L'INFORMATION GÉOGRAPHIQUE ET LES ENTREPÔTS DE DONNÉES

par Sandro Bimonte

INSA de Lyon, Laboratoire d'InfoRmatique en Image et Systèmes d'information
UMR CNRS 5205, INSA, 7 avenue Capelle
69621 Villeurbanne Cedex, France
E-mail : sandro.bimonte@univ-lyon2.fr

Cet article résume le travail de thèse intitulé « Intégration de l'information géographique dans les entrepôts de données et l'analyse en ligne : de la modélisation à la visualisation », effectué par l'auteur à l'INSA de Lyon au sein du Laboratoire d'InfoRmatique en Image et Systèmes d'information (LIRIS). Ce travail concerne l'introduction de la composante sémantique de l'information géographique et la flexibilité de l'analyse spatiale dans les systèmes Spatial OLAP.

Contexte général de l'étude

Les entrepôts de données associés à des outils d'analyse On Line Analytical Processing (OLAP), représentent une solution effective pour l'informatique décisionnelle (Immon 1992). Les données dans les hypercubes sont organisées en axes d'analyse nommés « dimensions ». Les sujets d'analyse, appelés « faits », sont caractérisés par des mesures, qui sont pré-calculées à l'aide de fonctions d'agrégation selon les différentes granularités définies par le schéma hiérarchique de chaque dimension. Dans le cas classique, une mesure est une valeur numérique qui décrit quantitativement le fait. Ainsi une analyse multidimensionnelle portant sur un fait « ventes » d'un ensemble de magasins pourra être réalisée en définissant comme mesures « le volume de la vente » et « le montant de la vente ». Le processus d'analyse est conduit par la navigation dans le cube multidimensionnel. Les opérateurs OLAP (roll-up, drill-down, slice, etc.) permettent de visualiser les mesures pour des ensembles de membres à des niveaux de granularité sélectionnés par l'utilisateur. Les opérations de forage (roll-up, drill-down) sont fondées sur des fonctions d'agrégation appliquées aux mesures, par exemple la somme appliquée au volume de produits vendus. Des interfaces orientées navigation (tableau de bord, tableau multidimensionnel, graphes) complètent le panel des outils décisionnels.

Un système d'informations géographique permet d'acquérir, de structurer, de mémoriser, d'analyser et de visualiser les données géographiques. Les utilisateurs potentiels d'un SIG sont tous les spécialistes qui ont besoin d'analyser d'importants volumes de données géographiques dans différents domaines. Les systèmes d'aide à la décision, en particulier les systèmes OLAP, n'offrent aucun instrument pour la gestion des données spatiales. Des solutions connues sous le terme d'OLAP Spatial, qui visent à intégrer la donnée spatiale dans l'OLAP, ont donc été développées. L'OLAP Spatial (SOLAP) a été défini par Yvan Bédard comme « une plateforme visuelle conçue spécialement pour supporter une analyse spatio-temporelle rapide et efficace à travers une approche multidimensionnelle qui comprend des niveaux d'agrégation cartographiques, graphiques et tabulaires » (Bédard 1997). Le SOLAP enrichit les capacités d'analyse des systèmes OLAP classiques car la visualisation des mesures sur une carte permet de comprendre la distribution géographique d'un phénomène et de mettre en relation les différents phénomènes spatiaux par rapport aux axes d'analyse alphanumériques, et de comparer ces phénomènes à diverses granularités géographiques. De plus, la composante cartographique dans l'OLAP représente une interface vers l'entrepôt de données spatiales. En d'autres termes, l'utilisateur peut accéder aux opérations de navigation multidimensionnelle à travers la simple interaction avec la composante cartographique.

L'intégration des données géographiques dans l'analyse en ligne est un enjeu majeur. La modélisation des entrepôts de données géographiques tout comme l'adaptation des fonctionnalités des systèmes d'entrepôt de données classiques pour les données géographiques, sont des problématiques ouvertes.

