce mémoire.

Table des matières

Préambule ____ *3*

Chapitre I. Introduction ____ *5*

I.1. Problématique de recherche et contributions ____ **5**
I.2. Organisation de ce mémoire ____ **10**

Chapitre II. Modélisation avancée des données spatiales ____ *11*

II.1. Problématique et limites de l'état de l'art ____ **11**
II.2. Contributions ____ **13**
II.3. Conclusion ____ **17**
II.4. Références ____ **18**

Chapitre III. Entrepôt de données spatiotemporelles ____ *21*

III.1. Problématique ____ **21**
III.2. Etat de l'art ____ **22**
III.3. Contributions ____ **27**
III.4. Conclusion ____ **37**
III.5. Références ____ **38**

Chapitre IV. Analyse en ligne des données spatiotemporelles ____ *41*

IV.1. Problématique ____ **41**
IV.2. Etat de l'art ____ **42**
IV.3. Contributions ____ **43**
IV.4. Conclusion ____ **48**
IV.5. Références ____ **48**

Chapitre V. Fouille de données spatiales ____ *51*

V.1. Spécificités du problème et approches ____ **51**
V.2. Etat de l'art ____ **54**
V.3. Contributions ____ **59**
V.4. Conclusion ____ **71**

V.5. Références 73

Chapitre VI. Fouille d'autres données complexes *79*

VI.1. Fouille de données séquentielles 80
VI.2. Fouille de données textuelles 85
VI.3. Conclusion 89
VI.4. Références 90

Chapitre VII. Conclusion et perspectives *93*

VII.1. Bilan 93
VII.2. Perspectives 94

Annexe - Bases de données spatiales *97*

VII.3. Notions de base 97
VII.4. L'évolution vers les SGBD spatiaux 99
VII.5. Problèmes soulevés par les SIG aux bases de données 100
VII.6. Références 101

PREAMBULE

Ce mémoire présente mes principaux travaux de recherche tout au long de ma carrière de maître de conférences. J'ai effectué ces travaux au sein du thème « Systèmes de Bases de Données » du laboratoire PRISM.

Dans la continuité de ma thèse, mes travaux ont porté entre 1992 et 1995 sur la modélisation des données spatiales 3D, d'abord dans le contexte d'un projet sur la gestion des données géographique 3D de l'axe A2 du PSIG (Programme sur les Sciences de l'Information Géographique soutenu par le CNRS et l'IGN), puis lors de la thèse CIFRE de Thomas Delpy en collaboration avec l'EDF. Ensuite, j'ai débuté mes travaux sur l'intégration des données spatiales en abordant l'aspect sur l'intégration de modèles. A partir de 1998, j'ai orienté mais travaux vers la fouille de données spatiales. Ces travaux ont été développés d'abord dans le cadre de projets PSIG, puis lors de collaborations industrielles et ont abouti à la thèse de Nadjim Chelghoum soutenue fin 2004. Pendant cette même période, nous avons constitué en 2002 l'équipe "Intégration et fouille de données complexes" au sein du thème SBD. J'ai alors participé, au côté du Professeur Georges Gardarin, à un projet RNTL sur la fouille de données textuelles concrétisé par la thèse de Huaizhong Kou. Depuis 2003, j'ai développé la thématique des entrepôts de données spatiotemporelles. Ces derniers travaux avaient comme cadre le projet Européen HEARTS. Ils ont abouti à deux thèses dont celle de Lionel Savary soutenue fin 2005 et celle de Tao Wan dont la soutenance est prévue fin 2006.

Les thématiques développées sont souvent liées et certains aspects comme la modélisation de données ou la fouille de données se retrouvent dans des recherches sur les entrepôts de données. Dans ce mémoire, je présente mes principaux travaux dans un ordre thématique plutôt que chronologique.

Une grande partie de ces travaux ayant porté sur l'analyse et l'extraction des connaissances depuis des bases de données spatiotemporelles, j'ai intitulé ce mémoire ainsi.

CHAPITRE I. INTRODUCTION

De nombreux domaines d'applications utilisent des représentations spatiales. C'est le cas de la géographie, de l'environnement, du transport, de la santé ou de l'architecture. Ces dernières années ont vu une croissance phénoménale dans la production et la diffusion des données spatiales. Ceci a généré divers problèmes pour représenter, manipuler et interroger de manière fiable et efficace ces données. Il y a fort longtemps que des systèmes ad hoc dits d'information géographique (SIG) ont été développés à cet effet. Mais ils sont limités à certains types d'applications géographiques et n'offrent pas les fonctionnalités essentielles des bases de données telles que la séparation des niveaux logique et physique, l'intégrité et la fiabilité des données, la concurrence d'accès et le passage à l'échelle. Ceci a amené au développement de la recherche en bases de données spatiales. Aujourd'hui, les systèmes de gestion de bases de données (SGBD), comme Oracle, DB2 ou Postgres, ont intégré la gestion des données spatiales et prennent le pas sur les SIG traditionnels, du moins en tant que serveurs de données. De plus, l'émergence de standards, à l'initiative du consortium Open GIS regroupant des industriels et des chercheurs, puis de l'ISO, a eu un impact décisif sur le développement des produits et sur leur interopérabilité.

Malgré ces avancées, les recherches se poursuivent encore aujourd'hui sur de nombreux sujets en bases de données spatiotemporelles. La principale raison est le caractère spécifique de la spatialité à prendre en compte à tous les niveaux de la gestion et de l'exploitation des données.

I.1. Problématique de recherche et contributions

Une grande partie de nos travaux se situe dans le domaine des bases de données spatiales. Récemment, nous avons élargi nos recherches à d'autres bases de données complexes.

Nous donnons en annexe un aperçu du domaine des bases de données spatiales et soulignons le lien entre les principaux sujets de recherche et les fonctions (les 5A) des systèmes d'information géographiques. Le lecteur peut se référer à cette annexe pour les notions générales de ce domaine.

Ces recherches sont nombreuses, portent sur divers aspects et sont étalées dans le temps. Les sujets traditionnels comme la modélisation, l'indexation et le langage, s'ils sont aujourd'hui maîtrisés pour une large gamme d'applications SIG, restent d'actualité dans des problématiques spécifiques. Parmi les nouveaux sujets, certains concernent la nature des données comme la gestion de données 3D, spatiotemporelles ou en flux, d'autres concernent le type d'application comme les applications distribuées ou en pair à pair. Il s'agit principalement de la fonction d'Archivage des SIG (cf. Annexe).

Les problèmes d'intégration, de multi-représentation et d'interopérabilité sont au cœur des recherches d'aujourd'hui. L'émergence de standards a généré des travaux sur les architectures et notamment sur les architectures de médiation. La mouvance du web sémantique a développé un thème récent sur les ontologies spatiales. Ces travaux portent donc sur la fonction d'Assemblage des SIG.

Par ailleurs, l'analyse globale et l'extraction de connaissances comme la fouille de données, la recherche par similarité et la génération automatique de méta-données (descripteurs) sont des thèmes très actifs. Cela concerne la fonction d'Analyse des SIG

Enfin, l'interaction avec la télématique a amené de nouvelles recherches sur les réseaux de capteurs géographiques, les bases de données mobiles et les services localisés. Ces sujets impactent notamment l'Acquisition dans les SIG.

Nos travaux se situe dans les trois catégories que sont l'archivage, l'assemblage et l'analyse des données spatiales. Dans la première catégorie, nous avons étudié la modélisation des données spatiales, puis la modélisation des données spatiotemporelles. Concernant l'assemblage, nous nous sommes intéressés à la gestion de formats hétérogènes et à l'intégration dans une architecture d'entrepôt de données spatial. Pour ce qui est de l'analyse, nos contributions ont concerné d'un côté l'analyse en ligne de bases de données spatiotemporelles et de l'autre la fouille de données spatiales. Nous présentons ci-dessous un aperçu de ces travaux

I.1.1. Modélisation avancée des données spatiotemporelles

Du niveau conceptuel au niveau physique, la modélisation des données spatiales est toujours d'actualité. Ainsi, une lacune des SIG, des SGBD spatiaux et des normes, encore aujourd'hui, est le manque de support des données tridimensionnelles (cf. chapitre II). De même, la prise en compte de la variabilité temporelle est absente ou peu satisfaisante dans les systèmes existants. Ces lacunes sont probablement liées au fait qu'il était difficile d'acquérir à l'échelle géographique des données en 3D et que la problématique spatiotemporelle était souvent limitée à la gestion de versions de données historiques. Aujourd'hui, les progrès dans les moyens d'acquisition redynamisent les recherches dans ces domaines.

Contributions

Au milieu des années quatre-vingt-dix, la modélisation spatiale 2D dans les bases de données était bien avancée et les normes étaient en préparation. Toutefois, la troisième dimension n'était guère prise en

compte. C'est à cette époque que nous nous sommes intéressés à la modélisation 3D dans les bases de données et nous avons traité successivement deux aspects. Le premier a été l'intégration des données planimétriques et des données 3D pour la description géographique en milieu extérieur. Le second a concerné la modélisation topologique en milieu intérieur, où un modèle topologique 3D de bâtiments a été proposé au cours de la thèse de Thomas Delpy en 1995. Il a été appliqué à la navigation en Réalité Virtuelle dans le milieu hostile des centrales nucléaires.

Nous avons été amenés à étudier la modélisation spatiotemporelle lors de nos projets récents sur l'analyse de la mobilité urbaine. Les déplacements des individus forment des objets mobiles. Le manque de support des objets mobiles dans les systèmes existants nous a amenés à réaliser des solutions pour la modélisation et pour l'interrogation. La première idée est de décrire grossièrement les trajectoires et de les représenter de manière symbolique par des zones et des intervalles de temps existants. La seconde est une représentation précise, mais qui projette les trajectoires dans un espace réduit (2D) permettant son implémentation dans un SGBD spatial quelconque.

Aujourd'hui, les données spatiales sont passées de la phase de gestion au jour le jour à la phase d'exploitation décisionnelle. Nos travaux les plus récents s'inscrivent justement dans le domaine de l'aide à la décision. Ils visent la découverte de connaissances à partir des bases de données spatiales impliquant des techniques d'intégration de données dans des entrepôts, d'analyse en ligne et de fouille de données spatiales.

I.1.2. Intégration des données spatiotemporelles dans les entrepôts de données

Les applications décisionnelles des bases de données spatiales ont souvent recours à des sources hétérogènes et organisées de manière inadaptée à l'analyse. Les techniques d'entrepôts de données visent justement l'intégration de sources hétérogènes et la transformation dans un modèle approprié facilitant l'application des outils d'analyse et de fouille de données. Seulement, s'agissant d'un entrepôt de données spatiales, ces techniques doivent d'abord être étendues pour prendre en compte les problématiques spécifiques des données spatiales ou spatiotemporelles. Ces spécificités impactent l'intégration de sources et la modélisation multidimensionnelle.

L'intégration des sources spatiales se confronte, entre autres, au problème de l'hétérogénéité des formats. En effet, chaque système possède un modèle spatial propriétaire. Echanger ou intégrer les données de sources différentes revient généralement à les convertir dans un format particulier, ce qui est fastidieux et mène souvent à des pertes d'information. Ces difficultés ont été à l'origine des travaux de standardisation du consortium Open GIS (OGC). Le développement récent du langage XML et la spécification par OGC de GML (Geography Markup Language) ont facilité l'échange des données géographiques.

Le second problème d'intégration est dû à la précision et à l'échelle de représentation géométriques. Les mêmes objets peuvent avoir des définitions géométriques différentes selon la source. C'est un problème connu dit de *conflation* ou de *map matching*. Dans ce processus, il est difficile de fixer une mesure de

similarité « spatiale » sur laquelle pourra se baser l'intégration. Il a fallu étudier le processus d'intégration et l'améliorer. Cette étude était focalisée sur les géométries de type linéaire.

Contributions

Nos contributions sur l'intégration de données spatiales multi-sources ont concerné deux périodes. L'une avant l'ère XML où nous avons proposé un modèle interne de SGBD pouvant s'adapter à différents formats de données spatiales, puis nous avons développé un prototype GEOS sous le SGBD objet O2.

La seconde est basée sur XML. Nous avons proposé une architecture d'entrepôt basée sur un SGBD spatial XML/GML natif. Elle s'appuie sur la flexibilité du langage XML pour intégrer des sources hétérogènes. Le stockage natif au format GML entraîne un problème d'évaluation et d'optimisation de requêtes spatiales dans un modèle de stockage XML. Notre contribution a été principalement la proposition d'un cache sémantique adapté à GML dans le cadre de la thèse de Lionel Savary.

Concernant le problème de *map matching*, nous avons proposé une nouvelle méthode d'appariement d'objets basée sur la similarité de leurs géométries. Cette méthode est flexible dans le sens qu'elle permet à l'utilisateur de spécifier une pondération entre la similarité des formes et celle basée sur la distance. Elle tient compte également de critères de filtrages sémantiques. Elle a été testée avec succès dans l'appariement de deux représentations du réseau routier : l'une détaillée (comprenant tous les tronçons et leurs géométries précises) et l'autre partielle (certains axes et carrefours principaux) et simplifiée (seul les carrefours sont localisés). Ces sujets ont été traités en particulier dans la thèse de Lionel Savary soutenue en 2005 et ont été utilisés dans le projet Européen HEARTS[1].

I.1.3. Modélisation et analyse en ligne des données spatiotemporelles

L'analyse en ligne ou l'OLAP est une technique essentielle dans les applications décisionnelles. Elle se base sur la modélisation multidimensionnelle des entrepôts de données. Son application aux données spatiotemporelles soulève des problèmes dus à la complexité des dimensions spatiales et temporelles. De plus en plus de travaux portent sur les techniques OLAP pour les données spatiales et/ou spatiotemporelles [109]. Quelques travaux récents ont traité particulièrement le cas des objets mobiles [111].

Contributions

Nous avons étudié dans ce volet la modélisation multidimensionnelle des données mobiles. Nous avons proposé successivement deux approches. La première se base sur une représentation symbolique de la trajectoire mobile par référence à des unités spatiales et temporelles prédéfinies. Le modèle peut être implémenté dans les modèles classiques et peut répondre à des requêtes ne nécessitant pas une précision élevée. La seconde solution capture réellement la variation continue de l'espace et du temps. Elle permet

[1] HEARTS (Health Effects and Risk of Transport Systems) est un projet du programme énergie, environnement et développement durable du 5ème PCRD de la Commission Européenne (contrat n°: QLK4-CT-2001-00492).

de s'adapter à tout type d'agrégats et de requêtes spatiotemporelles non connus à l'avance et retourne des résultats avec plus de précision que la solution précédente. Pour ce faire, nous avons étendu les concepts de modélisation multidimensionnelle aux faits et aux dimensions continus. Nous avons redéfini le modèle, le mode de stockage, l'indexation et les requêtes d'agrégation pour ces concepts. Ce travail a débuté dans le cadre du projet Européen HEARTS et fait l'objet de la thèse de Tao Wan en cours.

I.1.4. Fouille de données spatiotemporelles

Les premiers travaux sur la fouille de données spatiales étaient ceux de Koperski et Han en 1995 [150] et de Ester et al. en 1997 [143]. Le problème est d'intégrer le raisonnement spatial dans un processus de fouille de données et principalement de prendre en compte l'influence du voisinage. Depuis, de nombreuses recherches ont concerné ce domaine et des implémentations commencent à apparaître dans des produits.

Contributions

Nous avons proposé une approche en deux étapes, où la première matérialise les relations de voisinage et la seconde les intègre dans la construction du modèle induit. A l'issue de la première étape, la fouille de données spatiale se ramène à la fouille de données multi-relations. Etant donné que la fouille de données classique est mono-relation, nous avons dû développer différentes solutions pour l'étendre au cas multi-tables. Nous avons appliqué notre approche à deux méthodes de fouille de données spatiales : la classification et la découverte de règles d'association. Ces travaux ont fait l'objet de la thèse de Nadjim Chelghoum soutenue en décembre 2004. Ils ont été appliqués à l'analyse de l'accidentologie routière, dans le but d'intégrer les propriétés du voisinage dans l'analyse des occurrences d'accidents.

I.1.5. Fouille de données séquentielles et textuelles

En plus de la fouille de données spatiales, nous nous sommes intéressés à d'autres problèmes de fouille de données complexes, dont le texte et les séquences temporelles. La fouille de données textuelles est un domaine de recherche très vaste à la frontière de la recherche d'information. Elle permet d'enrichir par la sémantique (et parfois une structure) des données à l'origine non structurées. La catégorisation fait partie des méthodes les plus utilisées. Les travaux visent essentiellement à améliorer la qualité de l'apprentissage. Les algorithmes réputés les plus efficaces sont k-NN (ou les k plus proches voisins), la méthode des centroïdes et SVM [204].

A l'inverse, les données séquentielles sont bien structurées et la fouille de séquences est un problème plus restreint. Il s'agit d'étendre les méthodes de fouille de données envers la prise en compte de l'ordre séquentiel et d'optimiser les performances d'exécution.

Contributions

Nous avons étudié principalement la méthode de catégorisation et proposé une nouvelle approche basée sur une combinaison des approches k-NN et centroïde. Celle-ci a permis d'améliorer les performances en

termes de qualité et de rapidité du classement de texte. Ces résultats ont d'abord été appliqués dans le cadre du projet RNTL Contexte Bourse pour la classification de dépêches financières, puis récemment dans le cadre du projet RNTS RHEA pour découvrir les codes maladie depuis des comptes-rendus hospitaliers.

Dans le volet sur la fouille de données séquentielles, nous avons proposé une méthode utilisant un index bitmap qui a été prouvée supérieure aux méthodes existantes. Cette méthode a été étendue à l'analyse de séquences multidimensionnelles. Ces deux méthodes ont été appliquées à la recherche de motifs dans les séquences d'activités et à la découverte de profils de la population selon ces motifs. Ils ont été exploités également au sein du projet HEARTS.

I.2. Organisation de ce mémoire

Hormis l'introduction et la conclusion, ce mémoire comporte cinq chapitres. Quatre chapitres sont liés à des problématiques de bases de données spatiales. Le chapitre II concerne la modélisation avancée ; le chapitre III est consacré aux entrepôts de données ; le chapitre IV porte sur l'analyse en ligne des données spatiotemporelles et enfin ; le chapitre V développe la thématique de fouille de données spatiales. Le chapitre VI relate les autres domaines de fouille de données abordés dont la fouille de données séquentielles et la fouille de textes. Nous terminons par la conclusion qui résume les contributions et introduit les multiples perspectives pour la gestion opérationnelle et décisionnelle des données spatiotemporelles et complexes en bases de données.

CHAPITRE II. MODELISATION AVANCEE DES DONNEES SPATIALES

Dans ce chapitre, nous relatons nos travaux sur certains aspects de la modélisation. Une partie de ces travaux porte sur la modélisation spatiale 3D. Nos recherches dans ce domaine avaient pour cadres, d'un côté le projet PSIG de l'axe A2 sur la gestion des données géographique 3D et de l'autre, la thèse de Thomas Delpy [7] soutenue fin 1995. La seconde partie porte sur la modélisation spatiotemporelle dans le cadre de travaux récents et de la thèse de Tao Wan en cours. Pour chacune de ces questions, nous décrivons précisément la problématique, l'état de l'art actualisé et nos contributions.

II.1. Problématique et limites de l'état de l'art

Les SIG ainsi que les SGBD spatiaux les plus courants se limitent à la représentation et la manipulation des données en 2D. Les principaux concepts correspondants sont résumés en annexe. Cependant, la gestion de la troisième dimension ou des données spatiotemporelles est essentielle pour de nombreux domaines d'application.

Ainsi, la géologie, le génie civil, l'urbanisme ou la robotique requièrent la possibilité de stocker des descriptions géométriques 3D, de les interroger et de les manipuler. La gestion environnementale, la gestion du trafic (routier, maritime aérien, …) sont des exemples d'applications nécessitant le support de données spatiotemporelles. Dans toutes ces applications, l'accès et l'analyse par requêtes, notamment spatiales, sont très demandés et portent sur des « collections » généralement volumineuses et organisées d'objets. D'où l'intérêt des techniques de bases de données.

La modélisation 3D pour les SIG ne peut pas simplement hériter des modèles développés dans d'autres domaines comme en CAO, car le raisonnement spatial en géographie combine souvent des données décrites dans des dimensions 2D et 3D. Par exemple, les applications archéologiques nécessitent de prendre en compte à la fois une carte 2D pour localiser la zone de fouilles, le relief (2,5D) pour les analyses et les simulations d'érosion ainsi que la représentation 3D des couches géologiques ou des ruines de bâtiments. Cela nécessite de **placer** la représentation 3D sur une carte qui elle est décrite en 2D ou en 2,5D (lorsqu'elle décrit le relief).

De plus, les objets ont une composante spatiale et une description par des attributs classique et peuvent être interrogés par critère spatial et /ou selon ses attributs. Par exemple, en archéologie, la requête peut porter sur le type, les époques des sites, la zone d'exploration ou les caractéristiques des couches géologiques. En outre, les données spatiales 3D sont organisées dans l'espace et possèdent des relations topologiques. Dans le domaine de la robotique, la description topologique facilite la navigation 3D.

Au moment où ont débuté nos travaux, la modélisation 3D se basaient sur la modélisation des solides dans les applications de CAO. La topologie était vue comme un moyen de vérification de cohérence grâce à la formule d'Euler [17], des k-simplex de Pigot [26]. Néanmoins, Molenaar [20] avait proposé une extension d'un modèle 2D géographique en définissant un type 3D donné par ses contours. La spécification des relations topologiques la plus connue est celle des 9-intersections de Egenhofer [9] [10] qui se base essentiellement sur des formes 2D. Aucun de ces modèles ne distingue l'orientation et le type de connexité et aucun n'intègre les types complexes composés de différentes formes géométriques.

Concernant les données spatiotemporelles, nous nous focalisons sur le cas particulier des objets mobiles. Ce sont les objets dont la localisation varie continuellement dans le temps. La gestion de bases d'objets mobiles a reçu une attention particulière durant ces dernières années en raison des avancées et de la banalisation des technologies mobiles et de géo-localisation, telles que les téléphones cellulaires, le GPS (*Global Positionning Systems*) et récemment le RFID (*Radio Frequency Identification*). Les travaux dans ce domaine portent principalement sur la modélisation d'objets mobiles [12] [14] [16] [24] [33], sur les méthodes d'accès [105][25][27], et enfin sur les requêtes prédictives et l'optimisation des mise à jour [94][32]. Un ouvrage sur ce sujet est paru dernièrement [13]. On distingue deux types d'objets mobiles : ceux ayant une trajectoire libre dans l'espace et ceux dont la trajectoire est contrainte [25][30]. Un exemple de trajectoire contrainte est le déplacement des véhicules dans le réseau routier ou celui des trains dans un réseau ferré. Assez souvent, les trajectoires mobiles sont représentées en 3D : 2D pour la localisation spatiale et 1D pour le temps.

II.2. Contributions

II.2.1. Modélisation 3D

La principale spécificité des applications géographiques 3D par rapport à celles de la CAO est que la localisation est toujours perçue par rapport à la surface terrestre et qu'en plus de la topologie planaire, les objets spatiaux sont naturellement agencés et perçus selon l'axe vertical. En effet, la force de gravité fait que, soit ils épousent le sol, soit ils s'appuient les uns sur les autres. Notre objectif était donc de pouvoir établir un modèle topologique 3D consistant qui capture cette information. Ce modèle devait être une généralisation de la topologie 2D afin de permettre l'intégration des représentations surfaciques (2D) ou planimétriques (2,5D).

Nous avons proposé un modèle topologique 3D (3DGT) spécialement adapté au raisonnement et à la navigation dans l'espace géographique [37] [38]. L'idée est d'étendre les formes de bases planimétriques 2D ou 2,5D (relief) par des formes verticales (ligne, surface ou volume vertical) comme le montre la Figure 1. Dès lors, des liens topologiques caractérisés par l'orientation verticale (au-dessus ou en dessous) sont définis entre ces formes de base en plus des liens existants. Ces formes de bases sont appelées primitives géométriques. La Figure 2 donne les primitives et leurs liens.

Des objets complexes sont également définis comme étant composés d'une ou de plusieurs primitives géométriques (Figure 3). Ce modèle à deux niveaux est dans le prolongement du modèle proposé durant nos travaux antérieurs [36] et est une généralisation de la norme AFNOR EDIGEO. Par conséquent, aucune conversion n'est nécessaire pour les données planimétriques fournies dans ce format d'échange. De plus, la traduction d'un autre format ou standard revient à une traduction au format EDIGEO, pour lequel des convertisseurs existent déjà.

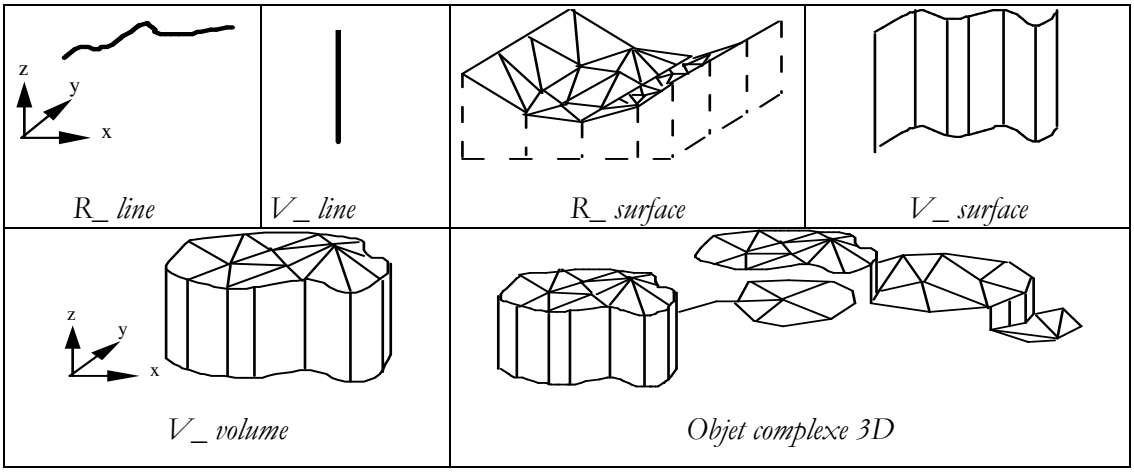


Figure 1. Extension des formes planimétriques R_xx par des formes verticales V_xx

Auparavant, nous avons proposé un modèle topologique 3D basé sur un graphe de voisinage. Ce modèle était restreint aux données volumétriques comme la modélisation de bâtiments en architecture. Nous

avons introduit la notion d'espace vide afin de formaliser les délimitations réelles ou virtuelles des pièces dans le bâtiment et développé des algorithmes de navigation 3D basés sur une structure de graphe [5] [6]. Une démonstration de ces algorithmes a été appliquée à la simulation en réalité virtuelle de la navigation 3D dans des bâtiments, en l'occurrence de centrale nucléaire.

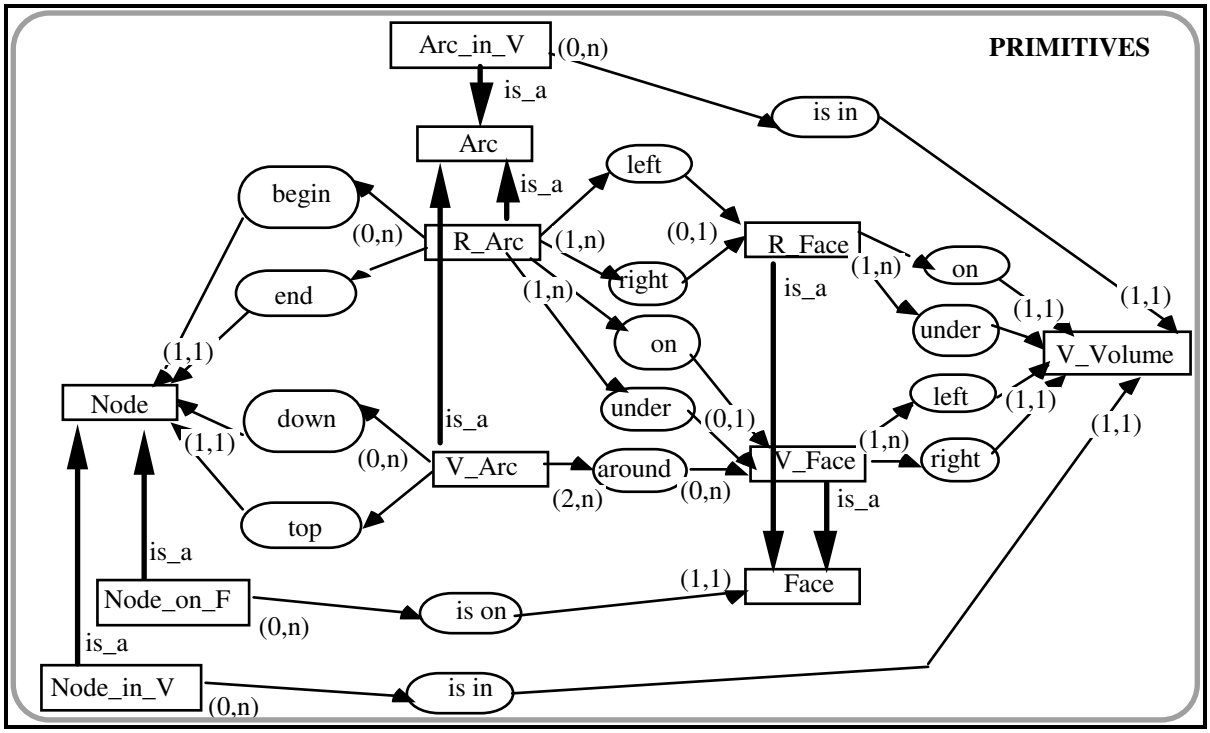


Figure 2.Primitives et relations topologiques 3D

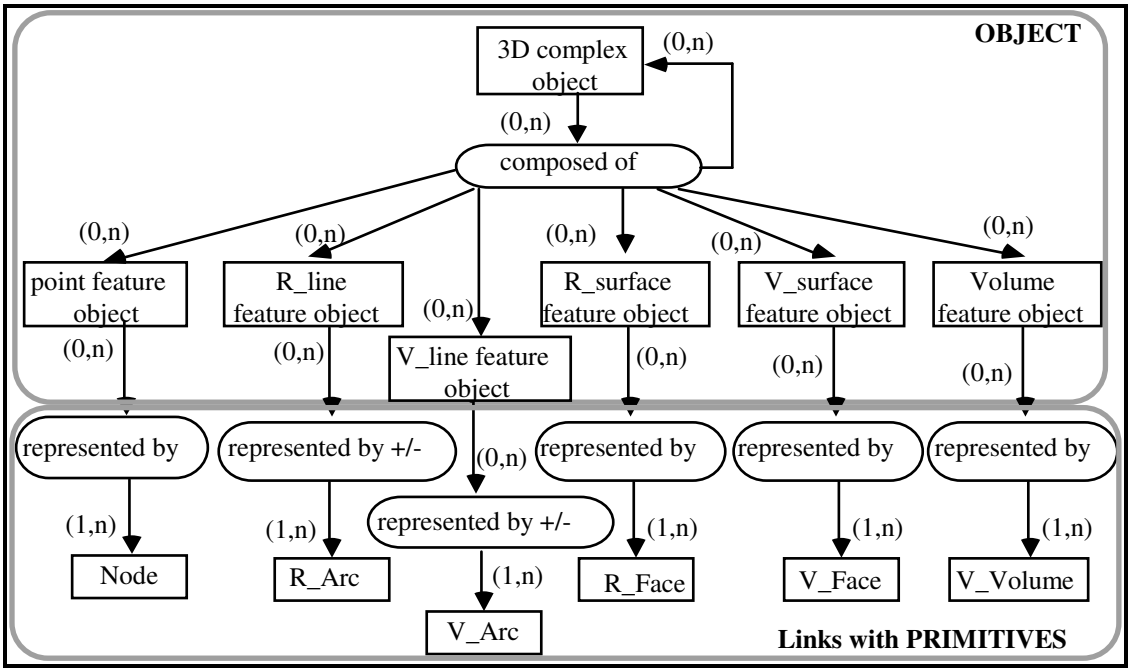


Figure 3. Relations entre objets et primitives

II.2.2. Modélisation spatiotemporelle

Nous avons été amenés à étudier la modélisation spatiotemporelle dans nos recherches sur l'analyse en ligne des données relatives aux objets mobiles. Nous avons commencé par développer le modèle de représentation pour objets mobiles. Le premier objectif est de pouvoir interroger ce type de données. Les requêtes peuvent porter sur des critères spatiaux, temporels et/ou sur des attributs.

Nous considérons qu'un objet mobile est défini par :

ObjetMobile (#ID, A1, A2, …An, TR) où #ID est un identifiant, Ai sont des attributs et TR une trajectoire [33]. Le modèle spatiotemporel concerne la représentation de la trajectoire. La représentation 3D, largement répandue dans les travaux, ne convient pas aux trajectoires contraintes car elles ne maintiennent pas leur relation avec les zones qui les contraignent en sont coûteuse en stockage et dans leur traitement.

Nous avons donc proposé deux modèles qui tirent profit de la mobilité contrainte. L'un est un modèle discret ou par référence. L'autre est un modèle continu décrivant des trajectoires contraintes en s'inspirant de travaux existants. Ces modèles ont été exploités dans l'application et l'adaptation des techniques OLAP aux objets mobiles, comme exposé dans le chapitre IV.

II.2.2.1. *Modèle de représentation discret d'objets mobiles*

L'idée de ce modèle est de transformer la représentation complexe des objets mobiles en une représentation symbolique. Pour ce faire, nous utilisons un découpage en unités spatiales et temporelles prédéfini ou qui est construit à cet effet. Nous représentons ensuite les trajectoires par référence à ces unités spatiales et temporelles. Une trajectoire est alors définie comme un ensemble :

TRef = {(#S1, #T1), …} où #Si référence une unité spatiale et #Ti une unité temporelle telles que la trajectoire a traversé Si à un instant de l'intervalle Ti.

Une unité spatiale peut être une cellule d'une grille raster, une zone d'un découpage polygonal ou une polyligne d'un réseau. Très souvent, ces unités de référence existent au préalable et sont parfois organisées en une hiérarchie spatiale. Le choix de l'unité spatiale de référence est alors naturel.

Au niveau des requêtes spatiale et/ou temporelles, soit que le critère utilisé désigne des unités prédéfinies, auquel cas, sa résolution se fait simplement en relationnel ; soit qu'il exprime une fenêtre géométrique et/ou un intervalle temporel, auquel cas elle nécessite la sélection dans le découpage prédéfini des unités (spatiales / temporelle) correspondantes et ensuite la résolution sa réécriture par rapport à ces unités.

L'avantage de ce modèle est sa simplicité et la possibilité d'utiliser les systèmes existants. Le seul inconvénient est la limitation des opérateurs et le manque de précision. En effet, on ne peut ni déterminer l'intersection de trajectoires différentes ou les comparer, ni déterminer si elle inclut une localisation. On ne peut pas non plus résoudre avec précision l'intersection avec une fenêtre spatiotemporelle qui ne se conforme pas aux unités prédéfinies.

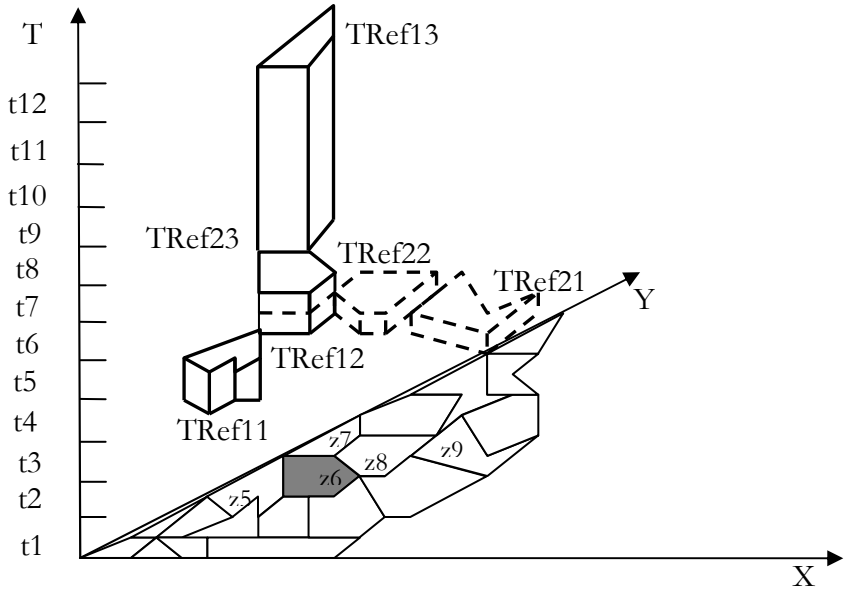


Figure 4. Modèle de représentation d'objets mobiles par référence

II.2.2.2. *Modèle de représentation continu d'objets mobiles*

Les objectifs de ce modèle sont :

(i) de représenter les trajectoires des objets mobiles de manière plus fidèle que dans la représentation discrète ;

(ii) de répondre à tout type de requêtes sur ces trajectoires ;

(iii) d'offrir une représentation compacte ;

(iv) de pouvoir être manipulé facilement et efficacement.

