Introduction générale

La fouille de données, également connue sous l'expression anglo-saxonne *data mining*, est une jeune discipline apparue au début des années 90. Son émergence est principalement due d'une part à des avancées technologiques notamment dans les domaines du stockage informatique et de la vitesse de transmission des réseaux et d'autre part d'un contexte socio-politico-économique mettant en avant le besoin de valoriser les données autrefois uniquement gérées dans un but d'accès ou de restitution.

L'Extraction de Connaissances à partir de Données (ECD) est un processus complexe qui se déroule suivant une série d'opérations (FAYYAD *et al.*, 1996). Nous pouvons regrouper ces opérations en trois étapes majeures. Elles sont la préparation des données, la fouille de données à proprement parler qui est l'étape centrale de l'ECD et enfin la validation des modèles ainsi élaborés (ZIGHED et RAKOTOMALALA, 2002) :

- Les étapes de pré-traitements portent sur l'accès aux données en vue de construire des tables bidimensionnelles, des corpus de données spécifiques. En fonction du type des données (e.g. numérique, symbolique), des méthodes de pré-traitements mettent en forme les données, les nettoient, traitent les données manquantes et sélectionnent les attributs ou les individus lorsque ceux-ci sont trop nombreux : choix des attributs les plus informatifs dans le premier cas, et échantillonnage dans le second. Cette phase est loin d'être anodine puisque c'est elle qui va conditionner la qualité des modèles établis lors de la fouille de données. En effet, ces choix ont pour objectif de faire émerger l'information contenue au sein de cette masse de données.
- La fouille de données fait appel principalement aux disciplines de l'intelligence artificielle, de la statistique et de l'analyse de données. Elle s'effectue généralement sur des tables bidimensionnelles et se décompose essentiellement en trois grandes familles de méthodes :
  1. Les méthodes descriptives qui sont principalement issues de la statistique descriptive et de l'analyse des données, et adaptées ou augmentées à l'aide de techniques de visualisation graphique pouvant faire appel à la réalité virtuelle et à des métaphores de représentation assez élaborées ;
  2. Les méthodes de structuration qui regroupent toutes les techniques d'apprentissage non supervisé et de classification automatique provenant des domaines de la reconnaissance de formes, de la statistique, de l'apprentissage automatique et du connexionisme ;
  3. Les méthodes explicatives qui cherchent à établir un modèle décrivant un phénomène, défini à partir d'une variable endogène, à l'aide d'un ensemble de descripteurs appelés variables exo-

gènes. Ces méthodes proviennent de la statistique, de la reconnaissance de formes, de l'apprentissage automatique et du connexionisme, voire du domaine des bases de données dans le cas de la recherche de règles d'association.

- L'obtention de ces modèles est une étape intéressante et riche en informations. Cependant, ces informations n'acquièrent le stade de connaissance qu'après leur validation. Au regard des résultats en validation, une méthode (ou un ensemble de méthodes) est jugée meilleure que d'autres. Mais malheureusement, aucune méthode de validation ne permet de déterminer quel est l'outil que l'on doit employer pour répondre à tel problème. Un consensus général semble se dégager pour reconnaître qu'aucune méthode ne surpasse les autres car elles ont toutes leurs forces et faiblesses spécifiques les rendant ainsi plus ou moins performantes pour un problème donné. Il semble plus avantageux de faire coopérer des méthodes comme nous le ferions avec une équipe de spécialistes. Enfin, le choix des méthodes dépend également et surtout de l'utilisateur et de la compréhension qu'il peut avoir des modèles élaborés.

Les techniques de fouille de données ont été employées avec beaucoup de succès dans de grands secteurs d'application tels la gestion de la relation client ou encore la gestion des connaissances (LEFEBURE et VENTURI, 2001). Par ailleurs, face à l'explosion des technologies de l'information, de nouveaux types de documents se sont massivement répandus. Par exemple, le Web est un vecteur de forte diffusion de documents multimédia comme le texte, l'image et la vidéo (CLECH et HASSAS, 2003). Les usages ont conséquemment ou parallèlement évolué, et l'échange de documents numériques est chaque jour plus important, faisant ainsi émerger des besoins nouveaux. Les bases de données s'adaptent en proposant un cadre plus générique cherchant à intégrer ces données fortement hétérogènes et pour la plupart non structurées.

