Chapitre 8

Conclusion générale

“ C’est là en effet un des grands et merveilleux caractères des beaux livres que pour l’auteur ils pourraient s’appeler ‘Conclusions’ et pour le lecteur ‘Incitations’. ”

Marcel Proust, “*Sur la lecture*”

8.1 Bilan et contributions

Dans le cadre de cette thèse, nous avons essayé d’apporter des solutions au problème de l’analyse des données complexes. Pour y parvenir, nous nous sommes basés sur le couplage entre l’analyse en ligne et la fouille de données. Nous avons montré que ces deux domaines sont complémentaires et peuvent évoluer dans le cadre d’un processus décisionnel unique. Leur association est capable d’enrichir et de rehausser le processus décisionnel. De plus, la fouille a déjà avancé des solutions pour l’extraction des connaissances à partir des données complexes. Par conséquent, la fouille de données est capable d’étendre les capacités de l’OLAP pour analyser les données complexes.

À partir de la fin des années 90, le couplage de l’analyse en ligne et de la fouille de données a suscité beaucoup d’intérêts. Plusieurs travaux ont abordé le sujet en proposant des approches variées selon différents types de motivations. Néanmoins, nous avons pu distinguer trois grandes approches où chacune se caractérise par une manière d’opérer le couplage entre les deux domaines. La première consiste à transformer les données multidimensionnelles en données tabulaires exploitables par les algorithmes classiques de fouille. La deuxième approche repose sur une extension des outils OLAP et des langages de requêtes des SGBDMs aux techniques de fouille. Enfin, la troisième approche adapte les techniques classiques de fouille au contexte des données multidimensionnelles.

Nos travaux s’articulent autour de ces trois approches. Dans chacune de ces dernières, nous avons avancé une proposition combinant l’OLAP avec une technique

de fouille des données en vue d'améliorer certains aspects de l'analyse en ligne.

Notre première proposition consiste à réorganiser un cube de données en se basant sur l'analyse des correspondances multiples (ACM). Selon la première approche du couplage, nous avons eu recours à une transformation des données du cube en un *tableau disjonctif complet*. À partir de ce tableau, l'ACM est capable de fournir un nouvel ordonnancement des modalités dans les dimensions. Cet ordonnancement engendre une nouvelle représentation du cube et un point de vue intéressant homogénéisant au mieux son nuage des faits. Notre proposition offre ainsi une solution au problème de la visualisation des données engendré par la volumétrie et l'éparsité des ces dernières. Afin de fournir à l'utilisateur une évaluation de la pertinence de notre réorganisation des cubes de données, nous avons également proposé un indice pour la qualité des représentations multidimensionnelles.

Notre deuxième proposition établit une nouvelle agrégation dans un cube de données en utilisant la classification ascendante hiérarchique (CAH). Dans cette proposition, les opérateurs OLAP ont servi comme outils pour extraire les données nécessaires à la classification à partir des données multidimensionnelles. Nous avons mis en place une formalisation permettant de définir, dans un cube de données, les individus et les variables de la classification. Cette dernière fournit des classes de modalités qui sont par la suite exploitées afin de construire de nouveaux agrégats de données *sémantiquement plus riches* que les agrégats classiques de l'OLAP. De plus, afin d'aider l'utilisateur dans le choix du meilleur nombre d'agrégats, nous avons proposé un nouveau critère d'évaluation de la qualité des partitions de la CAH.

Notre troisième proposition est dédiée à la recherche des règles d'association dans les cubes de données. Selon la troisième approche du couplage, nous avons adapté un algorithme, de type Apriori, pour extraire des règles d'association *inter-dimensionnelles* directement à partir d'un cube de données. Notre algorithme repose sur un cadre général dans lequel nous avons défini la notion d'une *méta-règle inter-dimensionnelles* en vue de piloter le processus de fouille vers des contextes d'analyse ciblés par l'utilisateur. Nous avons également adapté les règles d'association au contexte de l'analyse en ligne et redéfini le support et la confiance d'une règle en y intégrant les mesures du cube étudié. Selon cette nouvelle définition, une règle d'association n'est pas évaluée selon la fréquence des faits qui la supportent mais plutôt selon la somme des mesures de ces faits. En plus du support et de la confiance, nous avons intégré le *Lift* et l'indice de *Loevinger* pour évaluer l'intérêt des règles d'association découvertes. Afin de valoriser les connaissances extraites, nous avons proposé une visualisation des règles inter-dimensionnelles dans un espace de représentation multidimensionnel.

Ainsi, à l'image des trois familles des techniques de fouille de données, avec nos trois propositions, nous avons pu étendre l'analyse en ligne à de nouvelles capacités. En effet, (i) la réorganisation des cubes par l'ACM offre à l'OLAP une capacité de *description* et de *visualisation*, (ii) l'agrégation par la CAH dans les cubes de données

lui donne une capacité de *classification* et (iii) l'extraction des règles d'association l'enrichit avec une capacité d'*explication*.

