Découverte interactive de règles d'association via une interface visuelle

Pascale Kuntz, Rémi Lehn, Fabrice Guillet, Bruno Pinaud

Découverte interactive de règles d'association via une interface visuelle

Pascale Kuntz, Rémi Lehn
Fabrice Guillet, Bruno Pinaud

Laboratoire d'informatique de Nantes Atlantique (LINA)
Site Ecole Polytechnique
La Chantrerie - rue Christian Pauc
BP 50609, 44306 Nantes Cedex 3
{pascale.kuntz, remi.lehn, fabrice.guillet, bruno.pinaud}@univ-nantes.fr,

Résumé. En nous appuyant sur des hypothèses majoritairement empruntées à des travaux sur les systèmes anthropocentrés d'aide à la décision, nous décrivons dans cet article un environnement interactif de fouille de règles d'association dans lequel l'utilisateur pilote le processus, en jouant le rôle d'une heuristique dans un environnement de recherche complexe. Afin de permettre à la fois une représentation visuelle accessible et une instanciation aisée des outils d'interactivité le modèle choisi est ici un graphe en niveaux - les niveaux étant associés aux cardinaux des sous-ensembles d'attributs des prémisses. Le processus a été déployé dans un logiciel prototype dont l'analyse des résultats ouvre de nouvelles perspectives sur l'analyse comportementale d'un utilisateur en situation de fouille.

1 Introduction

Si les efforts les plus remarqués de la première décennie de l'Extraction de Connaissances dans les Données (ECD) ont principalement porté sur le développement d'algorithmes automatiques performants, le rôle de l'utilisateur est peu à peu devenu un sujet de préoccupation majeur. Ce besoin d'intégration a conduit à l'émergence de nombreux outils de visualisation et d'interaction [Fayyad *et al.*, 2002], [Grinstein, 1996]. Il s'agit d'apporter à l'utilisateur un substrat artificiel qui transcrive un grand nombre d'informations et qui soit un support à ses connaissances et à son intuition pour lui permettre de découvrir de nouvelles relations et d'imaginer de nouvelles questions.

1.1 Visualisation et interaction en ECD

Traditionnellement, le recours à des techniques de visualisation précède et suit les étapes de traitement automatiques. En amont du processus d'ECD, elles assistent les tâches de sélection et de pré-traitement des données en renforçant la convivialité des interfaces. En aval, à l'issue de l'application des algorithmes de découverte de structures, elles visent à présenter des résultats sous des formes intelligibles facilitant leur interprétation. Dans ce contexte, la plupart des outils ont été définis pour des objectifs spécifiques (recherche de règles d'association, classification de données spatiales, etc.).

En parallèle, dans la lignée de l'analyse exploratoire des données [Cleveland, 1985], [Tukey, 1977], la "fouille visuelle de données" s'est développée pour mettre en lumière les propriétés intrinsèques des données et permettre la génération d'hypothèses (*data-driven hypothesis generation* v.s. *task-oriented data mining*) [Keim et Kriegel, 1996]. L'accent est alors mis sur la conception de représentations de données nombreuses adaptées aux capacités remarquables du système de perception visuelle [Spence, 2001]. Dans cette démarche, dont l'origine est souvent attribuée aux travaux précurseurs de J. Bertin [Bertin, 1977] et E. Tufte [Tufte, 1983], il s'agit d'utiliser l'espace pour organiser les données.

En complémentarité étroite avec la visualisation, l'interaction est devenue un mot clé des logiciels d'ECD. Dans la phase de fouille, ce concept est mis en œuvre à deux niveaux : pour les langages de requêtes spécifiques (e.g. [Goethals et van den Bussche, 2000]), et surtout, pour les interfaces permettant à l'utilisateur d'interagir avec les représentations. Dans ce dernier cas, l'interaction matérialise la boucle de rétroaction entre l'utilisateur et le support graphique [Keim et Kriegel, 1996]. Avec un recours massif aux opérations de "*roll-up*" et "*drill-down*", l'interaction permet alors d'obtenir sur des vues complémentaires sur les données selon le leitmotiv de B. Shneiderman [Shneiderman, 1996] "*Overview first, zoom and filter, then details -on demand*". Bien que les modèles théoriques sous-jacents aux représentations les plus populaires soient encore bien souvent assez simples (nuages de points, tableaux de contingence, distributions de fréquences, etc.), les modes de visualisation-interaction peuvent être sophistiqués, le recours récent à la réalité virtuelle étant un exemple extrême.