Nous nous sommes basés sur l'idée de réduction de dimensionnalité de [25]. En effet, lorsqu'une trajectoire est contrainte par un réseau, les coordonnées absolues de l'espace 2D peuvent être remplacées par la position relative dans les tronçons traversés du réseau.

Si en plus on transforme le réseau, à l'origine 2D, en intervalles 1D connexes, il devient possible de projeter ensuite la position relative de l'objet sur cet axe pour obtenir une position absolue en 1D. Cette position combinée avec le temps permet d'exprimer la trajectoire en 2D.

Nous décrivons ci-dessous les étapes de cette transformation.

Transformation 1D du réseau routier : elle désigne une fonction :
Tr $(rid_i) = [i', i'+1[$ dans un espace 1D cible borné appelé ***TR***, où rid_i est l'identifiant d'un tronçon et $i, i' \in \{1\ldots, n\}$.

Cette transformation revient à numéroter les tronçons de routes et les aligner les uns après les autres sur des intervalles de l'axe TR. Toute position relative *pos* à un tronçon rid_i du réseau peut être transformée en position absolue *p* dans la dimension TR par l'expression : $p = i' + pos$.

Dès lors, la transformation de trajectoire contrainte par le réseau peut se baser sur les positions absolues dans TR. Il en découle la définition suivante de la transformation en 2D d'une trajectoire.

Transformation 2D de trajectoire : elle correspond à sa représentation par :
TT $(tid, ((p_i^d, t_i^d), \ldots, (p_f^f, t_f^f)))_{i=1..k}$ où p_i^d (respectivement, p_f^f) est une position sur l'axe *TR* correspondant au point du début (respectivement, de la fin) de la trajectoire dans le tronçon rid_i à l'instant t_i^d (respectivement à l'instant t_f^f) et tel que $t_{i+1}^d = t_f^f$ et $t_f^f > t_i^d$**.** (…) désigne une séquence.

Ainsi, une trajectoire est transformée en séquence de lignes 2D, pas forcément connectés (Figure 5).

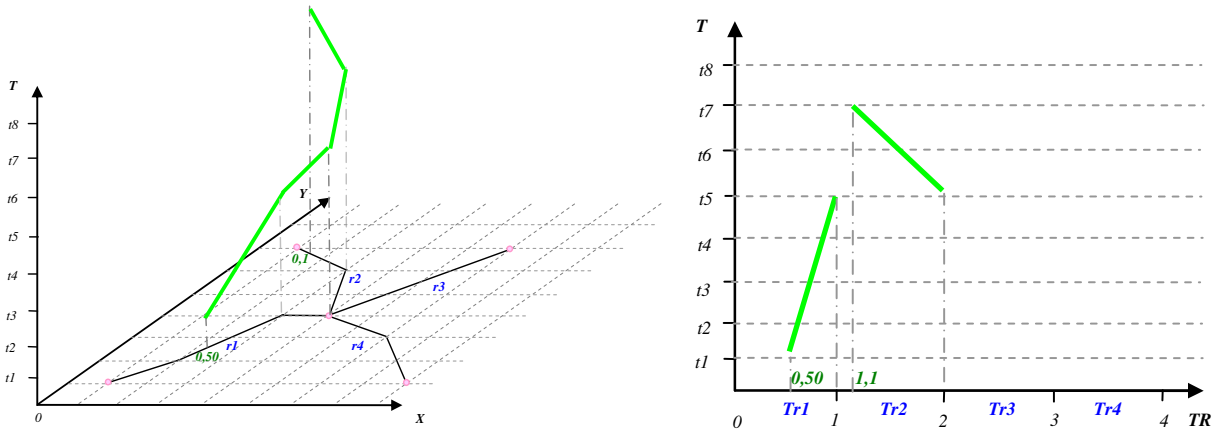


Figure 5. Transformation 2D d'une trajectoire.

Au moment de la requête, celle-ci est d'abord réécrite dans l'espace 2D (TR x T) en un ensemble de sous requêtes. Celles-ci sont ensuite résolues comme s'il s'agissait de requêtes spatiales 2D. Son implémentation peut se faire à l'aide d'un SGBD spatial standard.

Ce modèle permet de résoudre tous les opérateurs et retourne des résultats précis.

II.3. Conclusion

Peu développé dans les produits SIG et bases de données spatiales, la gestion des données spatiales 3D suscite encore l'intérêt des chercheurs. Le modèle de base proposé par le consortium Open GIS et normalisé par l'ISO [15] modélise bien la 3D, décrit la topologie selon le modèle des 9-intersections [10] et permet aux systèmes de spécialiser le modèle selon différents profils avec ou sans topologie. Mais les systèmes se limitent à certaines classes et la 3D est rarement implémentée. A notre connaissance, seule la librairie GeoToolkit [1] implémente un type abstrait spatial 3D basé sur les simplex (réseau de tétraèdres). GML 3 [22] permet aujourd'hui d'échanger des données spatiales 3D.

Malgré ces avancées, la recherche continue sur les SIG 3D. Récemment, un numéro spécial de la revue de géomatique [28] et un ouvrage [39] lui ont été consacrés. Des articles sur le sujet paraissent régulièrement dans les principales conférences en géomatique. Parmi les travaux récents, citons [31] qui propose un type abstrait Spatial3D proche de notre proposition et décrit une algèbre d'opérateurs, puis [40] et [3] qui

concluent sur le manque de consensus pour la représentation de la topologie en 3D et constatent l'inexistence d'implémentation réelle de modèles 3D dans les systèmes. Dans les environnements intérieurs, un modèle spécifique proche de [38] est proposé dans [19].

Concernant la gestion d'objets mobiles dans les bases de données, l'abondance et l'actualité des travaux montre que ce domaine ne cesse de se développer. Nous reviendrons sur ce domaine dans le chapitre IV où nous verrons d'autres aspects que la modélisation de base.

II.4. Références

[1] Balovnev O., Bode T., Breunig M., Cremers A.B., W. Muller, G. Pogodaev, S. Shumilov, J. Siebeck, Siehl A., and Thomsen A.., *The Story of the GeoToolKit – An Object-Oriented Geodatabase Kernel System*. GeoInformatica, 8(1):5–47, 2004.

[2] Barraud F., Zeitouni K., "GEOS : Towards an Open Spatial Data Model", 8th International Symposium on Spatial Data Handling, SDH 98, July, 11-15, 1998, Vancouver, Canada, pp 408-417.

[3] Breuning, M. and S. Zlatanova, 3D Geo-DBMS, Chapter 4 in S. Zlatanova & D. Prosperi (Eds.) *Large-scale 3D Data Integration: challenges and opportunities*, Taylor & Francis, A CRC Press book, pp. 88-113, 2005.

[4] CEN (1996) Geographic Information. European Prestandard final draft. CEN Technical Committee 287 Geographic Information. Comité Europeéen de Normalisation, Brussels.

[5] Delpy, T., K. Zeitouni (1993) A Graph Model to Describe Topological Relationships in a Three-Dimensional World, in Proceedings of the Third International Conference on Computers in Urban Planning and Urban Management, Atlanta, USA, Vol. 1, 151-166.

[6] Delpy, T., B. Brillault, K. Zeitouni (1994) A Topological Database to Simulate 3D Robotics Navigation, in Proceedings of the Third International Conference on Automation, Robotics and Computer Vision, ICARCV'94, Singapore, Vol. 1, 50-55.

[7] Delpy Thomas, Modélisation et Conception d'une Base de Données Environnementale pour de la simulation de navigation 3D en Environnement Complexe, Doctorat de l'Université Pierre et Marie Curie, Novembre 1995.

[8] Donikian, S., G. Hégron (1993) A declarative design method for 3D scene sketch modeling, in Eurographics'93, Barcelona.

[9] Egenhofer, M. J., J. R. Herring (1990) A mathematical framework for the definition of topological relationships, in Proceedings of the 4th International Symposium on Spatial Data Handling, Switzerland, Vol. 2, 803-813.

[10] Egenhofer M.J., "Reasoning about Binary Topological Relations", Proc. 2nd Int. Symp. on Large Spatial Databases, SSD'91, Zurich, Switzerland, August 1991, LNCS n° 525, 143-160.

[11] de Floriani, L., A. Maulik and G. Nagy (1990) Manipulating a modular boundary model with a face-based graph structure, in M. J. Wozny, J. U. Turner and K. Press (ed) Geometric modeling for product engineering, Elsevier Science Publishers, 131-143.

[12] Güting. R., Böhlen . M., Erwig . M., Jensen .C.: Lorentzos. N., Schneider. M., Vazirgiannis . M. Z.. A Foundation for Representing and Querying Moving Objects. Int. ACM TODS, (2000) 25(1):1–42.

[13] Güting R., Schneider M.. Moving Objects Databases, Morgan Kaufmann, ISBN 0-12-088799-1, 2005

[14] Grumbach S., Rigaux P., Segoufin L., Spatio-Temporal Data Handling with Constraints, GeoInformatica, 5(1), 2001

[15] ISO 19107:2003, Geographic Information – Spatial Schema, WG 2.

[16] Koubarakis M., Pernici B., Schek H.J., Scholl M., Theodoulidis B., Tryfona N., Sellis T., Frank A.U., Grumbach S., Güting R.H., Jensen C.S., Lorentzos N., Manolopoulos Y., and Nardelli E. (Eds.), Spatio-Temporal Databases: The CHOROCHRONOS Approach. Springer-Verlag, Lecture Notes in Computer Science 2520, 2003.

[17] Lienhardt P. (1991) Topological models for boundary representation: a comparison with n-dimensional generalized maps, in Computer-Aided Design, 23 (1), 59-82.

[18] Longley P.A., M.F. Goodchild, D.J Maguire, D.W Rhind, Geographical Information Systems, Principles and Technical Issues, John Wiley & Sons, Inc., 2nd Edition, 1999.

[19] Meijers, M., S. Zlatanova and N. Preifer, 3D geoinformation indoors: structuring for evacuation, in: Proceedings of Next generation 3D city models, Bonn, Germany, 21-22 June, 2005

[20] Molenaar M. (1990) A formal data structure for the three dimensional vector maps, in Proc. of the 4th Int. Symp. on Spatial Data Handling, Switzerland, Vol. 2, 830-843.

[21] OGC, 1999, OpenGIS Consortium. The OpenGIS Abstract Specification, Topic 1: Feature Geometry

[22] OGC 03-105r1, OpenGIS Geography Markup Language (GML) Implementation Specification, Version 3.1.1, April 2004.

[23] Oracle Corporation, Oracle Spatial Topology and Network Data Models, 10g Release 2 (10.2) B14256-02, January 2006.

[24] Pfoser. D., Jensen. C. S. Theodoridis. Y.: Novel Approaches in Query Processing for Moving Object Trajectories. Proc of VLDB, Cairo, Egypt (2000) 395-406.

[25] Pfoser D. Jensen C.S., Indexing of Network-constrained MOs. ACM-GIS, 2003.

[26] Pigot S. (1992), A topological model for 3D spatial information system, in Proceedings of the 5th Int. Symp. on Spatial Data Handling, Charleston, USA, Vol. 1, 344-360.

[27] Rigaux P., Scholl M., Segoufin L., and Grumbach S., Building a Constraint-Based Spatial Database System: Model, Languages, and Implementation. Information Systems, 28(6): 563-595, 2003.

[28] Saux E. et Billen R. Information géographique tridimensionnelle, Revue internationale de géomatique, Edition Hermès, Vol 16 N°1/2006.

[29] Savary L., Zeitouni K., "Spatio-Temporal Data Warehouse - a Prototype", In Second Electronic Government (EGOV) Conference, Joint conference to DEXA , Prague Czech Republic, Lecture Notes in Computer Science, Springer-Verlag, Volume 2739, September 1-5, 2003, pp. 335 - 340.

[30] Speicys L., Jensen C.S., Kligys A., Computational Data Modeling for network-Constrained MOs. In ACM-GIS, 118-125, 2003.

[31] Schneider M. and Weinrich B. E.. An Abstract Model of Three-Dimensional Spatial Data Types. Proc. of the 12th ACM Int. Symp. on Advances in Geographic Information Systems (ACM GIS 2004), pp. 67–72, 2004.

[32] Tao, Y., Papadias, D.: Time-Parameterized Queries in Spatio-Temporal Databases. Proc. of ACM SIGMOD, Madison, Wisconsin (2002) 334-345.

[33] Vazirgiannis. M. and Wolfson. O.: A Spatiotemporal Query Language for Moving Objects. Proc of SSTD , Los Angeles, CA (2001) 20-35.

[34] Weibel, R., M. Heller (1992) Digital Terrain Modelling, in D. J. Maguire, M. F. Goodchild (ed.) Geographical Information Systems, vol.1, Principles, Longman Scientific & Technical, 269-297.

[35] Wilson P. (1985) Euler Formulas and Geometric Modelling, in IEEE Transactions on Computer Graphics and Applications, 5 (8), 24-36.

[36] Zeitouni K., (1991) GéoGraph : un modèle de représentation et d'organisation physique des bases de données spatiales, Thèse d'Université Paris 6, Juillet 1991.

[37] Zeitouni K. and de Cambray B. (1995a) Supporting the Semantics of 3D Geographical Space in GIS, in Proceedings of Joint European Conference on Geographical Information, The Hague, Netherlands.

[38] Zeitouni K., de Cambray B., and Delpy T. (1995b) Topological Modelling for 3D GIS, 4th Int. Conf. on Computers in Urban Planning and Urban Management, 1995.

[39] Zlatanova, S. and D. Prosperi (Eds.), Large-scale 3D Data Integration: challenges and opportunities, Taylor & Francis, A CRC Press book, 2005.

[40] Zlatanova S., On 3D Topological Relationships, Int. Workshop on Database and Expert System Applications, pp. 913–919, 2000.

CHAPITRE III. ENTREPOT DE DONNEES SPATIOTEMPORELLES

Ce chapitre présente nos contributions dans le domaine des entrepôts de données spatiotemporelles. Il correspond principalement aux travaux de thèse de Lionel Savary [230] et rentre dans le cadre du projet Européen HEARTS [41].

Les techniques d'intégration des données sont souvent un maillon important de la chaîne de l'extraction de connaissances depuis des données spatiales et spatiotemporelles. En effet, les sources de données sont souvent hétérogènes et nécessitent au préalable une phase de nettoyage et d'intégration. De plus, leur organisation est souvent inadaptée au sujet d'analyse visé et nécessite une modélisation appropriée et une transformation. Les techniques d'entrepôts de données visent justement l'intégration de sources hétérogènes et la transformation dans un modèle approprié facilitant l'application des outils d'analyse et de fouille de données. Seulement, ces techniques sont insuffisantes dans un entrepôt de données spatiales car elles ne prennent pas en compte les problématiques spécifiques des données spatiales ou spatiotemporelles lors de l'intégration des sources.

Ce chapitre est organisé comme suit. La première section détaille les problématiques. La deuxième section fait une synthèse de l'état de l'art pour les sujets étudiés. Ensuite, les sections suivantes relatent nos travaux. Un bilan et des perspectives de ces travaux concluent ce chapitre.

III.1. Problématique

Une des problématiques des entrepôts de données spatiotemporelles est la construction même de l'entrepôt, c'est-à-dire l'architecture du système, et le processus d'intégration des sources.

Si dans les entrepôts de données classiques, les architectures sont majoritairement basées sur les SGBD relationnels, les formats de plus en plus variés des sources (comme le web) et la popularité grandissante de

XML amènent à des architecture basées sur XML. Les sources de données spatiales se caractérisent justement par la diversité de leurs formats. Une architecture basée sur XML est intéressante pour un entrepôt de données spatiales. Cet entrepôt peut utiliser un SGBD XML, comme il en existe plusieurs aujourd'hui. Ceci soulève néanmoins un problème quant au support des requêtes spatiales dans ce gestionnaire. Non seulement ces fonctionnalités doivent être intégrées, mais elles doivent être optimisées afin de répondre efficacement aux requêtes spatiales sur des documents XML.

Le second problème de construction de l'entrepôt est l'intégration de sources. Ce problème n'est pas uniquement lié aux entrepôts, car les applications géographiques se caractérisent par le recours fréquent à des données externes (cf. tableau 6 en annexe). Ces sources ont souvent des formats incompatibles ou inadaptés au modèle utilisé par le SIG. Il devient alors nécessaire de convertir les données entre les deux formats. A moins que des convertisseurs soient prévus, cette tâche s'avère souvent longue et fastidieuse et constitue un obstacle pour certains projets en SIG. De plus, la conversion systématique mène parfois à une perte d'informations car certains modèles capturent certaines données et pas d'autres. La convergence vers des standards du monde XML facilite néanmoins la conversion sans pallier la perte d'information. Outre la conversion de format, les entrepôts en général nécessitent la fusion de données et détectent les objets dont la description est redondante ou complémentaire dans différentes sources. La difficulté de cette détection provient du fait que les objets décrits dans des sources différentes ont souvent des identifiants différents et des niveaux sémantiques différents. Les outils d'intégration utilisent des procédures d'appariement (de *matching*) de données basées sur la similarité du contenu. S'agissant des sources spatiales, l'appariement ne se limite pas à la description attributaire classique, mais se fait également selon la description géométrique. Or, les mêmes objets peuvent avoir des définitions géométriques différentes selon la source, selon le niveau de précision et parfois même selon l'objectif des applications pour lesquelles ces sources ont été créées. C'est un problème connu dit de *conflation* ou de *map matching*. Dans ce processus, il est difficile de fixer une mesure de similarité « spatiale » sur laquelle peut se baser l'intégration. Il faut donc étudier le processus d'intégration et l'améliorer. Notre étude s'est focalisée sur les géométries de type linéaire.

III.2. Etat de l'art

Sur la base de la classification précédente, la synthèse de l'état de l'art comprend trois parties :

1. l'architecture d'entrepôts de données géographiques ;
2. l'optimisation des requêtes dans ce contexte ;
3. l'intégration de formats hétérogènes ;
4. l'appariement des données géométriques.

III.2.1. Architectures d'entrepôts de données géographiques

Il n'existe pas, à notre connaissance, d'architecture d'entrepôts de données géographiques basée sur GML. Les travaux relatifs à l'accès aux sources géographiques hétérogènes se basent plutôt sur une architecture de type médiation.

Ainsi, Corocoles *et al.* [54] proposent une architecture de médiation basée sur GML. Le médiateur est composé d'un analyseur, d'un optimiseur et d'un exécuteur. Sa particularité est de traduire les requêtes en langage SQL, puis de retourner les résultats de ces requêtes dans un document GML.

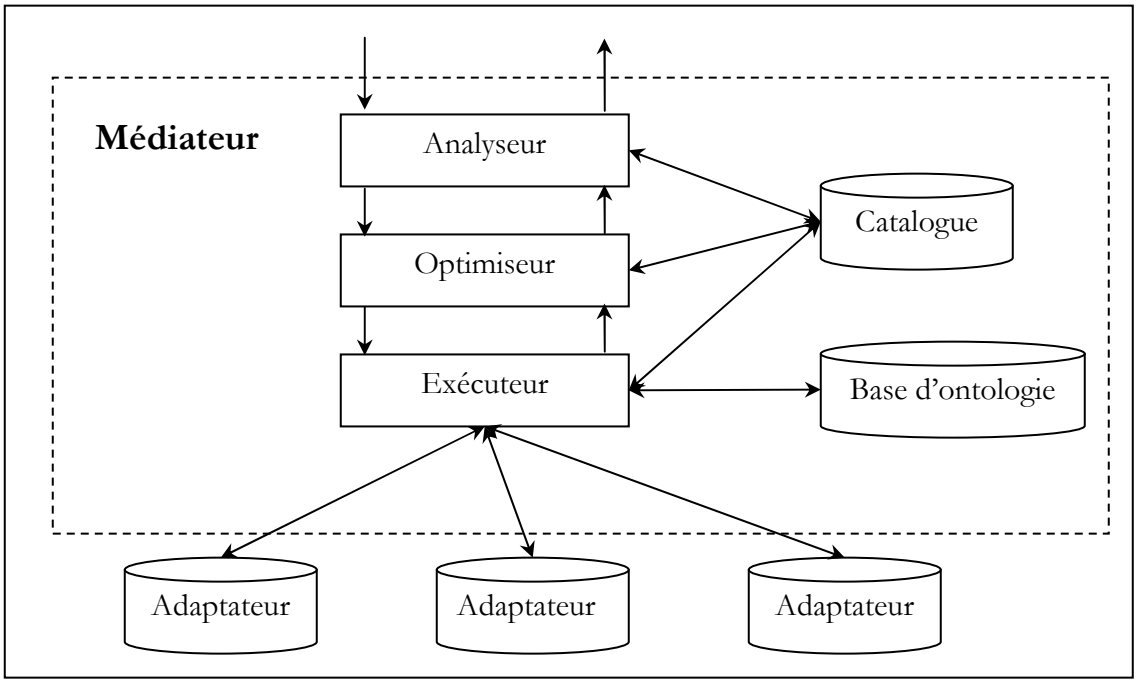


Figure 6. Architecture dans [54]

Un second exemple d'architecture de médiation est celle proposée par Boucelma *et al.* [45]. Celle-ci est basée sur WFS (Web Feature Services) [258]. Le médiateur est composé de trois modules : un analyseur, un optimiseur et un exécuteur. Le plan d'exécution décompose les requêtes en sous-requêtes et les dirige vers les WFS associées aux sources. Les requêtes et sous-requêtes sont envoyées au format XQuery étendu aux données spatiales en incluant des opérateurs spatiaux. Les résultats des sous-requêtes sont retournés par le WFS dans des documents GML à l'exécuteur. Si une source n'implémente pas d'opérateurs spatiaux, c'est l'exécuteur qui se charge d'exécuter la requête en utilisant des opérateurs spatiaux. Les résultats des sous-requêtes sont ensuite intégrés dans un même document GML.

Hormis ces travaux, une architecture hybride médiation / entrepôt a été proposée par Voisard et Juergens [80]. Si toutes les données sources sont entreposées dans l'entrepôt de données, alors toutes les requêtes peuvent être satisfaites par celui-ci. Dans le cas contraire, les requêtes sont traitées par la partie médiation. Le modèle pivot est XML.

Contrairement aux architectures de type médiation, les architectures de type entrepôt de données permettent d'exploiter efficacement les données et offrent des fonctionnalités de préparation et de nettoyage de données. Les quelques travaux sur les architectures d'entrepôt de données géographiques

sont généralement limités à des plateformes très peu évolutives et dépendantes de la base de données de l'entrepôt. Ainsi, les auteurs de [75] proposent une architecture d'entrepôt de données géographiques liées à l'agriculture. L'architecture est de type distribué et les données sont répliquées de manière à assurer leur disponibilité et l'efficacité des requêtes par la répartition des charges sur différentes copies de la base. Les données sont extraites des sources et placées dans une zone de préparation (staging area) où elles sont nettoyées et intégrées (Figure 7) avant leur chargement dans l'entrepôt. Mais la redondance de données ralentit les processus de chargement et de rafraîchissement qui doit se faire pour chaque copie.

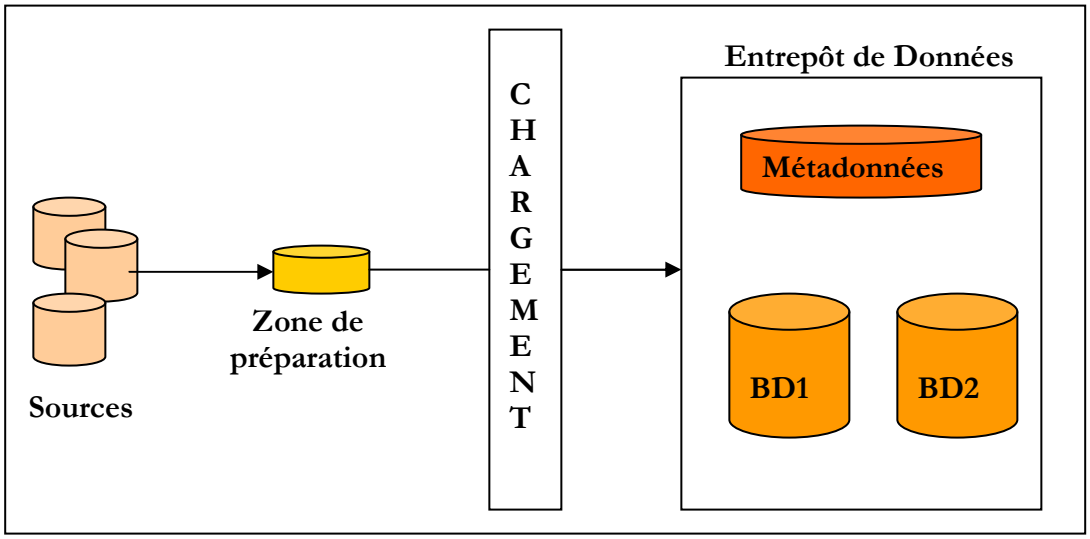


Figure 7. Entrepôt de données géographiques distribué

III.2.2. Optimisation des requêtes par le gestion de cache

Les bases de données géographiques sont souvent volumineuses. De plus, les requêtes spatiales sont coûteuses et ce coût augmente avec le volume de données et le nombre et la complexité des opérateurs spatiaux utilisés dans une requête. Ajouté à cela, le stockage sous forme de documents GML entraîne un surcoût de stockage et d'accès dû à la structure arborescente typique en XML. C'est ce qui nous a amenés à étudier l'optimisation de requêtes pour GML. L'objectif est de réduire globalement le coût d'un flot de requêtes sur des documents GML, d'éviter l'accès aux données sur disque autant que possible et de limiter l'exécution (ou la ré-exécution) des opérateurs spatiaux.

Une technique d'optimisation connue en bases de données est la gestion de cache dont le principe est de mémoriser les résultats de certaines requêtes en vue de pouvoir les réutiliser ultérieurement. Il existe différentes techniques pour la gestion des caches en général dont les trois plus connues sont :

(i) la mémorisation de pages retrouvées par adresse physique [50] ;

(ii) la mémorisation de tuples retrouvés par identifiants [59];

(iii) les caches sémantiques où les données sont retrouvées suivant leur description (requêtes, formules de contraintes) [55][61][62].

D'après [55], un cache sémantique est défini comme une collection de *régions sémantiques* où chaque région contient une formule de contrainte, un compteur sur les tuples satisfaisant la contrainte, un pointeur sur la

liste des tuples en cache et des informations additionnelles qui sont utilisées par la politique de remplacement. La manipulation d'un cache sémantique s'effectue au niveau de la description des requêtes. En effet, on se base sur l'analyse du contenu de la requête plutôt que sur le parcours des identifiants de tuples ou pages pour déterminer si une requête utilisateur peut être extraite du cache. Les caches sémantiques sont les plus performants.

Mais, les caches sont des zones de stockage en mémoire et de tailles limitées. Il est donc nécessaire de définir une politique de remplacement de cache. Il existe pour cela différentes techniques (voir le tableau ci-dessous) [51].

Tableau 1 : *Politiques de remplacement de cache*

POLITIQUE	**COUT**	**TAILLE**	**FREQUENCE**	**PROBABILITE**
LRU & LRU-Threshold	Non	Non	Oui	Non
LFU & LRU-k	Non	Non	Oui	Non
LRV	Oui	Oui	Oui	Oui
Size	Non	Oui	Non	Non
Log (size)+LRU	Non	Oui	Oui	Non
GD-Size	Oui	Oui	Non	Non
GDSF	Oui	Oui	Oui	Non
GDSF + Taylor	Oui	Oui	Oui	Oui

De nos jours, la technique LRU (Least Recently Used) est encore beaucoup utilisée du fait de sa simplicité. Cependant, l'inconvénient de LRU est qu'elle ne prend pas en compte les fréquences d'accès aux pages. Sa variante LRU-Threshold [53] ne place en cache que les objets de taille en deçà d'un seuil donné.

La technique de LFU (Least Frequently Used) considère que plus une page est utilisée pour répondre à une requête, plus elle aura de chance d'être utilisée pour des requêtes futures. L'inconvénient de cette technique est qu'une page peut être utilisée très souvent au début, mais pas du tout par la suite. La méthode LRU-K [67] prend en considération l'instant des k derniers accès de chacune des pages et supprime une page si sa k-distance est supérieure à celle de toutes les autres pages du cache.

L'algorithme Size supprime simplement le document de plus grande taille. Sa variante Log(size)+LRU [42] supprime les documents selon les valeurs de log(taille) et applique LRU en cas d'égalité.

L'algorithme LRV (Lowest Relative Value) proposé dans [65] prend en compte le coût et la taille d'un document. Il estime la probabilité qu'un document soit lu une nouvelle fois et supprime le document de plus petite valeur du cache.

Dans le contexte WEB, l'algorithme GD-Size (Greedy Dual Size) introduit par Cao et Irani [49] est l'un des plus connus. GD-Size augmente la valeur associée à un objet à insérer dans le cache par la valeur de l'objet supprimé. Arlitt *et al.* [43] vont plus loin en ajoutant le facteur de fréquence dans GDSF (Greedy Dual Size Frequency). Basé sur GDSF, Yang *et al.*[87] introduisent en plus le facteur temps. Ils prédisent le temps du prochain accès à l'objet en utilisant les séries de Taylor.

Enfin, dans le domaine des bases de données spatiales, des politiques de remplacement de cache ont été développées uniquement pour des bases relationnelles ou objet-relationnelles. Elles reposent généralement sur LRU et ses variantes comme dans [44] et [47]. Ce dernier gère un cache de pages mais donne plus de priorité pour le maintien en cache aux pages d'index par rapport aux pages de données et des objets géométriques associés.

III.2.3. Intégration de formats hétérogènes

L'accès aux sources géographiques hétérogènes avait été pris en charge par les groupes de normalisation dont le Comité Européen de Normalisation CEN/TC287 dont la prénorme [4] propose un modèle ouvert et paramétrable pour l'expression et l'échange de n'importe quel format source de données géographiques.

Aujourd'hui, des logiciels spécialisés dans cette conversion, comme Feature Manipulation Engine (FME2), ont été développés. Par Ailleurs, le consortium Open GIS a depuis été moteur pour réduire cette hétérogénéité ou son impact, mais visait au départ plutôt l'interopérabilité des SIG. Il a défini un modèle géométrique pour l'accès à des serveurs hétérogènes [21]. Le comité ISO/TC211 a ensuite édité le schéma spatial [15] sur la base de la prénorme du CEN et des travaux d'Open GIS. L'échange des données géographiques passe de plus en plus par le langage GML basé sur XML. Une autre avancée vient des services Web géographiques WFS [258] par Open GIS[3], permettant des architectures plus flexibles. Cependant, aucun système ne permet de s'adapter aux modèles des sources.

III.2.4. Appariement des données géométriques

L'intégration de données géographiques est une tâche complexe, car on doit prendre en compte les conflits et hétérogénéités portant sur les données de type spatiales et non spatiales [46] [57] [88]. Les conflits sur les données spatiales sont bien plus complexes à résoudre que les conflits sur les données non spatiales [68] principalement à cause de l'hétérogénéité d'échelle et de modèle de représentation. On distingue deux types d'intégration :

- l'intégration de schémas ;
- l'intégration de données.

---

[2] Outil basé sur le standard canadien SAFE (www.safe.com)

[3] Toutes les spécifications du consortium OGC sont disponibles sur : www.opengeospatial.org/spec

L'intégration de schémas a pour but de résoudre les conflits [68] de noms (synonymes, homonymes), d'identifiant et d'isomorphisme de schéma, alors que l'intégration de données permet de résoudre les inconsistances et l'incohérence de granularités dans le domaine spatial. Assez souvent, les relations entre des objets géographiques de sources différentes ne peuvent être déduites que par leurs composantes spatiales. Cela implique de déduire ces relations à partir de leurs positions géographiques et leurs géométries respectives.

Une solution serait l'application d'opérateurs spatiaux tels que l'intersection géométrique pour retrouver les correspondances de géométrie. Seulement, cela ne suffit pas en raison de l'imprécision et de la différence d'échelles des différentes sources. On a alors recours aux méthodes dites d'*appariement* ou de *conflation* [79]. On s'intéresse ici aux méthodes d'appariement d'objets linéaires. Elles peuvent être classées en quatre catégories :

- Celles qui définissent une zone d'appariement [60],
- Celles qui définissent des mesures de distances entre objets [58],
- Celles qui se basent sur la ressemblance de formes entre objets [66],
- Celles qui se basent sur la forme et la distance entre objets [69][86]

III.3. Contributions

III.3.1. Architecture d'entrepôt de données géographiques

Afin de fournir une architecture flexible, évolutive, indépendante de la base de données de l'entrepôt et permettant d'intégrer facilement des sources hétérogènes, nous proposons une architecture d'entrepôt de données géographiques entièrement basée sur XML. Son implémentation utilise des outils libres d'accès conformes aux normes ISO, aux propositions d'OpenGIS et du W3C.

Cette architecture comprend différents niveaux : les sources, les adaptateurs, un module d'intégration, la base de l'entrepôt de données et l'interface utilisateur. Durant le processus d'extraction - transformation – chargement (ETL) les données sont :

- Extraites des sources.
- Converties dans un format standard par les adaptateurs associés à chaque source.
- Nettoyées dans la zone de préparation (Data Staging Area ou DSA) et intégrée par l'intégrateur associé. Les conflits entre les données sont identifiés et résolus. Puis les données sont restructurées via des règles de correspondances.
- Converties dans un format standard et chargées dans la base de données de l'entrepôt.

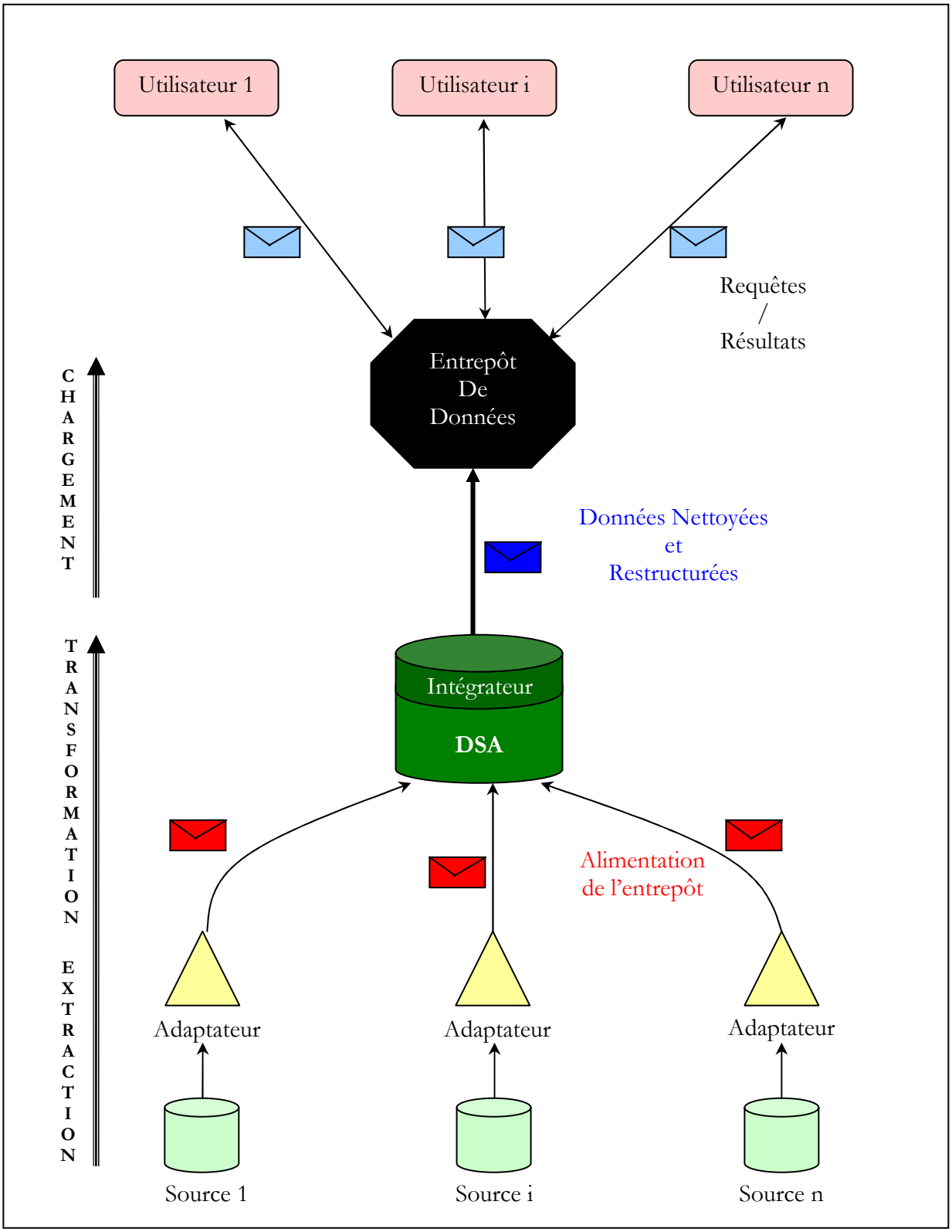


Figure 8. Architecture proposée

Afin de pouvoir manipuler à la fois les données spatiales et non spatiales, l'architecture intègre des technologies standard diverses :

- XQuery [83][78] pour la manipulation des données non spatiales.
- JTS [78] (Java Topolgy Suite) pour la manipulation de données spatiales.
- SAX [74] comme parseur pour les documents XML.
- GML-QL [77] comme langage de requêtes entre les utilisateurs et l'entrepôt de données
- SVG [84] (Scalable Vector Graphic) pour l'affichage graphique des données géographiques.