Problématiques

L'information géographique est la représentation d'objets ou de phénomènes réels, localisés dans l'espace. Cette information est caractérisée par sa localisation dans l'espace, par sa forme et par ses aspects sémantiques, c'est-à-dire par ses attributs descriptifs alphanumériques et ses relations avec d'autres objets (Denègre et Salgé 1997). Les modèles SOLAP existants se concentrent généralement sur la composante spatiale de l'information géographique. Ils définissent une mesure spatiale comme une collection d'objets spatiaux, et une dimension spatiale comme une dimension dont les membres contiennent un attribut spatial (Malinowsky et Zimányi 2004 ; Rivest et al. 2005). Ces modèles de données comportent une limite importante, liée à la prise en compte de la composante sémantique de l'information géographique dans les mesures et dans les dimensions. D'un point de vue mesure, nous pensons que les attributs descriptifs des objets géographiques peuvent être utiles au processus décisionnel pour expliquer un phénomène ou caractériser un ensemble de faits. Lorsque l'information spatiale est utilisée comme axe d'analyse, les modèles SOLAP existants utilisent les dimensions spatiales. Une dimension spatiale est décrite par des hiérarchies dont les membres sont des objets géographiques liés par les relations topologiques d'inclusion ou d'intersection (Malinowsky et Zimányi 2005). Cette définition ne reflète pas la sémantique sous-jacente aux liens hiérarchiques. En effet, les objets géographiques peuvent être en relation avec d'autres objets à travers des relations spatiales, des relations de généralisation (Weibel et Dutton 2001) et des relations non spatiales. La prise en compte de ces types de relations est, selon nous, fondamentale dans l'analyse multidimensionnelle car à chaque type de hiérarchie correspond une analyse différente, qui peut se traduire par différentes politiques d'agrégation et de navigation.

Un deuxième point faible des solutions SOLAP existantes concerne les opérateurs d'analyse spatio-multidimensionnelle. Nous pensons que ces opérateurs devraient pouvoir modifier les dimensions géographiques selon l'exigence de l'utilisateur. En effet, au contraire de l'approche utilisée dans le processus décisionnel OLAP, l'analyse spatiale est flexible et itérative (Longley et al. 2001). Les données géographiques, grâce aux méthodes de transformation d'analyse spatiale, peuvent être modifiées ou remplacées tout au long du processus d'analyse. Or les opérateurs SIG fournis par les différents systèmes SOLAP sont des opérateurs orthogonaux aux opérateurs multidimensionnels, autrement dit ils ne sont pas utilisés dans la navigation OLAP. L'introduction et l'adaptation des opérateurs d'analyse spatiale dans un contexte OLAP afin d'intégrer la flexibilité de l'analyse spatiale au modèle multidimensionnel sont fondamentales pour une analyse spatio-multidimensionnelle satisfaisante. De plus, les modèles SOLAP existants utilisent des modèles de données différents pour la représentation des membres et des mesures spatiales. En termes d'analyse multidimensionnelle, cela se traduit par une asymétrie entre la mesure et la dimension spatiale, ce qui oblige l'utilisateur à concevoir des hypercubes différents pour changer le point de vue de l'analyse. Ce manque de symétrie et de flexibilité apporte une limite importante aux solutions SOLAP existantes, car l'information géographique doit pouvoir être aussi bien utilisée en dimension qu'en mesure. Enfin, une mesure en étant un objet géographique peut appartenir à des schémas hiérarchiques. L'utilisation de cette information pour analyser les mesures à différentes granularités reste un défi important.

L'OLAP géographique

Les travaux de recherche portent sur l'intégration de l'information géographique dans l'analyse multidimensionnelle. Une analyse détaillée des concepts principaux du SOLAP, de l'information géographique et de l'analyse spatiale, nous a mené à la définition du concept d'OLAP géographique (Tchounikine et al. 2005 ; Bimonte et al. 2005 ; Bimonte et al. 2006c ; Bimonte et al. 2007b). L'OLAP géographique reformule les concepts du Spatial OLAP pour prendre en compte la composante spatiale et sémantique de l'information géographique et la flexibilité de l'analyse spatiale. L'OLAP géographique définit les concepts de mesure, de dimension et de hiérarchie géographiques, ainsi que des nouveaux opérateurs multidimensionnels.

La mesure géographique étend le concept de mesure spatiale aux attributs alphanumériques d'un objet géographique, nécessaires et complémentaires du processus d'analyse. Une dimension géographique est décrite par trois différents types de hiérarchies : descriptive, spatiale et de généralisation. Ces trois types de hiérarchies reflètent la sémantique des relations entre les membres de niveaux différentes. L'OLAP géographique reformule également les opérateurs spatio-multidimensionnels pour introduire dans un contexte multidimensionnel les opérateurs d'analyse spatiale, pour lever l'asymétrie entre mesure et dimension spatiale et pour permettre d'exploiter les relations spatiales et aspatiales entre les mesures géographiques.