1.2 Vers un système de fouille centré sur l'utilisateur

Le rôle de l'utilisateur dans un processus d'ECD avait été mis en avant dès le début des années 90 dans l'article fondateur de W.J. Frawley et al. [Frawley *et al.*, 1992] : dans le paragraphe consacré aux directions futures de l'ECD, ces derniers insistaient sur la nécessité d'intégrer l'utilisateur dans la "boucle de découverte" afin de combiner les potentialités humaines de jugement avec les capacités de calcul de la machine. Dans cette perspective, l'ambition de l'ECD est de dépasser l'assistance technologique à la manipulation et l'interprétation des données et des connaissances en considérant explicitement l'utilisateur comme une composante à part entière du processus dynamique de découverte ; "*The process of knowledge discovery in databases : a human centered approach*" ainsi que l'énoncent R.J. Brachman et T. Anand [Brachman et Anand, 1996]. Cependant, si cet objectif d'intégration est de plus en plus revendiqué lors du développement d'outils de visualisation, peu de travaux sont à notre connaissance consacrés à la caractérisation de la tâche de l'utilisateur, comme processus cognitif, dans la démarche menant à la découverte de connaissances potentiellement utiles [Frawley *et al.*, 1992].

S'appuyant sur des hypothèses empruntées à des travaux sur les systèmes anthropocentrés d'aide à la décision (e.g. [Barthélemy et Mullet, 1992], [Bisdorf, 2002]), l'objectif qui sous-tend les travaux présentés dans cet article, est le développement d'un environnement interactif de fouille de règles d'association dans lequel l'utilisateur pilote le processus, en jouant en quelque sorte le rôle d'une heuristique dans un environnement de recherche complexe modélisé ici par un graphe de règles.

Dans une première partie, nous discutons des hypothèses sur le comportement de l'utilisateur en phase de fouille sur lesquelles se fondent notre travail. Puis, dans la seconde partie, nous présentons un modèle de représentation, spécifiquement adapté à la fouille interactive de règles d'association. L'architecture du prototype que nous avons développée est synthétique-

ment présentée dans la troisième partie. Nous concluons par une discussion sur les premières applications qui ouvre sur des perspectives sur l'analyse comportementale d'un utilisateur en ECD.

2 Hypothèses sur le comportement de l'utilisateur

Bien qu'il ne s'agisse vraisemblablement pas au stade de la fouille d'une tâche explicite de résolution de problème mais plutôt d'une tâche d'acquisition des connaissances pour une prise de décision ultérieure non nécessairement verbalisée préalablement, il nous semble que, d'un point de vue épistémologique, un parallèle fécond peut être établi avec l'aide à la décision.

L'évolution de la recherche opérationnelle classique vers l'aide multicritère à la décision [Roy et Bouyssou, 1993] est caractérisée par un passage du problème mathématique d'optimisation vers un problème qualifié par certains de "logique ou sémiotique" où l'attention se décale vers les préférences pragmatiques du décideur [Bisdorf, 2002]. Associé au développement des sciences cognitives, le nouveau glissement plus récent vers les systèmes anthropocentrés d'aide à la décision [Barthélemy *et al.*, 2002] vise à introduire dans la méthodologie les stratégies cognitives du décideur.

Du fait probablement de l'histoire plus récente de l'ECD en tant que telle et des pressions des besoins exprimés dans le contexte marchand pour développer des outils opérationnels rapidement les stratégies cognitives de l'utilisateur n'ont guère encore été étudiées. Par conséquent, nous nous sommes appuyés d'une part, sur des recherches plus abouties en aide à la décision, et d'autre part sur l'analyse de quelques retours d'expérience.

2.1 L'utilisateur-décideur

Dans ses commentaires de l'article de synthèse de D. Hand et al. [Hand *et al.*, 2000] sur les spécificités de l'ECD, un chercheur en économie rappelle la citation d'un ouvrage de statistique des années 40 : "*The purpose of collecting data is to provide a basis for action*" [Deming, 1943]. Une illustration concrète de cet adage en ECD est donnée par le "cercle vertueux" de l'exploitation des données en marketing [Berry et Linoff, 1997], un des champs d'applications à l'origine de nombreux problèmes majeurs du domaine, dont en particulier celui de la découverte des règles d'association pertinentes : (1) Identifier l'opportunité commerciale. Cette étape arrive lorsqu'un apport d'informations permet l'amélioration d'exécution de tâches, (2) Utiliser des techniques de fouille pour transformer les données en informations permettant des opérations concrètes, (3) Agir sur les informations en les incorporant dans le processus commercial pour que les actions qui en découlent soient parties intégrantes du cercle vertueux, (4) Mesurer les résultats.

A l'étape 2, sur laquelle se focalise cet article, l'objectif est donc bien celui d'une acquisition d'informations à valeur ajoutée qui doit guider aux étapes suivantes une décision et sa mise en œuvre. Plus généralement, bien que l'ECD soit orientée *in fine* par des buts, ceux-ci ne sont pas toujours explicitement exprimés et peuvent mettre en jeu des échelles de temps variées.