En conclusion, cette architecture offre plus de souplesse que les architectures existantes :

- Elle ne dépend pas de la base de données de l'entrepôt grâce à l'utilisation des standards.
- Le format d'échange de données étant basé sur GML (proposition d'OpenGIS) [22], il devient plus facile d'intégrer de nouvelles sources sans recours à des traducteurs spécifiques.
- Le langage de requêtes est basé sur une extension du langage standard XQuery, adapté aux données géographiques, permettant ainsi d'effectuer des requêtes spatiales sur des documents GML.

III.3.2. Optimisation de requêtes GML

D'après notre synthèse de l'état de l'art, les techniques de cache sémantique sont les plus prometteuses. Nous avons donc étudié et expérimenté les techniques de cache sémantique pour l'optimisation de requêtes dans le contexte spécifique de données GML.

Quant à la politique de remplacement de cache, le coût et la taille sont des facteurs très importants car les données spatiales stockées dans des documents semi-structurés peuvent être de tailles importantes et le coût des requêtes spatiales peut vite devenir prohibitif. Il est donc important de considérer des politiques de remplacement de cache prenant en compte au minimum le coût des opérateurs spatiaux et la taille des objets en cache. Par ailleurs, la solution proposée par Brinkhoff [47] est un cache physique et sa politique de remplacement suppose l'utilisation d'index spatial. Or, un tel index n'existe pas dans des bases de données XML. Le parcours du document entier est, a priori, nécessaire pour résoudre une requête spatiale. Il faut donc étudier d'autres politiques de gestion et de remplacement de cache pour les données spatiales.

Enfin, comme dans toute base de données spatiale, la description géométrique est la partie la plus coûteuse dans le stockage des données GML. De plus, la même description géométrique est susceptible d'appartenir aux résultats de plusieurs requêtes. Afin de permettre une bonne exploitation de l'espace réservé au cache, un des objectifs est l'optimisation globale du coût de stockage du résultat des requêtes en cache.

Nous avons alors proposé dans [73][230] l'utilisation d'un cache sémantique pour une base de données en XML/GML. Une représentation spécifique du résultat des requêtes a été proposée afin de réduire l'espace mémoire des documents GML en cache. Une nouvelle politique de remplacement de cache a également été proposée. Cette dernière prend en compte le coût des opérateurs spatiaux et offre de meilleures performances que les algorithmes existant grâce à une meilleure exploitation du cache.

III.3.2.1. *Organisation du cache*

La représentation des données en cache est une structure facilitant l'identification d'objets. Cette structure est divisée en deux parties : la première est réservée aux caractéristiques non spatiales des données géographiques ; la seconde contient des fragments spatiaux dont les géométries sont distinctes.

Un cache sémantique comprend un *descripteur de région* décrivant chaque résultat de requête enregistré en cache et une *région physique* où sont enregistrées les données [52]. Dans le cas de requêtes géographiques, nous introduisons deux types de *régions physiques* : une *région physique non spatiale* et une *région physique spatiale*. La *région physique non spatiale* contient uniquement la partie non spatiale des résultats des requêtes géographiques. La *région physique spatiale* contient la partie spatiale de ces résultats. De plus, nous introduisons un *descripteur de région géographique* qui mémorise des informations sur les deux régions (spatiale et non spatiale).

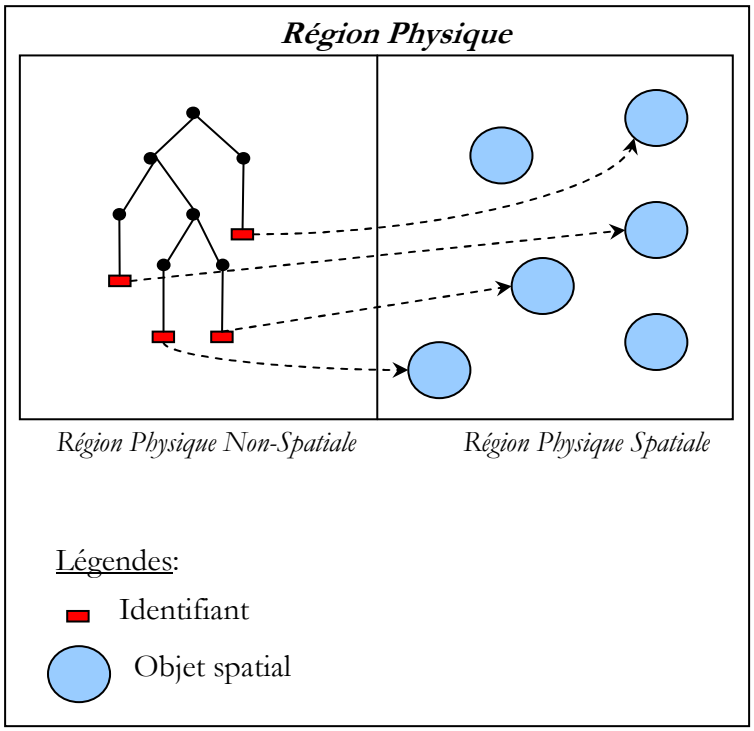


Figure 9. Organisation du cache

La partie non spatiale des résultats de requêtes est stockée dans le cache sous forme d'un arbre DOM [85]. Afin de relier la partie non spatiale à la partie spatiale du cache, un identifiant est généré. Chaque fragment spatial dans les données GML initiales est remplacé dans l'arbre DOM par l'identifiant associé dans (Figure 9).

La partie spatiale du cache contient une structure spécifique permettant d'une part de ne stocker que des géométries distinctes évitant ainsi les redondances, et d'autre part, d'accéder rapidement aux fragments spatiaux référencés par les identifiants. Plus exactement, l'accès à ces fragments est optimisé par l'utilisation d'une table de hachage (Figure 10). Cette table est constituée d'un identifiant ID, d'un

compteur et d'un pointeur vers le nœud racine de l'arbre DOM encodant un fragment spatial. Le compteur indique le nombre de références à un fragment spatial et permet de gérer son stockage en cache : le fragment est maintenu en cache tant qu'il est référencé, puis supprimé lorsque son compteur devient nul

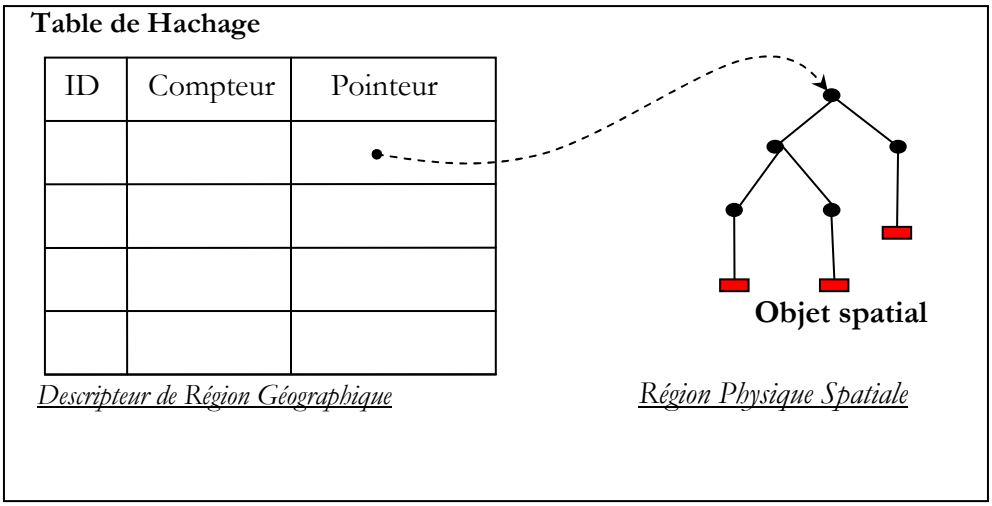


Figure 10. Partie spatiale du cache

III.3.2.2. *Traitement des requêtes*

Soit Q une nouvelle requête à traiter et C l'ensemble des requêtes contenues dans le cache. D'après Lee [64], il existe cinq cas généraux de configuration de Q par rapport à C : *exact match (exacte)*, *containing match (contenue)*, *contained match (contient)*, *overlapping match (recouvrement)* et *disjoint match (disjointe)*. La Figure 11 résume ces configurations où les parties en gris représentent les requêtes en cache et les blanches, la nouvelle requête Q à traiter. Dans les cas d'*exact match* et de *containing match*, Q peut être entièrement satisfaite depuis le cache. Mais dans les autres cas, un accès à la base de données est nécessaire afin de récupérer les données manquantes non présentes en cache (*remainder query*) [55].

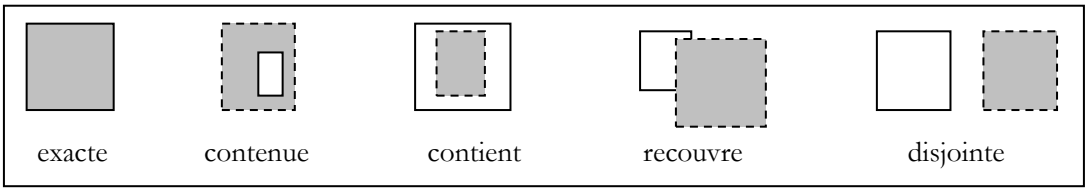


Figure 11. Configurations des requêtes

S'agissant de requêtes spatiales, déterminer la configuration entre Q et le cache revient à analyser les liens entre les critères spatiaux utilisés dans Q et ceux des requêtes présentes en cache. Pour ce faire, nous exploitons les propriétés d'inclusion entre les prédicats spatiaux. Ainsi, en nous basant sur les opérateurs du standard Open GIS [21], nous définissons les règles d'inférence illustrées dans la Figure 12 où les flèches traduisent les implications.

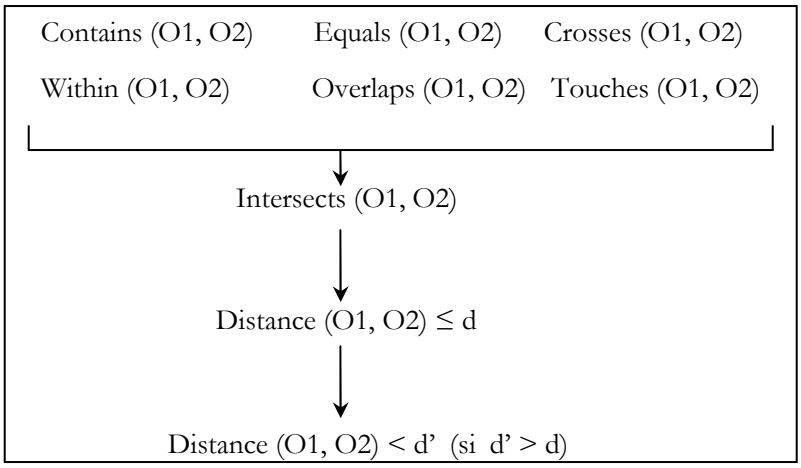


Figure 12. Règles d'inférences spatiales

Par exemple, la règle :

Intersects (O1, O2) => Distance (O1, O2) ≤ d ,

signifie que si O1 et O2 ont une intersection non vide, ils vérifient forcément le prédicat de distance quelque soit cette distance. De même, chacun des six prédicats topologiques placés en haut de la Figure 12 implique le prédicat d'intersection.

Le processus général de traitement de requêtes géographiques est décrit dans la Figure 13. L'analyseur *de requêtes* détermine la configuration de la requête soumise Q par rapport au cache. Le *reformulateur* de requêtes réécrit la requête Q en deux requêtes : une requête pour récupérer les données du cache (*requête en cache*) et une autre pour récupérer les données manquantes dans la base de données (*requête sur disque*).

Dans le cas de requêtes géographiques, nous introduisons deux types de *requête en cache* : une *requête en cache standard* permettant de récupérer les données contenues en cache ; puis une *requête en cache optimisée* permettant de récupérer des *données spatiales temporaires* pour une optimisation ultérieure.

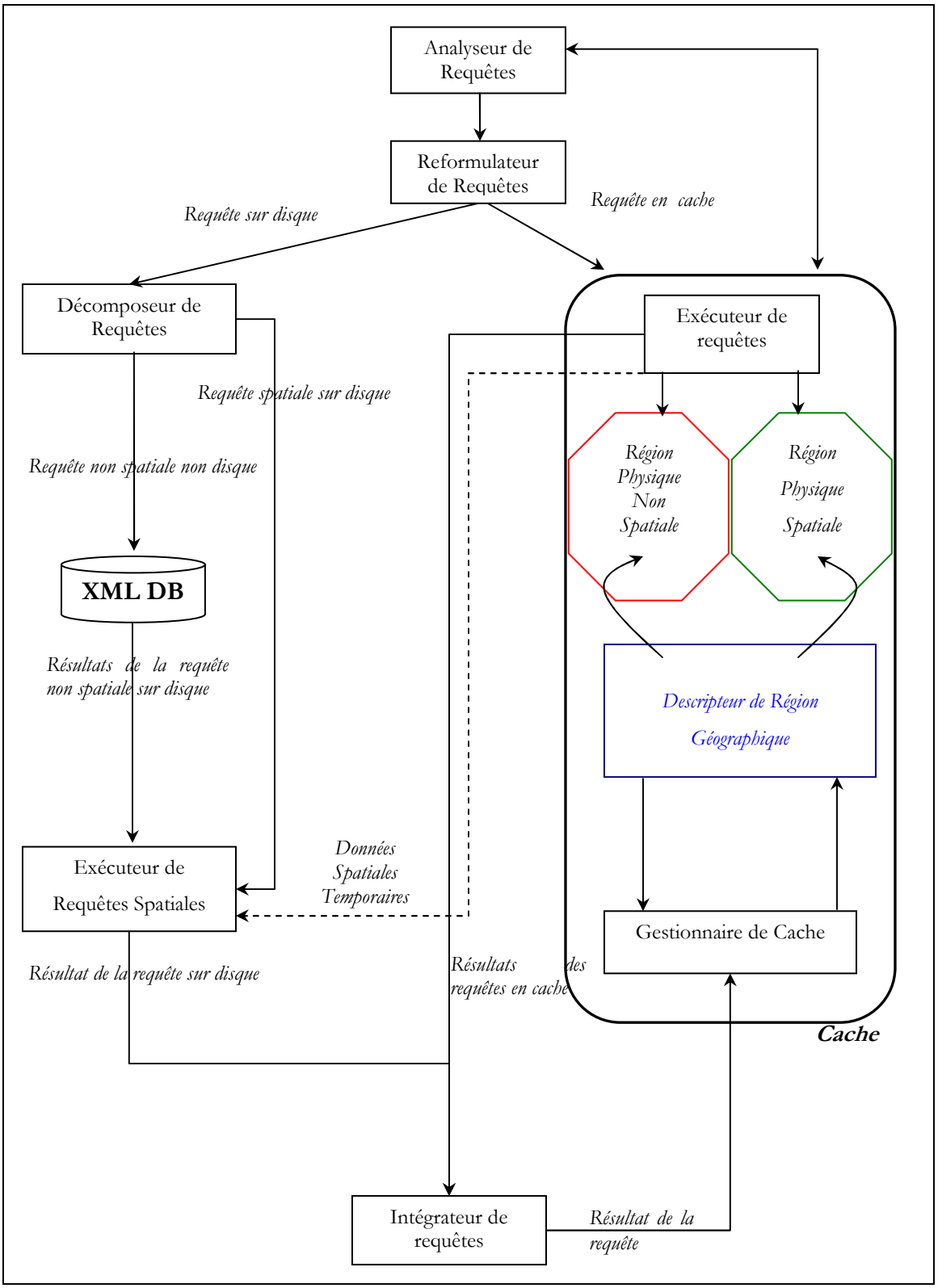


Figure 13. Processus de traitement de requêtes géographiques

La *requête sur disque* est décomposée à son tour par le *décomposeur de requêtes* en deux sous-requêtes : une *requête non spatiale sur disque* et une *requête spatiale sur disque*.

La *requête non spatiale sur disque* permet de récupérer les données non spatiales spécifiées par la requête Q et qui ne sont pas contenues en cache. La *requête spatiale sur disque* inclut les prédicats spatiaux spécifiés dans la requête Q et est traitée par *l'exécuteur de requêtes spatiales*. Si une optimisation est possible, alors les *données spatiales temporaires* sont utilisées afin d'optimiser le temps de traitement de la *requête spatiale sur disque*.

L'*intégrateur de requêtes* a pour rôle de fusionner les résultats des *requêtes sur disque* et en *cache* en un unique document GML résultat de la requête. Ce résultat est alors transmis au *gestionnaire de cache* qui, en fonction de la politique de remplacement, détermine son insertion ou non en cache et les objets éventuels du cache à supprimer.

III.3.2.3. *Politique de replacement : B&B$_{GDSF}$*

Lorsque le résultat d'une requête doit être stocké dans un cache saturé, les résultats de requêtes en cache les moins utiles doivent être supprimés. Il est important de prendre en compte les contraintes sur les tailles, mais aussi sur les coûts des requêtes géographiques et les fréquences d'accès aux objets du cache. Pour cela, nous avons proposé une politique de remplacement combinaison de l'algorithme GDSF [43] et de l'algorithme Branch and Bound [63]. Baptisée B&B$_{GDSF}$ [230], celle-ci comprend deux étapes :

1. Durant la première étape, le poids (la rentabilité) de chaque résultat de requête contenu dans le cache est calculé en utilisant la formule d'Arlitt (formule 1).

$$C_{Cache\ i} = F_i * C_i / T_i + L \qquad \text{(formule 1)}$$

   avec :

   - F_i la fréquence d'utilisation de l'objet à insérer *i*.
   - C_i le coût associé à l'objet *i* du cache.
   - T_i la taille du document *i*.
   - *L* est un facteur d'inflation qui est mise à jour à chaque fois q'un objet est supprimé.

2. Dans la deuxième étape, l'algorithme Branch and Bound (B&B) de recherche opérationnelle est utilisé pour supprimer les objets du cache les moins utiles. Il permet la recherche de solution optimale satisfaisant deux contraintes : la première est que l'ensemble des objets à supprimer a un coût minimal en cache et un coût minimal pour être retrouvé en dehors du cache (formule 2) ; la seconde contrainte est que la taille de cet ensemble est la plus petite possible, mais égale ou supérieure à la taille T du résultat de la requête à insérer (formule 3).

$$MIN_{i=1..n} \left(\sum_{j=1}^{n} X_j * C_{Cache\ j} + \sum_{j=1}^{n} (1-X_j) * C_{Disk\ j}\right) \qquad \text{(Formule 2)}$$

   où :

   - $(Xi)_{i=1..n} \in \{0 ; 1\}$ tel que $X_i = 1$ si le document i est conservé dans le cache, 0 s'il est supprimé.
   - $C_{Disk\ j}$ est le coût d'accès au document j sur disque (lorsqu'il n'est pas en cache).

$$\sum_{i=1}^{n} X_i * T_i \geq T \qquad \text{(Formule 3)}$$

Pour l'estimation de C_i, nous avons proposé un modèle de coût prenant en compte traitement des données non spatiales et spatiales, ce qui n'est pas le cas des politiques de remplacement de cache existantes généralement été étudiées pour des requêtes classiques.

Les résultats expérimentaux effectués montrent que l'algorithme $B\&B_{GDSF}$ permet une meilleure gestion de cache que les algorithmes existants. La comparaison a concerné les politique LRU, LFU, B&B seul et GDSF seul. Les résultats de ces comparaisons sont donnés dans [230].

III.3.3. Intégration de formats hétérogènes

Notre but est d'importer différentes sources de données spatiales tout en minimisant la tâche de conversion et sans altérer la richesse du modèle spatial.

L'intérêt de la prénorme CEN/TC287 [4] est de définir un modèle général qui peut être spécialisé et se décliner en différents modèles spatiaux. Partant de cette idée de modèle général et de spécialisations, nous avons développé un modèle ouvert (Figure 14) au sein même d'un serveur de données appelé GEOS.

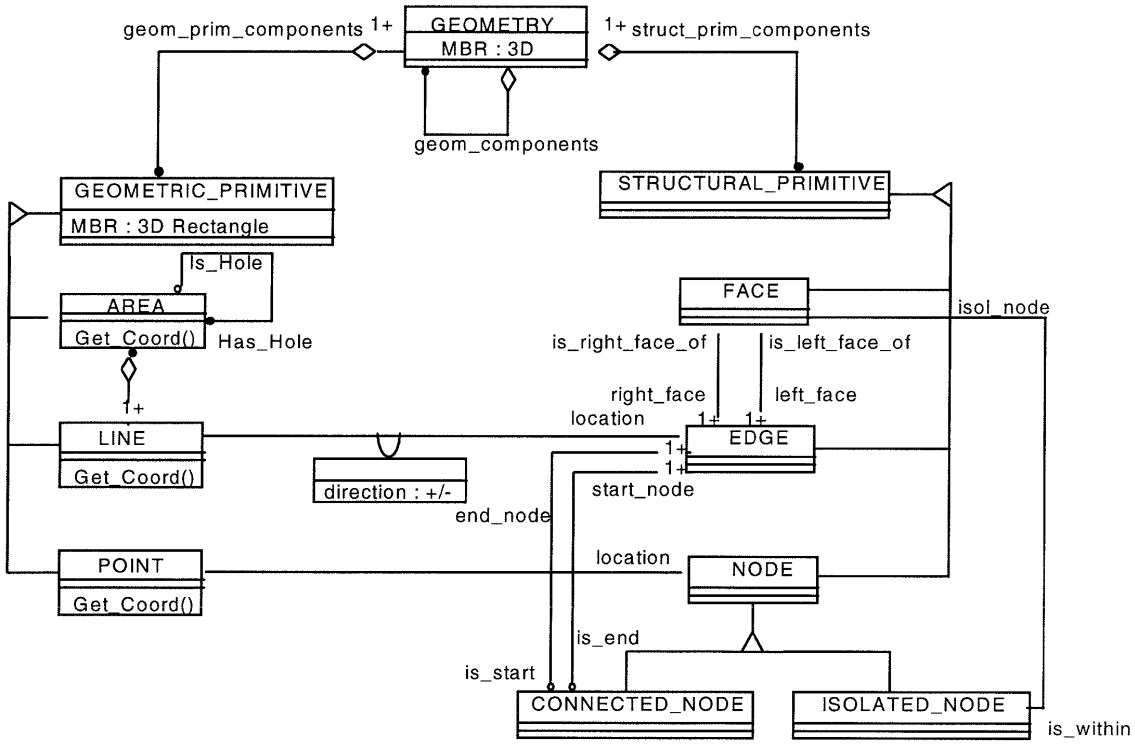


Figure 14. Représentation multiple de la composante spatiale dans GEOS

Afin de pouvoir paramétrer le modèle, nous proposons un ensemble de profils prédéfinis et un méta modèle permettant au développeur de dériver ses propres profils. GEOS permet l'adaptation à un format du plus structuré ou topologique au moins structuré et facilite l'intégration de formats hétérogènes grâce à la puissance du méta modèle spatial. Ce serveur est muni d'opérateurs spatiaux dont l'implémentation diffère selon le modèle spatial (avec ou sans topologie par exemple), ce qui permet de fournir les seules

données intéressant l'utilisateur et non tout le lot de données. Ce travail a donné lieu à une publication [2] et à un prototype basé sur un serveur d'objets persistants.

Aujourd'hui un serveur XML natif serait le plus adapté pour le développement d'un tel système ouvert, en se basant sur GML. Celui-ci intègre de fait les méta-données et permet leur spécialisation.

III.3.4. Appariement des données géographiques

La qualité d'un entrepôt de données géographique dépend grandement du processus d'intégration. Si cette intégration est basée sur la géométrie des objets et particulièrement dans le cas de cartes de réseaux, les écarts entre les lignes et les nœuds de sources différentes rendent difficile leur appariement. Ces écarts résultent des diverses formes et échelles de représentation d'un même réseau selon la source et le type d'exploitation visé. L'exemple le plus commun est un plan de métro où les stations sont localisées avec une relative précision mais les liaisons sont simplifiées par des segments de droite. Mais, pour l'exploitation et l'entretien des voies, les courbures précises ont de l'importance. De même, dans un réseau de transport routier, la gestion de l'infrastructure et la navigation GPS exigent la connaissance détaillée du réseau routier, tandis que la gestion du trafic se limite aux principaux carrefours et leurs liaisons simplifiées en segments de droites. Une liaison correspond à chemin reliant ces carrefours.

Ce problème d'appariement est assez fréquent. Sa spécificité, par rapport aux méthodes de l'état de l'art, est de considérer l'hypothèse de correspondance 1:N où un segment d'une source (ici Trafic) est en correspondance avec une séquence de (N) segments de l'autre source (réseau détaillé).

L'algorithme que nous proposons [230] [72] permet de résoudre ce genre de situations en combinant la similarité de la distance et de la forme. Il utilise le filtrage et le lissage afin de réduire le nombre de candidats avant le calcul de similarité selon la mesure proposée. Ces phases préliminaires sont essentielles à la fois pour améliorer la qualité des résultats et pour l'optimisation des performances d'exécution. L'algorithme se compose ainsi des cinq phases suivantes :

1. filtrage sémantique : par exemple, sachant que la constitution de la carte de trafic exclut les axes secondaires et les voies de desserte, l'utilisateur peur filtrer le réseau détaillé en écartant les voies correspondantes ;
2. appariement des nœuds selon un périmètre paramétré par l'utilisateur : tous les nœuds des polylignes du réseau détaillé se trouvant à l'intérieur des nœuds de début et de fin sont sélectionnés ;
3. délimitation de l'espace de recherche à une zone tampon (ou *buffer* en terme SIG) autour du segment à apparier ;
4. lissage des polylignes par réduction de leur sinuosité ;
5. appariement des segments avec les polylignes résultant de la phase précédente en utilisant la mesure proposée DF.

Mesure proposée DF

Pour pallier les limites des méthodes d'appariement de l'état de l'art, nous avons proposé une mesure adaptée *DF* pour Distance et Forme, car elle prend en compte à la fois la distance entre les objets à apparier et leur forme générale. Elle est définie comme la somme pondérée entre la distance métrique et la distance angulaire (de forme).

L'algorithme a été implémenté et testé sur des données réelles représentant des segments de trafic et le réseau routier de l'agglomération lilloise. Ces données comprennent 36709 polylignes pour le réseau routier et 4308 segments de trafic. Les résultats ont montré que :

- après les trois premières phases de filtrage, l'algorithme résout l'appariement de 67% du réseau ;
- l'étape de lissage et la mesure de distance définie permettent de résoudre l'appariement de 18% d'objets supplémentaires ;
- le reste (près de 15%) ne sont pas appariés mais correspondent à des cas non décidables à l'oeil nu, souvent liés à une des données erronées.

La méthode proposée s'avère efficace pour une classe de problèmes de plus en plus courante pour des données géo localisées multi-échelles. Il permet ainsi de résoudre efficacement des cas complexes qui n'étaient pas résolus par les travaux antérieurs. Il permet en outre de pointer les anomalies dans les données en rejetant certains objets. Dans notre exemple, nous avons détecté que certains segments de trafic ne correspondaient à aucun chemin du réseau routier. Pour plus de détail, le lecteur peut se référer à la thèse de Savary [230] ou l'article présentant cette contribution [72].

III.4. Conclusion

Nous avons proposé une architecture d'entrepôt basée sur XML/GML et étudié l'optimisation de requêtes spatiales sur GML par la gestion d'un cache sémantique. L'architecture propose le stockage natif de l'entrepôt en GML, des adaptateurs associés pour des sources SIG ou SGBD, des requêtes XQuery avec une intégration d'opérateurs spatiaux du standard Open GIS et des communications par flux SAX. Cette nouvelle architecture nous a amenés à étudier la résolution efficace de requêtes sur des documents GML. Nous avons contribué par la proposition d'un cache sémantique adapté à GML.

Nous avons étudié l'intégration de sources de formats hétérogènes et proposé un modèle spatial paramétrable qui peut s'adapter aux formats des sources. Concernant le problème de *map matching*, nous avons proposé une nouvelle méthode d'appariement d'objets basée sur la similarité de leurs géométries. Cette méthode est flexible dans le sens qu'elle permet à l'utilisateur de spécifier une pondération entre la similarité des formes et la distance. Elle tient compte également de critères de filtrages sémantiques. Elle a été testée avec succès dans l'appariement de deux représentations du réseau routier : l'une détaillée (comprenant tous les tronçons et leurs géométries précises) et l'autre partielle (certains axes et carrefours principaux) et simplifiée (seul les carrefours sont localisés).

III.5. Références

[42] Abrams M, Standbridge C.R, Adbulla G, Williams S, Fox E.A. (1995). Caching Proxies: Limitations and Potentials. World Wide Web Conference (WWW'95), Boston, December, 1995.

[43] Arlitt M, Friedrich R, Cherkasova L, Dilley J, Jin T. (1999). Evaluating Content Management Techniques for Web Proxy Caches. In HP Technical report, Palo Alto, April, 1999.

[44] Beckmann N, Kriegel H.P, Schneider R, Seeger B. (1990). An Efficient and Robust Access Method for Points and Rectangles. In: Proceeding ACM SIGMOD International Conference on Management of Data. Atlantic City, NJ. 1990, pp. 322-331.

[45] Boucelma O, Messid M, Lacroix Z. (2002). A WFS-based Mediation System for GIS Interoperability. 10th ACM International Symposium on Advances in Geographic Information Systems (GIS), McLean, Virginia, November, 2002.

[46] Branki T, Defude B. (1998). Data and Metadata: Two-Dimensional Integration Heterogeneous of Spatial Databases. Spatial Data Handling Conference Proceedings, Vancouver, BC, Canada, July, 1998, pp. 172-179.

[47] Brinkhoff T. (2002). A Robust and Self-tuning Page-Replacement strategy for Spatial Database Systems. In proceedings of the 8th International Conference on Extending Database Technology (EDBT'02), Prague, Czech Republic, 2002. Lecture Notes in Computer Science, Vol. 2287, Springer-Verlag, pp.533-552.

[48] Brown J, Rao A, Baran J. (1995). Automated GIS Conflation: Coverage Update Problems and Solutions. Proc. of Geographic Information Systems for Transportation Symposium (GIS-T'95). American Association of State Highway and Transportation Officials, Sparks, Nevada, 1995, pp. 220-229.

[49] Cao P, Irani S. (1997). Cost-Aware WWW Proxy Caching Algorithms. Proceedings of USENIX Symposium on Internet Technologies and Systems (USITS), Monterey, CA, December, 1997, pp. 193-206.

[50] Carey M.J, Franklin M.J, Zaharioudakis M. (1994). Fine-Grained Sharing in a Page Server". In Object Oriented Database Management System (SIGMOD OODBMS'94), Minneapolis, Minnesota, 1994, pp. 359-370.

[51] Chen L, Wang S, Rundensteiner E.A. (2004). Replacement Strategies for XQuery Caching Systems. Data Knowledge Engeneering, 2004, pp. 145-175.

[52] Chidlovskii B, Roncancio C, Schneider M-L. (1999). Semantic CacheMechanism for Heterogeneous Web Querying. Proceedings of the 8th World-Wide WebConference (WWW8), 1999.

[53] Chou H, DeWitt D. (1985). An Evaluation of Buffer Management Strategies for Relational Database Systems. Proceedings of the 11th VLDB Conference, 1985. DeWitt 90];

[54] Corocoles J.E, Gonzalez P. (2003). Querying Spatial Resources. An Approach to the Semantic Geospatial Web. In the 15th Conference on Advanced Information Systems Engineering (CAiSE '03) Workshop, Web, e-Business, and the Semantic Web WES: Foundations, Models, Architecture, Engineering and Applications, LNCS, Springer-Verlag, 2003.

[55] Dar S, Franklin M.J, Jonsson B.T, Srivastava D, Tan M. Semantic Data Caching and Replacement. In Proceedings of the 22 VLDB Conference, Mumbai (Bombay), India, 1996.

[56] Devogele T. (1997). Processus d'intégration et d'appariement de bases de données géographiques. Application à une base de données routières multi-échelles. Thèse de doctorat, Université de Versailles Saint-Quentin-en-Yvelines, 12 décembre 1997.

[57] Devogele T, Parent C, Spaccapietra S. (1998). On Spatial Database Integration. International Journal of Geographic Information Systems, Special Issue on System Integration, Vol. 12, No 3, 1998.

[58] Devogele T. A new Merging Process for Data Integration Based on The Discrete Fréchet Distance. 10th International Symposium on Spatial Data Hangling, ISBN 3-540-43802-5 Springer-Verlag. Ottawa, Canada, July 9-12, 2002, pp. 167-181.

[59] DeWitt D, Futtersack P, Maier D, Velez F. (1990). A study of Three Alternative Workstation-Server Architectures For Object-Oriented Database Systems. In VLDB, Queesnland, Australia, August, 1990, pp. 107-121.

[60] Gabay Y, Doytsher Y. (1994). Automatic Adjustment of Line Maps. In proceedings of GIS/LIS 94. Bethesda: ACSM-ASPRS-AAG-URISA-AM/FM, 1994, pp. 332-340.

[61] Godfrey P, Gryz J. (1997). Semantic Caching for Heterogeneous Databases. In Proceedings of KRDB at VLDB, Athens, Greece, August, 1997, pp. 61-66.

[62] Haas L.M, Kossmann D., Ursu I. (1999). Loading a Cache with Query Results. In proceedings of the 25th VLDB Conference, Edinburgh, Scotland, 1999, pp. 351-362.

[63] Kellerer H, Pferschy U, Pisinger D. (2004). Knapsack Problems. Springer, ISBN: 3-540-40286-1, 2004, 546 pages.

[64] Lee D, Chu W.W. (1999). A Semantic Caching via Query Matching for Web Sources. In Proceeding of the 8th ACM International Conference on Information and Knowledge Management (CIKM'98), Kansas City, Missouri, USA, November, 1999.

[65] Lorenzetti P, Rizzo L, Vicisano L. (1996). Replacement Policies for a Proxy Cache. Report, Universita di Pisa, December, 1996.

[66] Mustière S., "Measure of linear generalisation quality". Rapport de satge, DESS Cartographie, Université Paris I, COGIT, 1995.

[67] O'Neil E. J, O'Neil P. E, Weikum G. (1993) . The LRU-K Page Replacement Algorithm for Database Disk Buffering. In SIGMOD, 1993, pp. 297-306.

[68] Park J. (2001). Schema Integration Methodology and Toolkit for Heterogeneous and Distributed Geographic Databases. Journal of the Kore Industrial Information Systems Society, Vol.6, 3 September 2001, pp. 51-54.

[69] Pendyala R.M. (2002). Development of GIS-Based Conflation Tools for Data Integration and Matching. Final Report: Executive Summary. Research Center, Florida Department of Transportation, 605 Suwannee Street, MS 30 Tallahassee, FL 32399-0450, 2002.

[70] Savary L., Zeitouni K., "Spatio-Temporal Data Warehouse - a Prototype", In Second Electronic Government (EGOV) Conference, Joint conference to DEXA , Prague Czech Republic, Lecture Notes in Computer Science, Springer-Verlag, Volume 2739, September 1-5, 2003, pp. 335 - 340.

[71] Savary L., Gardarin G., Zeitouni K., "Entrepôt de données spatiales basé sur GML : politique de gestion de cache", Poster aux 5èmes Journées d'Extraction et de Gestion des Connaissances, EGC 2005, Edition CEPADUES, Paris, Janvier 2005.

[72] Savary L., Zeitouni K., "Automated Linear Geometric Conflation for Spatial Data Warehouse Integration Process", 8th AGILE International Conference on Geographic Information Science, ISBN 972-8093-13-6, May 26-28, 2005, Estoril, Portugal, pp. 23-32.

[73] Savary L., Gardarin G., Zeitouni K., GeoCache: A Cache for GML Geographical Data, International Journal of Data Warehousing and Mining (IJDWM), Idea Group Publishing, to appear.

[74] Simple API for XML (SAX 1.0). Released on May 11, 1998. http://www.saxproject.org.

[75] Service Center Implementation. (2001). Implementation of Geospatial Data Warehouses II. Prepared by Science Applications International Corporation (SAIC) For the Service Center Initiative, Data Management Team, United Sates Department of Agriculture (USDA), 2001.

[76] Theodoridis Y, Sellis T. (1996). A Model for the Prediction of R-Tree Performance. Proc. 15th ACM Symp. Principles of Database Systems, 1996, pp. 161-171.

[77] Vatsavai R. (2002). GML-QL: A Spatial Query Language Specification for GML. Department of Computer Science and Engineering, University of Minnesota, 2002,

[78] Vivid Solutions (2006). Java Topology Suite (JTS), http://www.vividsolutions.com/jts/jtshome.htm (accessed in September 2006).

[79] Vivid Solutions (2006). Java Conflation Suite (JCS), http://www.vividsolutions.com/products.asp (accessed in September 2006).