Cet ensemble de données fortement hétérogènes et non structurées semble consensuellement s'appeler données complexes, regroupant à la fois des données usuelles (numériques, qualitatives à valeur discrète), des données moins élémentaires (intervalle, distribution, floues, imprécises), des données temporelles, ou encore des données à contenu sémantique riche pour l'humain comme les données à support média. La qualification des données en données complexes est davantage qu'une simple généralisation des familles de données, et l'extraction de connaissances à partir de données complexes nécessite une modélisation spécifique et des méthodes d'accès très avancées.

Il est difficile de définir ce que sont les données complexes dans le cadre de l'ECD. Nous retiendrons que la découverte de connaissances à partir de telles données nécessite de considérer leur interprétation. A la différence d'une donnée « simple » (e.g. numérique), la lecture d'une donnée complexe re-

quiert une analyse de la part du lecteur. Ce dernier doit généralement se baser sur sa connaissance du domaine dont est issue la donnée. D'autre part, cet effort d'interprétation lié à une donnée complexe peut demander un besoin de la contextualiser à l'aide de données « voisines » afin de faire émerger les particularités intrinsèques de ladite donnée complexe.

Ainsi, la fouille de données complexes concerne l'extraction de connaissances implicites, les relations entre données, ou d'autres structures qui ne sont pas explicitement stockées dans les bases de données. En ce sens, nous proposons de caractériser les données complexes comme étant un ensemble de données regroupant des modalités de type fortement éloigné et dont la similarité entre deux individus nécessite une contextualisation ainsi qu'un point de vue subjectif. Dans un contexte médical, le dossier d'un patient nous semble un parfait exemple pour illustrer un tel type de données, car contenant des données usuelles comme celles décrites dans les analyses biologiques, des données temporelles indiquant l'évolution du patient comme l'électrocardiogramme, des clichés et des compte-rendus médicaux synthétisant la situation du patient.

A notre sens, ces caractéristiques supplémentaires nécessitent leur prise en compte au sein du processus ECD. Dans le but de mettre en exergue ces caractéristiques à chacune des étapes du processus, nous proposons ici de spécifier ce dernier par Extraction de Connaissances à partir de Données Complexes (**ECDC**). En outre, par opposition à la fouille de données usuelles, nous proposons d'appeler **objet** ou **document** ce qui est classiquement défini par individu, un objet étant un agrégat de « sous objets » ou de documents. De même, nous proposons d'appeler **corpus**, et non jeu de données, un ensemble d'objets complexes liés a une même thématique.

Cette thèse s'inscrit dans le cadre général du traitement de ces données complexes. Nous avons plus particulièrement utilisé les données textuelles pour la mise en application de ces traitements. Nos contributions personnelles se déclinent en trois volets : méthodologique, applicatif et développement logiciel.

L'aspect méthodologique a consisté à adapter ou proposer de nouvelles méthodes intervenant dans les étapes du processus d'ECDC. Dans le cadre de la problématique de la représentation des données, nous nous sommes intéressés à la sélection supervisée de variables textuelles, et nous avons proposé une méthode multivariée (CLECH *et al.*, 2003a). Cette méthode originale se démarque des méthodes usuelles en sélection de termes par l'utilisation d'une information plus riche. En effet, elle est fondée sur la fréquence des termes et non sur la simple présence ou absence d'un terme dans les documents

composant le corpus. En outre, elle prend en compte les interactions entre les termes ainsi que les interactions entre les termes et les catégories.

Toujours dans la problématique de la représentation des données complexes, nous nous sommes aussi intéressés à la qualité de l'étiquetage lors d'un apprentissage supervisé (CLECH et ZIGHED, 2004). L'étiquetage consiste à assigner les catégories d'appartenance d'un document. Par essence, l'apprentissage supervisé nécessite que chaque document du corpus d'apprentissage soit préalablement étiqueté. Cette opération est réalisée par un expert et peut être perçue comme subjective puisque basée sur l'interprétation du document. La qualité de cette opération est primordiale car impliquant directement sur les performances du modèle élaboré. Pour améliorer cette qualité, nous utilisons une méthode de relaxation visant à ré-étiqueter automatiquement certains documents en fonction de documents « voisins ». Cette approche est appliquée sur un corpus de données textuelles.