Si il dispose d'un système d'extraction interactif, une part de la démarche de cet utilisateur-décideur peut être associée à une stratégie de navigation dans un environnement complexe, qui se situe dans un continuum entre la requête qui correspond à un but précis et bien défini, et

le butinage qui correspond à un besoin mal exprimé initialement qui conduit l'utilisateur à poursuivre son chemin jusqu'à une certaine satisfaction. Ainsi, comme pour certains systèmes interactifs d'aide à la décision [Vincke, 1992], la procédure ne s'arrête donc pas avec un test de convergence, mais s'arrête parce que l'utilisateur a le sentiment d'avoir obtenu suffisamment d'informations utiles pour son problème.

2.2 Cadre méthodologique

Les hypothèses précédentes nous ont conduit à adapter pour la découverte interactive de connaissances les quatre étapes bien connues de la résolution de problème proposées par A. Newel et H. Simon dans leur ouvrage *Human Problem Solving* [Newell et Simon, 1972] :

1. La caractérisation de l'environnement. Il s'agit d'extraire les propriétés et opérations pertinentes et accessibles à l'utilisateur pour la découverte de connaissances ;
2. Le choix d'une représentation formelle pour définir l'espace de navigation ;
3. L'implémentation informatique de cette représentation formelle : son codage et sa représentation visuelle associée ;
4. L'implémentation de la procédure de découverte de connaissances : l'interactivité.

Notons, qu'à l'étape 4, le traitement automatique de l'information de l'hypothèse cognitiviste du modèle initial de Newel et Simon, est remplacé ici par une paire {utilisateur, heuristique locale} où, lors d'un processus itératif, l'utilisateur par son action sur un support de représentation déclenche un algorithme qui transforme les informations présentées sur le support. Ces transformations peuvent être de différentes natures, en particulier, le changement de point de vue, et l'ajout ou le retrait d'informations.

Dans la suite nous nous attachons à décliner les différentes étapes de ce modèle pour la problématique spécifique de la fouille de règles d'associations.

2.3 Hypothèses spécifiques à la fouille de règles d'association

Introduites en ECD au début des années 90 par Agrawal et al. [Agrawal *et al.*, 1993] pour exprimer simplement des tendances implicatives entre les attributs d'une table relationnelle, les règles d'association ont rapidement connu une utilisation intensive. Rappelons qu'une règle d'association $A_i \rightarrow A_j$ entre deux sous-ensembles A_i et A_j d'un ensemble d'attributs A signifie que les individus qui présentent les attributs de A_i ont tendance à avoir également ceux de A_j. Contrairement aux approches initiales de l'analyse combinatoire des données et de la classification conceptuelle, on relaxe ici la condition d'inclusion $I(A_i) \subset I(A_j)$ entre le sous-ensemble d'individus $I(A_i)$ décrits par A_i et le sous-ensemble d'individus $I(A_j)$ décrit par A_j. En effet, cette inclusion n'est généralement pas vérifiée dans les grandes bases de données où il est courant d'observer quelques rares individus possédant les attributs de A_i mais pas tous ceux de A_j sans que ne soit contestée la tendance générale à posséder A_j quand on a A_i. Si l'on instancie le programme initial de l'ECD [Frawley *et al.*, 1992] pour les règles d'association, il s'agit donc d'extraire des règles non triviales et potentiellement utiles pour l'utilisateur-décideur.

Pour ce faire, l'approche interactive que nous avons développée s'appuie d'une part sur quelques retours d'expériences [Bandhari, 1994], [Lehn, 2000], et d'autre part sur des principes cognitifs mis en évidence dans le cadre décisionnel [Montgomery, 1983], [Shanteau, 1988].

Tous soulignent la tendance de l'utilisateur à focaliser son attention sur un sous-ensemble de faible dimension de l'ensemble des données potentiellement intéressantes. Plus précisément, en décision, des algorithmes de sélection [Barthélemy et Mullet, 1992] s'appuient sur les deux principes suivants dont la validité semble confirmée en ECD :

- le principe de parcimonie. Du fait de ses capacités limitées en traitement de l'information et des limites de sa mémoire à court-terme, l'utilisateur-décideur ne manipule à chaque instant qu'une quantité restreinte de l'information ;
- le principe de fiabilité. Ce principe module le précédent. Pour justifier sa décision, un sous-ensemble d'informations suffisamment grand et significatif doit être considéré.

Appliquées à l'extraction interactive de règles d'association, ces caractéristiques nous ont conduit à construire un processus autour des deux contraintes suivantes. A chaque étape de la fouille, l'utilisateur peut se focaliser sur la représentation d'un sous-ensemble restreint de règles associées à des descriptions (conjonctions d'attributs) de sous-populations. Cette représentation peut évoluer dynamiquement selon les requêtes de l'utilisateur, ce dernier devant disposer à terme du processus d'extraction d'un ensemble de règles suffisamment pertinent pour justifier les conséquences induites à plus ou moins long terme par les informations découvertes.

3 Représentation de l'espace de navigation

Selon le cadre méthodologique dans lequel nous nous plaçons, il s'agit de définir un modèle formel de l'espace de recherche dans lequel l'utilisateur navigue qui permette, d'une part une représentation visuelle facilement accessible, et d'autre part une instanciation aisée des opérateurs d'interaction.

