Conclusion générale

La fouille de données complexes se propose de fournir un modèle d'analyse permettant d'intégrer de larges variétés de données, structurées et non structurées locales ou distantes. Le point de vue retenu est de dire que, face à une tâche d'extraction des connaissances à partir des données, l'utilisateur doit être libéré des contrainte liées à l'organisation, le codage, le format, la représentation,… des données. Pour lui, l'important c'est le contenu. Il doit donc disposer des outils qui lui permettent d'accéder à ce contenu car c'est celui-ci qui doit être traité pour en extraire la quintessence. Après tout, on peut décrire la réalité par différents modes : en image, en texte, graphiques, ou d'autres moyens encore. Sans être exactement identiques, ces contenus se recouvrent dès lors qu'ils se rapportent à une même réalité. Le cerveau humain réalise cette opération de façon naturelle. Nous écoutons un interlocuteur et nous intégrons naturellement sa gestuelle dans la compréhension du message. Une fusion entre les données s'opère pour que nous ayons accès à un signifiant. C'est dans cette direction que nous avons lancé ces travaux.

Nous sommes conscient que dans le cadre de cette thèse nous n'avons abordé que quelques aspects de ce vaste chantier. Par exemple, nous n'avons pas abordé l'intégration des données temporelles, ni celle des données spatiales. Nous avons tenté de montrer la faisabilité d'une approche unique en fouille de données capable de prendre en compte des données textuelles, images, numériques, etc.

Nous avons tout au long de cette thèse repris les étapes du processus ECD, en tentant de mettre en valeur les éléments permettant d'intégrer le traitement des données complexes. Ainsi, au Chapitre 1 nous nous sommes intéressés à l'extraction de caractéristiques de données complexes afin de permettre l'utilisation des méthodes usuelles de fouille de données. Nous avons décrit la difficulté de cette étape de mise en forme qui doit absolument préserver le contenu informationnel des objets complexes. Nous avons illustré cette problématique à travers le cas des données textuelles pour lesquelles nous avons proposé une nouvelle méthode de sélection de termes.

Le Chapitre 2 nous a permis de poser un cadre formel sur les méthodes à base d'instances. L'intérêt majeur que nous y voyons est leur capacité à contextualiser un objet dans son espace de représenta-

tion. En ce sens, nous avons montré que l'exploitation de la structure reliant les individus du graphe de voisinage permet de définir facilement cette notion de contextualisation.

En nous basant sur ces méthodes, nous avons pu proposer au Chapitre 3 un environnement de visualisation et de navigation au sein de données complexes. En effet, nous pouvons visualiser les documents qui composent un objet complexe particulier. En outre, en mettant en œuvre le principe de contextualisation, nous permettons la visualisation des objets voisins. Par ailleurs, nous autorisons la navigation au sein de ces données soit à l'aide d'une navigation contextuelle en nous déplaçant de voisin en voisin, soit à l'aide d'une navigation globale permettant à l'utilisateur de sélectionner l'objet qui l'intéresse.

Enfin, nous avons lors du Chapitre 4 abordé la question de l'exploitation de telles données à travers la problématique de la recherche d'information. Nous avons proposé de nouveaux systèmes de recherche prenant en compte les caractéristiques des données complexes. Le principe de contextualisation a permis de définir un système où l'utilisateur soumet une requête par l'exemple et peut affiner sa recherche en se déplaçant de voisin en voisin. Le besoin d'interprétation de l'utilisateur a été mis en œuvre dans un autre système qui modélise l'interprétation faite par l'utilisateur de la requête initiale qu'il a soumis au système.

Concrètement, la formalisation de la problématique du traitement des données complexes nous a permis de réaliser un ensemble d'applications exposé au Chapitre 5.

Les perspectives de ces travaux sont multiples. Dans l'immédiat cela peut concerner l'ensemble des perspectives évoquées dans les diverses applications du Chapitre 5.

A moyen terme, nous pensons développer notre environnement de visualisation et de recherche au sein de données complexes. Notre outil de visualisation offre l'information nécessaire pour permettre l'interprétation d'un objet complexe. Cependant, il nous semble pertinent d'intégrer des méthodes automatiques (ou semi-automatiques) facilitant ainsi la compréhension du document par l'utilisateur. Par exemple, nous pouvons envisager d'intégrer des méthodes de résumés automatiques de documents textuels en nous basant sur des documents voisins. Nous pouvons également envisager d'extraire automatiquement les caractéristiques communes des documents issus d'un même sommet d'un arbre plylogénétique, et mettre en avant ce qui rend différent un document d'un autre.

Enfin, sur une durée plus longue, nous allons approfondir la notion de similarité entre documents complexes. En effet, pour définir la similarité entre deux objets complexes, nous nous sommes basés ici sur une simple mesure comparant les caractéristiques deux à deux. Il nous semble que cette approche est insuffisante car elle ne tient pas compte de toute l'information que peuvent contenir de tels documents. Une des pistes que nous allons étudier est les mesures de similarités capables de prendre en compte des hiérarchies d'objets.