[80] Voisard A, Juergens M.(1999). Geographic Information Extraction: Querying or Quarrying? In Interoperating Geographic Information Systems, Goodchild M and Egenhofer M and Fegeas R and Kottman C (Eds.), Kluwer Academic Publishers, New York, 1999.

[81] World Wide Web Consortium (W3C). (2000). Simple Object Access Protocol (SOAP) 1.1. W3C Note 08 May 2000. http://www.w3.org/TR/2000/NOTE-SOAP-20000508/.

[82] Extensible Markup Language (XML) 1.0 (Second Edition), W3C Recommendation 6 October 2000, http://www.w3.org/TR/2000/REC-xml-20001006.

[83] World Wide Web Consortium (W3C). XQuery 1.0: An XML Query Language W3C Working Draft 20 December 2001.

[84] World Wide Web Consortium (W3C). (2002). Scalable Vector Graphics (SVG). http://www.w3.org/Graphics/SVG/ (accessed in March 2002).

[85] World Wide Web Consortium (W3C). (2004). Document Object Model (DOM). Level 3 Validation Specification Version 1.0 W3C Recommendation, 27 January, 2004.

[86] Xiong D, Sperling J. (2004). Semiautomated matching for network database integration. ISPRS Journal of Photogrammetry & Remote Sensing 59. pp. 35-46, 2004.

[87] Yang Q, Zhang H, Zhang H. (2003). Taylor Series Prediction: A Cache Replacement Policy based on Second-Order Trend Analysis. 34th Annual Hawaii International Conference on System Sciences (HICSS-34)- Maui, Hawaii, January 03-06, 2003, Vol.5.

[88] Zhang M.S, Javed A, and Gruenwald L. (2000). A Prototype for Wrapping and Visualizing Geo-Referenced Data in Distributed Environments Using the XML Technology. ACM GIS 2000, USA, 2000, pp.27-32.

CHAPITRE IV. ANALYSE EN LIGNE DES DONNEES SPATIOTEMPORELLES

Ce chapitre présente nos contributions pour l'adaptation et pour l'extension des techniques OLAP aux données spatiotemporelles et particulièrement pour des objets mobiles. Il correspond aux travaux de thèse de Tao Wan qui ont servi, en partie, dans le projet Européen HEARTS.

Ce chapitre est organisé comme suit. La première section détaille les problématiques. La deuxième section fait une synthèse de l'état de l'art focalisé sur l'analyse multidimensionnel des objets spatiotemporels et mobiles. Ensuite, les sections suivantes relatent nos travaux. Un bilan et des perspectives de ces travaux concluent ce chapitre.

IV.1. Problématique

La modélisation multidimensionnelle dans les systèmes OLAP (On-Line Analytical Processing) est indéniablement un facteur clé du succès des entrepôts aujourd'hui. En effet, elle permet l'analyse en ligne interactive, flexible et multi-niveaux des données et produit efficacement de nombreux tableaux statistiques. Comme dans les entrepôts de données classiques, il serait intéressant d'explorer toutes les dimensions des données spatiales et spatiotemporelles et de générer des agrégats à différents niveaux de détail. Or, les modèles actuels ne permettent d'explorer que des dimensions qualitatives classiques et ne peuvent prendre en compte la dimension ni le raisonnement spatial et spatiotemporel.

Prenons comme exemple une application type en santé environnementale. Dans cette application, l'expert en santé environnementale cherche à estimer l'exposition aux risques de pollution sous différents angles : risque individuel, par groupe (ex. d'âge), par lieu quelconque et/ou par période. Cela mène à des requêtes du style :

Q1 : Quel est le nombre de personnes exposées au cours du temps au champ de pollution ?

Q2 : Quel est le nombre de personnes par catégorie d'âge exposées au champ de pollution ?

Ces requêtes impliquent principalement des agrégations groupées par localisation, intervalle de temps et/ou attributs d'objets mobiles (ex : âge). Ceci est analogue à l'analyse multidimensionnelle dans les systèmes OLAP. Mais, contrairement au modèle traditionnel, ce type de requêtes concerne des dimensions spatiales et temporelles qui ne sont pas de simples données qualitatives et se basent sur des critères d'intersection spatiotemporelles (ici entre la trajectoire des personnes et les zones de pollution). De plus, les trajectoires mobiles forment une variation continue selon les dimensions temps et espace. Or, les modèles multidimensionnels conventionnels sont tous basés sur des faits et des dimensions discrets. Il est donc nécessaire d'étendre les modèles OLAP pour le support des applications spatiotemporelles décisionnelles. Cette extension doit en outre offrir des techniques d'implémentation et de requêtes d'agrégation de données efficaces.

IV.2. Etat de l'art

Traditionnellement, les entrepôts de données sont conçus dans un modèle multidimensionnel [96] et définissent des cubes de données que les outils OLAP permettent d'explorer aisément [89]. Le paradigme multidimensionnel est intéressant pour l'exploration des données spatiotemporelles, mais il ne peut s'appliquer tel quel car il n'intègre pas le raisonnement spatiotemporel supporté par les requêtes spatiales et/ou temporelles.

On distingue trois types de travaux liés. Le premier concerne les entrepôts de données spatiales et spatiotemporelles. Le second type correspond aux travaux sur les bases d'objets mobiles. Enfin, le troisième correspond à une forme d'entrepôts d'objets mobiles.

Il existe des travaux tant sur les entrepôts de données spatiaux, [109] étant pionnier, que sur les entrepôts spatiotemporels [101]. Bien qu'ils proposent l'intégration des caractéristiques spatiales ou spatiotemporelles dans un entrepôt, seules les attributs des objets peuvent varier dans le temps. Par conséquent, le cas des objets mobiles n'y est pas considéré.

La plupart des travaux sur les bases d'objets mobiles s'inscrivent dans un contexte transactionnel (cf. chapitre II) et ne considèrent pas l'exploitation d'historiques d'objets mobiles et encore moins les requêtes de type OLAP.

Le premier travail lié à l'entreposage d'objets mobiles celui de Papadias et al. [104]. Les auteurs traitent un modèle multidimensionnel se référant aux objets mobiles où une cellule mesure l'effectif d'objets mobiles dans une unité espace-temps. Ils proposent un index, combinaison du *R-tree* et du *B-tree*, pour stocker l'historique de l'effectif par zone. Ces deux articles partent de l'hypothèse que les objets mobiles sont agrégés à l'origine et ne permettent pas de représenter d'autres dimensions que l'espace et le temps. Limités par cette hypothèse, ces modèles ne peuvent ni représenter les trajectoires d'objets individuels ni les attributs descriptifs (ex : âge, sexe). De plus, le résultat de leurs requêtes OLAP ne produit que des

statistiques approximatives, bien que l'article de Tao et al. [111] tente d'améliorer l'approximation. Dans cette meme catégorie, Jensen et al. [98] traitent le cas particulier d'inclusion partielle entre hiérarchies spatiales. Ils proposent différentes techniques pour réduire l'imprécision des agrégats.

IV.3. Contributions

Motivés par les limitations des travaux existants, nous avons proposé de nouvelles approches de modélisation des objets mobiles dans un entrepôt de données pour une exploration de type OLAP.

Dans un premier temps, nous avons proposé une modélisation multidimensionnelle des objets mobiles basée sur des découpages de référence pour l'espace et le temps [107][112]. Le modèle représente de manière discrète la mobilité d'un objet et permet d'intégrer ses attributs (ex : l'âge) dans l'analyse OLAP. L'autre avantage de cette approche est de s'adapter aux modèles d'entrepôts classiques comme le modèle en flocon et de permettre l'utilisation d'outils OLAP classiques. Cependant, cette modélisation ne permet pas de répondre avec précision aux agrégats par requêtes spatiales et/ou temporelles à cheval sur les découpages de l'espace et du temps prédéfinis. Nous donnons une brève description de ce modèle dans la prochaine section.

Dans un deuxième temps, nous avons proposé un nouveau modèle d'entrepôt introduisant des notions de dimensions et de faits continus. Ce modèle permet la définition d'un fait mobile sans discrétisation des dimensions spatiale et temporelle. Basés sur cette extension de la modélisation multidimensionnelle, nous avons proposé un modèle de représentation et d'indexation permettant le stockage et la résolution efficaces des requêtes OLAP et spatiotemporelles sur ces données.

IV.3.1.1. *Modèle multidimensionnel discret pour objets mobiles*

Une trajectoire forme, a priori, une variation continue dans l'espace et dans le temps. Or, dans un modèle multidimensionnel traditionnel, les valeurs des dimensions doivent se limiter à quelques modalités et être largement partagées par les faits représentés. Une représentation exacte de l'espace et du temps ne peut donc former une dimension dans les systèmes OLAP existants.

En se basant sur le modèle discret d'objet mobile (cf. chapitre II), il suffit de considérer un fait spatiotemporel comme l'association des dimensions attributaires D_i et de l'unité spatiale S et temporelle T correspondant :

FaitSpatioTemporel (#D1, #D2, …#Dn, #S, #T, MS, MT, MX, …)

Cette table de faits est associée aux tables de dimensions attributaires *Di (i=1..n)* de la dimension spatiale discrétisée et de la dimension temporelle discrétisée. Ces dimensions sont éventuellement décomposées à leur tour en d'autres tables décrivant des hiérarchies de dimensions. Des mesures peuvent être associées selon les applications. Par exemple, MS et MT peuvent décrire respectivement le taux de recouvrement de la trajectoire avec l'unité spatiale S et la durée effective par rapport à l'intervalle T. D'autres mesures MX peuvent être définies sur les autres dimensions.

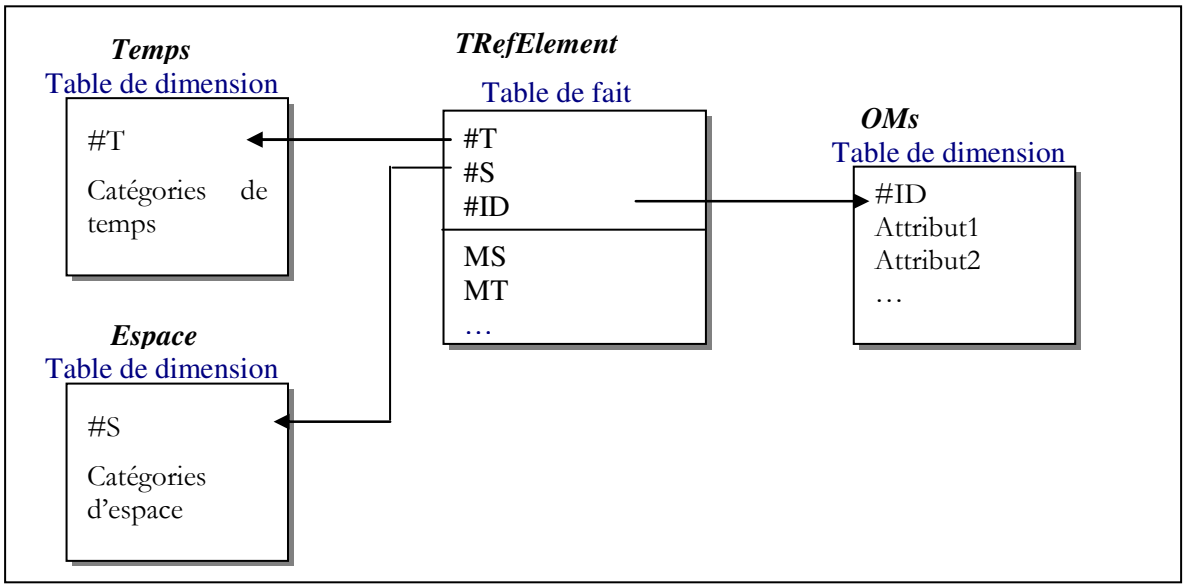


Figure 15. Schéma multidimensionnel spatiotemporel discret (modèle général)

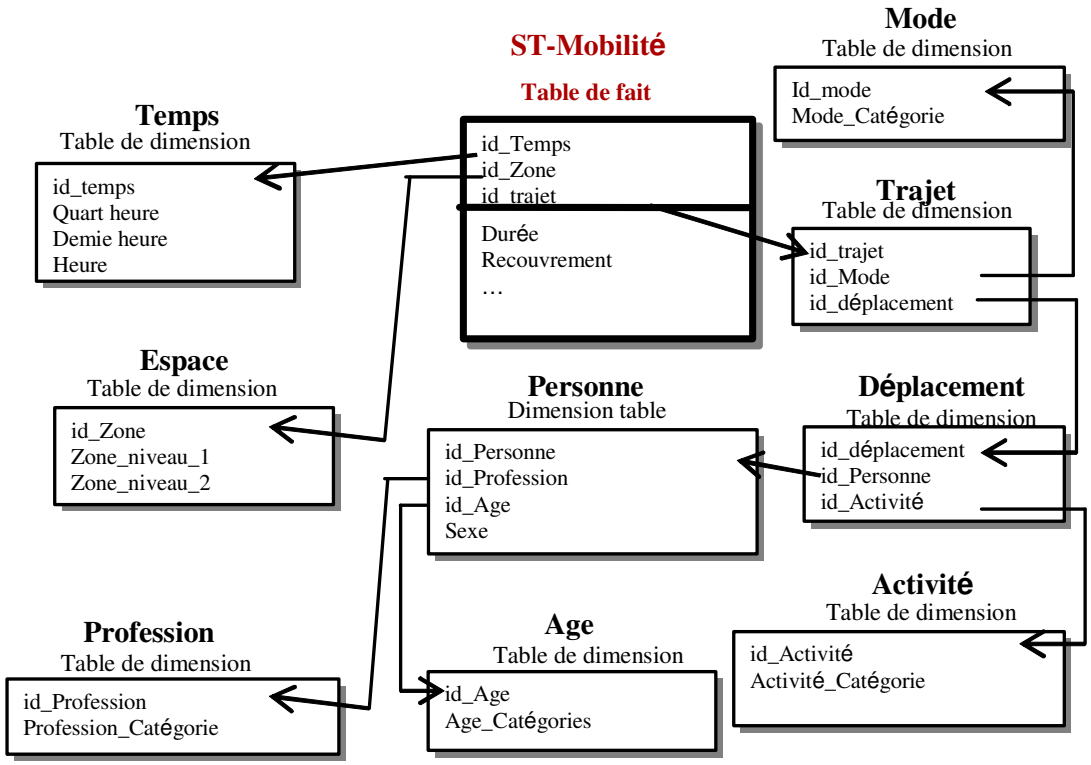


Figure 16. Schéma multidimensionnel spatiotemporel discret (application dans HEARTS)

Ce modèle correspond à un schéma OLAP classique, comme illustré dans la. Par conséquent, cette approche offre la puissance de l'algèbre OLAP sur toute combinaison de données classiques, spatiales et/ou temporelles et mobiles.

Cette approche a été validée par un prototype et a été appliquée à l'analyse de la mobilité urbaine. La Figure 16 illustre cette application. Les résultats de l'expérimentation montrent la validité de l'approche et les tests de performances son efficacité. Les détails de cette approche sont donnés dans [112] et dans le rapport de contrat PRISM-ONADA [116].

IV.3.1.2. *Modèle multidimensionnel continu pour objets mobiles*

Les requêtes Q1 et Q2, citées en exemple dans la problématique, exigent principalement des agrégations groupées par localisation, intervalle de temps et/ou attributs d'objets mobiles (ex : âge). Si l'on suppose les cartes de pollution données par des mesures par localisation et par intervalle de temps, ces requêtes correspondent, a priori, à un calcul d'agrégat d'objets mobiles, ici un comptage du nombre de personnes distinctes. En notant INTERSECT l'intersection spatiotemporelle, on pourrait exprimer ces requêtes en notation SQL comme suit :

```
SELECT      p.localisation, p.temps, COUNT (DISTINCT id_personne) AS nombre_personnes
FROM        population m, pollution p
WHERE       m.trajectoire INTERSECT (p.localisation, p.temps)
GROUP BY    p.localisation, p.temps
```

et :

```
SELECT      p.localisation, p.temps, COUNT (DISTINCT id_personne) AS nombre_personnes
FROM        population m, pollution p
WHERE       m.trajectoire INTERSECT (p.localisation, p.temps)
GROUP BY    m.âge, p.localisation, p.temps
```

Les entrepôts de données et les systèmes OLAP actuels ne s'appliquent pas à ce type d'analyse en raison des spécificités suivantes :

1. Les requêtes spatiotemporelles OLAP comme ci-dessus exigent le calcul d'agrégats par intervalles de temps et zones de l'espace. Ce résultat d'agrégation dépend de la trajectoire des objets mobiles qui forment une variation continue selon les dimensions temps et espace. Or, les modèles multidimensionnels conventionnels sont tous basés sur des faits ponctuels et n'intègrent pas des critères spatiotemporels ;

2. Les critères de regroupement (ici les champs de pollution) ne sont pas forcément connus à l'avance et ne peuvent se référer à un découpage spatial ou temporel prédéfinis. Par exemple, la mobilité à caractère périodique (stockée une fois pour toute) est croisée avec les champs de pollution relevés quotidiennement. Par conséquent, aucune discrétisation préalable de l'espace ou du temps ne serait satisfaisante et donc, leur caractère continu doit être pris en compte dans les dimensions et dans l'analyse ;

3. Il est indispensable d'intégrer dans le modèle les dimensions continues avec les dimensions discrètes comme le type d'activité pouvant être combinées lors de l'analyse. L'identité de l'objet mobile et ses attributs sont en effet des dimensions d'analyse essentielles, comme pour analyser les risques individuels ou les risques agrégés par âge.

C'est ce qui a motivé notre proposition d'un modèle d'entrepôt spécifique. Notre contribution se résume en trois points :

1. Pour décrire la variation continue de trajectoires d'objets mobiles, nous étendons les notions de fait et de dimension. Les dimensions sont étendues aux valeurs continues (ici, le temps ou l'espace sont des dimensions continues). Les faits sont étendus aux faits continus et ne sont plus obligatoirement rattachés à un évènement comme dans les modèles classiques. Un fait mobile est continu car c'est une variation continue du temps dans l'espace.
2. Nous dérivons une structure de données basée sur des intervalles pour représenter des faits mobiles.
3. Nous proposons et mettons en œuvre une structure d'index de type R-tree étendu et l'utilisons pour optimiser des requêtes spatiotemporelles OLAP.

De plus, comme la grande majorité des objets se déplacent dans un environnement géographique contraint par un réseau (les routes, les chemins de fer ou les couloirs aériens), nous focalisons notre travail sur la mobilité restreinte au réseau. Cette propriété de localisation relative au réseau nous permet de réduire la dimensionnalité et d'optimiser le stockage [25].

A notre connaissance, il n'existe dans la littérature aucune véritable approche pour l'analyse en ligne d'objets mobiles.

Extension du modèle multidimensionnel aux objets mobiles

Nous proposons d'étendre les notions de fait et de dimension à la variabilité continue. Pour cela, nous extensions suivantes.

Dimension continue : c'est une dimension variant dans un domaine continu sans limite de granularité.

Fait continu : c'est un fait qui varie continuellement dans l'espace des dimensions. Il dépend d'une ou de plusieurs dimensions continues et peut aussi dépendre des autres dimensions.

Un fait mobile est un cas particulier de fait continu comme décrit ci-dessous :

Fait mobile : c'est un fait qui varie selon une fonction $F_{OM}(t) = s$ continue du temps dans l'espace pour un objet mobile donné OM.

En fait, la fonction F_{OM} dépend des deux dimensions continues « espace » et « temps ». Elle correspond aux trajectoires de l'objet. Dans le cas des objets mobiles contraints, ces trajectoires peuvent être représentées en 2D avec les dimensions *TR* et le temps. Cette représentation du cube est illustrée dans Figure 17. Soient les quatre objets mobiles associés à Marie, Fred, Tao et Jack. Dans ce modèle, les trois dimensions représentent l'objet par son identité, la dimension spatiale continue « réseau routier » et la dimension temporelle continue.

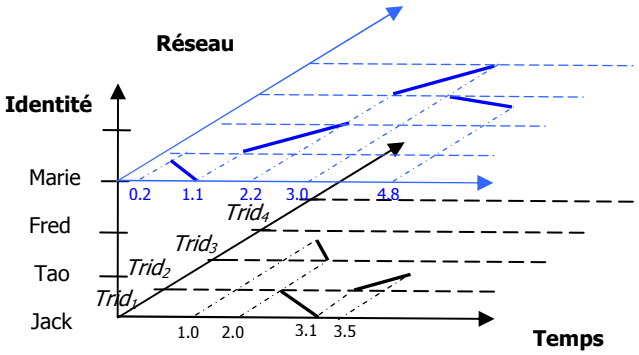


Figure 17. Schéma multidimensionnel spatiotemporel continu

Implémentation et optimisation des requêtes OLAP pour objets mobiles

Au vu des exemples de requêtes Q1 et Q2 donnés en référence, le problème posé maintenant est de traiter efficacement l'agrégation par intervalle ou par combinaison d'intervalles et de dimensions discrètes. La combinaison d'intervalles traduit ici des requêtes spatiotemporelles, tandis que les dimensions discrètes expriment l'exploration par les attributs de l'objet. De plus, ces requêtes par intervalles portent sur des segments de droite et nécessitent le support de requêtes spatiales 2D.

Pour ces requête OLAP, dites « *range query* » ou « *requêtes par intervalles* », des travaux ont été proposés exploitant le stockage d'agrégats dans des index *R-tree* comme dans [97] et dans [99]. Mais ces travaux se limitent aux fonctions agrégats algébriques ou distributives. Or, les requêtes, telles Q1 et Q2 données en exemple, utilisent une fonction agrégat « *count distinct* » holistique et combinent parfois les attributs de l'objet. Stocker simplement le résultat de la fonction agrégat dans l'index comme dans ces propositions ne suffit pas car, comme dans [104], il ne maintient pas les références aux objets pour adjoindre leurs propriétés et engendre le double comptage des objets dans l'agrégat.

Afin d'optimiser tout type de requêtes agrégat et spatiotemporelles, nous proposons donc une structure d'index appelée *TTR-tree* (ou *Transformed Trajectory R-tree*) extension du *R-tree*. En effet, les nœuds comprennent, en plus des informations normales d'un R-tree, un bitmap représentant les objets mobiles indexés par ce nœud. La Figure 18 illustre un exemple de *TTR-tree* pour les objets mobiles. Chaque objet mobile est représenté par une position dans le bitmap.

Les requêtes d'agrégats spatiotemporelles sont résolues par une recherche dans le TTR-tree. Nous avons implémenté les algorithmes correspondants et effectué des tests et des comparaisons avec les méthodes d'accès spatiales les plus proches. L'expérimentation a prouvé la validité et l'efficacité de notre approche.

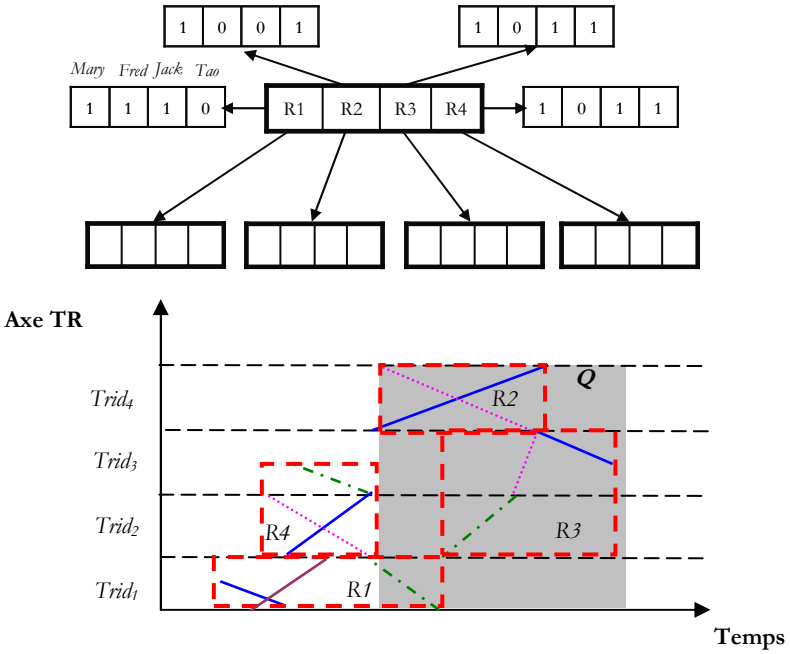


Figure 18. Index TTR-tree

IV.4. Conclusion

Nous avons proposé et implémenté deux approches pour l'analyse en ligne des données spatiotemporelles. L'une pragmatique se base sur un modèle spatiotemporel discret et l'autre, plus précise, capture la variation continue des données spatiotemporelles. Ces deux solutions ont été conçues pour permettre l'agrégation des objets selon l'espace, le temps et/ou leurs attributs Nous avons expérimenté les deux solutions. Le modèle discret se révèle simple à implémenter et très efficace dès lors que l'analyse ne met pas en jeu des requêtes spatiotemporelles ou que des résultats approximatifs suffisent. La modélisation continue permet d'analyser les données selon n'importe quelle granularité de l'espace et du temps sans perte de précision. Elle exploite la localisation sur le réseau pour réduire d'une dimension la représentation des trajectoires mobiles, réduisant également l'espace de stockage.

IV.5. Références

[89] Agrawal R., Gupta A., Sarawagi S., Modelling multidimensional databases. ICDE, April, 1997, 232-243.

[90] Ahmed T., Miquel M. Laurini R. Continuous Data Warehouse: Concepts, Challenges and Potentials. 12th Int. Conf. on Geoinformatics, Geospatial Information Research: Bridging the Pacific and Atlantic, Sweden, 2004.

[91] Beckmann, N., Kriegel, H., Schneider, R., Seeger, B. The R*-tree: an Efficient and Robust Access Method for Points and Rectangles. SIGMOD, 1990.

[92] Brinkhoff T. Network-based Generator of Moving Objects. http://fh-oow.de/institute/iapg/personen/brinkhoff/generator/

[93] Chaudhuri S., Dayal U., An Overview of Data Warehousing and OLAP Technology. ACM SIGMOD, vol. 26, n° 1, 1997.

[94] Chon, H. D., D. Agrawal, A. El Abbadi (2002). Query Processing for Moving Objects with Space-Time Grid Storage Model. Mobile Data Management, 121.

[95] Fernando, I., V. Lopez, R. T. Snodgrass, and B. Moon (2004). Spatiotemporal Aggregate Computation: A Survey, TIMECENTER TR-77.

[96] Gray J., Bosworth A., Layman A., Pirahesh H., Data cube: A relational aggregation operator generalizing group-by, cross-tab, and sub-total. ICDE, 152–159, 1996.

[97] Ho C., Agrawal R. Megiddo N., Srikant R., Range Queries in OLAP Data Cubes. ACM SIGMOD, 73-88, 1997.

[98] Jensen C. S., Kligys A., Pedersen T.B., Timko I., Multidimensional data modeling for location-based services. The VLDB Journal, vol. 13 n° 1, 1-21, 2004.

[99] Jurgens, M. and H. J. Lenz. The R a –tree (1998). An Improved R-tree with Materialized Data for Supporting Range Queries on OLAP-Data. DEXA Workshop.

[100] Kimball R., Factless fact tables. In Data Warehouse Architecture of DBMS on Line. http://www.dbmsmag.com/9609d05.html.

[101] Marchand P., Brisebois A., Bédard Y. Edwards G.. Implementation and evaluation of a hypercube-based method for spatio-temporal exploration and analysis, Journal ISPRS, vol. 59, n°1-2, 6-20, 2004.

[102] Mudu R. et al., Health effects and risks of transport systems: the HEARTS project, Collective publication authored by all the partners (2006), ISBN 92 890 2294 7, WHOLIS number E88772, World Health Organisation. http://www.euro.who.int/document/E88772.pdf

[103] Ooi B.C., Mcdonell K.J., Sack-davis R.. Spatial kd-tree: An indexing mechanism for spatial databases. IEEE COMPSAC Comp. Software&Applications Conf. Tokyo. 1987.

[104] Papadias D., Tao Y., Kalnis P., Zhang J.,. Indexing Spatio-Temporal Data Warehouses. ICDE, 166-175, 2002.

[105] Saltenis. S., Jensen. C. S., Leutenegger. S. T., Lopez. M. A.: Indexing the Positions of Continuously Moving Objects. Proc of ACM SIGMOD, Dallas, Texas (2000) 331-342,.

[106] Savary L. et Zeitouni K., "Modélisation et analyse spatio-temporelle de la mobilité urbaine", 7ème conférence du GDR SIGMA, Cassini'04, Grenoble, juin 2004, pp. 27-32.

[107] Savary L., Wan T., Zeitouni K., "Spatio-Temporal Data Warehouse Design for Activity Pattern Analysis", Int. DEXA Workshop on Geographic Information Management, September 2-3, 2004, Zaragoza, Spain.

[108] Shanmugasundaram J., Fayyad U., Bradley P.. Compressed data cubes for OLAP aggregate query approximation on continuous dimensions. KDD, 223--232. New York, 1999.

[109] Stefanovic N., Han J., Koperski K.. Object-based selective materialization for efficient implementation of spatial data cubes. IEEE Transactions on Knowledge and Data Engineering, 938-958, 2000.

[110] Sun, J., Papadias, D., Tao, Y., Liu, B. Querying about the Past, the Present and the Future in Spatio-Temporal Databases. Proceedings of 20th IEEE International Conference on Data Engineering (ICDE), Boston, MA, 2004.

[111] Tao Y., Kollios G., Considine J., Li F., Papadias D., Spatiotemporal Aggregation Using Sketches. ICDE, 214-225, 2004.

[112] Wan T. and Zeitouni K., "Modélisation d'objets mobiles dans un entrepôt de données", 5èmes Journées d'Extraction et de Gestion des Connaissances, EGC 2005, Edition CEPADUES, Paris, Janvier 2005, pp. 343-348.

[113] Wan T., Zeitouni K., "Vers un entrepôt d'objets mobiles contraints par le réseau", Conférence INFORSID 2006, Hammamet, Tunisie, pp. 1055-1070.

[114] Wan T., Zeitouni K., "Représentation et indexation d'objets mobiles contraints par le réseau dans un entrepôt de données", 2ème Conférence Entrepôts de Données et Analyse en ligne, EDA 2006, Revue Nouvelles Technologies de l'Information (RNTI), Cépaduès Edition, 2006, pp. 139-154.

[115] Wu K., Otoo Ekow J., Shoshani A., An efficient compression scheme for bitmap indices. Lawrence Berkeley National Laboratory. Paper LBNL-49626, 2004.

[116] Zeitouni K., "Lot 1 - Conception de la base de données mobilité" & "Lot 2 - Prototype et résultats", Rapport de contrat PRiSM – ONADA, Septembre 2005.

[117] Zhang D., Tsotras V. J., Gunopulos D., Efficient Aggregation over Objects with Extent, 21th ACM International SIGMOD PODS, Wisconsin, 2002

CHAPITRE V. FOUILLE DE DONNÉES SPATIALES

Ce chapitre étudie la fouille de données spatiales en soulignant ses spécificités par rapport à la fouille de données tabulaires. Ces travaux se sont déroulés principalement entre 2000 et 2004 dans le cadre de collaborations scientifiques ou industrielles. Ils ont été testés sur des problèmes réels et ont donné lieu à la thèse de Nadjim Chelghoum [130], soutenue fin 2004 et à des publications référencées à la fin de ce chapitre.

Avec le développement de la cartographie numérique, le volume de données dans les bases de données spatiales ne cesse d'augmenter. Le développement d'outils de géocodage permettant la localisation par l'adresse d'un côté et la banalisation des moyens de localisation de l'autre génère de nouvelles sources de données spatiales. Ces données sont de plus en plus utilisées dans des applications décisionnelles. Seulement, la nature et le volume de données de base dépassent les capacités humaines d'analyse.

La fouille de données est reconnue comme un moyen très efficace d'analyse avancée de données, permettant d'extraire des connaissances cachées depuis des grandes masses de données. Etant donné le volume croissant des données spatiales, la fouille de données spatiales permet d'extraire des propriétés spatiales cachées dans ces données et présente donc un intérêt certain pour les applications spatiales décisionnelles. La fouille de données spatiale (FDS) est aujourd'hui identifiée comme un domaine de la fouille de données à part entière. Elle résulte de la combinaison de la fouille de données et des bases de données spatiales.

Nous présentons ici une synthèse de l'état de l'art et proposons une nouvelle classification. Ensuite nous décrivons nos contributions, avant de conclure par un bilan et des perspectives.

V.1. Spécificités du problème et approches

La fouille de données spatiales est définie comme l'extraction de connaissances implicites. Celles-ci peuvent être des relations spatiales ou d'autres propriétés non explicitement stockées dans la base de données spatiales [157], [193], [168], [178].

Un exemple historique est souvent cité pour illustrer l'extraction de connaissances depuis des données spatiales. Il remonte à 1854 où une épidémie de choléra avait touché Londres. Grâce à la cartographie des cas de choléra (figure ci-dessous), un certain Dr. Johns Snow découvre des concentrations de cas et les met en relation avec des points d'eau. Il ordonna alors leur fermeture, ce qui a stoppé l'épidémie. En quelque sorte, le Dr. Snow a fouillé des données spatiales par l'extraction des relations entre les puits et les cas de choléra. Toutefois, cette fouille était manuelle et donc limitée aux capacités humaines d'analyse.
Par analogie à ce processus, la fouille de données spatiales désigne la génération automatique de ce type de connaissances, de manière exploratoire et ce quelque soit le volume de données. Elle explore à la fois les relations spatiales et les propriétés des objets liés et révèle automatiquement les sélections pertinentes d'objets formant un modèle remarquable. Le tableau ci-dessous illustre cette analogie et ses différences.

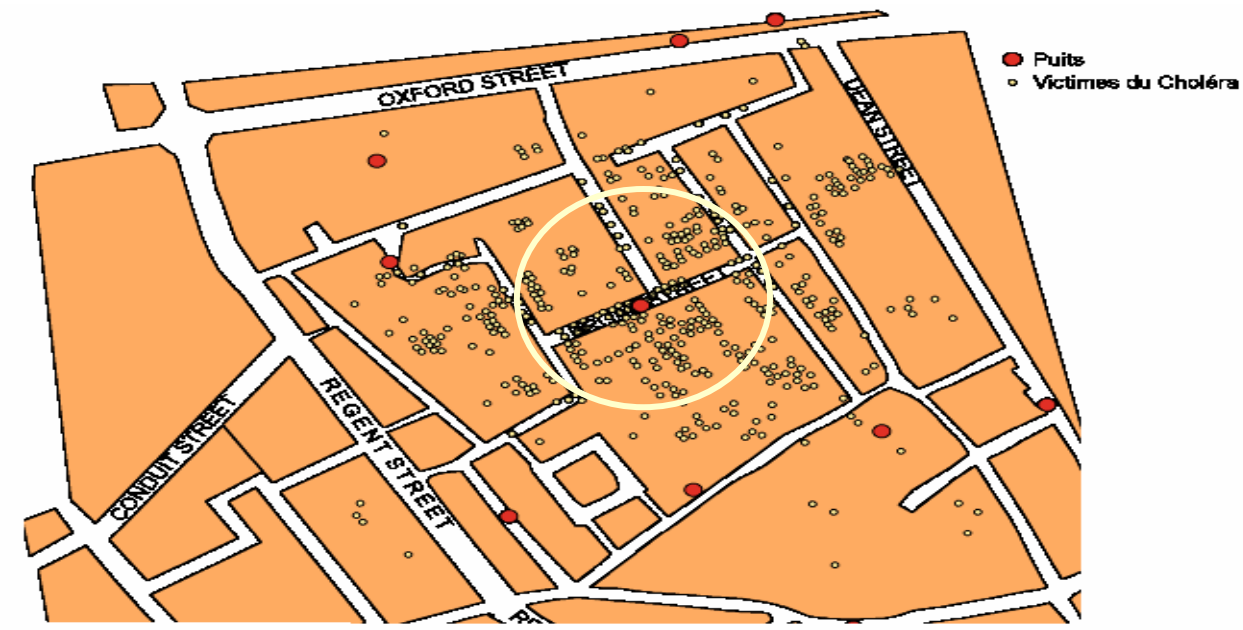


Figure 19. Exemple d'analyse spatiale

Tableau 2 : *Fouille de données spatiales versus analyse cartographique*

Fouille de données spatiales	**Analyse cartographique**
Découverte automatique de connaissances	Découverte visuelle de connaissances
Exploratoire (génère des hypothèses)	Confirmatoire
Opère sur de gros volumes de données	Valable pour un volume de données limité

V.1.1. Spécificités

Contrairement aux données traditionnelles, les données spatiales sont interdépendantes. En effet, elles décrivent des phénomènes avec de fortes interactions dans l'espace. Cette propriété est à l'origine de la

première loi en géographie de Tobler [180] "*Everything is related to everything else but nearby things are more related than distant things*". Analyser les données spatiales sans tenir compte de cette propriété est définitivement incorrect [121].

Ce qui caractérise la fouille de données spatiales, comme toute analyse spatiale, est la prise en compte des relations spatiales ente objets. Les méthodes classiques de fouille de données sont insuffisantes pour ce type d'analyse. C'est ce qui a motivé les recherches en fouille de données spatiales.