Dans le cas des règles d'association, les modèles basés sur des graphes semblent parmi les plus pertinents. Ils peuvent être effectivement utilisés à la fois comme des modèles théoriques puissants en tant qu'objets combinatoires, et comme des supports visuels efficaces qui permettent à un utilisateur d'accéder à des structures complexes sans se perdre dans un formalisme touffu [Buntine, 1996].

3.1 Un modèle de graphe

Dans la représentation la plus simple d'un ensemble Ω de règles d'association, les sommets du graphe sont les prémisses et les conclusions des règles, et les arcs représentent les implications (e.g. [Rostam, 1981]). Un tel graphe, permet sans légende, d'appréhender les classes d'équivalence de la relation sur Ω "avoir une prémisse ou une conclusion commune", et de déduire une partition sur Ω des composantes connexes du graphe. En revanche, hormis la transitivité quand elle existe, il permet plus difficilement de déduire d'autres relations. Cette limitation a conduit au développement d'autres modèles de graphes permettant notamment des inférences déductives sous forme de chemins [Horschka et Klösgen, 1991, Lehn, 2000].

Le modèle que nous avons privilégié est similaire à celui proposé par P. Horschka et W. Klögsen [Horschka et Klösgen, 1991] dans un autre cadre. Notons $A = \{a_1, a_2, \ldots, a_p\}$ l'ensemble des attributs (descripteurs) de la base qui sont ici binaires. Une règle d'association $\{a_i, a_j\} \rightarrow \{a_k\}$, notée pour simplifier $a_i \wedge a_j \rightarrow a_k$, est modélisée par un arc ayant pour origine le sommet $a_i \wedge a_j$ décrivant la prémisse de la règle, et pour extrémité le sommet $a_i \wedge a_j \wedge a_k$

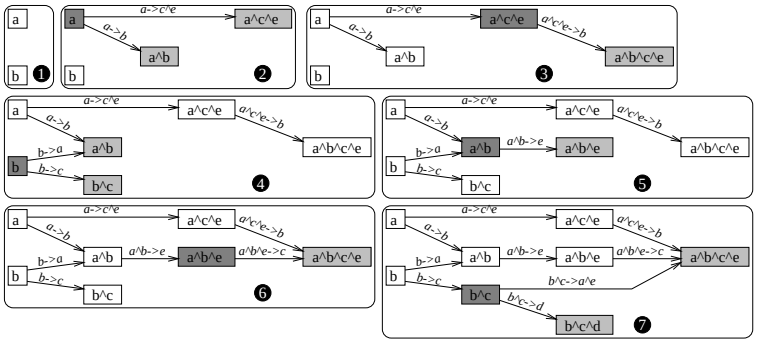

FIG. 1 – Séquence de requêtes.

décrivant tous les attributs intervenant dans la règle. La relation portée par les arcs est une relation d'inclusion sur les sous-ensembles d'attributs. Ainsi, ce graphe est par construction orienté sans circuit.

Du point de vue interprétatif, ce graphe permet facilement à l'utilisateur de se focaliser sur des conjonctions d'attributs décrivant des sous-populations. Dans le processus de fouille interactif, la sélection d'une conjonction c d'attributs entraîne, selon le désir de l'utilisateur, l'affichage de deux types de règles : les règles pertinentes plus spécifiques avec c en prémisse, et les règles pertinentes plus générales dont la prémisse est strictement inclue dans c. Un scénario de requêtes est illustré sur la figure 1. Partant d'attributs ayant un support suffisant selon un seuil fixé par l'utilisateur (étape 1), l'utilisateur sélectionne l'attribut a et une heuristique calcule les règles plus spécifiques pertinentes qui sont ici $a \rightarrow b$ et $a \rightarrow c \wedge e$ (étape 2). Puis, l'utilisateur sélectionne la description $a \wedge c \wedge e$ (étape 3) ; la règle plus spécifique est ici $a \wedge c \wedge e \rightarrow b$ (étape 4), et ainsi de suite.

D'un point de vue logique, ce modèle permet, dans le cadre du modèle classique "support-confiance" d'évaluation des règles [Agrawal *et al.*, 1993], de représenter des inférences déductives sous la forme de chemins dans le graphe. Considérons par exemple les règles $a_i \rightarrow a_j$ et $a_i \wedge a_j \rightarrow a_k$ représentées respectivement par les arcs $(a_i, a_i \wedge a_j)$ et $(a_i \wedge a_j, a_i \wedge a_j \wedge a_k)$. La confiance de la règle $a_i \rightarrow a_j \wedge a_k$, représenté par l'arc transitif $(a_i, a_i \wedge a_j \wedge a_k)$ est simplement le produit des confiances des règles du chemin $((a_i, a_i \wedge a_k \wedge a_k), (a_i \wedge a_j, a_i \wedge a_j \wedge a_k))$. Cependant, du fait des limites associées à l'utilisation seule de la confiance pour l'évaluation de la pertinence des règles, nous l'avons complété avec d'autres indices (voir [Gras *et al.*, 2004] pour plus de détails). Notons que l'extension de notre démarche à des attributs autres que binaires dépend essentiellement du choix de ces indices.