Le présent travail de thèse propose un modèle formel (GeoCube) (Bimonte et al. 2005 ; Bimonte et al. 2006) et une algèbre associée qui prend en compte les concepts principaux de l'OLAP géographique. L'originalité de GeoCube est la modélisation de toutes les données de l'univers d'analyse à travers les concepts d'objet complexe et/ou géographique. Les mesures, comme les membres de dimension, sont des objets géographiques décrits par un ensemble d'attributs descriptifs et un attribut spatial. La symétrie entre mesure et dimension se traduit notamment par l'appartenance des mesures à des schémas hiérarchiques. GeoCube, en s'appuyant sur cette modélisation des données particulière, propose une algèbre spatio-multidimensionnelle qui reformule et étend les opérateurs spatio-multidimensionnels classiques. L'algèbre fournit des opérateurs de forage et de coupe. L'opérateur de forage permet l'agrégation des objets géographiques, contrairement aux approches classiques où l'agrégation porte sur des mesures qui sont de simples valeurs quantitatives. L'opérateur de coupe permet de couper l'hypercube en utilisant des prédicats alphanumériques et aussi des prédicats spatiaux. Pour exploiter la symétrie entre mesures et dimensions, GeoCube introduit deux nouveaux opérateurs multidimensionnels qui permettent de naviguer dans la hiérarchie de la mesure et un opérateur qui permet d'intervertir mesure et dimension Enfin, une autre innovation apportée par GeoCube est un opérateur qui permet de changer dynamiquement la structure de l'hypercube, en ajoutant dans les données décisionnelles les résultats obtenus par des opérateurs d'analyse spatiale.

Cette approche formelle a été implémentée dans le prototype GeWOlap (Bimonte et al. 2006b ; Bimonte et al. 2007 ; Bimonte et al. 2007b). GeWOlap est une solution web OLAP-SIG intégrée qui gère les mesures géographiques et complexes, les dimensions géographiques, et implémente un ensemble d'opérateurs de l'algèbre de GeoCube. GeWOlap permet une analyse spatio-multidimensionnelle satisfaisante grâce à une interface qui synchronise une composante tabulaire, une carte interactive et de affichages graphiques.

Enfin, la thèse s'intéresse aussi à la visualisation des mesures géographiques dans un outil SOLAP. Elle propose un nouveau paradigme de visualisation et d'interaction pour l'analyse des mesures géographiques (Bimonte et al. 2006c ; Bimonte et al. 2007a), qui combine les techniques OLAP classiques et une méthode de géovisualisation (Andrienko et al. 2003). L'innovation consiste en particulier dans l'adaptation de la table de pivot et du Space-Time Cube pour l'analyse des données spatio-multidimensionnelles.

Cette thèse s'inscrit dans le cadre d'une collaboration internationale avec l'organisation italienne CORILA (Consorzio per la Gestione del Centro di Coordinamento delle Attività di Ricerca inerenti il Sistema Lagunare di Venezia). Le but de cette organisation est la sauvegarde environnementale, architecturale et économique de la lagune de Venise. L'équipe de l'auteur intervient dans l'axe de recherche « gestion des données ». L'objectif est de fournir aux spécialistes environnementaux un système pour l'analyse spatio-multidimensionnelle de données environnementales (Bimonte et al. 2006d). Ces données concernent la pollution des eaux de la lagune. Elles représentent les mesures de la pollution en 25 différentes zones (appelées unités), à 2826 moments différents (3 années, 28 mois et 207 jours), et pour 100 polluants. Nous utilisons principalement les données environnementales de CORILA pour illustrer nos contributions.

Bibliographie

Andrienko N. V., Andrienko G. L. et Gatalsky P., 2003, "Exploratory spatio-temporal visualization: an analytical review", *Journal of Visual Languages and Computing,* vol. 14, n° 6, p.503-541.

Bédard Y., 1997, *Spatial OLAP. 2e Forum annuel sur la R-D, Géomatique VI: Un monde accessible,* 13-14 Novembre, 1997, Montréal

Bimonte S., Tchounikine A. et Miquel M., 2005, « Towards a spatial multidimensional model », dans Song Il-Yeol et Trujillo Juan, *8th International Workshop on Data Warehousing and OLAP, 4-5 Novembre, 2005 Bremen, Allemagne*., New York, NY, USA : ACM press, p. 39-46.

Bimonte S., Tchounikine A. et Miquel M., 2006, « GeoCube, a Multidimensional Model and Navigation Operators Handling Complex Measures: Application in Spatial OLAP", dans Yaknho Tatyana M. et Neuhold Erich J., *Advances in Information Systems, 4th International Conference, 18-20 October, 2006, Izmir, Turquie,*. Berlin-Heidelberg, Springer, p. 100-109 (Lecture Notes in Computer Science 4243).