Retour sur les relations spatiales

Comme souligné dans le chapitre I, les relations spatiales sont couramment exploitées dans les requêtes et les analyses spatiales. Elles sont de différents types - topologiques ou métriques – et sont généralement implicites – autrement dit cachées dans les données -. Une des premières tâches de la fouille de données spatiales est de les exhiber. Ces relations traduisent l'influence du voisinage que nous classons en deux types : intra-thème comme l'auto-corrélation spatiale des mesures de phénomènes géographiques (la température de deux lieux proches est proche) ; et inter-thèmes comme l'influence du trafic routier sur le phénomène de pollution et de bruit.

V.1.2. Approches existantes en fouille de données spatiales

De la même façon qu'on distingue deux perceptions de l'information géographique (se référer à la modélisation en chapitre I), on peut distinguer deux approches respectives en fouille de données spatiales. Elles dépendent de la manière de considérer les relations spatiales. Certaines méthodes ne considèrent que les relations intra-thème. Elles constituent la première approche, que nous appellerons fouille de données monothématique. Elles sont souvent apparentées aux statistiques et à l'analyse de données. Citons par exemple la régression spatiale [178], l'analyse de données sous contrainte spatiale [162] et le regroupement par densité [171]. L'idée est d'incorporer un paramètre de contiguïté dans le modèle ou d'effectuer une pondération des variables par les valeurs du voisinage. Ceci est faisable car s'agissant d'un même thème, les données sont décrites avec les mêmes variables et sont comparables.

En revanche, lorsque l'on considère le voisinage entre thèmes, ces méthodes ne peuvent plus s'appliquer. L'objectif même est différent car on cherche à décrire ou à expliquer un phénomène par un autre ayant lieu au même endroit ou proche de cet endroit. En effet, analyser un thème sans sa relation avec d'autres couches thématiques est souvent insuffisant. La seconde approche est désignée par fouille de données multi-thèmes. Les méthodes de fouille de données multi-thèmes sont généralement basées sur des prédicats spatiaux interprétés comme des propriétés à prendre en compte dans le modèle à induire [157], [159], [143], [165]. La fouille de données multi-thèmes a été explicitement définie dans [134], puis récemment dans [183]. Pour cela, ces méthodes distinguent un thème cible de l'analyse et explorent les autres thèmes (ou phénomènes) susceptibles de l'influencer. Enfin, certains travaux distinguent des catégories d'objets, comme la recherche de co-localisation de phénomènes différents [152], mais dans ce cas, seule la catégorie et la localisation sont utilisées dans la description de l'objet, sans pouvoir tenir compte d'autres attributs. Or, un thème est généralement décrit par plusieurs attributs et ces attributs sont

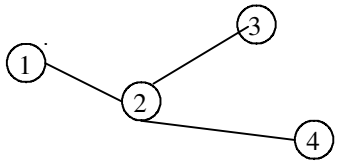
différents d'un thème à l'autre. Par conséquent, nous considérons la méthode de co-localisation comme monothématique.

V.2. Etat de l'art

Au début de nos travaux, nous avons publié un état de l'art sur la fouille de données spatiales [193], [197]. Cette section est une nouvelle synthèse mise à jour selon le caractère monothématique ou multi-thématique des méthodes et selon leur prise en compte ou non d'attributs.

V.2.1. Méthodes de fouille de données monothématique

Historiquement, la communauté analyse spatiale de données s'est focalisée sur une représentation de l'espace continu, soit par cellule, par ensemble de points ou par zonage. Les relations de voisinages considérées sont définies par une matrice de contiguïté ou de distance intra-thème. La matrice de contiguïté est binaire définie par M[i,j]=1 si l'objet i est voisin de l'objet j et M[i,j]=0 dans le cas inverse, comme illustré Figure 20. La matrice de distance est plus précise. Elle est pondérée et est définie par : M[i,j]= distance ($objet_i$, $objet_j$).



$$M = \begin{vmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{vmatrix}$$

Figure 20. Matrice de voisinage

De nombreuses méthodes exposées ici utilisent les matrices de voisinage, parfois rebaptisées graphes de voisinage dans [143].

Analyse de localisations sans attributs

Parmi les approches monothématiques, certaines sont basées uniquement sur les localisations. Elles explorent généralement un ensemble de localisations (ensemble de points) pour révéler des tendances ou des concentrations [146], [170], [171]. L'approche de regroupement (clustering) spatial proposée dans [142] peut être classée dans cette catégorie.

Analyse de tendances par la méthode de densité

Lorsque le nombre d'objets analysés est élevé et que leur distribution spatiale est hétérogène, de nombreuse localisation se superposent. Par conséquent, la visualisation cartographique de ces objets ne reflète pas l'intensité locale, ni la tendance générale de leur localisation. L'idée est de quantifier les localisations ponctuelles par des mesures de densités dans leur voisinage puis, par interpolation, d'attribuer une « intensité » au phénomène en tout lieu de l'espace. La visualisation cartographique du champ de mesures ainsi obtenu montre l'intensité du phénomène de manière plus claire. Les densités sont calculées par un balayage de l'espace par des fenêtres mobiles circulaires [148]. Une analyse de tendance globale est

obtenue par des fenêtres mobiles de rayon élevé, tandis qu'une analyse locale révélant les contrastes correspond au choix de rayon de faible taille.

Clustering

Le clustering est une méthode de classification automatique bien connue en fouille de données permettant le reproupement d'objets par classes homogènes. Pour ce faire, elle cherche à maximiser la similarité intra-classe et à minimiser la similarité inter-classes. Les principales méthodes sont celles par agrégation autour de centres mobiles, comme les k-means, les nuées dynamiques, la classification automatique hiérarchique (CAH) et enfin, les méthodes par densité comme DBSCAN [142], BIRCH [198] et OPTICS [120].

La transposition au domaine spatial s'appuie sur une mesure de similarité d'objets localisés suivant leur distance métrique. Néanmoins, l'application de ces méthodes au domaine spatial vise moins à classifier qu'à détecter les concentrations ou les points chauds d'un phénomène. Par exemple, dans l'étude de criminalité ou des zones accidentogènes en sécurité routière.

Les travaux récents sur le clustering spatial sont surtout axés sur l'optimisation des algorithmes. DBSCAN utilise un index spatial R*tree afin d'optimiser l'accès aux données indexées. Openshaw prône l'utilisation de machine parallèle pour résoudre le clustering dans la machine GAM/K [171].

Analyse de localisations munies de mesures numériques

Cette catégorie s'intéresse aux mesures relevées sur un domaine spatial, souvent couvrant l'espace par un découpage surfacique. Il s'agit souvent d'un seul attribut numérique. L'analyse vise à caractériser la variation spatiale de cette ou de ces mesures.

Auto-corrélation spatiale globale et locale

L'auto-corrélation spatiale permet de mesurer la ressemblance entre voisins. Elle est souvent utilisée comme technique exploratoire indiquant si une modélisation spatiale est nécessaire [139]. Elle s'applique à des données quantitatives localisées et dont une relation de voisinage a été calculée. L'indice de Geary - datant de 1954 - teste si la variabilité d'une variable entre voisins (donnée par la notion de variance locale) est significativement différente de celle attendue d'un modèle aléatoire (donnée par la variance). Etant donnée une matrice de voisinage M, la variance locale d'une variable $X = \{x_1,x_2,...., x_n\}$ est définie comme suit :

$$V_{loc} = 1/2m \, \Sigma_i \Sigma_j M(i,j) \, (x_i-x_j)^2 \qquad où \; m = \Sigma_i \Sigma_j M(i,j) \;\; pour \;\; i \, , j = 1..n$$

L'indice de Geary est défini comme le rapport c de la variance locale V_{loc} et de la variance V.

$$c = V_{loc} \; / V = [1/2m \, \Sigma_i \Sigma_j M(i,j) \, (x_i-x_j)^2] \; / \; [\, 1/2n(n-1) \, \Sigma_i \Sigma_j (\, x_i - x_j)^2]$$

L'absence d'auto-corrélation spatiale se traduit par c=1. A l'inverse, une forte auto-corrélation spatiale (indiquant une ressemblance des valeurs d'objets voisins) correspond à une valeur de c proche de 0.

L'auto-corrélation se décline en indice local LISA (Local Indice of Spatial Association) [122], [172]. Les indices locaux mettent en évidence des situations locales particulières par comparaison à la valeur de l'indice global et peuvent parfois montrer des structures qui auront échappé à la mesure globale.

Analyse de tendances par régression linéaire

Dans le même ordre d'idée que l'analyse des tendances dans les séries temporelles, Ester et al. [143] proposent une approche de mesure de tendance spatiale pour une mesure donnée partant d'un sous-ensemble de centres choisis par l'utilisateur. L'algorithme utilise un graphe de voisinage et calcule la régression linéaire de la mesure concernée par extension progressive du voisinage en partant de ces centres et selon chaque orientation (nord, nord-est, etc.) de cette extension.

Analyse de localisations munies de catégories

Dans cette catégorie, les localisations sont supposées être décrites par des attributs catégoriels. L'analyse porte sur la présence simultanée de catégories dans l'espace ou sur les propriétés caractéristiques s'étendant au voisinage.

Co-localisation

Une méthode développée dans l'équipe de Shashi Shekhar [177] est la co-localisation de phénomènes dans l'espace. Cette méthode part d'un ensemble de localisations relatives à des catégories et recherche les catégories fréquemment proches dans l'espace. L'algorithme de co-localisation étend l'algorithme APRIORI de recherche de règles d'associations et redéfinit le support et la confiance. Pour chaque combinaison de catégories, il mesure une corrélation spatiale inter-catégories. La seule difficulté est d'améliorer cette méthode pour un passage à l'échelle, ce qui a motivé la poursuite de ces travaux jusqu'à aujourd'hui [153], [189]. Cette méthode est inspirée de la méthode statistique cross K-fonction de Ripley [140], qui elle ne passe pas à l'échelle.

Caractérisation

La caractérisation correspond à l'induction de propriétés (attribut = valeur) caractéristiques d'un sous-ensemble de données relativement à l'ensemble des données. Son adaptation à la fouille de données spatiales dans [144] correspond aux propriétés caractéristiques qui s'étendent jusqu'à un degré de voisinages.

Plus précisément, les paramètres de la caractérisation sont :

(i) un sous-ensemble S d'objets à analyser dans la base de données,

(ii) un *seuil* de fréquence relative à la base,

(ii) une confiance définie comme la proportion d'objets de S qui satisfont le seuil de signifiance dans leur voisinage et

(iii) un degré de voisinage *n_max* .

Elle découvre les propriétés *pi = (attribut, valeur)*, la fréquence relative (*freqi*> *seuil*) de *pi* dans le voisinage et le nombre (*ni* < *n_max*) de voisins auxquels s'étend *pi* . La caractérisation peut s'exprimer par une règle :

S -> p1(n1 , freq1) et ... pk(nk , freqk)

Par ailleurs, le module `Geo-Characterizer` de `GeoMiner` [150] permet la découverte de règles caractéristiques à l'issue d'une généralisation des données spatiales et non spatiales [164].

Analyse d'individus munis de localisations

Un autre point de vue sur l'analyse spatiale des données est de considérer que la localisation n'est pas le principal objet de l'analyse, mais plutôt les variables classiques de l'individu. La localisation rajoute à ces individus une contrainte de contiguïté. L'analyse doit donc pondérer ces variables classiques par le voisinage.

Auto-régression spatiale

L'auto-régression spatiale [178] est une extension de la régression linéaire $y = \beta*X+\varepsilon$ par l'ajout d'un terme correctif $\rho*W*y$ où W est la matrice de voisinage. Le poids ρ dépend de l'indice d'auto-corrélation spatiale de y.

Clustering spatial multi-attribut

Alors que le clustering à l'origine traite des tableaux à plusieurs variables, l'application aux données spatiales exposée précédemment est basée uniquement sur la localisation, souvent de type ponctuelle. L'extension aux attributs non spatiaux et aux objets de forme linéaire ou surfacique a été proposée dans [175] par une redéfinition de la mesure de similarité. Une autre approche de clustering multi-varié spatial a été proposée dans Neighbourhood EM (NEM) [119] [149] par l'ajout dans le critère d'optimalité d'un terme correctif pondéré par le voisinage.

V.2.2. Méthodes de fouille de données multi-thématique

Dan la fouille de données mono-thématique, l'espace (ou la localisation) est l'objet de l'analyse avec peu d'attributs, généralement une mesure ou une catégorie unique. Or, les bases de données spatiales et la majorité des SIG organisent les données en couches thématiques (cf. chapitre II), chacune avec une description ou schéma propre. Les méthodes précédentes ne prennent pas en compte cette organisation et par conséquent, ne peuvent révéler des relations inter-thèmes cachées.

Le but de la fouille de données multi-thématique est de considérer, en plus de la description de l'objet par ses propres attributs, sa relation de voisinage ainsi que la description des objets voisins.

Parmi les pionniers de la fouille de données spatiales, Koperski et Han [156] avaient défini des méthodes de règles d'association et de classification impliquant explicitement plusieurs couches thématiques. La méthode de classification proposée par Ester et al. [143] considère les liens et les types de voisinage, mais la notion de thème n'y est pas explicite. Plus récemment, le travaux de Malerba et Lisi [165] ont appliqué la fouille de données multi-relationnelle basée sur la programmation logique inductive aux données spatiales

multi-thèmes. Les méthodes les plus courantes sont les règles d'association spatiales et la classification supervisée des données spatiales.

Règles d'association spatiales

L'extension de la découverte de règles d'association de [118] aux données spatiales a été proposée par Koperski et al. [156], [159] et permet de générer des règles de type :

$X \rightarrow Y$ (s, c) avec un support s et une confiance c

telles que X et Y sont des ensembles de prédicats, dont au moins un est un prédicat spatial. En d'autres termes, ces règles correspondent à des associations entre des propriétés des objets et celles de leurs « voisins ». Cette méthode génère pour chaque objet d'un thème cible de l'analyse, une liste de prédicats comprenant sa description, la description d'objets liés d'autres thèmes et leurs relations spatiales avec l'objet. Cette liste constitue des items d'une transaction sur laquelle un algorithme de type APRIORI peut opérer pour générer des règles d'association spatiales de type :

$is\_a(x, school) \wedge close\_to(x, sport\_center) \rightarrow close\_to(x, park)$ (s, c) Ici, le premier prédicat est non spatial tandis que les trois autres prédicats sont spatiaux.

Cette méthode permet, en outre, de générer des règles multi-niveaux en exploitant à la fois la généralisation des attributs sur la base d'une hiérarchie de concepts et la généralisation de la relation spatiale de voisinage. Un avantage est de permettre la découverte de règles qui n'apparaissent qu'à un niveau général. De plus, cette approche permet d'optimiser la recherche car elle utilise la propriété d'anti-monotonicité qu'une combinaison de prédicats détaillés n'est fréquente que si le niveau général correspondant l'est.

La transformation en prédicats mène naturellement à l'application de la programmation logique. C'est ainsi qu'une méthode similaire a été développée dans le cadre de la fouille de données multi-relationnelles [165], [129] par une adaptation de l'algorithme de [156] aux données spatiales exprimées en logique du premier ordre.

Classification spatiale

Contrairement au clustering qui cherche à identifier des classes, la classification supervisée permet d'affecter les objets à une classe parmi celles prédéfinies. Elle peut être utilisée pour prédire les classes de nouveaux objets ou simplement pour décrire ou expliquer les liens entre les propriétés de l'objet et sa classe. Parmi les méthodes les plus utilisées, car fournissant des règles interprétables par l'analyste, on trouve les arbres de décision [162], [200]. Un arbre de décision est construit par l'application successive de critères de subdivision sur une population d'apprentissage afin d'obtenir des sous-populations plus homogènes. Le critère de subdivision est déterminé au niveau de l'attribut dans l'arbre ID3 [174] et au niveau d'une valeur d'attribut dans CART [128].

Ester et al. [143] ont proposé une méthode de classification spatiale basée sur ID3 et utilisant le concept de graphe de voisinage. Pour intégrer les propriétés du voisinage, les auteurs définissent une notion d'attribut généralisé par l'ajout d'un degré de voisinage. Ils intègrent cet attribut dans l'algorithme ID3

comme critère possible de subdivision. Cette méthode ne distingue pas explicitement des couches thématiques et considèrent que tous les objets possèdent les mêmes attributs. Néanmoins, les auteurs supposent dans l'exemple donné en illustration l'existence d'un attribut « type d'objet », qui implicitement reflète la notion de thème.

Comparativement la méthode de classification de Koperski [158] considère des thèmes de référence et des relations de voisinage précises. Elle représente toutes les propriétés par une liste de prédicats. Ainsi le prédicat noté *attribut (x, valeur)* est l'équivalent de "x.attribut = valeur" et close_to(x, catégorie) représente le voisinage à une catégorie d'objets. De plus, elle intègre des mesures agrégées par extension au voisinage (ex : la population totale aux alentours) et dont le périmètre d'extension est déterminé algorithmiquement. Il suffit ensuite de calculer le gain d'information pour chacun des prédicats. Afin d'optimiser les traitements, ce calcul est fait sur une généralisation des données selon des hiérarchies données par l'utilisateur. D'autres optimisations sont proposées pour le calcul approximatif des prédicats spatiaux et pour le filtrage préalable des prédicats pertinents de l'objet.

V.3. Contributions

Nous avons développé de nouvelles approches pour la fouille de données spatiales et nous les avons expérimentées et publiées [130]. Nous avons tout d'abord défini les objectifs, puis nous avons cherché le moyen le plus flexible et le plus efficace pour les réaliser. Nous sommes parvenus ainsi à plusieurs propositions qui ont été implémentés, testées et dont les performances ont été évaluées. Nous décrivons ces différents résultats et situons nos travaux par rapport à l'état de l'art.

V.3.1. Limites des travaux existants

La synthèse de l'état de l'art montre des limitations multiples :

(i) Dans la première catégorie de méthodes, c'est-à-dire mono-thématiques :

- L'analyse de localisations sans attributs est insuffisante dans l'analyse de bases de données spatiales.
- L'analyse de localisations munies de mesures ou de catégories est limitée car elle est mono-attribut.
- L'analyse d'individus munis de localisation est une méthode empirique, limitée aux données numériques.
- Pour la plupart, ces approches ne sont pas étudiées ni testées pour des données volumineuses et se basent sur des algorithmes en mémoire centrale. Elles nécessitent donc des adaptations et des optimisations par l'intégration de techniques de bases de données telles que l'indexation.
- De plus, toutes ces méthodes ont pour inconvénient de ne considérer que les relations entre objets d'un même thème excluant les relations spatiales pouvant exister entre objets de thèmes différents.

(ii) Dans la seconde classe de méthodes, c'est-à-dire de fouille de données multi-thématique :

- Malgré leur intérêt, la formalisation du voisinage reste confuse dans les méthodes de Koperski et de Ester. En effet, Ester [143] et Koperski [159] considèrent le voisinage entre l'objet et des objets d'un « type » donné, sans pour autant que cette notion de « type d'objet » ne soit explicitée. Les approches d'Ester ne parlent que d'objets sans distinction aucune entre les thèmes.
- La deuxième limitation majeure est de n'utiliser que les prédicats spatiaux (binaires) plutôt que les relations spatiales pondérées comme la distance. L'approche d'Ester limite ces prédicats à une relation définie dans le graphe de voisinage. Toutefois, l'algorithme prend en compte plusieurs degrés de voisinage. Koperski se base sur des relations spatiales hiérarchisées par l'utilisateur et construit le modèle en exploitant cette hiérarchie du plus général au plus spécifique. Cependant, le choix des quelques prédicats à prendre en compte reste à la charge de l'utilisateur. Or, les liens de voisinage (en considérant la distance) sont nombreux, voire infinis. Le choix par l'utilisateur devient difficile. De plus, il risque de biaiser le modèle.
- Concernant la classification, la méthode de Koperski commence par généraliser les données avant d'appliquer l'algorithme de classification. Bien que ce soit vu comme une technique d'optimisation, cela risque d'engendrer une perte d'information. De plus, cette méthode passe par une transformation des données en prédicats qui empêche l'utilisation d'algorithmes de classification existants. En outre, cette réécriture de la base sous forme de prédicats a un coût qui s'ajoute au coût de construction du modèle.
- L'algorithme de classification de Ester [143] manque de précision. En effet, pour chaque objet cible ou exemple d'apprentissage, il peut y avoir plusieurs voisins de degré n ayant chacun sa valeur pour l'attribut considéré. De tels attributs multi-valués faussent le calcul des critères de division des nœuds de l'arbre. L'algorithme ne donne pas une modification de la formule du gain d'information pour ce qu'il appelle « des attributs généralisés ». De plus, des valeurs différentes pour un attribut généralisé ne signifient pas que l'objet à classer est différent et donc la classification n'est pas discriminante [130].
- Les méthodes basées PLI de [165], [129] présentent les mêmes inconvénients que les précédentes. De plus, contrairement à celles-ci, elles ne considèrent pas le passage à l'échelle et ne proposent aucune technique d'indexation ou d'optimisation dans ce sens.

V.3.2. Objectifs

Notre vision de la fouille de données spatiales est de considérer, en plus de la description des objets à analyser, leurs relations de voisinage et la description des objets voisins. Cela nous place clairement dans la catégorie de fouille de données multi-thématique où la base de données spatiale est organisée en plusieurs couches, chacune étant une collection d'objets. Par rapport aux méthodes précédentes, nos objectifs sont :

(i) de décrire de manière explicite l'objet de l'analyse et la notion de voisinage ;

(ii) de filtrer automatiquement la relation pondérée de voisinage ;

(iii) de s'appliquer quelque soit le volume de données et d'offrir des performances acceptables en terme de temps de réponse ;

(iv) d'offrir des solutions flexibles et de s'appliquer à différents modèles de fouille de données.

V.3.3. Approche proposée

Notre approche apporte une réponse aux quatre objectifs ci-dessus. Concernant le premier objectif, notre approche distingue clairement le thème cible de l'analyse et le ou les thèmes à considérer comme objets voisins. La notion de voisinage est étendue à une valeur pondérée pouvant être la distance, le temps de parcours ou toute autre quantification et cette valeur est exploitée dans la construction du modèle. Toute propriété des thèmes voisins – hormis les restrictions exprimées par l'utilisateur – est susceptible d'être intégrée dans le modèle, au lieu du seul « type d'objet » des travaux précédents.

Concernant le deuxième objectif, l'utilisateur peut, s'il le souhaite, restreindre son analyse à des relations de voisinage particulières, mais cela n'est nullement obligatoire. Le modèle filtre automatiquement la « bonne relation spatiale ». Le périmètre de voisinage permettant de générer une règle est choisi par l'algorithme.

Les algorithmes développés sont couplés avec un système de bases de données spatiales qui gère la persistance sur disque et ne supposent pas les données en mémoire. Ainsi, on atteint le troisième objectif.

L'utilisation intensive des relations spatiales en fouille de données spatiale entraîne des jointures coûteuses sur des critères spatiaux. Afin de répondre au troisième objectif de performances, nous avons proposé de rendre ces relations explicites en utilisant un « index de jointure spatial ». L'idée est de pré-calculer la relation spatiale exacte entre les localisations de deux collections d'objets spatiaux et de la stocker dans la base. Nous proposons d'exploiter cet index et de l'intégrer dans la fouille de données spatiales.

Outre l'optimisation du coût des jointures spatiales, cette table d'index ramène la fouille de données spatiales à la fouille de données multi-tables. Dès lors, différentes solutions deviennent possibles. Ainsi, l'application de méthodes de fouille de données multi-relationnelle, étudiée en dehors des données spatiales, devient possible. Cela offre plus de flexibilité, répondant à l'objectif (iv).

En définitive, nous avons proposé une démarche en deux étapes principales et trois options pour la fouille de données spatiales. La Figure 21 illustre cette démarche.

La première étape calcule les relations spatiales et les matérialise dans une table d'indexes. Elle ramène ainsi le problème à une fouille de données multi-relationnelles. Seulement, cette organisation de données ne peut pas être directement analysée par les méthodes classiques de fouille de données, car celles-ci considèrent en entrée une relation unique où chaque tuple est une observation. La seconde étape propose alors trois options de fouille de données multi-relationnelles dans le cadre de fouille des données spatiales. La première adapte les algorithmes existants aux données multi-relationnelles. La deuxième transforme les données multi-relations en une relation unique avant d'appliquer les algorithmes existants. Enfin, la dernière option transforme les données en prédicats avant d'appliquer les méthodes de programmation logique inductive.

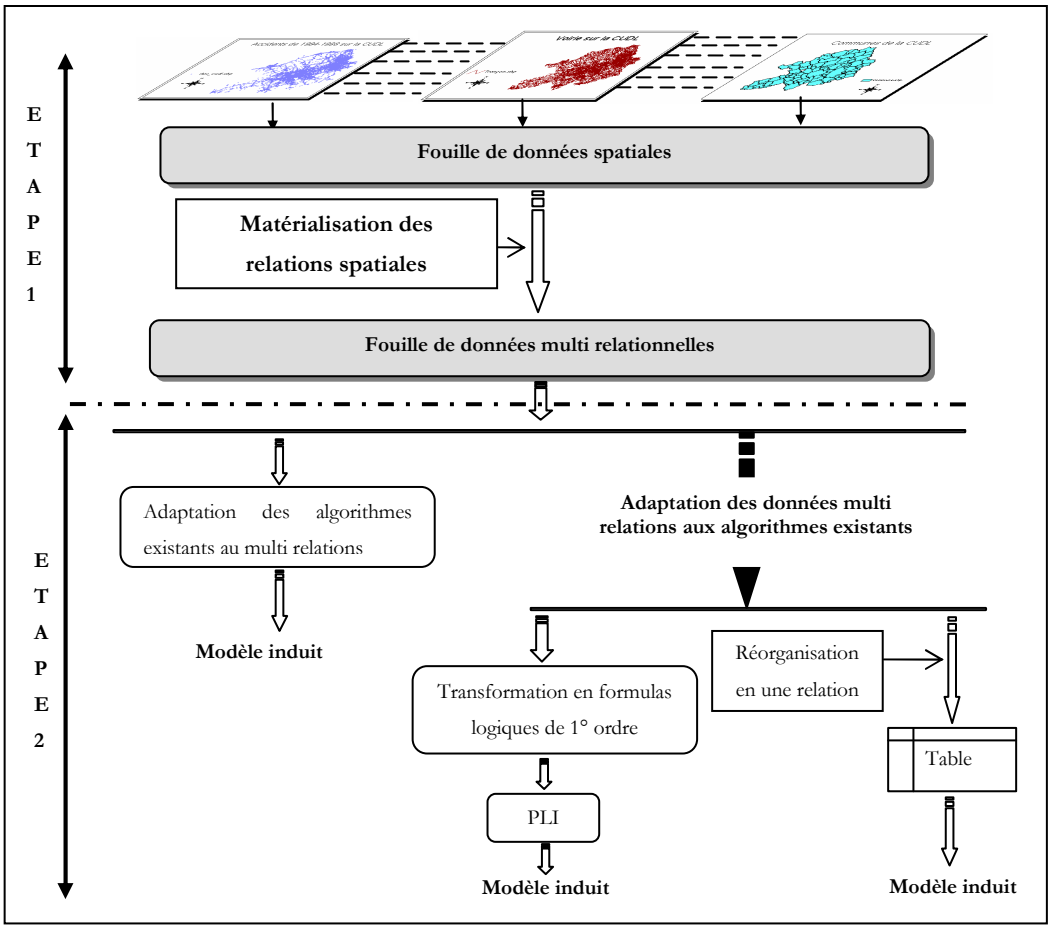


Figure 21. Approche générale de fouille de données spatiales

Ces trois solutions ont été mises en oeuvre sur la méthode de classification supervisée par arbre de décision. Nous avons ainsi proposé une nouvelle méthode de classification par arbre de décision spatial baptisée SCART (Spatial Classification And Regression Tree). Nous avons également testé notre approche dans l'application aux règles d'associations spatiales.

Nous exposons ci-dessous brièvement l'ensemble de ces propositions en commençant par l'index de jointures spatiales, puis nous explicitons la fouille de données spatiales par la fouille multi-relations. Nous terminons par une synthèse des solutions proposées et une description de leur mise en œuvre.

V.3.4. Première étape : Index de jointures spatiales

L'index de jointure est une structure secondaire qui stocke les références des tuples qui joignent permettant l'optimisation des requêtes de jointure [186]. A la différence des jointures les plus courantes en relationnel, les jointures spatiales ne sont pas des équi-jointures, mais des jointures sur critère spatial. Ce critère peut être topologique, métrique ou directionnel. Différentes extensions de cette structure ont été proposées pour des données spatiales. Ainsi, une application directe pour un critère spatial spécifique avait été proposée dans [185] et récemment optimisée dans [179]. Lu et Han [163] visent les jointures sur le critère de distance et ont proposé une extension de l'index par l'ajout de l'attribut distance. Enfin, une structure appelée indice de voisinage a été d'abord introduite dans [143] comme une table indexée par

l'identifiant de chaque objet et renseignant la liste des Identifiants des objets voisins. Comme cette table se limitait à une seule relation de voisinage et son coût de construction était élevé, elle a été remplacée dans [144] par l'extension de l'index de jointure à trois attributs renseignant la distance, l'orientation et la topologie.

A la même période, nous avons proposé dans [195] un index de jointure spatial qui a un double but. Le premier est d'accélérer toutes les jointures spatiales de proximité (topologique ou métrique). La seconde est d'enrichir par la matérialisation de la relation spatiale la base de données et de ramener le raisonnement spatial à des requêtes relationnelles (cf. Figure 22).

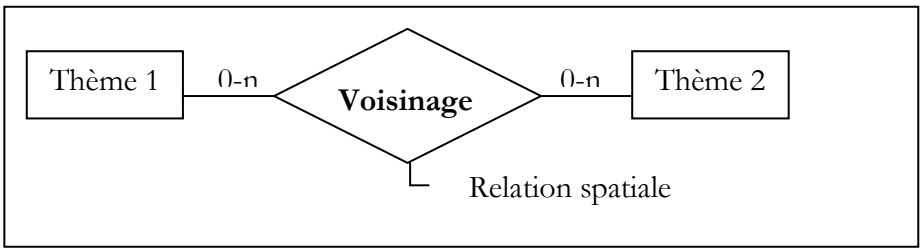


Figure 22. Enrichissement sémantique par la relation spatiale

L'idée est de générer une table relationnelle de schéma (ID1, Relation_Spatiale, ID2) où ID1 et ID2 sont respectivement des identifiants d'objets et Relation_Spatiale est une valeur numérique correspondant à leur relation spatiale de proximité (cf. Figure 23). Par proximité, nous entendons une distance métrique dans un périmètre utile (défini par le concepteur) ou toute relation topologique. Ainsi, la valeur est 0 si les objets sont adjacents, elle devient négative pour différencier des relations topologiques.

La définition du périmètre utile permet de réduire l'espace de stockage de l'index et la durée nécessaire à son calcul. Dès lors, le coût de construction de cet index équivaut à celui de la jointure spatiale sur critère de distance, mais son avantage est de permettre de gagner ce coût lors des multiples utilisations ultérieures des relations spatiales pré-calculées. Bien que ce ne soit pas essentiel, la méthode SPOT [188] a été utilisée pour optimiser la construction de cette table.

Cet index de jointure spatiale a plusieurs avantages :

(i) le stockage des relations spatiales évite de les recalculer à chaque application ;

(ii) le même index permet d'accélérer les jointures selon tout prédicat spatial dans la limite du périmètre défini ;

(iii) il représente de manière sémantique toute relation spatiale entre objets ;

(iv) il ramène le problème de la fouille de données spatiale à la fouille de données relationnelles basée sur trois tables.

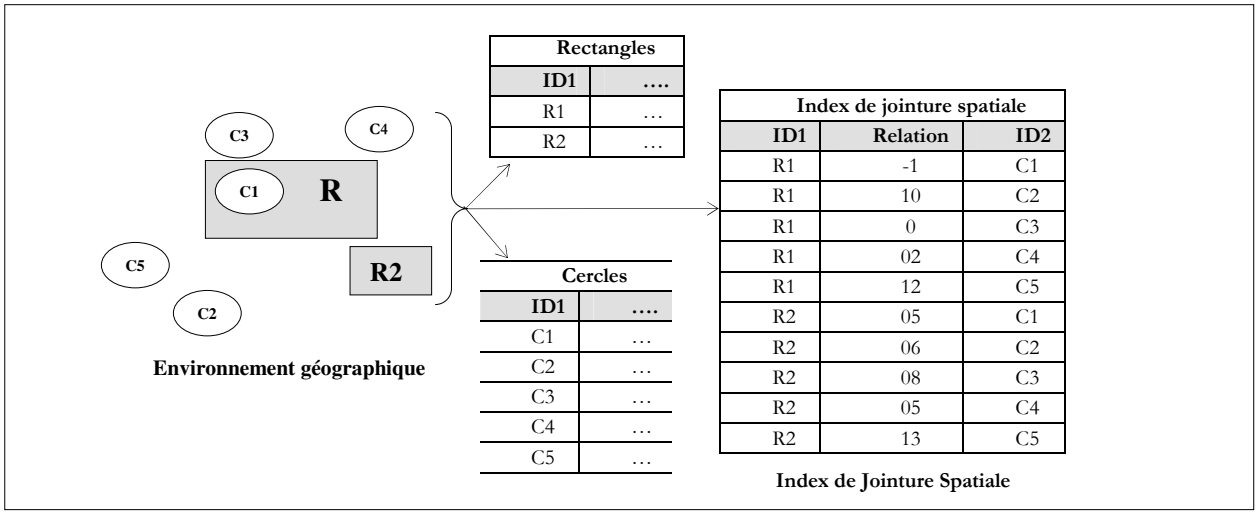


Figure 23. Du data mining spatial au data mining multi-relations

V.3.5. Deuxième étape - Formulation du problème en fouille multi-relations

Par la matérialisation des relations de voisinage, les prédicats spatiaux utiles à la fouille de données sont représentés en relationnel classique. Seulement, la fouille de donnée traditionnelle ne s'applique que pour une relation unique où une observation est représentée par un tuple unique. Ni l'organisation en plusieurs relations, ni la jointure de ces relations ne satisfait ces deux conditions. L'autre problème est que les relations considérées n'ont pas le même rôle pour la fouille de données. En effet, en accord avec les objectifs énoncés ci-dessus, on distingue une relation (un thème) cible, une ou plusieurs relations liées (thèmes de référence) à considérer et la ou les relations de voisinage qui les relie à la relation cible. Une observation est vue comme l'enrichissement d'un objet de la relation cible par les objets liés via la relation de voisinage. Par conséquent, s'il est essentiel de maintenir l'identité d'un objet de la relation cible, il n'est pas pertinent de considérer l'identité des objets liés. Ce sont plutôt leurs attributs et leur relation à l'objet cible qui sont utiles à l'analyse. Il est donc indispensable d'intégrer ces interprétations dans la seconde étape de la fouille de ces données.

V.3.5.1. *Adaptation des algorithmes de fouille de données*

Cette solution consiste à adapter les méthodes de fouille de données à cette organisation et à l'interprétation particulière des exemples d'apprentissage. D'un côté, l'algorithme prend en entrée les données multi-relations. De l'autre, il différencie entre les propriétés et celles de son voisinage : celles-ci sont combinées à la pondération de la relation spatiale correspondante. Cette adaptation diffère selon la méthode. Le cas de la classification est exposé ci-dessous.

Application à la classification spatiale par arbre de décision

La construction de l'arbre de décision spatial est une adaptation de CART [128]. Elle se traduit par la modification du critère de division d'un nœud dans l'arbre et de l'évaluation du gain informationnel. Dans

SCART, la partition d'un nœud peut se faire sur une propriété du voisinage auquel cas la relation précise la plus pertinente sera retournée.

L'algorithme prend en entrée les trois relations, les attributs explicatifs, l'attribut à expliquer et les conditions de saturation. Il commence par effectuer une jointure externe gauches sur clés entre la table cible, l'index de jointure spatiale puis la table des objets voisins (étape 1 de la Figure 24). Les tuples qui n'ont pas de voisins seront ainsi complétés par des valeurs nulles.

La construction de l'arbre se fait par division successive du résultat de cette double jointure (étape 2 à 6). Cette division évalue le gain d'information pour chacun des critères de division possibles. C'est là qu'est intégrée l'interprétation du voisinage comme une combinaison d'une propriété du voisinage et d'une relation spatiale. En effet, la condition testée pour les attributs du thème cible est comme dans CART « cible.attribut *comparateur* valeur », mais pour un attribut provenant du voisinage, ce critère est remplacé par : « voisin.attribut *comparateur* valeur et voisin *à distance R de* cible [4]». Parmi l'ensemble de ces divisions, on choisit celle qui donne le meilleur gain informationnel. On trouve ainsi automatiquement la valeur de la relation R la plus discriminante pour chaque propriété.