3.2 Représentation visuelle du modèle

La qualité de la représentation visuelle du graphe est décisive pour son appropriation par l'utilisateur. Pour préciser cette notion délicate, on retient généralement dans la communauté "Graph Drawing" quatre concepts de base [Battista *et al.*, 1999] :

– la convention de tracé qui spécifie les règles géométriques de lecture du tracé et qui dépend étroitement du domaine d'application ;

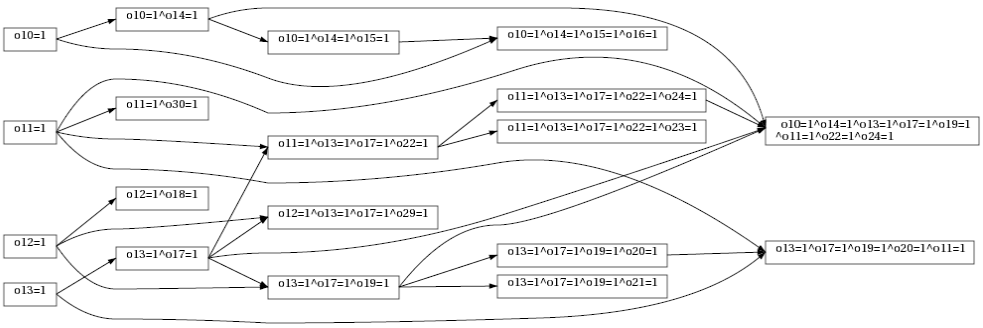

FIG. 2 – Tracé d'un graphe de règles en niveaux

- les contraintes de résolution du support et de l'œil humain qui imposent par exemple des écarts minima entre les sommets ;
- les critères esthétiques qui définissent les propriétés à satisfaire pour la lisibilité. Des travaux récents en psychologie cognitive ont montré que la réduction des croisement d'arêtes est le critère le plus important pour la lisibilité et la mémorisation [Purchase, 1998].
- les contraintes sémantiques associées à l'interprétation des composantes du graphe.

Ces caractéristiques nous ont amenés à privilégier pour le modèle de graphe décrit ci-dessus un tracé en niveaux (figure 2). Chaque niveau correspond à un degré de précision défini par le nombre d'attributs mis en jeu dans les prémisses des règles.

Lorsque l'on introduit les opérateurs d'interaction avec l'utilisateur, des sommets et des arcs peuvent être ajoutés ou supprimés, et la lecture des tracés dans un contexte dynamique conduit à de nouvelles contraintes. Le tracé présenté à chaque instant doit non seulement être intelligible selon les critères précédents, mais doit également préserver la carte mentale de l'utilisateur en limitant les perturbations apportées par le nouveau tracé par rapport aux précédents [Eades *et al.*, 1991]. La stabilité est une notion complexe qui dépend des caractéristiques géométriques et combinatoires du tracé, et aussi des facultés de perception et de mémorisation de l'utilisateur. Cependant, deux facteurs prédominants semblent se dégager [Battista *et al.*, 1999]. Les positions des sommets doivent bouger le moins possible ; ils servent de repères spatiaux et leur stabilité prime sur celle des arcs qui sont essentiellement utilisés pour découvrir des relations entre des sommets précédemment localisés. L'ordre relatif des composantes du graphe doit être conservé tant que cela est possible puisque l'utilisateur repère les composants les uns par rapport aux autres.

4 La problématique du tracé

Les différentes contraintes énoncées ci-dessus, nous ont conduit à poser notre problème de représentation visuelle dynamique comme un problème d'optimisation bi-objectif. Notons Π_t le tracé du graphe G_t obtenu à l'instant t. La contrainte statique retenue est la minimisation,

pour chaque t, du nombre de croisements $c(\Pi_t)$ sur Π_t. L'affectation des sommets à chaque niveau étant ici fixée par une contrainte sémantique (nombre d'attributs dans la conjonction), le problème revient à trouver une permutation pour chacun des niveaux qui minimise globalement les croisements. Au problème statique, il faut ajouter la contrainte dynamique qui consiste à maximiser une similarité $s(\Pi_{t-1}, \Pi_t)$ entre deux tracés consécutifs qui est une fonction des inversions dans les permutations associées à chaque niveau.

La minimisation du nombre de croisements dans ce cas peut sembler plus simple que le problème plus général de minimiser le nombre de croisement d'arêtes sur un graphe quelconque puisque le choix de coordonnées géométriques pour les sommets est ici remplacé par le choix d'un ordre des sommets sur chaque niveau. Le problème reste néanmoins *NP*-complet [Garey et Johnson, 1983].