Dans le cadre de l'apprentissage, nous avons proposé l'amélioration (CIAMPI *et al.*, 2000) d'une méthode à base d'arbre nommée RECPAM. Nous montrons que les graphes d'induction peuvent traiter des problèmes de prédiction plus généraux que la prédiction d'une variable catégorielle ou d'une variable continue. Cette extension propose de traiter la prédiction d'une réponse multivariée et continue. En généralisant le cadre de notre étude, nous montrons qu'avec la même démarche nous pouvons traiter des problèmes d'apprentissage supervisé ou non. Cette approche est appliquée sur des données de diététique alimentaire.

Nous avons ensuite abordé la problématique de la visualisation d'objets complexes et de leurs interactions et la problématique de la navigation au sein d'un corpus. Nous proposons un modèle topologique basé sur les graphes de voisinage permettant la visualisation du contexte d'un objet, des interactions inter-objets et de la globalité du corpus. Notre outil permet également de naviguer facilement au sein du corpus et de visualiser les documents composant un objet complexe particulier.

Enfin, nous nous sommes intéressés aux méthodes d'interrogation ou de recherche d'information au sein de telles données. Nous proposons, comme pour la visualisation, de nous baser sur un modèle topologique (SCUTURICI *et al.*, 2003a ; SCUTURICI *et al.*, 2003b ; SCUTURICI *et al.*, 2004) car ce dernier contient des propriétés intéressantes pour faciliter la recherche effectuée par l'utilisateur, à la différence des systèmes basés sur les méthodes des k plus proches voisins. Nous proposons également une approche basée sur les techniques de fouille de données afin de prendre en compte par apprentissage la partie non exprimée de la requête de l'utilisateur.

Sur le plan applicatif, nous avons traité trois applications dans le cadre des données textuelles. La première concerne le traitement des curriculums vitæ (CLECH et ZIGHED, 2003). La problématique que nous abordons est l'élaboration de modèles permettant la détection automatique des curriculum vitæ (CV) de cadres au sein d'un corpus constitué à plus de 90% par des CV de non cadres. Le CV est un document textuel singulier : faible structure, informations éparses, contenu fortement symbolique, …, d'où la difficulté de traitement de ces documents.

La seconde concerne des données médicales avec l'étude de compte-rendus médicaux de mammographie (CLECH *et al.*, 2003b). Nous exprimons les motivations et la démarche nous ayant conduit à utiliser un corpus de compte-rendus de mammographies élaborés par des radiologues, l'objectif étant de déterminer un ensemble de caractéristiques fréquentes et discriminantes d'une mammographie permettant de prédire la malignité ou la bénignité d'un cas.

La dernière concerne l'exploitation de données textuelles multilingues (JALAM *et al.*, 2004). Nous proposons un cadre pour la catégorisation de textes multilingue. L'objectif est de pouvoir, à partir d'un modèle de prédiction construit par apprentissage sur un corpus de textes rédigés dans une langue donnée, inférer sur une série de textes qui sont rédigés dans une langue quelconque. Cette phase d'inférence, par rapport à la généralisation classique, comprend deux étapes supplémentaires : la détection de la langue du texte, puis sa traduction automatique vers la langue de référence. Nous avons mis en œuvre cette approche sur un ensemble d'articles de journaux composés d'articles rédigés en langues allemande, anglaise et française.

Enfin, dans le cadre du développement logiciel, nous avons produit la plate-forme RECPAM permettant le traitement de données dont les variables ont une forte interaction les unes avec les autres. Ce développement a été réalisé en collaboration avec le Professeur CIAMPI, A. de l'Université MAC GILL, CANADA. Nos rencontres régulières ont permis de faire évoluer cette plate-forme en lui ajoutant d'autres éléments d'apprentissage ou de validation.

Le second développement logiciel, et certainement le plus conséquent, a été la réalisation de l'ensemble des outils permettant les différents tests effectués durant ces années de thèse. Il est à noter que la quasi totalité des méthodes présentées dans cette thèse a été développée d'un point de vue logiciel.