Ces nouvelles capacités, acquises suite au couplage de l'analyse en ligne et la fouille de données, nous ont permis d'aborder le problème de l'analyse des données complexes. Ainsi, nous avons proposé un cas d'application de l'agrégation par classification dans des données complexes relatives au domaine du dépistage du cancer du sein. Ces données représentent des dossiers de patientes où chaque dossier comprend des sources hétérogènes et éparpillées sur plusieurs types de supports.

En vue de préparer ces données médicales pour les fins décisionnelles de notre cas d'application, nous les avons représentées dans des documents XML. Ensuite, nous avons défini un contexte d'analyse sur ces données. Pour cela, dans des travaux annexes, nous avons formulé une méthodologie d'entreposage des données complexes basée sur XML. Avec cette méthodologie, appelée X-Warehousing, il nous a été possible de concevoir et de créer un *cube XML de mammographies*. Suite à l'agrégation par classification dans les données de mammographies, les résultats obtenus nous ont montré l'intérêt de notre proposition pour l'analyse des données complexes.

Sur un plan technique, pour valider nos contributions, nous avons mis en place une plateforme logicielle générale d'analyse, appelée MiningCubes. Il s'agit d'une application Web, dotée d'interfaces ergonomiques, faciles à utiliser et adaptées au contexte de l'analyse en ligne. Dans cette application nous avons implémenté des modules pour nos précédentes propositions. Nous avons également implémenté un module de connexion aux cubes XML afin de prendre en compte l'analyse des données complexes.

Pour terminer, l'ensemble de nos travaux nous ont amenés à réfléchir à un cadre formel général pour le couplage de l'analyse en ligne et de la fouille de données. À cet effet, nous avons proposé un modèle multidimensionnel et une algèbre OLAP reposant sur un *noyau minimal fermé* d'opérateurs de *structuration* et de *navigation*. En se basant sur nos différentes expériences, nous avons formulé une première tentative pour étendre cette algèbre à une nouvelle génération d'opérateurs de *fouille de données en ligne*.

8.2 Perspectives de recherche

Les travaux réalisés dans cette thèse ouvrent diverses perspectives de recherche. Tout d'abord, nous continuons à croire que le couplage de l'analyse en ligne et de la fouille de données est une solution adéquate pour l'analyse des données complexes. Nous projetons la généralisation des cas d'application aux données complexes de nos différentes propositions basées sur le couplage. Nous pensons que, par analogie à l'agrégation par classification, la réorganisation par l'ACM et l'explication par les règles d'association peuvent aussi fournir des connaissances pertinentes dans les

données de mammographies, en particulier, et dans les données complexes, en général.

Nous croyons aussi que XML est une solution adaptée à la modélisation multidimensionnelle des données complexes. Au vu des divers efforts dans le domaine des entrepôts de données XML, nous pensons que, dans un avenir proche, XML sera un nouveau standard pour un processus d'entreposage particulièrement adapté aux données complexes. Cette évolution, va naturellement engendrer une redéfinition des mécanismes d'interrogation des données au niveau de l'analyse en ligne. Parallèlement, l'extension de l'analyse en ligne à la fouille doit aussi tenir compte de cette nouvelle représentation des données complexes. D'une manière similaire aux données multidimensionnelles, nous pensons que nous serons amenés à réfléchir à un nouveau type de couplage entre l'analyse en ligne et la fouille de données qui *adapterait les algorithmes de fouille aux données XML.*

Dans nos travaux réalisés, nous avons exploité le couplage de l'analyse en ligne et de la fouille de données afin d'étendre les capacités de l'OLAP. Ces capacités ont porté principalement sur la *description* et la *visualisation*, la *classification* et l'*explication*. Cependant, il est encore important d'étendre l'analyse en ligne à des capacités de *prédiction*. En effet, dans un processus décisionnel, un utilisateur observe les faits OLAP dans un cube afin d'extraire des informations intéressantes au regard du contexte d'analyse. Ces informations permettent à l'utilisateur de comprendre des relations ou des phénomènes existants dans les données. Ils permettent aussi à l'utilisateur d'anticiper, intuitivement, la réalisation de phénomènes futurs selon un certains nombre de conditions. Nous pensons que, avec une technique de prédiction appropriée au contexte des données multidimensionnelles, il est possible d'assister l'utilisateur dans cette tâche. La combinaison de l'analyse en ligne avec une technique de prédiction est capable de fournir, par exemple, des estimations des valeurs des mesures d'un fait inexistant ou d'un fait qui va se réaliser dans l'avenir.

Enfin, nous sommes convaincus de la nécessité de la mise en place d'un cadre formel général pour le couplage de l'analyse en ligne et de la fouille de données. Nous avons déjà mis en place une première base théorique à cet effet. Nous projetons une formalisation complète de ce cadre afin de fournir une algèbre générale incluant à la fois les opérateurs classiques de l'OLAP et la nouvelle génération des opérateurs de *fouille de données en ligne*. À l'image de nos réalisations existantes et futures, notre objectif est d'étendre le noyau minimal de notre algèbre actuelle à un nouveau noyau dédié, non seulement à la *structuration* et la *navigation* dans les données multidimensionnelles, mais aussi à la *description*, la *classification*, l'*explication* et la *prédiction* dans les données complexes.