De nombreuses heuristiques ont été développées pour ce problème du fait de son importance dans de multiples applications. Beaucoup suivent le principe du balayage successif des différents niveaux : les sommets de chaque niveau sont réordonnés de façon à réduire le nombre de croisements d'arcs. Des stratégies très variées ont été proposées pour le réordonnancement (*e.g.* [Laguna *et al.*, 1997] pour plus de détails). Les plus utilisées sont basées sur des méthodes de tris qui utilisent le nombre de croisements d'une façon proche des tris classiques et des heuristiques basées sur le principe selon lequel le nombre de croisements diminue si un sommet se trouve proche de la position moyenne de ses voisins sur les niveaux adjacents ([Sugiyama *et al.*, 1981]). Plus récemment différentes méthaheuristiques ont été développées pour ce problème : recherche tabou ([Laguna *et al.*, 1997]), GRASP ([Marti, 2001]), algorithme génétique hybridé ([Kuntz *et al.*, 2004]).

L'intégration de la contrainte dynamique conduit à un problème ouvert à notre connaissance. Partant des résultats obtenus pour le problème statique nous avons adapté un algorithme génétique hybridé par une recherche locale [Pinaud *et al.*, 2004]. L'idée de base consiste à relaxer la première condition $\min c(\Pi_t)$ et à rechercher parmi les meilleurs tracés de la population construite en t par l'algorithme génétique celui qui maximise la ressemblance par rapport au tracé retenu à l'étape précédente. En sus des performances des résultats expérimentaux obtenus, le développement actuel des AG pour les problèmes d'optimisation multi-objectif [Coello *et al.*, 2002] s'avère très prometteur pour ce type de problème.

5 Architecture du processus

L'architecture du processus de fouille de règles interactif qui a été implémenté dans le prototype Felix 0.42 (http ://193.52.110.95/oasis/EGC/logiciels/felix/index.html) s'articule autour de trois composantes fortement connectées et évoluant dynamiquement en fonction des requêtes de l'utilisateur décrites dans le paragraphe 3.1 (Figure 3) :

1. Une base de données qui contient, outre les données de l'étude, des tables de travail permettant d'optimiser le temps de réponse d'une requête [Lehn, 2000] ;
2. Une heuristique calculant les sous-ensembles de règles d'association pertinentes associés aux requêtes de l'utilisateur. Cette heuristique est une version locale de l'algorithme A Priori [Agrawal *et al.*, 1993]. Les mesures choisies pour la qualité des règles sont le support, la confiance et l'intensité d'implication. Les seuils de sélection peuvent être ajustés interactivement.

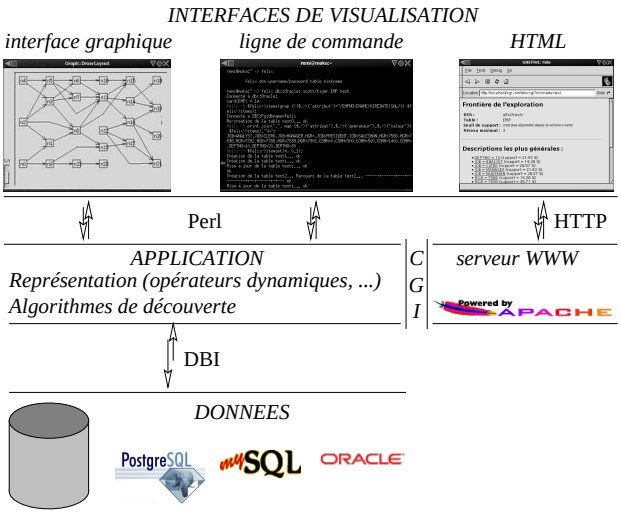

FIG. 3 – Architecture du processus de fouille interactif

Trois interfaces de visualisation : l'interface de visualisation dynamique des règles par un graphe sur laquelle nous nous sommes focalisées ici, une interface d'interrogation en ligne de commande, et une interface de parcours de la représentation sous forme de pages HTML générées dynamiquement.

6 Discussion

En nous référant aux travaux sur les systèmes anthropocentrés menés en Aide à la Décision, nous avons présenté ici un processus dynamique d'extraction de règles d'association centré sur l'utilisateur *via* une interface graphique interactive. Le prototype développé a été testé sur des données réelles en ressources humaines pour rechercher des associations entre des traits comportementaux dans un corpus de tests de personnalité, et en marketing pour rechercher des profils de clients. Le travail à venir est l'analyse de toutes les traces des utilisateurs en situation d'extraction.

En effet, ayant pris soin à chaque étape du processus de stocker les actions de la paire {utilisateur, support visuel} nous pouvons disposer de la trace "objective" du processus en situation réelle. Notons, qu'il est beaucoup plus délicat d'obtenir par verbalisation des informations fiables et exploitables sur les variables internes qui permettent d'apprécier les conditions subjectives dans lesquelles l'utilisateur effectue sa fouille. Ainsi, pour analyser les stratégies de fouille, nous pouvons nous placer uniquement sous une hypothèse behavioriste en nous appuyant sur les observables uniquement.