Bimonte S., Wehrle P., Tchounikine A. et Miquel M., 2006**, «** GeWOlap: A Web Based Spatial OLAP Proposal", dan Meersman Robert, Tari Zahir et Herrero Pilar, *Workshop on Semantic-Based Geographical Information Systems, 29-30 Octobre, 2006 Montpellier, France,* Berlin-Heidelberg , Springer, p. 1596-1605 (Lecture Notes in Computer Science 4278).

Bimonte S., Di Martino S., Ferrucci F. et Tchounikine A., 2006**, «** GeOlaPivot Table: a Visualization Paradigm for SOLAP Solutions », dans *Visual Languages and Computing Workshop, 30 Août – 1 Septembre, 2006, Grand Canyon, USA.,* K. S. Institute, p. 181-186.

Bimonte S., Tchounikine A., Miquel M., Laurini R. et Ahmed T., 2006, « Spatial Online Analytical Processing for Environmental Data », dans Corila, *Scientific Research and Safeguarding of Venice. Research Programme 2004-2006, Vol. 4, 2005 results*, Venise, Italie, Corila, p. 393-400.

Bimonte S., Tchounikine A. et Miquel M., 2007, « Spatial OLAP: Open Issues and a Web Based Prototype", dans Wachowicz Monica et Bodum Lars, *10th AGILE International Conference on Geographic Information Science, 8-11 Mai 2007, Aalborg, Danemark.*

Bimonte S., Di Martino S., Ferrucci F. et Tchounikine A., 2007**, «** Supporting Geographical Measures Through A New Visualization Metaphor In Spatial OLAP", dans Cardoso Jorge, Cordeiro Jose et Filipe Joaquim, 9th International Conference on Enterprise Information Systems, 12-16 Juin, 2007, Funchal, Madeira, Portugal. INSTICC, p. 19-26.

Bimonte S., Tchounikine A., Miquel M., et Laurini R., 2007**, «** Vers l'intégration de l'analyse spatiale et multidimensionnelle », dans *Colloque International de Géomatique et d'Analyse Spatiale, 18 -20 juin 2007, Clermont-Ferrand, France* .

Denègre J. et Salgé F., 2004, *Les systèmes d'information géographique. 2nde éd.*, Paris, Presses Universitaires de France, 128 p. (Que sais-je?)

Inmon W.H., 1996, *Building the Data Warehouse. 2nd ed.*, New York, Wiley, 401 p.

Longley P., Goodchild M., Maguire D. et Rhind D., 2001, *Geographic Information Systems and Science,* New York , John Wiley & Sons, 517 p.

Malinowski E. et Zimanyi E., 2004, « Representing spatiality in a conceptual multidimensional model", dans Pfoser Dieter, Cruz Isabel F. et Ronthaler Marc, *12th ACM International Workshop on Geographic Information Systems, 12-13 Novembre, 2004, Washington, DC, USA,* New York, USA, ACM Press, p. 12-22.

Malinowski E. et Zimanyi E., 2005, « Spatial Hierarchies and Topological Relationships in SpatialMultiDimER model", dans Jackson Mike, Nelson David et Stirk Sue, *22nd British National Conference on Databases, 5-7 Juillet 2005, Sunderland, UK,* Berlin Heidelberg, Springer, p. 17-28 (Lecture Notes in Computer Science 3567).

Rivest S., Bédard Y., Proulx M.-J., Nadeaum M., Hubert F. et Pastor J., 2005, "SOLAP: Merging Business Intelligence with Geospatial Technology for Interactive Spatio-Temporal Exploration and Analysis of Data", *Journal of International Society for Photogrammetry and Remote Sensing,* vol. 60, n° 1, p. 17-33.

Tchounikine A., Miquel M., Laurini R., Ahmed T., Bimonte S., et Baillot V., 2005, « Panorama de travaux autour de l'intégration de données spatio-temporelles dans les hypercubes », *Revue des nouvelles technologies de l'Information - entrepôts de données et l'analyse en ligne,* p. 21-33.

Weibel R. et Dutton G., 2001, "Generalizing Spatial Data and Dealing with Multiple Representations", dans Longley Paul, Goodchild M., Maguire David et Rhind David, *Geographic Information Systems and Science*, New York , John Wiley & Sons, p. 125-155.