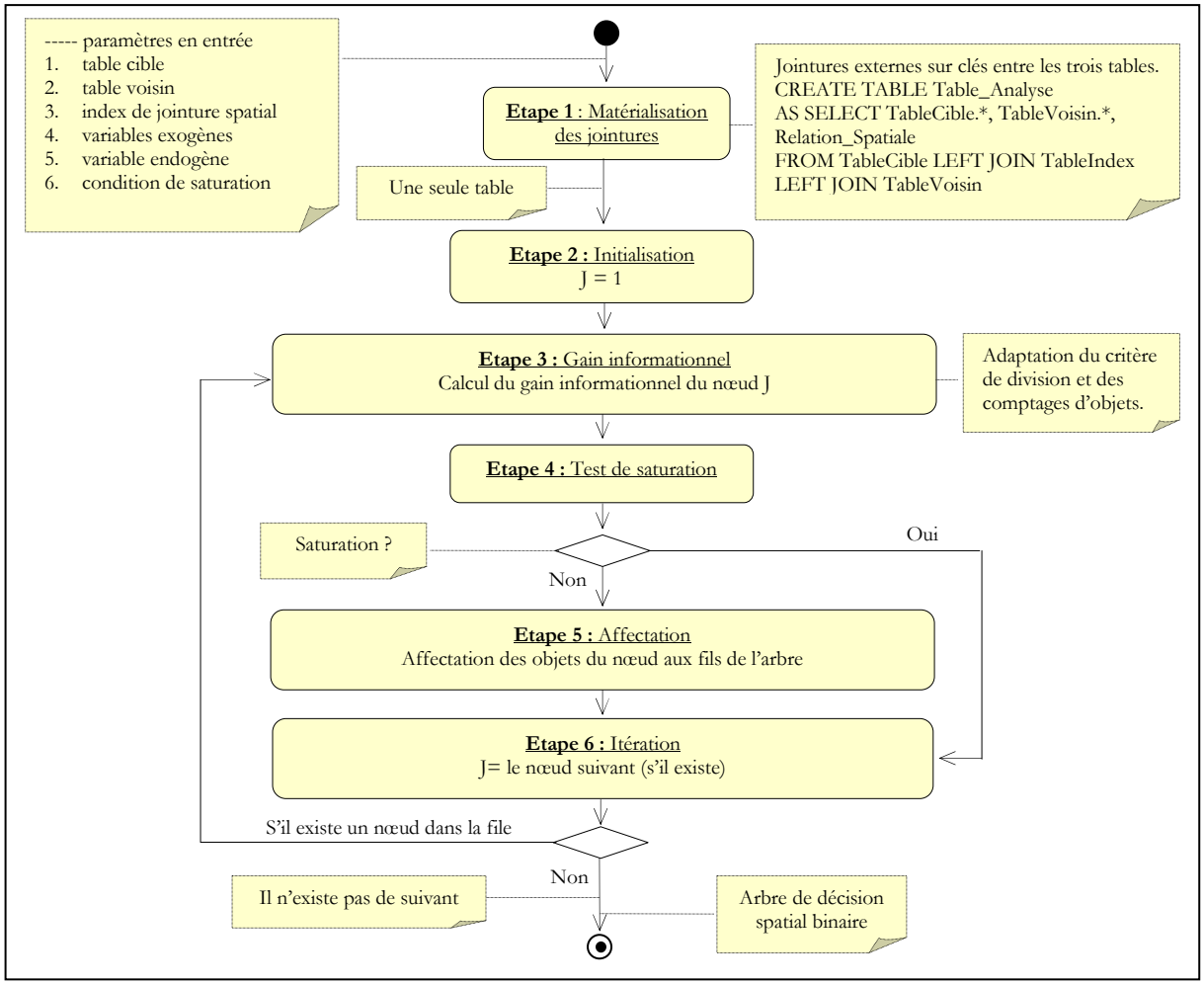


Figure 24. Description de l'algorithme SCART en utilisant l'option 1

[4] « voisin *à distance R de* cible » est vrai si la relation spatiale correspondante a un poids inférieur à *R*.

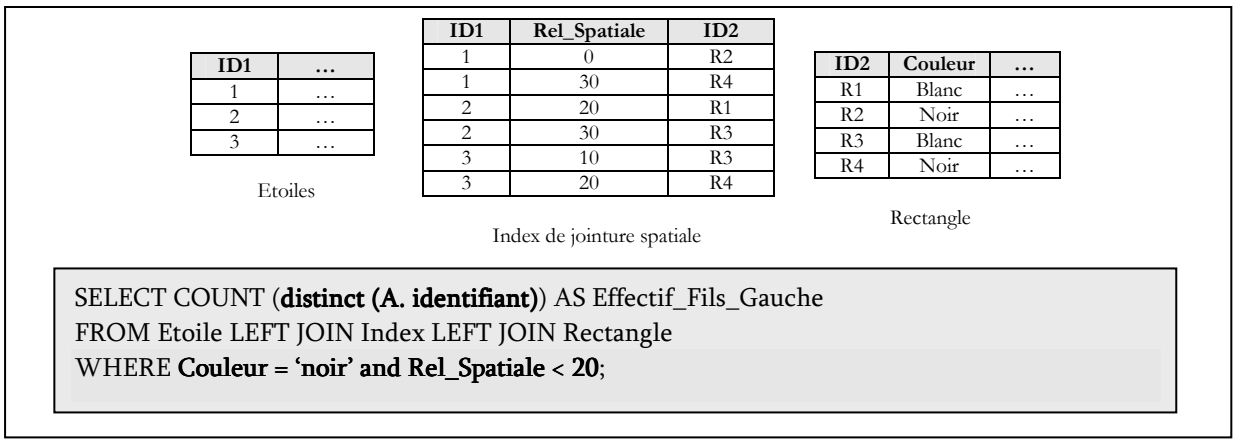


Figure 25. Exemple d'adaptation du critère et des statistiques (partie en gras)

La seconde modification pour l'évaluation du gain d'information est d'éviter le double comptage dû à la duplication des objets à classer suite à la jointure. On ne compte plus simplement les tuples, mais les objets cibles distincts. Ces modifications sont illustrées par l'exemple de la Figure 25. On procède à la division (étapes 4 et 5), puis on répète le même processus (étape 6) pour chaque nœud jusqu'à saturation de tous les nœuds.

V.3.5.2. *Transformation des données pour la fouille de données classique*

L'inconvénient de la solution précédente est de ne pouvoir utiliser les outils existants de fouille de données, car elle implique une modification « en dur » de chaque algorithme. Nous avons ensuite proposé une autre option que nous présentons ci-dessous.

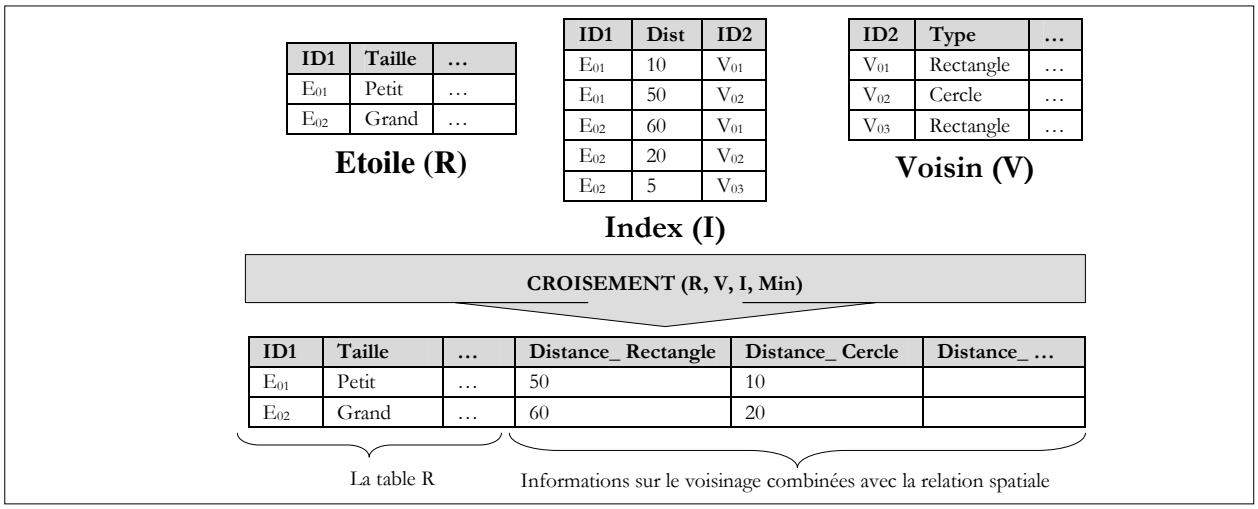


Figure 26. Illustration de l'opérateur CROISEMENT

Nous avons introduit un opérateur – appelé CROISEMENT – permettant la transformation de la structure multi-relations en une relation unique sans dupliquer les observations. Le principe de cet

opérateur est de compléter, et non pas de joindre, la table d'analyse par des données présentes dans les autres tables - comme illustré dans la Figure 26.-.

Cet opérateur est générique et peut être appliqué chaque fois qu'on a une relation pondérée de cardinalité N-M entre deux types d'objets. Il peut être vu comme un moyen de préparation des données multi-relationnelles à la fouille de données. Il permet d'adjoindre les propriétés des relations liées et de les combiner avec la pondération du lien. Les tuples n'ayant pas de liens sont complétés par des valeurs nulles tout comme une jointure externe gauche. L'application de cet opérateur nous ramène à une relation unique où chaque observation est présentée par un tuple unique. Il devient alors possible d'appliquer tout algorithme classique de fouille de données sans modification.

Application à la classification spatiale par arbre de décision

La validation de cette option a été réalisée également pour SCART, mais cette fois-ci en commençant par l'opérateur CROISEMENT (comme dans la Figure 26), puis en utilisant l'algorithme classique. Le résultat est strictement identique à la solution précédente.

Par ailleurs, cela a permis à nos partenaires du projet d'utiliser un outil de fouille de données commercial (Alice de la société Isoft) offrant plus de fonctionnalités et d'ergonomie.

V.3.5.3. *Transformation des données pour la PLI*

La "programmation logique inductive" (PLI) est née du croisement de l'apprentissage automatique et de la programmation logique. À l'inverse de la programmation logique déductive, qui dérive des conséquences à partir des théories, la programmation logique inductive a pour but de trouver des hypothèses H à partir d'un ensemble d'observations E. En cela, elle est similaire à la fouille de données traditionnelle. Cependant, elle se différencie de celle-ci par le fait que les données en entrée ainsi que les modèles extraits sont exprimés en logique du premier ordre [161].

La fouille de données multi-relationnelle est largement basée sur la programmation logique inductive (PLI). Elle consiste à transformer les données provenant des différentes relations en logique des prédicats et d'appliquer ensuite les méthodes de PLI [141] pour l'extraction des connaissances.

Une dernière solution au problème de fouille de données spatiales est donc d'exprimer ces données en logique de premier ordre, puis d'appliquer les méthodes générales de PLI. Lors de cette transformation, il est tout à fait possible d'intégrer l'interprétation du voisinage comme des connaissances implicites exprimées également en prédicats.

Application à la classification spatiale par arbre de décision

La méthode proposée S-TILDE (Spatial Top-down Induction Logical DEcision tree) [138] est basée sur l'algorithme TILDE de PLI [124]. La première phase de cette méthode est la transformation des données en logique de premier ordre.

La transformation des données relationnelles en logique de premier ordre se fait de manière classique selon les règles ci-dessous :

(i) Chaque table T devient un prédicat P ;

(ii) Chaque attribut Att de la table T devient un argument *arg* du prédicat P ;

(iii) Chaque tuple (Att_1, …, Att_n) de T devient un fait ou un modèle P (arg_1, …., arg_n).

Ainsi, la transformation de l'exemple de la Figure 23 est donnée dans la Figure 27.

Transformation des données de l'exemple en logique du 1er ordre			
Begin(model (rectangle1)).	Cercle (C1, blanc, …).	**Begin**(model (rectangle2)).	Cercle (C1, blanc, …)
Rectangle (R1, grand, …).	Cercle (C2, blanc, …).	Rectangle (R2, petit, …).	Cercle (C2, blanc, …)
Index (R1, -1, C1).	Cercle (C3, blanc, …).	Index (R2, 05, C1).	Cercle (C3, blanc, …)
Index (R1, 10, C2).	Cercle (C4, blanc, …).	Index (R2, 06, C2).	Cercle (C4, blanc, …)
Index (R1, 0, C3).	Cercle (C5, blanc, …).	Index (R2, 08, C3).	Cercle (C5, blanc, …)
Index (R1, 12, C5).	**End**	Index (R2, 05, C4).	**End**

Figure 27. Exemple de transformation des données en logique de 1er ordre.

Seulement, ces règles générales de transformation sont insuffisantes dans le cas de la classification spatiale et ce pour trois raisons. La première est qu'elles ne distinguent pas les valeurs de la classe. Pour y remédier, nous proposons l'ajout de la règle :

(iv) chaque valeur de la classe devient un prédicat dont l'argument est l'identifiant de l'objet à classer.

La deuxième raison est qu'elles ne permettent pas la combinaison des relations de voisinage et des propriétés des voisins. Pour ce faire, on modifie la règle (i) par :

(v) la transformation des tables "Index " et "voisin" est remplacée par la génération de prédicats ″voisinage (Id, relation spatiale, attributs du voisinage)″ où Id est l'identifiant de l'objet à classer.

Enfin, il faut rajouter les règles du domaine. Ici, on rajoute la règle d'inclusion des relations de voisinage :

(vi) Voisinage (id, Rel, X, Y, …, Z) ∧ (Rel < r) ⇒ Voisinage (id, r, X, Y, …, Z)

L'adaptation de cette transformation appliquée à l'exemple de la Figure 23 donne maintenant l'ensemble des prédicats suivants :

Transformation des données de l'exemple en logique du 1er ordre			
Begin(model (rectangle1)).	Voisinage (R1, 0,…).	Rectangle (R2, , …).	Voisinage (R2, 08,…).
Rectangle (R1, …).	Voisinage (R1, 02, …).	Petit (R2) « issue de R1 »	Voisinage (R2, 05,…).
Grand (R1) « issue de R1 »	Voisinage (R1, 12, …).	Voisinage (R2, 05, …).	Voisinage (R2, 13,…).
Voisinage (R1, -1, …).	**End**	Voisinage (R2, 06, …).	**End**
Voisinage (R1, 10, …).	**Begin**(model (rectangle2)).	Voisinage(R,T)∧(R < r)⇒Voisinage(r,T)	

Figure 28. Adaptation des règles de transformation des données en logique des prédicats.

Le résultat obtenu est encore une fois équivalent (non identique car TILDE utilise une formule de gain d'information légèrement différente de CART) aux deux précédentes solutions. Seule la phase de transformations en prédicats est adaptée. Les méthodes de PLI sont appliquées ensuite sans modification.

V.3.6. Applications et tests

C'est principalement dans le cadre de l'analyse de l'accidentologie en sécurité routière que nous avons testé notre approche. L'analyse du risque routier permet d'identifier les problèmes de sécurité sur le réseau routier en vue de proposer des mesures de sécurité pour y remédier. Le risque est estimé à partir du retour d'expérience sur les accidents corporels de la circulation. En effet, on dispose aujourd'hui d'une immense quantité de données numérisées sur les accidents, sur le réseau routier et sur le flux de véhicules. Ces données sont recueillies par les administrations ou par les services de police et de gendarmerie. Par ailleurs, d'autres bases de données fournissent des informations complémentaires sur l'environnement géographique comme le découpage administratif en communes, quartiers ou îlots, le bâti, la population, etc. Ces données constituent une base de données spatiales et renferment une mine d'informations utiles pour l'analyse du risque d'accidents routiers. Conscients de ce fait, les organismes d'état en charge de la sécurité routière au ministère de l'équipement et dans les collectivités locales commencent à s'intéresser à l'application des techniques de fouille de données spatiales.

L'expérimentation a été menées dans le cadre de deux projets : d'abord le projet PSIG [131], puis la collaboration avec le Conseil Général des Hauts de Seine et le CERTU [130]. Ces études ont porté sur des données réelles, la première relative aux 86 communes gérées par la communauté urbaine de Lille, la seconde concerne deux communes du département des Hauts de Seine.

Hormis la sécurité routière, nous avons appliqué ce modèle dans l'analyse de la contamination des coquillages pour le test de la méthode STILDE. Puis, dans le cadre d'une collaboration avec l'université Hassan II, nous avons étudié les règles d'association spatiales en appliquant l'algorithme CLOSE [173], [127], [126]. Le test a porté sur l'analyse de l'implantation des cybercafés à Mohammedia au Maroc.

V.3.7. Expérimentations et résultats de SCART

Les solutions proposées pour la classification par arbre de décision ont été implémentées dans un prototype nommé ADS (Arbre de Décision Spatial). Pour plus de détail, le lecteur pourra se référer aux rapports du projet [136]. Un des objectifs était d'explorer des relations cachées entre les accidents et les autres couches comme le réseau ou le tissu urbain. Ceci revient à appliquer la classification par arbre de décision en intégrant le caractère spatial des accidents et leur interaction avec l'environnement géographique. L'analyse part d'une base de données spatiale comprenant des données sur les accidents de la route et de l'environnement géographique (Enceintes publiques, voierie).

Le premier exemple vise à classer les segments de route en accidentogènes (plus de deux accidents) ou non. Les attributs explicatifs sont, soit liés aux sections de route (ex : sens de circulation), soit liés à

l'environnement urbain (ex : école, marché, administration, etc.) combinés avec la relation spatiale « distance » (ex : distance_école ≤ 425m). L'arbre de décision spatial obtenu est illustré Figure 29.

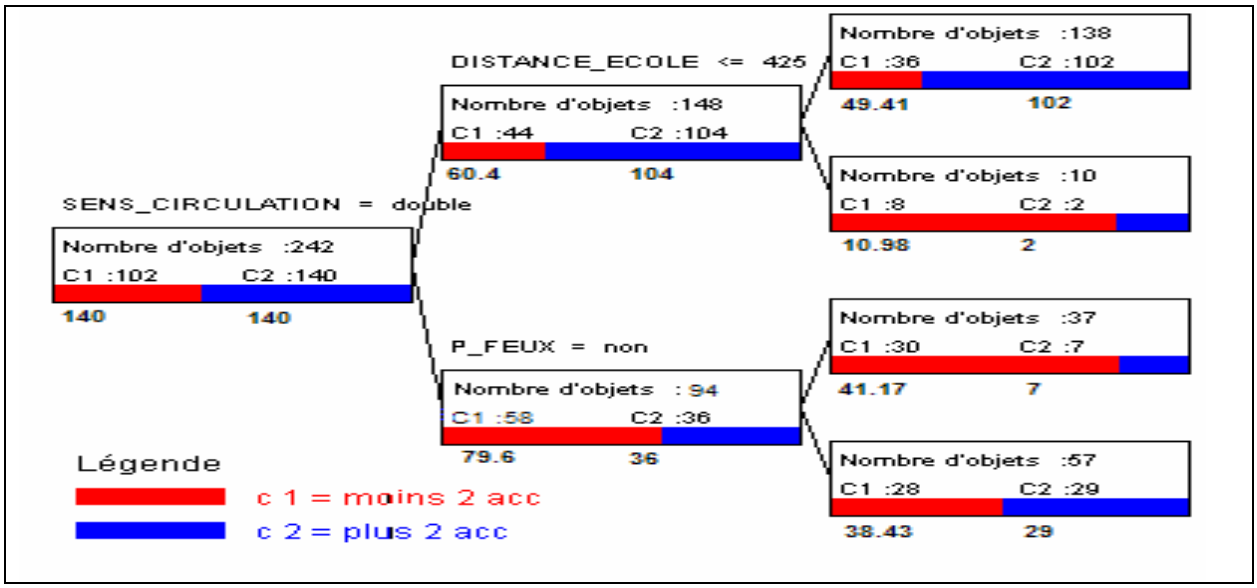


Figure 29. Arbre de décision spatial – Exemple 1

Un autre exemple a concerné, non pas la classification des segments, mais des occurrences d'accidents (localisations ponctuelles) afin de décrire ou d'expliquer le lien entre le type d'accident (impliquant au moins un piéton, au moins un 2 roues, ou uniquement des véhicules –autres-) avec le voisinage de l'accident. Un exemple de résultat est donné Figure 30.

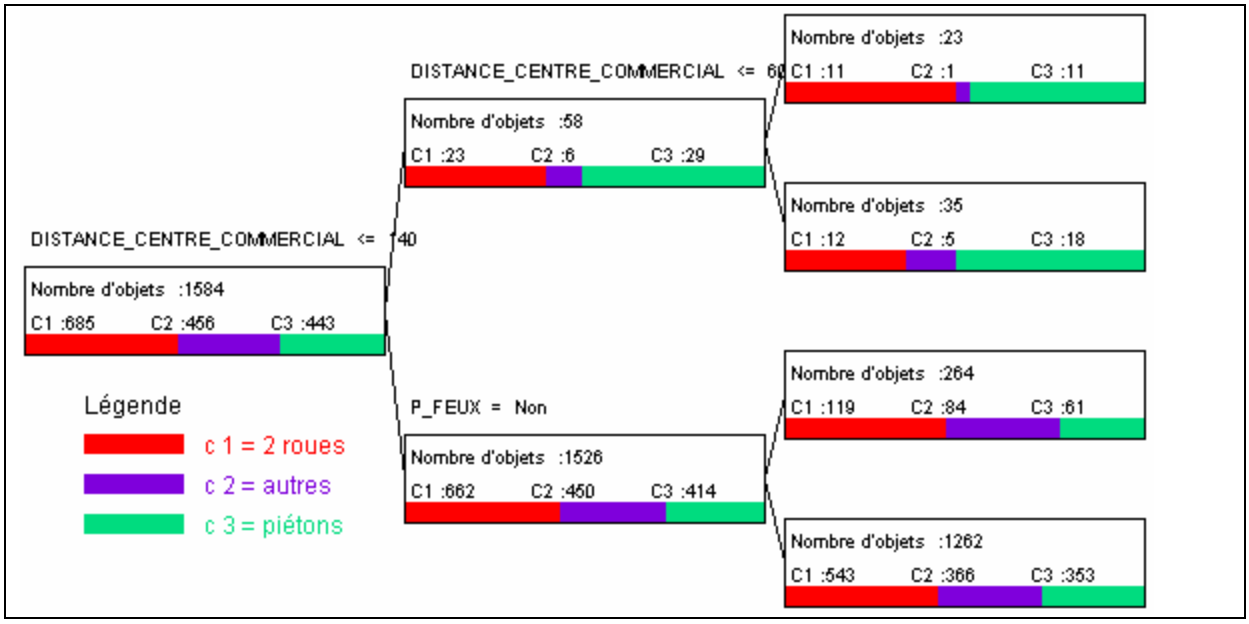


Figure 30. Arbre de décision spatial – Exemple 2

Remarque :

La règle (distance-centre-commercial < 60) -> plus d'accidents piétons ou 2 roues de la Figure 30 ne signifie pas que « la proximité des centres commerciaux génère des accidents piétons ou 2 roues », mais plutôt que « parmi les accidents à classer, ceux à 20 m des centres commerciaux impliquent

proportionnellement plus de piétons et 2 roues ». Ces accidents ne sont pas plus nombreux que dans le reste de la carte, car le nombre d'accidents n'est pas pris en compte. Il faut donc être prudent quant à l'interprétation des règles générées. A noter que les méthodes de classification proposées dans les autres travaux rencontrent le même problème car c'est une discrimination des classes (du phénomène accident) mais pas du nombre d'occurrences d'accidents.

V.3.7.1. *Tests des performances*

Les mesures de performances [130] [135] [137] ont été réalisées pour comparer les deux premières options et montrent que la construction du modèle est aussi efficace dans les deux solutions et que le coût est plutôt linéaire par rapport au volume de données. La solution 2 proposée suppose néanmoins une phase de préparation par l'opérateur CROISEMENT. Ce dernier a un coût élevé (environ 22 minutes pour près de 30000 objets d'analyse). Ce coût est toutefois négligeable - de 70 à 100 fois moins - par rapport à celui de la construction de l'index de jointure qui fait partie également de la phase de préparation. Ce coût reste acceptable car la préparation n'est faite qu'une seule fois mais sert à la construction de modèles différents d'arbre de décision ou de règles d'association en variant les paramètres.

V.4. Conclusion

La principale contribution que nous avons apportée dans ce domaine est une vision multi-thématique de la fouille d'objets localisés. L'espace n'est pas pris en compte en tant que tel, mais les relations de voisinage avec d'autres objets et les propriétés de ces objets sont considérées et ajoutées comme données d'analyse. Cette interprétation n'est pas très différente des travaux antérieurs, mais a le mérite d'être explicite. Fonctionnellement, l'avantage de notre approche, par rapport aux méthodes existantes, est la détermination automatique du périmètre de voisinage.

Après avoir testé les méthodes existantes [191] de généralisation [164] et de caractérisation [144], nous nous sommes focalisés sur la classification et les règles d'association. Nous avons développé de nouvelles approches et avons implémenté différents algorithmes. Cela a permis de valider des approches efficaces et flexibles, intégrées aux bases de données et de programmation logique pour les mêmes méthodes. Les solutions proposées, ont fait l'objet de prototypes et ont été testées dans le cadre d'applications dans différents domaines (analyse de l'accidentologie et contamination de coquillages). Le prototype ADS développé au cours de la thèse de Nadjim Chelghoum a été fourni au CERTU et à l'Université d'Alger et de Casablanca, où il a fait des émules [190] [167].

Une des propositions offrant le plus de flexibilité est de ramener le problème de fouille de données spatiales à la fouille de données classique, grâce à l'opérateur CROISEMENT proposé. Nous avons ainsi pu appliquer des méthodes traditionnelles ainsi que des méthodes de fouille de données relationnelles.

Les travaux se poursuivent dans ce domaine. La question clé reste la définition même du problème d'analyse. Ainsi, Rinzinvillo et Turini posent la question de la définition d'une transaction spatiale et proposent une interprétation similaire à la notre dans [182] et [183]. Shekhar et Chawla traitent de la co-

localisation et cherchent à l'optimiser en évitant les jointures [189]. Aujourd'hui, le fait marquant est que des éditeurs comme Oracle offrent des solutions pour la fouille de données spatiales [181]. Leur approche actuelle rejoint notre idée de prétraitement pour aboutir à une structure mono-relationnelle sur laquelle ils appliquent une méthode de fouille de données classique (Figure 31), mais l'enrichissement des données reste limité.

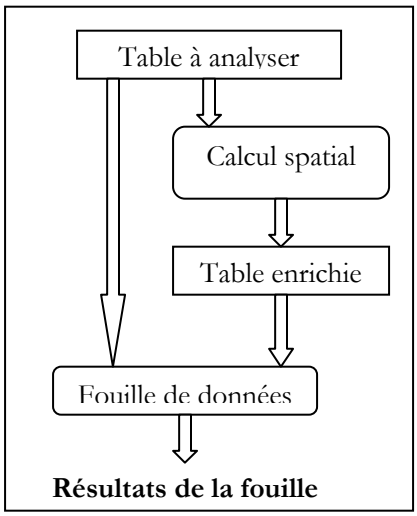


Figure 31. Approche d'Oracle pour la fouille de données spatiale [181]

Avec le recul, nous avons tiré de notre expérience les leçons suivantes :

- Il est impératif de définir, comme nous l'avons fait, le thème cible de l'analyse dans les modèles de fouille de données spatiales.
- La matérialisation des propriétés spatiales par des attributs facilite l'interprétation et permet l'utilisation d'une large gamme de méthodes existantes.
- Les prétraitements, tels que l'index de jointure spatiale ou l'opérateur CROISEMENT, sont essentiels pour l'optimisation des performances d'exécutions des algorithmes de fouille de données spatiales.
- Cependant, décrire individuellement des objets sans tenir compte de la densité (agrégation spatiale) du phénomène étudié est insuffisant. Une combinaison des méthodes d'agrégation spatiales (généralisation ou clustering en zones) et de notre vision par objet est probablement un bon compromis des méthodes numériques et logiques proposées. Une solution possible est de chercher d'abord les concentrations anormales formant des unités spatiales, ensuite de classer ces unités par type en tenant compte du voisinage. Seulement, les autres propriétés individuelles des accidents sont perdues et doivent être généralisées au niveau des clusters.
- Concernant l'analyse en sécurité routière, l'idée de départ était de trouver des liens avec des localités (école, commerces, gares, etc.) aux alentours qui peuvent expliquer le nombre, la gravité ou le type d'accidents routiers. Ce lien est ensuite analysé comme une influence sur le risque d'accident, car ces localités génèrent le trafic de personnes ou de véhicules à l'origine des accidents. Mettre en évidence ce lien n'est pas suffisant, car le trafic généré varie dans le temps et obéit à des modèles complexes qui ne

sont pas pris en compte. Or, une fois qu'on connaît le trafic (donné par la concentration et la vitesse) à chaque endroit du réseau et à chaque moment, on n'a plus besoin de chercher des informations de voisinage. Dans le projet HEARTS par exemple, la concentration et la vitesse sont estimées à partir des connaissances sur la mobilité. Il serait intéressant de partir plutôt de ces données, mais peut-on encore parler de fouille de données spatiales ?

V.5. Références

[118] Agrawal R., Imielinski T. & Swami A., Mining Association Rules between sets of items in large databases. Proceedings of the ACM SIGMOD. Washington, DC, pp. 207-216 (1993).

[119] Ambroise, C., Dang, M. V., Govaert, G., « Clustering of spatial data by the EM algorithm », In Soares, A., Gômez-Hernandez, J., and Froidevaux, F. (Eds), geoENV1-Geostatistics for Environmental Applications, volume 9, 1997, Quantitative Geology and Geostatistics Publisher, Kluwer Academic, pp. 493-504.

[120] Ankerst M., Breunig M.M., Kriegel H-P., Sander J., OPTICS: Ordering Points To Identify the Clustering Structure. SIGMOD Conference 1999, pp. 49-60

[121] Anselin L. (1989), What is special about spatial data? Alternative perspectives on spatial data analysis, Technical paper 89-4. Santa Barbara, NCGIA.

[122] Anselin, L., Local indicators of spatial association – LISA, Geographical Analysis, Vol. 27(2), 1995, pp. 93-115.

[123] Aufaure M-A, Yeh L., Zeitouni K., "Fouille de données spatiales", Chapitre dans "Le temps, L'espace et l'évolutif en sciences du traitement de l'information - Tome 2, Cépaduès éditions, France, Septembre 2000, pp 319-328.

[124] Blockeel H., L. De Raedt, (1998) Top-Down induction of first order logical decision trees, Artificial intelligence, 102(2-2)/ 285-297.

[125] Boulmakoul A., Zeitouni K., Chelghoum N., R. Marghoubi, "Fuzzy structural primitives for spatial data mining", In 2nd IEEE International Symposium on Signal Processing and Information Technology (ISSPIT 2002), Marrakech, Morocco, December, 2002, pp. 294-298.

[126] Boulmakoul A., Marghoubi R., Zeitouni K., *Utilisation des treillis de Galois pour l'extraction et la visualisation des règles d'association spatiales*, ", Conférence INFORSID, Hammamet, Tunisie, Juin 2006, pp. 703-718.

[127] Boulmakoul A., Zeitouni K., Marghoubi R., "The Use of the Galois lattice for the extraction and the visualization of the spatial association rules", In 6th IEEE International Symposium on Signal Processing and Information Technology (ISSPIT 2006), Vancouver, Canada, August, 2006.

[128] Breiman L., J.H. Friedman, R.A. Olshen, C.J. Stone, (1984), Classification and Regression Trees, Ed. Wadsworth & Brooks. Montery, California.

[129] Ceci M., A. Appice, D. Malerba, Spatial Associative Classification at Different Levels of Granularity: A Probabilistic Approach, in J.-F. Boulicaut, F. Esposito, F. Giannotti, & D. Pedreschi (Eds.), Knowledge Discovery in Databases: PKDD 2004, Lecture Notes in Artificial Intelligence, 3202, 99-111, Springer, Berlin, Germany, 2004.

[130] Chelghoum N. (2004), Fouille de données spatiales - Un problème de fouille de données multi-tables, Thèse de Doctorat de l'Université de Versailles-Saint-Quentin, Décembre 2004 (www.prism.uvsq.fr/~nchelg).

[131] Chelghoum N., Zeitouni K., "Arbre de décision spatial - Application en sécurité routière", 1ères Journées francophones d'Extraction et de Gestion des Connaissances, EGC 2001, Nantes, Janvier 2001.

[132] Chelghoum N., Zeitouni K., Boulmakoul A., "Arbre de décision spatial multi-thèmes", 8ème rencontres de la Société Francophone de Classification, SFC'01, Pointe-à-Pitre, Guadeloupe, Décembre 2001.

[133] Chelghoum N., Zeitouni K., Boulmakoul A., "Fouille de données spatiales par arbre de décision multi-thèmes", Secondes Journées sur l'Extraction et la Gestion des Connaissances, EGC'2002, Montpellier, Janvier 2002, pp. 281-286.

[134] Chelghoum N., Zeitouni K., Boulmakoul A., "A Decision Tree for Multi-layered Spatial Data", In 10th International Symposium on Spatial Data Handling (SDH'02), Ottawa, Canada, July 2002, pp 1-10.

[135] Chelghoum N. (2004), Zeitouni K., « Mise en oeuvre des méthodes du data mining spatial. Alternatives et performances », 4èmes Journées d'Extraction et de Gestion des Connaissances, EGC 2004, Clermont-Ferrand, France, 20-23 Janvier 2004, Edition CEPADUES, Volume I, pp. 211-217.

[136] Chelghoum N. (2004), Zeitouni K., "Extension du projet TOPASE par la prise en compte des interactions entre le réseau viaire et l'environnement urbain", Convention PRISM-CERTU, Juillet 2004. http://www.prism.uvsq.fr/users/nchelg/Rapport_Certu.zip

[137] Chelghoum N. (2004), Zeitouni K., "Spatial data mining implementation - Alternatives and performances", in the 9th Brazilian Symposium on GeoInformatics, GEOINFO'04, Novembre 2004, Campos do Jordão, Brazil.

[138] Chelghoum N., Zeitouni K, Laugier T., Fiandrino A., Loubersac L., "Fouille de données spatiales - Approche basée sur la programmation logique inductive", 6èmes Journées d'Extraction et de Gestion des Connaissances, EGC 2006, Edition CEPADUES, Lille, Janvier 2006, pp. 529-540.

[139] Cliff A.D., Ord J.K., "Spatial autocorrelation", Pion, London, 1973.

[140] Cressie N.A.C, Statistics for spatial data, Edition Wiley, New York, 1993.

[141] Dzeroski S., N. Lavrac, Relational Data Mining, Springer, Berlin, 2001.

[142] Ester M., Kriegel H.P., Sander J., Xu X., "A density-Based algorithm for discovering clusters in lager spatial databases with noise", In proceeding of second international conference on knowledge discovery and data mining, Portland, 1996, pp 226-231.

[143] Ester M., Kriegel H.P., Sander J., "Spatial Data Mining: A Database Approach", in proceedings of 5th Symposium on Spatial Databases, Berlin, Germany, 1997.

[144] Ester M., Frommelt A., Kriegel H.-P., Sander J., "Algorithms for Characterization and Trend Detection in Spatial Databases", Proc. 4th Int. Conf. on Knowledge Discovery and Data Mining, New York, NY, 1998.

[145] Fayyad U.M., Piatetsky-Shapiro G., Smyth P., and Ramasamy Uthurusamy, "Advances in Knowledge Discovery and Data Mining", AAAI Press / MIT Press, 1996.

[146] Fotheringham S., Zhan B., "A comparison of three exploratory methods for cluster detection in spatial point patterns", Geographical Analysis, Vol. 28, n° 3, 1996, pp. 200-218

[147] Gardarin G., "Internet, Intranet et bases de données : Data Web, Data Media, Data Warehouse, Data Mining ", Editions Eyrolles, 1999.

[148] Gatrell A., Bailey T., Diggle P., Rowlingson B., "Spatial point pattern analysis and its application in geographical epidemiology", Transactions of the Institute of British Geographers, n° 21, 1996, pp. 256-274.

[149] Govaert G., "Classification automatique et modèle de mélange: application aux données spatiales", Revue Internationale de Géomatique, Editions Hermès, Vol. 9, n° 4, 1999, Mai 2000, pp. 457-470.

[150] Han J., Koperski K., and Stefanovic N., "GeoMiner: A System Prototype for Spatial Data Mining", Proc. ACM-SIGMOD Int. Conf. on Management of Data (SIGMOD'97), Tucson, Arizona, May 1997.

[151] Han J., Kamber M., Data Mining – Concepts and Techniques, Morgan Kaufmann Publishers, 2000.

[152] Huang Y., Xiong H., Shekhar S., and Pei J., Mining Confident Co-location Rules without a Support Threshold. 18th ACM Symp. on Applied Computing, 2003.

[153] Huang Y., Shekhar Sh., and Xiong H., Discovering Co-location Patterns from Spatial Datasets: A General Approach, IEEE Transactions on Knowledge and Data Engineering (TKDE), 16(12), pp. 1472-1485, December 2004

[154] Huguenin-Richard F, Lassarre S, Yeh L, and Zeitouni K, "Extraction de connaissances des bases de données spatiales en accidentologie routière", Journées Cassini, La Rochelle, Septembre 2000.

[155] Knobbe. A.J., Siebes A., Wallen V., Daniel M.G., "Multi-relational Decision Tree Induction", In Proceedings of PKDD' 99, Prague, Czech Republic, Septembre 1999.