Par l'interactivité, l'association {utilisateur, support visuel} est devenue un système dynamique. L'évolution du comportement observable d'un utilisateur peut ainsi être modélisée par une séquence d'états de ce système. Et, par conséquent, l'analyse des comportements peut se

ramener à une analyse d'un ensemble de ses séquences. On trouve actuellement en ECD un intérêt croissant pour les problèmes liés au traitement des séquences symboliques dans l'analyse des usages de la Toile, en particulier pour l'analyse des parcours des utilisateurs en vue de la constitution d'une "identité virtuelle" [Srivastava *et al.*, 2000]. Comme dans le modèle que nous avons utilisé ici, l'environnement peut être dans ce cas modélisé par un graphe (hypertexte, réseaux de sites, ...). Et, bien souvent l'observation porte uniquement sur des chemins du graphe. Pourtant, s'agissant de systèmes interactifs, la réponse du système à l'action de l'utilisateur est également une donnée très informative, dès lors que l'on s'intéresse non pas tant aux réponses de l'utilisateur en elles-mêmes mais plutôt aux processus et états qui génèrent ces réponses -ce qui peut être qualifié de "comportementalisme cognitif" [Coppin, 1999]-.

Sur le plan plus spécifique de la visualisation, nous nous sommes focalisés ici sur un modèle de graphe ; cherchant à mettre en évidence une structuration d'un ensemble de relations, un tel modèle semblait un choix "naturel". Dans la lignée des travaux plus prospectifs en réalité virtuelle dans la communauté InfoViz, un travail est actuellement en cours sur une représentation visuelle basée sur une métaphore graphique [Blanchard *et al.*, 2003]. Le prototype développé reprend les opérateurs presentés ici mais propose une représentation tri-dimensionnelle fondée sur un paysage de boules qui peut s'activer selon la démarche célèbre de la recherche d'information "*Overview first, zoom and filter, then details -on demand"* [Shneiderman, 1996]. La comparaison de ces deux approches dans un même cadre expérimental pourra nous donner des préconisations précieuses sur l'appropriation des modèles visuels par les utilisateurs en ECD.

Références

[Agrawal *et al.*, 1993] R. Agrawal, T. Imielinsky, et A. Swami. Mining association rules between sets of items in large databases. In *Proc. of the ACM SIGMOD'93*, pages 207–216, 1993.

[Bandhari, 1994] I. Bandhari. Attribute focussing : machine-assisted knowledge discovery applied to software production process control. *Knowledge acquisition*, 6 :271–294, 1994.

[Barthélemy *et al.*, 2002] J.-P. Barthélemy, R. Bisdorf, et P. Lenca (editors). Special issue on human centered processes. *European J. of Operational Research*, 136(2), 2002.

[Barthélemy et Mullet, 1992] J.-P. Barthélemy et E. Mullet. A model of selection by aspects. *Acta Psychologica*, 79 :1–19, 1992.

[Battista *et al.*, 1999] G. Di Battista, P. Eades, R. Tamassia, et I.G. Tollis. *Graph drawing - Algorithms for the visualization of graphs*. Prentice Hall, 1999.

[Berry et Linoff, 1997] M. Berry et G. Linoff. *Data Mining - Techniques appliquées au marketing, à la vente et aux services clients.* InterEditions, 1997.

[Bertin, 1977] J. Bertin. *La graphique et le traitement graphique de l'information*. Flammarion - Réédition de l'Ecole des Hautes Etudes en Sciences Sociales, 1977.

[Bisdorf, 2002] R. Bisdorf. *Design and implementation of decision making checkers for industrial production scheduling, control and maintenance (Habilitation)*. PhD thesis, Centre Universitaire de Luxembourg, Luxembourg, 2002.

[Blanchard *et al.*, 2003] J. Blanchard, F. Guillet, et H. Briand. A user-driven and quality oriented visualization for mining association rules. In *Proc. of the* 3^{rd} *IEEE Int. Conf. on Data Mining (ICDM)*, pages 493–497. IEEE Computer Society Press, 2003.

[Brachman et Anand, 1996] R.J. Brachman et T. Anand. The process of knowledge discovery in databases : a human-centered approach. In U.M. Fayyad, G. Piatetsky-Shapiro, et P. Smyth, editors, *Advances in Knowledge Discovery and Data Mining*, pages 37–58. MIT Press, 1996.

[Buntine, 1996] W. Buntine. Graphical models for knowledge discovery. In U.M. Fayyad, G. Piatetsky-Shapiro, et P. Smyth, editors, *Advances in Knowledge Discovery and Data Mining*, pages 59–82. MIT Press, 1996.