Introduction générale

Cette thèse va donc présenter nos contributions méthodologiques au cours des quatre premiers chapitres selon les étapes du processus ECD appliqué aux données complexes, alors que le dernier chapitre est consacré aux applications réalisées :

- Le Chapitre 1 vise à poser les objectifs et besoins de la représentation de telles données. Nous approfondirons en seconde partie de ce chapitre le cas des données textuelles ;
- Le Chapitre 2, aborde les méthodes d'apprentissages à base d'instances en s'intéressant plus spécifiquement aux approches géométriques qui, comme nous le verrons, apportent un ensemble de propriétés très intéressantes pour le traitement des données complexes. En dernière partie nous traiterons de la problématique de la validation des modèles ;
- Le Chapitre 3 est consacré à la définition du problème de la visualisation des données en général. Puis, après exploration des éléments répondant partiellement aux contraintes de cette visualisation, nous proposerons un ensemble d'outils permettant la visualisation des données complexes ;
- Le Chapitre 4 est dédié à l'interrogation au sein de ces données. Nous aborderons les systèmes de recherche d'information usuels puis traiterons ceux basés sur le contenu ;
- Enfin, le Chapitre 5 rapporte un ensemble d'applications liées au traitement de données complexes.

Liste des publications réalisées dans le cadre de cette thèse.

- Revue internationale avec comité de lecture :

  SCUTURICI, M., CLECH, J., ZIGHED, D. A. et SCUTURICI, V. M., Topological Representation Model for Image Database Query. *Journal of Experimental & Theoretical Artificial Intelligence*: à paraître. **2003a**.

- Conférences internationales avec comité de lecture et actes :

  JALAM, R., CLECH, J. et RAKOTOMALALA, R., Un cadre pour la catégorisation de texte multilingues, *7th International Conference on the Statistical Analysis of Textual Data*, Louvain-la-Neuve, Belgique, à paraître. **2004**.

  SCUTURICI, M., CLECH, J. et ZIGHED, D. A., Topological Query in Image Databases. *8th Iberoamerican Congress on Pattern Recognition*, Havana, Cuba: 144-151. **2003b**.

  CIAMPI, A., ZIGHED, D. A. et CLECH, J., Trees and Induction Graphs form Multivariate Response. *Principles of Data Mining and Knowledge Discovery, Proceedings of the 4th European Conference on Principles and Practice of Knowledge Discovery in Databases*, Lyon, France, Springer-Verlag Berlin Heidelberg, 1910: 359-366. **2000**.

- Revues nationales avec comité de lecture :

  CLECH, J. et ZIGHED, D. A., Une technique de ré-étiquetage dans un contexte de catégorisation de textes. *Document Numérique – Fouille de Textes et Organisation de Documents*, Editions Lavoisier: à paraître. **2004**.

  CLECH, J., ZIGHED, D. A. et BREMOND A., Apport des techniques de Text Mining pour la définition de caractéristiques clefs d'une mammographie. *Revue des Nouvelles Technologies de l'Information*, Lyon, France, Cépaduès, 1: 183-192. **2003b**.

- Conférences nationales avec comité de lecture et actes :

  SCUTURICI, M., CLECH, J., ZIGHED, D. A. et SCUTURICI, V. M., Modèle topologique pour l'interrogation des bases d'images. *Conférence EGC*, Clermont-Ferrand, Editions Lavoisier: à paraître. **2004**.

  CLECH, J., RAKOTOMALALA, R. et JALAM, R., Sélection multivariée de termes. *XXXVèmes Journées de Statistique*, Lyon, France, 2: 933-936. **2003**.

  CLECH, J. et ZIGHED, D. A., Data Mining et Analyse des CV : Une Expérience et des Perspectives. *Conférence EGC*, Lyon, France, Editions Lavoisier, 17: 189-200. **2003a**.

- Tutoriel :

  CLECH, J. et HASSAS, S., Web Mining et Système Multi-Agents. *Tutoriel EGC 2003*, Lyon, France. **2003**.

- Séminaire :

  CLECH, J., Web Mining et applications. *Club SAS*, Paris, France. **2002**.