[156] Koperski K. and Han J., "Discovery of Spatial Association Rules in Geographic Information Databases", In Advances in Spatial Databases (SSD'95), p. 47-66, Portland, ME, August 1995.

[157] Koperski K., J. Adhikary, J. Han, Knowledge Discovery in Spatial Databases: Progress and Challenges, Proc. SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery (DMKD). Technical Report 96-08, University of British Columbia, Vancouver, Canada, 1996.

[158] Koperski K., Han J., Stefanovic N., "An Efficient Two-Step Method for Classification of Spatial Data", In proceedings of International Symposium on Spatial Data Handling (SDH'98), p. 45-54, Vancouver, Canada, July 1998.

[159] Koperski K., "A progressive refinement approach to spatial data mining", PhD Thesis, Simon Fraser University. April 1999.

[160] Koperski K., Han J., Marchisio G. B., "Mining Spatial and Image Data through Progressive Refinement Methods", Revue Internationale de Géomatique n° 4/99, Mai 2000, 425-440.

[161] Lavrac N., Dzeroski S., "Inductive logic programming: Techniques and applications", Edition Ellis Horwood, New York, 1994, pp 3-38.

[162] Lebart L. et al. "Statistique exploratoire multidimensionnelle", Edition Dunod, Paris, 2éme édition, 1997.

[163] Lu W., and Han J., "Distance-Associated join indices for spatial range search", Proceeding of Eighth International Conference on Data Engineering, Tempe, Arizona, Fabruary 1992, pp. 284-292.

[164] Lu W., Han J. and Ooi B. C., "Discovery of General Knowledge in Large Spatial Databases", in Proc. of 1993 Far East Workshop on Geographic Information Systems (FEGIS'93), Singapore, June 1993, pp. 275-289.

[165] Malerba D., F.A. Lisi, An ILP Method for Spatial Association Rule Mining. In A. Knobbe and D. van der Wallen (Eds.), Notes of the ECML/PKDD 2001 Workshop on Multi-Relational Data Mining, 18-29, Germany Freiburg, 2001.

[166] Marghoubi R., Boulmakoul A., Zeitouni K., "Spatial Mining with the Galois lattice for information technologies", Int. Conf. on Modeling and Simulation (ICMS'05), Marrakech, Morocco, November, 2005.

[167] Marghoubi R., « Fouille de données spatiales - Utilisation des treillis de Galois pour l'extraction des règles d'association spatiales. Application au domaine du service universel des télécommunications », Thèse de Doctorat de l'Université Hassan II, Faculté des Sciences et Techniques de Mohammedia, Mai 2006.

[168] Miller H.J., Han J., *Geographic Data Mining and Knowledge Discovery*, Taylor & Francis, 2001.

[169] Ng R., Han J., ″Efficient and effective clustering method for spatial data mining″, in proceeding of international conference on very large database, Santiago, Chile, September 1994, p- 144-155.

[170] Openshaw S., Charlton M., Wymer C., Craft A., 1987 : "A mark 1 geographical analysis machine for the automated analysis of point data sets", International Journal of Geographical Information Systems, Vol. 1, n° 4, pp. 335-358

[171] Openshaw S., 1995, "Developing automated and smart spatial pattern exploration tools for geographical information systems applications", The Statistician, Vol. 44, n° 1, pp. 3-16

[172] Ord J.K., Getis A., 1995, "Local Spatial Autocorrelation Statistics: Distributional Issues and an Application, Geographical Analysis", Ohio State University Press, Vol. 27, n° 4, pp. 287-306

[173] Pasquier N., Bastide Y., Taouil R., Lakhal L., "*Efficient Mining of Association Rules using Closed Itemset Lattices*", *Journal of Information Systems*, vol. 24, no 1, 1999, Elsevier Science, pp. 25-46.

[174] Quinlan J.R., Induction of Decision Trees, Machine Learning (1), 82 - 106, 1986.

[175] Sander J., Ester M., Kriegel H.-P., Xu X., Density-Based Clustering in Spatial Databases: The Algorithm GDBSCAN and its Applications, in Data Mining and Knowledge Discovery, An International Journal, Kluwer Academic Publishers, Vol. 2, No. 2, 1998.

[176] Shaw G., D. Wheeler, Statistical Techniques in Geographical Analysis, Edition David Fulton, London, 1994.

[177] Shekhar Sh. and Huang Y., Discovering Spatial Co-location Patterns : A Summary of Results, In Proc. of 7th Int. Symposium on Spatial and Temporal Databases (SSTD), Springer-Verlag, Lecture Notes in Computer Science, July 2001

[178] Shekhar Sh.., Chawla Sanjay (2003), Spatial Databases: A Tour, Prentice Hall.

[179] Shekhar Sh., Lu C.T., Chawla S., Ravada S., Efficient Join Index Based Join Processing; A Clustering Approach, IEEE Transactions on Knowledge and Data Engineering, 15(1), 2003.

[180] Tobler W.R., Cellular geography, In Gale S, Olsson G, In Phylosophy in Geography Edition, Dortrecht, Reidel, 379-86, 1979.

[181] Ravikanth V. K., Oracle Corporation, Oracle Database 10g - Empowering Applications with Spatial Analysis and Mining , Oracle Technical White Paper, February 2004.

[182] Rinzivillo S., Turini F., Classification in geographical information system, In 8th Conf. on Principles and Practice of Knowledge Discovery in Databases (PKDD), Pisa, Italy, 2004, pp 374-385.

[183] Rinzivillo S., Turini F., Extracting Spatial Association Rules from Spatial Transactions, 13th ACM International Workshop on Geographic Information Systems, ACM-GIS 2005, November 4-5, 2005, Bremen, Germany, Proceedings. ACM 2005, pp 79-86.

[184] Roddick, J.F, Spiliopoulou, M.: A Bibliography of Temporal, Spatial and Spatio-Temporal Data Mining Research, ACM SIGKDD Explorations, volume 1, Issue 1 (1999)

[185] Rotem D: Spatial join indices, Proc. of 7th Conf. on Data Engineering, Kobe, Japan (1991) pp. 500-509.

[186] Valduriez P., "Join Indices", ACM Transactions on Database Systems, 12 (2), June 1987, pp. 218-246.

[187] Van Laer W., L. De Raedt, How to Upgrade Propositional Learners to First Order Logic: a Case Study. Chapter in "Relational Data Mining", Dzeroski S., N. Lavrac Ed., Springer, Berlin, 235-261, 2001.

[188] Yeh T-S., "Spot: Distance based join indices for spatial data", ACM GIS 99, Kansas City, 5-6 Nov 1999.

[189] Yoo J. S.and Shekhar Sh., A Join-less Approach for Mining Spatial Co-location Patterns, to appear in IEEE Transactions on Knowledge and Data Engineering (TKDE), 2006.

[190] Zeghache S., Admane F., Elarabia Ziane K., "Outil de datamining spatial appliqué à l'analyse des risques liés au territoire", Poster, 6èmes Journées d'Extraction et de Gestion des Connaissances, EGC 2006, Edition CEPADUES, Lille, Janvier 2006, pp. 723-724.

[191] Zeitouni, K., "Etude de l'application du data mining à l'analyse spatiale du risque d'accidents routiers par l'exploration des bases de données en accidentologie", Rapport de contrat PRISM-INRETS, Décembre 1999.

[192] Zeitouni K., Data Mining Spatial - Numéro spécial de la Revue Internationale de Géomatique, Editions Hermès, Vol. 9, N° 4 / 1999, Mai 2000.

[193] Zeitouni K., Yeh L., "Le data mining spatial et les bases de données spatiales", Revue Internationale de Géomatique, Editions Hermès, Vol. 9, N° 4 / 1999, Mai 2000, pp. 389-423.

[194] Zeitouni K., "A Survey on Spatial Data Mining Methods Databases and Statistics Point of Views", 11th Information Resources Management Association International Conference (IRMA 2000), Data Warehousing and Mining Track, IDEA Group Publishing, May, 2000, Anchorage, Alaska, USA, pp. 487-491.

[195] Zeitouni K., Yeh L., Aufaure M.A, "Join indices as a tool for spatial data mining", International Workshop on Temporal, Spatial and Spatio-Temporal Data Mining, in conjunction with PKDD, TSDM 2000, LNAI n° 2007, Springer, September, 2000, Lyon, France, pp 102-114.

[196] Zeitouni K., Chelghoum N., "Spatial Decision Tree - Application to Traffic Risk Analysis", ACS/IEEE International Conference on Computer Systems and Applications, Beirut, Lebanon, June, 2001.

[197] Zeitouni K., "A Survey on Spatial Data Mining Methods - Databases and Statistics Point of Views", Book Chapter In "Data Warehousing and Web Engineering", Shirley Becker Editor, IRM Press, pp. 229-242, 2002

[198] Zhang T., Ramakrishnan R., Livny M., BIRCH: An Efficient Data Clustering Method for Very Large Databases. SIGMOD Conference 1996, pp. 103-114.

[199] Zhe L., Kenneth A. R., "Fast joins using join indices ", The VLDB Journal Vol 8, Springer-Verlag Ed., pp 1-24, 1998.

[200] Zighed A., Ricco R., "Graphes d'induction - Apprentissage et Data Mining", Edition Hermès Sciences, 2000.

CHAPITRE VI. FOUILLE D'AUTRES DONNEES COMPLEXES

Ce chapitre décrit nos travaux sur la fouille de données complexes, principalement les séquences et le texte. Ces travaux se sont déroulés principalement entre 2002 et 2005 dans le cadre du projet RNTL *Contexte Bourse et d'un volet du projet Européen HEARTS. Ils font partie de la thèse de Huaizhong Kou [215] pour la fouille de texte et de la thèse de Lionel Savary [230] pour la fouille de séquences.*

Tout comme la fouille de données spatiales et spatiotemporelles, la fouille des autres données complexes soulève souvent des problèmes similaires qui nécessitent des adaptations ou des extensions des méthodes de fouille de données traditionnelles. En général, on retrouve soit des adaptations en amont dans la phase de préparation de données, ou bien des extensions des algorithmes de fouille en y intégrant les spécificités des données complexes analysées.

Nous nous sommes intéressés à la fouille d'autres données complexes dont les données séquentielles, symboliques et textuelles. Dans chacun de ces domaines, nous avons proposé de nouveaux algorithmes ou des démarches complètes. Ainsi, un algorithme d'extraction de règles d'associations des données séquentielles a été développé dans la thèse de Lionel Savary [230]. Nous avons exploré avec Tao Wan la fouille de données symboliques et avons proposé une approche à support multiple pour les règles d'association dans ce contexte [238][239]. Enfin, la fouille de textes a fait l'objet de la thèse de Huaizhong Kou [215] où différentes techniques ont été mises en œuvre pour la catégorisation de textes.

Nous résumons les principales contributions en fouille de données séquentielles dans la première section, puis en fouille de données textuelles dans la seconde.

VI.1. Fouille de données séquentielles

Parmi les travaux portant sur la fouille de séquences, nous nous intéressons ici à l'extraction de sous-séquences fréquentes et à la fouille de séquences multidimensionnelles. Nous exposons ci-dessous notre contribution dans l'extraction de motifs séquentiels. Quant à la fouille de séquences multidimensionnelles, nous renvoyons sur la thèse de Savary [230] et l'article [231].

L'extraction de motifs fréquents a toujours pris une place importante dans le processus d'extraction de connaissances. La recherche de motifs séquentiels s'applique particulièrement dans le contexte de l'analyse de la consommation, à la découverte de co-occurrences dans des documents textuels, à l'analyse de séquences d'ADN ou encore à l'analyse des parcours de pages sur des serveurs Web. Dans nos travaux, nous l'avons appliquée à des séquences d'évènements temporels décrivant la mobilité, plus précisément aux séquences d'activités journalières d'une population. Cette technique se focalise donc sur l'aspect temporel de phénomènes spatio-temporels (ici, la mobilité).

VI.1.1. Problématique et aperçu de l'état de l'art

Les bases de données sont de plus en plus volumineuses entraînant des dégradations de performances. Aussi, la mise en place d'un algorithme pour l'extraction de motifs séquentiels soulève principalement les questions suivantes :

1. minimiser le nombre de lectures ou de parcours dans la base de données ;
2. réduire au mieux l'espace occupé par les données en mémoire principale ;
3. optimiser la recherche et la génération de candidats dans la structure de données de l'algorithme.

La plupart des travaux portant sur la recherche de séquences fréquentes se situent dans le domaine de l'analyse de la consommation. A l'origine l'algorithme bien connu *Apriori* [201] ne considérait que des transactions indépendantes. Trois algorithmes traitant les séquences de transactions, c'est-à-dire les listes de transactions par client, ont été développés et comparés dans [202] : *AprioriAll*, *AprioriSome* et *DynamicSome*. L'algorithme *AprioriAll* est une adaptation de l'algorithme *Apriori* pour les séquences où la génération de candidats et les calculs de supports sont modifiés par rapport à la méthode de base. *AprioriSome* et *DynamicSome* sont des versions optimisées retournant les séquences fréquentes maximales. Pour ce faire, ils opèrent un saut dans le treillis des partitions permettant d'atteindre plus rapidement les fréquents maximaux. L'inconvénient majeur des méthodes basées sur *Apriori* reste leur lenteur, car ils nécessitent plusieurs lectures de la base de données engendrant un coût en entrées/sorties élevé.

L'algorithme *SPAM* [211] charge la base de données dans un arbre lexicographique et représente les séquences par un bitmap [206]. L'algorithme n'effectue donc ici qu'un seul parcours de la base de données pour la charger en mémoire, à condition que celle-ci soit de taille suffisante. L'inconvénient est qu'il considère que ces structures de données résident entièrement en mémoire principale.

L'algorithme *GSP* [234] est similaire à *Apriori*, sauf que les candidats sont générés plus efficacement en mémoire. Dans [223], les auteurs proposent un algorithme de type incrémental où les candidats sont générés selon le principe de GSP et montrent qu'il est plus performant que GSP.

L'algorithme *SPADE* [243] recherche des sous-séquences fréquentes en utilisant un partitionnement vertical de la base de données et des opérations de jointures appropriées, pour transformer des séquences en listes ordonnées d'articles. Ces listes sont jointes pour former un treillis utilisé pour trouver des sous-séquences fréquentes.

L'algorithme *PrefixSpan*[226] permet de générer des sous-séquences fréquentes en deux lectures de la base et sans génération de candidats. Durant la première lecture, les articles fréquents sont déterminés et sont considérés comme préfixes pour les séquences de la base de données les contenant. Puis les suffixes de ces séquences sont projetés dans des bases de données intermédiaires. Les séquences ayant les mêmes préfixes sont utilisées pour former des sous-séquences de taille 2. L'algorithme est alors réitéré de manière récursive.

Tableau 3 : *Performances des algorithmes de recherches de sous-séquences fréquentes*

Algorithme	Génération de candidats	Parcours de la base de données	Structure de données	Temps d'exécution	Ressources en mémoire
Apriori	Oui	Max(Freq)	Arbre haché	+ + + + + +	+ + + + +
AprioriSome	Oui	Max(Freq)	Arbre haché	+ + + + +	+ + + + + +
GSP	Oui	Max(Freq)	Arbre haché	+ + + +	+ + + +
SPADE	Oui	3	Treillis	+ + +	+ + +
PrefixSpan	Non	2	Bases projetées	+ +	+
SPAM	Oui	1	Tableau binaire	+	+ +

Le Tableau 3 récapitule les caractéristiques et les performances des différents algorithmes. On constate que *AprioriSome* offre de meilleures performances d'exécution mais requière plus d'espace mémoire qu'*Apriori*. GSP offre de meilleurs résultats qu'*Apriori* et *AprioriSome* grâce à une génération de candidats optimisée basée sur un principe de fusion. *PrefixSpan* n'effectue aucune génération de candidats, contrairement aux autres algorithmes. *SPADE*, *PrefixSpan* et *SPAM* sont de loin les algorithmes les plus performants. En effet, ils n'effectuent qu'un à trois parcours de la base de donnée, là où *Apriori*, *AprioriSome* et *GSP* effectuent un nombre de parcours égale à Max(Freq), la longueur maximale de sous-

séquences fréquentes trouvées. *SPAM* offre de meilleures performances d'exécution que *PrefixSpan*, mais consomme plus de ressources mémoire que ce dernier.

VI.1.2. Contributions

L'algorithme proposé doit s'adapter aux grandes bases de données et être efficace en temps d'exécution, dans des configurations de taille mémoire standard. Nous avons proposé un nouvel algorithme dans [230] [232] qui ne nécessite qu'une seule passe de la base de données et optimise les ressources mémoire. Cet algorithme a été implémenté et les tests comparatifs ont été effectués avec les deux meilleurs algorithmes existants, à savoir PrefixSpan et SPAM. Les tests ont montré l'efficacité et les performances de cet algorithme.

L'avantage majeur de cet algorithme réside dans la compacité de sa structure de données réduisant l'espace mémoire utilisé. Il peut ainsi s'adapter à de très grandes bases de données. Cette structure est basée sur un tableau de bits indexé (TBI). A ce tableau est associé un vecteur de séquences (VS) permettant de coder toutes les séquences enregistrées dans la base. Ce vecteur est construit au fur et à mesure en fusionnant les séquences de la base en une séquence fictive qui servira de référentiel pour la représentation des autres (voir Figure 32). L'algorithme de construction opère des décalages afin de maintenir l'ordre et de minimiser l'extension du vecteur. Les détails sont donnés dans l'article joint à ce mémoire. La longueur du vecteur de séquences détermine donc la largeur du tableau. TBI ne représente que les séquences distinctes de la base de données, limitant ainsi sa longueur. A titre d'exemple, dans notre base de données, sur 13054 individus il n'existe que 3429 séquences distinctes. La longueur du VS générée n'est que 79, pour une longueur maximale de séquences dans la base de 20 articles distincts.

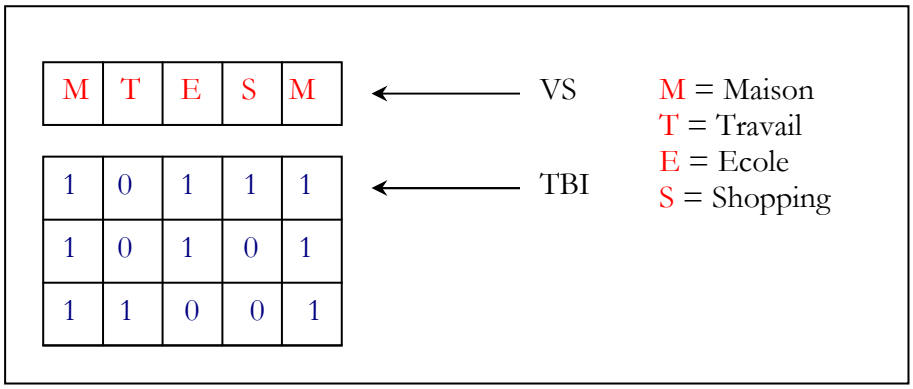


Figure 32: *Tableau de bit indexé*

Dans la Figure 32, le vecteur de séquences est constitué de cinq activités dans l'ordre (M, T, E, S, M). Dans cet exemple, nous supposons que la base de données est composée de trois séquences distinctes codées dans le TBI. Les bits du TBI placés à 1 indiquent les articles présents dans une séquence particulière, en correspondance avec le VS. Dans cet exemple, les trois séquences distinctes sont : (M, E, S, M), (M, E, M) et (M, T, M).

Les séquences sont classées par taille décroissante dans le TBI. Un tableau NB est associé au TBI et renseigne la fréquence des séquences dans la base. Un index associé au TBI permet un accès direct aux

séquences selon leur taille et donc de comparer efficacement les séquences d'une certaine taille directement avec les séquences de même taille ou de tailles supérieures (voir Figure 33). L'unique lecture de la base de données permet de créer cette structure.

Dans l'exemple de la Figure 33, la base contient des séquences de taille allant de 1 à 5. Comme indiqué dans la tableau NB, la séquence (M, T, R, T, M) se retrouve 15 fois dans la base de données. La première cellule de valeur 1 de l'index pointe sur la première séquence de taille 1 du tableau, la deuxième cellule de valeur 4 (n° de ligne dans TBI partant du bas) pointe sur la première séquence de taille 2 du tableau située à la quatrième ligne et ainsi de suite.

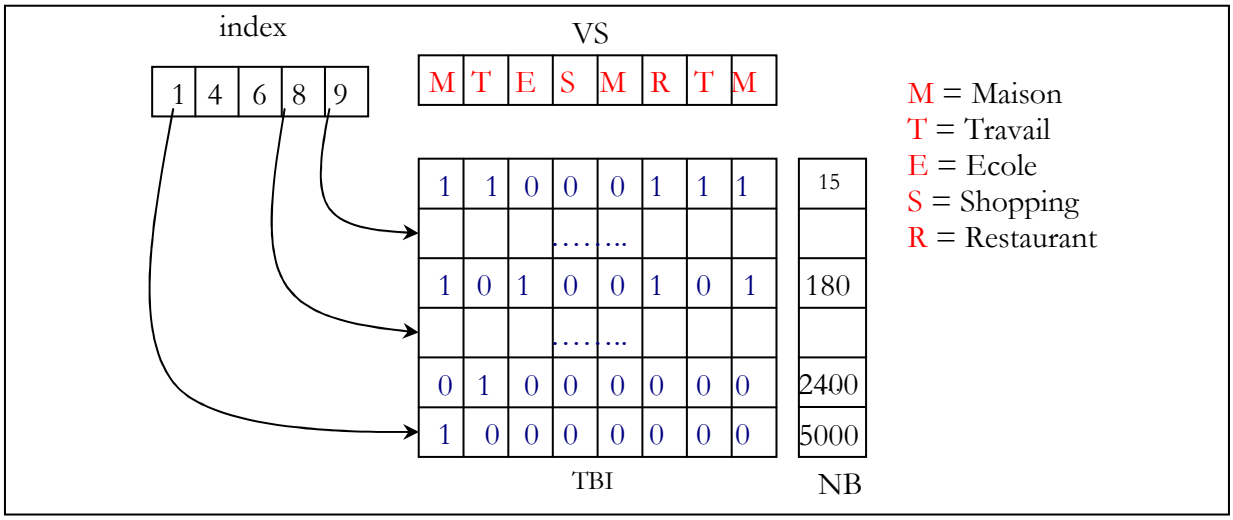


Figure 33: *Structure de données*

Nous avons implémenté plusieurs variantes d'algorithmes d'extraction de séquences basées sur cette structure de données. Le premier, appelé IBM, stocke réellement une matrice binaire offrant une très bonne compression, mais dont l'accès et la construction sont coûteux. En effet, des comparaisons et des opérations de décalage binaires sont nécessaires. Nous avons alors proposé une variante, appelée IBM2, où la matrice est stockée comme un tableau de booléens codées sur 8 bits. Si cette variante multiplie par huit la taille de stockage de la matrice, elle offre un accès direct et ne requiert aucune opération binaire. Elle est donc plus rapide en exécution. Par ailleurs, sachant que dans un contexte décisionnel, l'utilisateur procède souvent par essaie/erreur et teste successivement plusieurs seuils de support, nous avons proposé de rendre persistante notre structure de données et de la réutiliser dans les variantes IBM_Opt ou IBM2_Opt, en faisant varier le seuil de support. En effet, contrairement à *PrefixSpan*, notre structure de données ne dépend pas d'une valeur de support.

Nous avons utilisé les implémentations de PrefixSpan (PrefixSpan-0.4.tar.gz[5]) et de SPAM (Spam.1.3.1.tar.gz[6].) disponibles sur le Web et avons implémenté nos méthodes en Java.

[5] http://chasen.org/~taku/software/prefixspan/
[6] http://himalaya-tools.sourceforge.net/Spam/#download

Les résultats expérimentaux de ces quatre algorithmes ont été comparés, d'une part entre eux et d'autre part, avec les algorithmes de PrefixSpan et SPAM. Ces comparaisons montrent un gain de performances considérable, spécialement pour IBM2 et IBM2_Opt, par rapport aux autres méthodes, et ce lorsque l'alphabet des séquences est limité (entre 10 et 35 articles distincts). Ces cas sont assez courants dans plusieurs applications tels que le "Web Usage Mining" pour l'analyse des parcours de pages Web [223], l'analyse d'ADN ou l'analyse de l'activité humaine - dite "Time Use" - qui était notre exemple d'application.

Les tests confirment que les variantes IBM2 et IBM2_Opt sont plus performantes que IBM et IBM_Opt. Ils confirment également que la variante XX_Opt est une optimisation intéressante. L'écart constant entre les courbes de performances de IBM (respectivement IBM2) et IBM_Opt (respectivement IBM2_Opt) reflète le gain du temps de construction de la structure. Enfin, le temps d'exécution augmente généralement lorsque le support diminue. SPAM, IBM2 et IBM2_Opt sont toutefois moins sensibles à la variation du support que PrefixSpan, IBM et IBM_Opt.

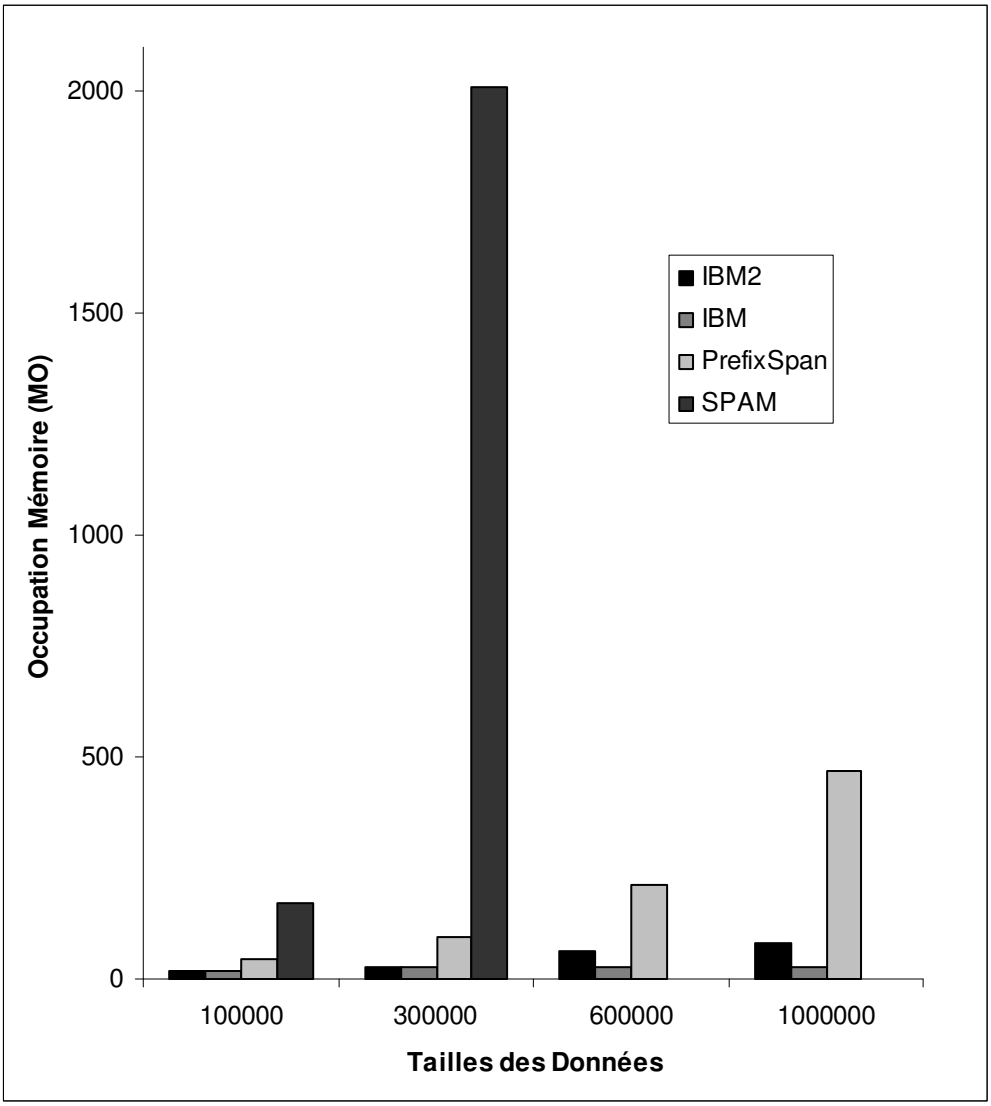


Figure 34: *Consommations mémoire*

Nous avons poussé ces tests avec une augmentation du nombre d'articles distincts (l'alphabet). Sur 100000 séquences, PrefixSpan présente de meilleurs résultats qu'IBM_Opt et IBM, mais reste moins performant qu'IBM2 et IBM2_Opt pour un alphabet allant 20 articles et que IBM2_Opt jusqu'à 35 articles. Mais, au-delà de 35 articles distincts, les performances d'IBM2 et IBM2_Opt se dégradent par rapport à PrefixSpan.

D'autres tests ont été effectués avec un jeu de données public du site www.kdnuggets.com avec le fichier *chess.dat*[7]. Les résultats ont confirmé la supériorité de nos méthodes, bien que le nombre d'articles distincts dans ces données soit supérieur à 35 (ici de 75). Ce test a aussi montré l'efficacité de la fusion via la structure VS. En effet, malgré les 75 articles distincts et une taille de séquences de 36 ou 37 articles, la taille de VS n'est que 91.

Quant à la consommation mémoire, IBM et IBM2 (de même que IBM_Opt et IBM2_Opt qui possèdent les mêmes structures) sont de loin moins coûteux que PrefixSpan et SPAM comme le montre la Figure 34. En utilisant une configuration machine de 1 GO de RAM, SPAM ne peut être exécuté au-delà de 300000 tuples, car il requiert plus de 2GO en mémoire principale. Les tests ont montré la validité et le passage à l'échelle des méthodes proposées grâce à la compacité de la structure de données.

Pour plus de détail, nous renvoyons le lecteur à la thèse de Savary [230] et à [232].

VI.2. Fouille de données textuelles

Avec l'extraordinaire développement des documents numériques et du web, l'organisation de grandes collections de documents en vue de faciliter leur exploitation est devenue une priorité dans le domaine de la gestion et de l'extraction des connaissances. En effet, c'est le seul moyen de faciliter l'accès aux documents numériques de plus en plus nombreux et de faire partager leur richesse sémantique par un large public. Ainsi, le web sémantique est aujourd'hui un thème majeur de recherche.

La fouille de textes fait partie des technologies facilitant l'exploitation de ces collections de documents. Elle se situe au carrefour des domaines de la recherche d'information (*Information Retrieval*) et des bases de données. Elle se caractérise par la combinaison de techniques linguistiques, statistiques et de fouille de données. Une des principales méthodes de fouille de textes est la catégorisation automatique. C'est précisément sur cette méthode qu'ont portée nos travaux.

VI.2.1. Problématique et aperçu de l'état de l'art

La catégorisation de documents correspond à la procédure d'affectation d'une ou de plusieurs catégories prédéfinies à un document de texte brut. Parmi ses applications, citons l'aide à l'indexation de documents [210], la personnalisation du filtrage de documents connaissant les préférences d'un internaute [220], le routage de mèl, l'amélioration de la recherche sur le web [203] et enfin l'organisation des sources textuelles sur le web.

---

[7] http://kdd.ics.uci.edu/

Le domaine de la catégorisation automatique de textes a été largement étudié, preuve en est la bibliographie consacrée [204]. Cette section donne un bref aperçu des différentes techniques et des étapes de la catégorisation de documents textuels.

VI.2.1.1. *Les étapes de la catégorisation*

Le processus de catégorisation de document comprend plusieurs phases : la préparation des données, la construction du modèle et le classement. Le tableau 5 résume les étapes de ce processus en indiquant les principales variantes pour leur réalisation La phase de préparation se caractérise par l'emploi de traitements linguistiques et statistiques spécifiques. Elle génère souvent une représentation vectorielle du document. La classification des documents proprement dite est souvent réalisée à l'aide d'une méthode d'apprentissage supervisée classique. La Figure 35 schématise le processus de catégorisation comprenant la phase d'apprentissage - comprenant la préparation et la construction du modèle (le classifieur) - et la phase de classement.

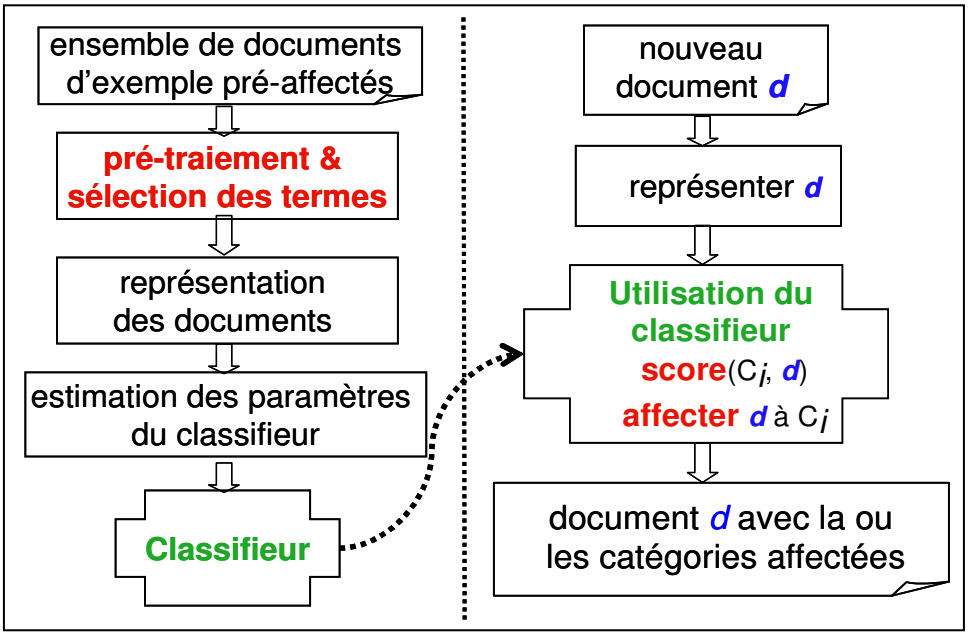


Figure 35. Processus de catégorisation

Tableau 4 : Résumé des étapes de la catégorisation

Phase	Etape	Observations et références
Préparation	Prétraitement du corpus	Liste des mots ; suppression des *stop words* ; canonisation des termes du même *concept*. Constitution d'un lexique. Utilise le traitement automatique de langues naturelles (TALN). Emploie un dictionnaire et parfois des thésaurus par domaine et un algorithme de canonisation des termes tels que *Ngram*, *lemme* et *stem* [227].
	Pondération des termes	Poids des termes dans les documents. Les variantes sont : le poids binaire reflétant la présence ou l'absence du terme dans le document ; la fréquence ; ou *tf-idf* [212] et ses variantes, largement utilisées car elles reflètent les termes caractéristiques du document par rapport au corpus.
	Filtrage des termes	Représenter le document par tous les termes conduit à une explosion de dimensions et n'améliore pas la qualité de l'apprentissage. Le filtrage utilise la réduction de dimension basée sur : un seuil de fréquences, le gain d'information ou le test de χ^2 [240]
	Représentation finale	Généralement, les documents sont représentés par un vecteur dans l'espace des termes filtrés [228]. La matrice de poids (*document X terme*) est parfois centrée réduite. La similarité de deux documents se traduit dans cet espace par l'angle des deux vecteurs et utilise le cosinus. Une autre représentation est le modèle probabiliste [224].
Modèle	Découpage de la BD	Consiste à choisir dans la base deux (voire trois) échantillons. Un pour l'apprentissage du modèle, un pour le test du modèle et éventuellement un troisième pour le calibrage du modèle. Ce découpage peut avoir un impact sur la qualité de l'indexation générée [215].
	Algorithme	Les principales sont : par arbre de décision [174], k-NN [241], centroïde [207], Naïve Bayes [222], SVM [213]. L'évaluation de ces méthodes montre que k-NN et SVM sont les plus performantes [241].
Classement	Calcul des scores	L'affectation de catégories aux documents passe d'abord par une mesure de leur score. Le score SSA (simple somme) est la somme des similarités du document à classer par catégorie, le MVA (Majorité Votant) pondère cette similarité par le nombre de documents par catégorie.
	Affectation	Le seuillage permet, après le calcul des scores des catégories, d'affecter les catégories aux documents. Les trois stratégies les plus courantes sont : RCut, PCut et SCut [242], avec quelques nuances.
	Evaluation des performances	Deux mesures sont couramment employées en recherche d'information, le *rappel (recall)* et la *précision (precision)*. Une troisième *F-mesure* combine les deux précédentes. Pour le cas multi-catégories, on utilise la *macro-moyenne* et la *micro-moyenne* des mesures précédentes : la *macro-moyenne* donne un poids égal à toutes les catégories ; tandis que la *micro-moyenne* est pondérée par leur fréquence et reflète les performances sur les catégories les plus fréquentes [241]. La *micro-moyenne* de F_1 est souvent utilisée dans les études comparatives de la catégorisation.

VI.2.1.2. *Les algorithmes de catégorisation*

Les algorithmes de catégorisation les plus couramment utilisés sont k-NN, centroïde, Naïve Bayes et SVM.

L'algorithme k-NN

k-NN (k-Nearest Neighbor ou k plus proches voisins) est un algorithme classique d'apprentissage automatique et un algorithme de base pour la catégorisation. En effet, il s'avère très efficace dans ce domaine tout en étant très simple à implémenter. Son principe est le suivant : pour un nouveau document à classer, il calcule sa similarité avec chaque document de la base d'apprentissage, puis détermine ses k plus proches voisins. Les catégories de ces documents permettent de déterminer la ou les catégories du nouveau document, et ce par le calcul de score et l'affectation décrits dans le tableau 5.

Plusieurs variantes de k-NN ont été étudiées [236][240][241][207]. Des produits de fouille de texte comme celui d'IBM l'ont intégré [235]. L'inconvénient principal de k-NN est son coût élevé en temps d'exécution, car le document à classer est comparé (pour le calcul de similarité) à l'ensemble des documents de la base.

L'algorithme centroïde

L'algorithme du centroïde est une variante de l'algorithme Rocchio [229] bien connu en théorie de l'information. Son idée est de calculer un centroïde pour représenter les catégories [212][207]. A la différence que k-NN, le calcul de similarité du nouveau document ne se fait qu'avec les vecteurs centroïdes des catégories. Par conséquent, il est plus rapide que k-NN. La catégorie dont le centroïde est le plus similaire est affectée directement ou bien après un calcul de score et un seuillage. Cet algorithme donne de bons résultats à condition que le vecteur centroïde soit représentatif de la catégorie.

L'algorithme Naïve Bayes

L'algorithme Naïve Bayes (NB) est également utilisé dans la catégorisation de documents. Il estime la probabilité conditionnelle d'une catégorie étant donné un document et affecte au document la (ou les) catégorie(s) la (les) plus probable(s). L'aspect naïf de ce modèle provient de l'hypothèse simplificatrice d'indépendance des mots. Ce type de catégorisation est parfois comparable aux autres méthodes [224]. Néanmoins, ses performances en qualité de classement se dégradent avec l'augmentation de la taille du lexique [221]. En effet, plus le nombre de termes augmente, plus les dépendances entre ces termes augmentent et moins l'hypothèse de NB est vérifiée.

Algorithme SVM

SVM (Support Vector Machine) est une méthode d'apprentissage relativement récente introduite par Vapnik pour résoudre un problème de reconnaissance de formes à deux classes [237]. Elle est basée sur la recherche d'un hyperplan dans un espace vectoriel qui sépare au mieux les points en deux classes. La qualité de l'hyperplan est déterminée par son écart avec les hyperplans parallèles les plus proches des points de chaque classe. Plus cet écart est important, meilleur est l'hyperplan. Il utilise des techniques de programmation quadratiques. SVM a été étendu pour les données ne pouvant être séparées de manière

linéaire, soit en relaxant la contrainte d'hyperplans, soit en transformant l'espace initial des vecteurs de données à un espace de dimension supérieure dans lequel les points deviennent séparables linéairement [233]. Une implémentation efficace de SVM est décrite dans [213]. Aujourd'hui, l'algorithme SVM semble le plus performant pour la catégorisation, seulement il est plus coûteux dans la phase d'apprentissage que les méthodes classiques.

VI.2.2. Contributions

Nos principales contributions ont été le développement d'une approche complète de catégorisation qui a été implémentée dans le système DocCat [215][205]. L'idée de base est d'améliorer le modèle en combinant deux espaces de représentation, généralement utilisés séparément dans les méthodes classiques : l'espace (*document X terme*) et l'espace (*catégorie X terme.*). Le filtrage des termes se base sur leur poids local vis-à-vis des catégories et sur le poids des catégories dans le corpus [217]. Cette approche de sélection de termes est dite *basée concept* (*Concept Based Approach* ou *CBA*). Par ailleurs, le modèle de similarité est également modifié [215] [216]. Il intègre une mesure d'association de termes estimée d'après leur poids dans les catégories et l'association des catégories dans le corpus. Enfin, le calcul des scores des catégories, contrairement à *MVA* et *SSA* (cf. Tableau ci-dessus), nuance les similarités des documents de chaque catégorie par la représentativité du document pour la catégorie (le centroïde est considéré comme le plus représentatif) [215] [216].

Concernant la méthode de classification, un nouvel algorithme hybride CKNN a été proposé comme la combinaison des algorithmes centroïdes et k-NN [215]. L'idée est de pré-sélectionner par les centroïdes les catégories candidates afin de limiter la recherche des k proches voisins aux documents appartenant à ces catégories. Il s'approche ainsi des bonnes performances de classement de k-NN tout en réduisant considérablement les temps d'exécution.

DocCat a été développé dans le cadre du projet RNTL CONTEXTE Bourse pour la fouille de dépêches financières et le balisage sémantique automatique. Toutes les approches proposées ont été testées à la fois sur le corpus fourni par les partenaires du projet CONTEXTE BOURSE [215] [219] et sur la collection standard Reuter-21578 de dépêches de l'agence Reuters.

Ces tests ont montré l'intérêt des modèles proposés (sélection, similarité et de calcul de score) et l'efficacité de l'algorithme hybride CKNN à réduire les temps de réponses par rapport k-NN pour la même qualité de classement. En effet, le nombre de similarités calculées a été divisé par 90 sur le corpus Reuters.

VI.3. Conclusion

Comme nous venons de le voir, chaque domaine d'application de la fouille de données génère des problèmes spécifiques en raison de ses caractéristiques structurelles ou sémantiques. Les travaux de recherche sur la fouille de ces données complexes proposent des solutions intégrant au mieux ces

caractéristiques. Nous avons contribué par la proposition de nouveaux algorithmes (IBM pour les séquences et CKNN pour le texte) ou de nouvelles approches générales (DocCat) qui améliorent les coûts algorithmiques et/ou les performances en qualité des modèles.

VI.4. Références

[201] Agrawal R., Srikant R.: Fast Algorithms for Mining Association Rules. In Proc. of the 20th Int. Conf. Very Large Data Bases (VLDB), Santiago, Chile, Septembre 1994.

[202] Agrawal R., Srikant R.: Mining sequential patterns. In Proc. of the 11th Int'l Conference on Data Engineering, Taipei, Taiwan, March 1995.

[203] Armstrong R., Freitag D., Joachims T., Mitchell T., «*WebWatcher: a Learning apprentice for the World Wide Web*», *AAAI Spring Symposium on Information Gathering from Heterogeneous*, Distributed Environments, March 1995, p.6-12.

[204] Gabrilovich E., "A Bibliography on Automated Text Categorization", http://liinwww.ira.uka.de/bibliography/Ai/automated.text.categorization.html, accessed July 2006.

[205] Gardarin G., Kou H., Zeitouni K., *DocCat : un composant logiciel de catégorisation de documents et de marquage sémantique XML*, revue RSTI-série ISI (Ingénierie des Systèmes d'Information), Editions Hermès, volume 8, n° 3/2003, « Systèmes d'information et langage naturel », pp. 33-54.

[206] Gardarin G., Pucheral P., Wu F.: Bitmap Based Algorithms for Mining Association Rules, BDA 98, pp157-175, Hammamet, Tunisie, Octobre 1998.

[207] Han E., Karpis G.,«*Centroid-Based Document Classification: Analysis & Experimental Results*», *In European Conference on Principles of Data Mining and Knowledge Discovery(PKDD),* 2000, p.424-431.

[208] Han E., Karypis G., Vipin K., «*Text Categorization Using Weight Adjusted k-nearest Neighbor Classification*», In Technical Report # 99-019, Department of CS, University of Minnesota.1999

[209] Han, J., Jamil, H. M., Lu, Y., Chen, L., Liao, Y. and Pei, J. DNA Miner: A system prototype for mining DNA sequences. In the proc. of the ACM SIGMOD International Conference on the management of data, Day 21-24, 2001, Santa Barbara, CA, USA.

[210] Hayes P., Weinstein S., «*CONSTRUE/TIS: a system for content-based indexing of a database of news stories*». *In second Annual conference on Innovative Applications of Artificial Intelligence*,1990.

[211] Jay A., Johannes .G, Tomi .Y, Jason F. Sequential Pattern Mining using A Bitmap Representation. SIGMOD, pp429-435, July 2002, Edmonton, Alberta, Canada. Isbn 1-58113-567-X.

[212] Joachims T., «*A Probabilistic Analysis of the Rocchio Algorithm with TFIDF for Text Categorization*», *In the 14th Int. Conf. Machine Learning*, 1997, p.143-151.

[213] Joachims T., «*Text categorization with support vector machines: Learning with many relevant features*». *In proceedings of the European conference on Machine learning*, 1998, p.137-142.

[214] Karypis G., Han E., «*Concept Indexing: A fast Dimensionality Reduction Algorithm with Applications to Document Retrieval & Categorization*», Technical Report TR-00-0016, University of Minnesota, 2000

[215] Kou H., "Génération d'adaptateurs web intelligents à l'aide de techniques de fouille de texte", Thèse de Doctorat de l'Université de Versailles-Saint-Quentin, Juillet 2003.

[216] Kou H, Gardarin G., «*Study of Category Score Algorithm for k-NN Classifier*», In the Proceedings of 25th ACM SIGIR , Tempere, Finland, 2002, p.393-394.

[217] Kou H, Gardarin G. and Zeitouni K., «*Approaches to Feature Selection for Document Categorizations*», In the proceedings of the 8th international conference NLDB, Germany, June, 2003

[218] Kou H., Gardarin G. «*Similarity Model and Term Association for Document Categorization*». In : the 7th International Workshop on Applications of Natural Language to Information Systems, NLDB 2002, June 27-28 2002, Stockholm, Sweden.

[219] Kou H., Gardarin G., d'Heygère A., Zeitouni K. «*Application of k-NN Classifier to Categorizing French Financial News*». In Proc of IASTED International Conference Artificial Intelligence and Applications (AIA 2002), September 9- 12, 2002, Malaga, Spain.

[220] Lang K., «*NewsWeeder: Learning to Filter Netnews*», *In Proc. of the 12th Int. Conf. on Machine Learning*, 1995, p.331-339.

[221] Lewis(92), D. D., «*An Evaluation of Phrasal and Clustered Representations on a Text Categorization Task*», *ACM 15th Ann Int'l SIGIR'92*, 1992, p.37-50.

[222] Lewis D. D., Ringuette, M., «*Comparison of two learning algorithms for text categorization*», *In Proceedings of the 3rd Annual Symposium on Document Analysis and Information Retrieval* (SDAIR'94),1994, p.81-93.

[223] Masseglia F., Poncelet P., Teisseire M. (2003) Incremental mining of sequential patterns in large databases. Data Knowledge Engineering 46(1): 97-121.

[224] McCallum A., Nigam K., «*A Comparison of Event Models for Naïve Bayes Text Classification*», *In AAA I-98 workshop on Learning for Text Categorization*, 1998.

[225] Mitchell, T., *Machine Learning*. Boston, MA, McGraw-Hill,1998.

[226] Pei J., Han J., Mortazavi-Asl B., Pinto H., Chen Q., Dayal U., and Hsu M-C. (2001). PrefixSpan: mining sequential patterns efficiently by prefix-projected pattern growth. Proceedings of the 17th International Conference on Data Engineering, 215-224.

[227] Porter, «*An algorithm for suffix stripping*», Program, Vol. 14, no. 3, 1980,p.130-137.

[228] Salton G., *Introduction to Modern Information retrieval*, 1983, McGraw-Hill Book Company

[229] Salton G., *Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer*. Addison-Wesley, Reading, Pennsylvania, 1989.

[230] Savary L., "Contributions à l'extraction des connaissances dans les bases de données spatio-temporelles", Thèse de Doctorat de l'Université de Versailles-Saint-Quentin, Décembre 2005.

[231] Savary L., Zeitouni K., "Mining Multidimensional Sequential Rules - A Characterization Approach", Proceedings of The First International Workshop on Mining Complex Data MCD'05, In conjunction with ICDM'05, IEEE Computer Society Press, November 2005, Houston, Texas, USA.

[232] Savary L., Zeitouni K., "Indexed Bit Map (IBM) for Mining Frequent Sequences", 9th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD 2005), Porto, Portugal, October 3-7, 2005. Lecture Notes in Computer Science n° 3721 / 2005, Springer-Verlag Ed, pp. 659 – 666.

[233] Shankar S., Karypis G., «*Weight adjustment schemes for a centroid based classifier*»,Computer Science Technical Report TR00-035, Department of Computer. Science, University of Minnesota, Minneapolis, Minnesota, 2000.

[234] Srikant R. and Agrawal R..(1996). Mining sequential patterns: generalization and performance improvements. Proceedings of the 15th International Conference on Extending Database Technology, 3-17.

[235] Tkach D., «*Text Mining: Turning Information Into Knowledge*». a white paper from IBM http://www-1.ibm.com/support/support_doc/swg/s0028/file/44354D1BBB8EEB438525 69D70047 CD89-whiteweb.pdf, February 17, 1998.

[236] Tuba Y., Guvenir H. A., «*Application of k-Nearest Neighbor on Feature Projections Classifier to Text Categorization*», *Proceedings of ISCIS-98, 13th International Symposium on Computer and Information Sciences*, 1998, p.135-142.

[237] Vapnik V., *The Nature of Statistical Learning Theory*. Springer New York, 1995.

[238] Wan T., Zeitouni K., « Extraction de motifs fréquents multi-supports Application aux données symboliques », 10èmes rencontres de la Société Francophone de Classification (10ème SFC), Edition PAN, Neuchâtel Suisse, Septembre, 2003, pp. 189-192.

[239] Wan T. and Zeitouni K., *Mining Association rules with Multiple Min-supports - Application to Symbolic Data*, Student Journal (ISSN 1420-1011), 5(1), 2004, pp. 59-74.

[240] Yang Y., Pederson J.O., «*A Comparative Study on Feature Selection in Text Categorization*», *In the 14th Int. Conf. On Machine Learning*,1997,412-420.

[241] Yang Y., «*An evaluation of statistical approaches to text categorization*». *Information Retrieval,*1(1), 1999,p.69-90.

[242] Yang Y., «A study on thresholding strategies for text categorization», Proceedings of ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'01), 2001.

[243] Zaki. M. J. (2001). SPADE: an efficient algorithm for mining frequent sequences. Machine Learning Journal, 42 (1/2): 31-60.

CHAPITRE VII. CONCLUSION ET PERSPECTIVES

Nous avons présenté dans ce mémoire les problèmes et les avancées dans l'intégration et la fouille des données spatiotemporelles et d'autres types de données complexes. Nous avons mis en exergue les problématiques spécifiques liées à nos travaux et nos contributions. Ces contributions ont concerné différents aspects allant de la modélisation à la fouille en passant par l'intégration dans un entrepôt de données, l'optimisation des requêtes et l'analyse en ligne.

VII.1. Bilan

Concernant la modélisation, nous avons proposé des modèles spatiaux 3D adapté au raisonnement géographique. Nous avons proposé récemment deux modèles spatiotemporels pour des objets mobiles contraints.

Concernant l'intégration, nous avons traité deux aspects : l'intégration de formats où dans un premier temps, un modèle ouvert a été défini, puis finalement le format GML plus flexible a été adopté comme format de stockage dans une architecture d'entrepôt de données ; puis l'intégration de données où une nouvelle méthode d'appariement géométrique a été développée. Le fait de baser l'entrepôt sur un SGBD XML natif nous a amenés à étudier l'optimisation de requêtes spatiales dans ce contexte. Nous avons alors proposé une technique d'optimisation basée sur un cache sémantique.

Pour ce qui est de l'analyse en ligne, la difficulté est d'explorer les données selon ses dimensions spatiales et/ou temporelles et/ou classiques. Le calcul d'agrégats selon des fenêtres spatiale et/ou temporelle pose des problèmes de performances et de précision. Nous avons successivement proposé et implémenté deux approches : la première pragmatique se ramène à une modélisation OLAP standard et des agrégats classiques au prix de résultats approximatifs ; quant à la seconde, elle étend les concepts de modélisation multidimensionnelle à la variation continue et donne des résultats précis. Elle met en œuvre un algorithme d'agrégat spatiotemporel basé sur un modèle de stockage compact et une méthode d'indexation appropriée.

Le problème de fouille de données spatiale a été largement étudié dans nos travaux. Nous avons relaté les problèmes et les avancées dans ce domaine et exposé nos contributions. Nous avons classé les méthodes en différentes approches et nous avons placé nos propositions dans la fouille de données multi-thématiques. Nous avons proposé trois options pour la fouille de données spatiales que nous avons expérimentées et testées, principalement pour deux méthodes : la classification par arbre de décision spatial et la recherche de règles d'associations spatiales.

Nous avons ouvert le spectre de nos recherches à la fouille d'autres données complexes, telles que les données séquentielles et textuelles. D'un côté, nous avons proposé un algorithme efficace de recherche de motifs séquentiels et étendu notre approche aux séquences multidimensionnelles. De l'autre, nous avons développé et testé une approche complète pour la catégorisation de données textuelles.

VII.2. Perspectives

Nos recherches ont concerné des domaines variés tout en étant connexes. Elles ouvrent de nombreuses perspectives, que ce soit dans le support des modèles 3D ou spatiotemporels ou dans l'intégration et la fouille de données complexes. En effet, comme nous l'avons souligné en conclusion de chaque chapitre, des recherches continuent dans ces domaines avec de plus en plus de travaux pour certains domaines comme la fouille de données complexes. Nous dégageons ci-dessous quelques axes de recherches qui nous intéressent pour les sujets traités dans ce mémoire. Nous donnons ces axes dans l'ordre de priorité que nous pensons suivre.

Bases et entrepôts d'objets mobiles

A court terme, nous nous focalisons sur le support des objets mobiles dans les bases de données, dans les entrepôts de données et en fouille de données. Ce thème a connu un regain d'intérêt, encouragé par l'émergence des systèmes de localisation mobiles et communicants. Il concentre la majorité de nos compétences (modélisation, indexation, intégration, *map matching* et prédiction) et s'inscrit dans la continuité de nos collaborations scientifiques et industrielles. En outre, il ouvre le champ à des problèmes intéressants comme la gestion des flux de données des capteurs mobiles, la prise en compte de systèmes complexes pour modéliser le trafic ou la confidentialité du traçage de la localisation. Une nouvelle thèse commence tout juste sur ce sujet. Le contexte applicatif est l'analyse du comportement des conducteurs en utilisant les traces des capteurs du véhicule et du GPS. Il repose sur des données réelles issues d'une campagne de test d'un équipement embarqué de sécurité routière. Ces données et ce type d'applications génèrent des problématiques nouvelles. Ainsi, nous disposons non seulement de trajectoires mobiles, mais également d'un ensemble de relevés (ou de séries) temporels et d'états. Ces relevés portent sur la période de déplacement mais ne sont ni synchrones ni à la même échelle. La représentation, le stockage (compressé) et l'interrogation de ces données sont des challenges que nous nous sommes fixés de relever. Le groupage et l'agrégation dans un système de type OLAP devront être revus avec ces nouvelles dimensions. L'idée générale de dimension et de faits continus peut probablement s'appliquer et la

représentation géométrique des différents phénomènes temporels permet de remédier au problème de synchronisation des relevés.

Fouille de données textuelles

Nos recherches sur la fouille de données textuelles sont toujours en cours dans le cadre du projet RNTS Rhéa où la base de données comporte des comptes-rendus hospitaliers (CRH). Le but est d'aider les praticiens hospitaliers à renseigner le code des pathologies par classification supervisée. Les méthodes de catégorisation classiques s'avèrent peu efficace pour ces documents. Ceci est probablement dû au fait que cette base se caractérise par un nombre très élevé (un millier) de catégories à apprendre, que l'affectation à de multiples catégories est fréquente et que certaines pathologies ne sont pas indépendantes. Nous expérimentons actuellement de nouvelles méthodes afin d'améliorer les performances de classement. Une idée est de repérer les concepts partagés par différentes catégories et l'autre est d'exploiter l'arborescente des pathologies qui existe et qui est normalisée afin de réduire l'apprentissage à des sous-ensembles de CRH.

Fouille de données spatiotemporelles

Les remarques données à la fin du chapitre V mènent à des perspectives de recherche sur la fouille de données spatiales et spatiotemporelle. La première est la combinaison de méthodes de fouille de données spatiales afin d'analyser des concentrations de phénomènes dans l'espace. La seconde est l'extension aux données spatiotemporelles. Concernant ce dernier point, nous distinguons deux types de données : d'un côté, les données à variation discrète comme l'état du trafic et de l'autre, les données à variation continue comme les objets mobiles. Parmi nos travaux en cours, nous explorons le clustering d'objets mobiles en collaboration avec l'Université Populaire de Chine.

Fouille de données complexes

Après avoir traité successivement la fouille de différent type de données complexes, nous constatons que les méthodes proposées traitent généralement d'un type de données complexes. Or, les bases de données comprennent aujourd'hui de plus en plus ce type de données combinés avec des données simples ou d'autres données complexes. Par exemple, dans le cas des CRH, le document est associé à des informations classiques sur le patient comme l'âge et le sexe et d'autres informations comme la date (temporel) et le service. On peut imaginer l'ajout d'images comme des radiographies ou d'attributs spatiaux comme le lieu d'habitation. Il est aujourd'hui rare que de telles informations hétérogènes soient combinées dans un processus de fouille de données. La combinaison de séquences et d'attributs peut être vue comme un pas dans ce sens.

Support des données 3D

Concernant le support des données 3D, un aspect intéressant de nos travaux précédents est d'avoir traité la modélisation de milieux intérieur et extérieur avec des approches similaires tout en capturant la sémantique des liens topologiques et directionnels. Il serait intéressant pour des applications comme le guidage ou la navigation 3D d'intégrer la modélisation de milieux intérieur et extérieur. Par ailleurs,

beaucoup reste à faire dans le modèle de traitement et le langage, car l'extension au spatial qui a été normalisée dans SQL3 est limitée au 2D

Perspectives à long terme

A plus long terme, tout tend à croire que la gestion des données ne se suffira plus des données simples, de la connaissance explicite ou des sources connues. On le voit bien aujourd'hui, le tout numérique transforme profondément les systèmes d'information qui s'orientent vers des données complexes (textuelles, sonores, spatiales, …). La recherche par le contenu de ces formes complexes est inévitable. La fouille de donnée est l'approche la plus prometteuse. La recherche est loin d'apporter les solutions adéquates pour l'intégration ou la fouille de ces données complexes. En effet, les méthodes existantes sont généralement limitées aux données tabulaires simples. Seulement, pour chaque type de données complexes, l'analyse doit prendre en compte l'interprétation de son contenu ou de sa structure.

Par ailleurs, l'information diffuse et hautement dynamique révolutionne les architectures distribuées traditionnelles de bases de données. Ces évolutions impacteront les systèmes décisionnels qui deviendront de plus en plus réactifs. En effet, généralement basés sur des données historiques plutôt statiques, les « entrepôts de données » devront d'adapter aux sources dynamiques, parfois inconnues à l'avance. En somme, comment gérer des entrepôts de données complexes et actifs ?

ANNEXE - BASES DE DONNEES SPATIALES

Dans le domaine des Systèmes d'Informations Géographiques (SIG), les progrès de ces dernières années ont permis de dégager un certain nombre de notions communément admises. Cette annexe donne un aperçu de ce domaine.

La première section introduit ces notions de base. Ensuite, nous résumons l'évolution de la gestion de ces données. La troisième section dresse un panorama des problèmes de recherche que soulève la gestion de l'information géographique en bases de données.

VII.3. Notions de base

L'information géographique est complexe. Sa modélisation s'appuie sur nombre de concepts [251], tels l'objet géographique, la couche thématique, la relation spatiale, la méta-information spatiale et le modèle spatial. Cette section se propose de développer ces concepts, mais tout d'abord, elle précise les notions de SIG et de base de données spatiale.

Système d'Information géographique

Un SIG est un Système d'Information au sens large capable de répondre à toute la chaîne de traitement des données géographiques. Il se caractérise par cinq fonctionnalités connues sous le nom des 5A, à savoir : Acquisition, Assemblage, Archivage, Analyse et Affichage des données géographiques [257].

Base de données géographique ou spatiale

Une base de données géographique [263] est définie par un ensemble d'objets géographiques organisé de manière à pouvoir être manipulé dans un SIG.

L'information géographique capturée dans les systèmes actuels est souvent une projection au sol d'entités du monde réel. On parle de coordonnées planimétriques.

Objet géographique ou spatial

Un objet géographique est la traduction d'une entité du monde réel dans une base de données géographique par une structure de donnée. Cette structure contient à la fois des données sémantiques et une donnée de localisation spatiale (comme l'exemple de la Figure 36). Les données sémantiques décrivent qualitativement ou quantitativement l'objet. La donnée de localisation décrit la morphologie de l'objet et sa position dans l'espace.

Couche thématique

Une couche thématique est un regroupement d'objets géographiques partageant les mêmes propriétés, les mêmes structures en une collection homogène. Par exemple, une couche thématique peut représenter l'hydrologie, le découpage administratif en communes, le bâti ou les routes. Les requêtes et la visualisation peuvent faire intervenir sélectivement les couches utiles, suivant la métaphore des papiers calques pour retourner une carte. Une base de données géographique est constituée d'un ensemble de couches thématiques.

Relation spatiale

Dans une base de données relationnelle, les liens structurels entre données sont représentés par des clés de jointures. Autrement, les objets sont supposés indépendants. Ce n'est pas le cas des bases de données géographiques où les relations spatiales sont souvent implicites. Ces relations sont nombreuses (adjacence, intersection, à distance de) et correspondent aux interactions avec le voisinage, soit au sein d'une même couche thématique, soit entre thèmes (intra ou inter thèmes). Elles sont couramment exploitées dans les requêtes et les analyses spatiales. Leur résolution lors des requêtes entraîne des calculs géométriques coûteux.

Méta-information spatiale

En plus des méta-données gérées habituellement dans le dictionnaire des bases de données, les bases de données géographiques comprennent des méta-informations spécifiques. Ce sont typiquement les informations d'échelle, d'emprise, de système de projection, de qualité, de datation, etc. Elles sont importantes aussi bien pour le système que pour l'utilisateur, pour leur description détaillée des sources. Cela a amené à leur normalisation assez tôt [253].

Modèle spatial

On distingue deux perceptions des données spatiales [261] [257]. La première est par étendu de mesures sur un maillage de l'espace (raster ou triangulation pour le modèle numérique de terrain) comme dans Figure 37 et la seconde par objet dont la localisation est représentée par le contour (dits vectoriels). Les modèles vectoriels sont les plus répandus et nos travaux se basent sur la perception par objet.

Modèle vectoriel

Il existe principalement trois catégories de formats vectoriels : spaghetti, réseau et topologique. Le format spaghetti (Figure 38) décrit les contours sans relations de contiguïtés et sans relations topologiques entre les objets. Le modèle réseau décrit la connexité des lignes. Ce modèle est principalement utilisé pour

effectuer des traitements de types parcours de graphes. Un modèle plus riche est le modèle topologique qui modélise les contiguïtés entre toutes les formes d'objets, que ce soit des lignes, des nœuds ou des surfaces. Il permet de maintenir la cohérence lorsque des contraintes portent sur la topologie. De plus, il permet une implémentation efficace des opérateurs utilisant la topologie. Les standards d'échange ont été parmi les premiers à implémenter ces trois types de formats comme la norme EDIGéo [244] inspirée du format de l'OTAN DIGEST [248] et le format DLG-E de l'USGS [250].

Figure 36. *Exemple d'objet spatial*

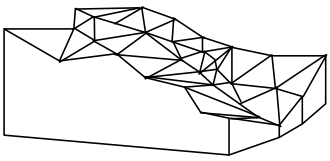


Figure 37. *.Modèle numérique de terrain*

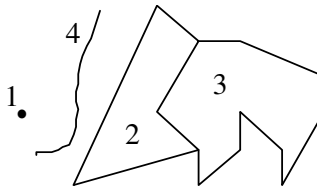


Figure 38. *Modèle spaghetti*

VII.4. L'évolution vers les SGBD spatiaux

Les SIG ont longtemps été basés sur des systèmes de fichiers ou sur un couplage faible séparant des fichiers propriétaires pour la composante géométrique et un SGBDR pour la composante spatiale. Cette situation tend à disparaître an faveur d'une séparation entre la fonction de serveur de données spatiales et sémantiques en même temps et d'une interface de visualisation. Cette section établit un parallèle entre les SIG et les SGBD Relationnels (SGBDR). Nous comparons dans le Tableau 5 les concepts bien connus dans le domaine des SGBDR (à gauche) avec les concepts spécifiques aux SGBD spatiaux (à droite du tableau). Malgré leurs spécificités, on constate qu'un SGBD spatial peut être vu comme une extension d'un SGBD relationnel.

Tableau 5 : *Parallèle entre les SGBD spatiaux et les SGBD relationnels*

	RELATIONNEL	**SPATIAL**
Données	Entier, Réel, Texte,…	Plus complexes : Point, Ligne, Région…, cellulaire
Prédicats et calculs	Tests : =, >,… Calculs : +, /,… et fonctions simples	Prédicats & calculs géométriques et topologiques: Tests : intersection, adjacence,… Fonctions : intersection, surface…
Manipulation	Opérateurs de l'algèbre : Sélection, Projection, Jointure Agrégats: Count, Sum, Avg…	Manipulation mono ou inter-thèmes Sélection et jointure sur critère spatial Agrégats : fusion d'objets adjacents
Liens entre objets	Par clés de jointures	Relations spatiales (souvent) implicites
Méthodes d'accès	Index B-tree, hachage	Index R-tree, Quad-tree, Grid-file, etc.
Origine des données	Majoritairement propriétaire Format d'échange simple	Majoritairement externes Formats d'échange complexe et très varié

Cette similitude entre les SGBDR et les SGBD spatiaux a été la clé pour concevoir aujourd'hui des SIG puissants qui héritent des qualités des SGBD.

La recherche au début des années quatre-vingt-dix était active sur l'intégration de l'information géographique au cœur des SGBD. Cela a eu des impacts sur les produits commerciaux qui sont en voie de conquérir les marchés des SIG. Par exemple, le SGBD Oracle qui a commencé à introduire les données spatiales depuis sa version 7 au début des années quatre-vingt-dix, s'est imposé aujourd'hui, comme un serveur de données géographique complet et efficace [260]. DB2 avec le rachat d'Informix [252] se place également sur ce créneau. Enfin, le logiciel libre Postgres/PostGIS [262] est de plus en plus utilisé. Cette évolution vers une architecture intégrée a apporté plus de performance et de fiabilité dans la gestion des données géographiques.

VII.5. Problèmes soulevés par les SIG aux bases de données

Les SIG ont suscité de nombreux travaux en bases de données. Nous les classons par rapport aux cinq fonctions de SIG cités plus haut (les 5A). Nous donnons quelques références à ces recherches, sans aller plus en détail. Les sujets qui nous ont intéressés sont présentés plus en détail dans les chapitres de ce mémoire.

– *Acquisition des données géographiques :* Cela va de la simple saisie ou mise à jour de quelques données spatiales jusqu'à la génération d'un lot de données complet par des processus complexes, variés et semi-automatiques. Les principaux problèmes soulevés en bases de données, si l'on met de côté les problèmes de numérisation et de photogrammétrie, sont : (i) la complexité des mise à jour qui doivent respecter des contraintes d'intégrité spatiales [255] ; (ii) le renseignement du processus d'acquisition et de la qualité dans les méta-données ; (iv) et plus récemment, la gestion en temps réel de flux de données de localisation comme les relevés GPS [245].

– *Assemblage des données géographiques* : Ceci consiste à importer des données de sources différentes et à les fusionner afin de constituer la base de données géographique adaptée aux besoins du SIG de l'entreprise. Le recours à des données externes est plus fréquent dans un SIG que dans un SI classique. Cette tâche se confronte à de nombreux problèmes dont : (i) le catalogage consistant à un assemblage minimal via les méta-données, où il est nécessaire néanmoins de les fusionner certaines méta-données en développant des ontologies spatiales [246] et en exploitant des normes comme [253] ; (ii) l'échange et l'intégration des sources de formats hétérogènes, l'intégration dans un modèle commun d'entrepôt de données spatiales ; (iii) l'intégration de données selon la composante géométrique en tenant compte de l'imprécision et plus généralement des représentations multiples ; (iv) l'accès distribué et la médiation, qui a motivé les propositions de services Web géographiques [258].

– *Archivage*. Ceci correspond aux fonctionnalités allant de simples systèmes de gestion de fichiers aux Systèmes de Gestion de Bases de Données (SGBD). Du point de vue bases de données, stocker cette information

complexe nécessite : d'enfouir un modèle spatial dans le systèmes en se basant sur une représentation spécifique (raster ou vecteur, spaghetti ou topologiques, 3D et spatiotemporel), en adaptant le modèle de stockage et d'accès [249] et d'étendre la gestion de transactions et la concurrence d'accès [256]. Par ailleurs, il faut gérer des méta-données spécifiques sans quoi la donnée géographique n'a pas la même valeur. La modélisation et l'archivage des données spatiales 2D n'est plus un problème de recherche et son implémentation dans les SGBD tend à se généraliser. Néanmoins, les aspects avancés tels que la gestion de la troisième dimension, du temps ou des flux de données restent du domaine de la recherche, de même que le stockage et l'accès efficace à ces formes de données.

– *Analyse des données géographiques*. L'analyse simple passe par des outils statistiques ou par des requêtes spatiales. Ce qui caractérise l'analyse des données géographiques est qu'elle utilise fréquemment les relations spatiales entre objets. En bases de données, cette fonctionnalité a engendré l'extension du langage SQL vers les opérateurs spatiaux [254]. Dans sa forme avancée, l'analyse spatiale des données correspond à l'extraction de connaissances par l'analyse en ligne (OLAP) ou la fouille de données spatiales.

– *Affichage les données géographiques*. C'est la fonction la moins en rapport avec les bases de données qui traditionnellement se consacrent au contenu plutôt qu'au contenant. Citons néanmoins les hyper média géographiques [247] et récemment le service WMS [259].

VII.6. Références

[244] AFNOR, Norme EDIGéO (Echanges de Données Informatisées dans le domaine de l'information Géographique), normalisation française, French standard, no : Z 13-150 (ISSN 0335-3931), Aug. 1992.

[245] Bai Y., Guo Y., Meng X., Wan T., Zeitouni K., "Efficient Dynamic Traffic Navigation with Hierarchical Aggregation Tree", Proceedings of APWeb 2006 will be published by Springer in its Lecture Notes in Computer Science series, The Eighth Asia Pacific Web Conference, January, 2006, Harbin, China

[246] Barde J., Libourel T., Maurel P., Ontologies et métadonnées pour le partage d'information géographique, Revue Internationale de Géomatique « Les ontologies spatiales », Vol 14/2 – 2004, pp.199-216.

[247] Caporal Julien, *Intégration de la cartographie numérique dans un système hypermedia*, Doctorat de l'Université de Versailles-Saint-Quentin, janvier 1999.

[248] DMA (Defense Mapping Agency) (1992) Military Switchboard Vector Product Format (VPF), THOUSAND-STD-600006.

[249] Gaede V., Gunther O., "Multidimensional access methods", ACM Computing Surveys, 30(2) , 1998, 170-231.

[250] Guptill S.C. (1990) Multiple Representations of Geographic Entities through Space and Time, in 4th Int. Symp. on Spatial Data Handling, Switzerland, Vol. 2, 859-868.

[251] Güting, R.H., An Introduction to Spatial Database Systems, *VLDB Journal 3 (1994),* 357-399.

[252] IBM Informix Spatial DataBlade : http://www-306.ibm.com/software/data/informix/blades/spatial/, accédé en Juillet 2006.

[253] ISO/DIS 19115 (2000), Geographic information — Metadata ISO TC 211/WG 3, 2000-12-12

[254] ISO/IEC 13249-3:2003, Information technology – Database languages – SQL multimedia and application packages – Part 3: Spatial

[255] Kadri Dahmani H., "Mise à jour des bases de données géographiques et maintien de leur cohérence", Doctorat Informatique de l'Université Paris 13, Novembre 2005.

[256] Kornacker M., Banks D., "High-Concurrency Locking in R-Trees", 21th Int. Conf. on Very Large Data Bases, VLDB'95, September 1995, Zurich, Switzerland, pp. 134-145.

[257] Laurini, L., D. Thompson (1992) Fundamentals of spatial information systems, The A.P.I.C. Series (n° 37), Academic Press.

[258] OGC 04-094, Web Feature Service Implementation Specification, Version: 1.1.0, Open Geospatial Consortium Inc., May 2005.

[259] OGC 06-042n, OpenGIS Web Map Server Implementation Specification, Version: 1.3.0, Open Geospatial Consortium Inc., March 2006.

[260] Oracle Corporation, Oracle Spatial User's Guide and Reference, 10*g* Release 2 (10.2), B14255-03, March 2006.

[261] Peuquet D.J., "*A Conceptual Framework and Comparison of Spatial Data Models*", Cartographica, vol. 21, no. 4, pp. 66--113, 1984.

[262] PostGIS web site : http://postgis.refractions.net , accédé en juillet 2006.

[263] Rigaux P., Scholl M., Voisard A.. Spatial databases with application to GIS. San Francisco, CA, Morgan Kaufmann Publishers Inc., 2002.