[Cleveland, 1985] W.S. Cleveland. *The elements of graphing data*. Hobart Press, 1985.

[Coello *et al.*, 2002] C.A. Coello Coello, D.A. Van Veldhuizen, et G.B. Lamont. *Evolutionary Algorithm for solving multi-objective problems*. Kluwer Academic Publishers, 2002.

[Coppin, 1999] G. Coppin. *Pour une contribution à l'analyse anthropocentrée des systèmes d'action*. PhD thesis, Ecole des Hautes Etudes en Sciences Sociales, Paris, 1999.

[Deming, 1943] W.E. Deming. *Statistical Adjustment of Data*. Wiley (republished in 1964 by Dover), 1943.

[Eades *et al.*, 1991] P. Eades, W. Lai, K. Misue, et K. Sugiyama. Preserving the mental map of a diagram. In *Proc. of Compugraphics*, pages 24–33, 1991.

[Fayyad *et al.*, 2002] U. Fayyad, G.G. Grinstein, et A. Wierse. *Information visualization in data mining and knowledge discovery*. Morgan Kaufmann, 2002.

[Frawley *et al.*, 1992] W.J. Frawley, G. Piatetsky-Shapiro, et C.J. Matheus. Kowledge discovery in databases : an overview. *AI Magazine*, Fall :57–70, 1992.

[Garey et Johnson, 1983] M.R. Garey et D.S. Johnson. Crossing number is np-complete. *J. Algebraic Discrete Methods*, 4(3) :312–316, 1983.

[Goethals et van den Bussche, 2000] B. Goethals et J. van den Bussche. On supporting interactive rule mining. In *Proc. of the 2nd Int. Conf. on Data Warehousing and Knowledge Discovery*, pages 307–316. LNCS 1874 - Springer, 2000.

[Gras *et al.*, 2004] R. Gras, R. Couturier, J. Blanchard, H. Briand, P. Kuntz, et P. Peter. Quelques critères pour une mesure de qualité de règles d'association. *Revue des Nouvelles Technologies de l'Information - Numéro spécial Mesures de qualité pour la fouille de règles*, E1 :3–33, 2004.

[Grinstein, 1996] G. Grinstein. Harnessing the human in knowledge discovery. In *Proc. of the 2nd Int. Conf. on Knowledge Discovery and Data Mining*, pages 384–385. AAAI Press, 1996.

[Hand *et al.*, 2000] D.J. Hand, G. Blunt, M.G. Kelly, et N.M. Adams. Data mining for fun and profit. *Statistical Science*, 15 :111–131, 2000.

[Horschka et Klösgen, 1991] P. Horschka et W. Klösgen. A support system for interpreting statistical data. In G. Piatetsky-Shapiro et W.-J. Frawley, editors, *Knowledge Discovery in Databases*, pages 325–345. AAAI Press, 1991.

[Keim et Kriegel, 1996] D.A. Keim et H.-P. Kriegel. Visualization techniques for mining large databases : a comparison. *IEEE Trans. on Knowledge and Data Engineering*, 8(6) :923–938, 1996.

[Kuntz *et al.*, 2004] P. Kuntz, B. Pinaud, et R. Lehn. Elements for the description of fitness landscapes associated with local operators for layered drawings of directed graphs. In M.G.C. Resende et J.P. de Sousa, editors, *Metaheuristics : Computer Decision-Making*, volume 86 of *Applied optimization*, pages 405–420. Kluwer Academic Publishers, 2004.

[Laguna *et al.*, 1997] M. Laguna, R. Marti, et V. Valls. Arc crossing minimization in hierarchical design with tabu search. *Computers and Operations Research*, 24(12) :1175–1186, 1997.

[Lehn, 2000] R. Lehn. *Un système interactif de visualisation et de fouille de règles pour l'extraction des connaissances dans une base de données*. PhD thesis, Université de Nantes, France, 2000.

[Marti, 2001] R. Marti. Arc crossing minimization in graphs with grasp. *IEE Trans.*, 33(10) :913–919, 2001.

[Montgomery, 1983] H. Montgomery. Decision rules and the search for dominance structure : toward a process model of decision making. In P.C. Humphreys, O. Svenson, et A. Vary, editors, *Analyzing and aiding decision making*, pages 471–483. North Holland : Amsterdam, 1983.

[Newell et Simon, 1972] A. Newell et H.A. Simon. *Human Problem Solving*. Prentice Hall, 1972.

[Pinaud *et al.*, 2004] B. Pinaud, P. Kuntz, et R. Lehn. Dynamic graph drawing with a hybridized genetic algorithm. In I.C. Parmee, editor, *Automatic Computing in Design and Manufacture VI*, pages 365–375. Springer, 2004.

[Purchase, 1998] H.C. Purchase. Which aesthetic has the greatest effect on human understanding ? In *Proc. Symp. Graph Drawing' 95*, pages 248–261. LNCS, 1998.

[Rostam, 1981] H. Rostam. *Construction automatique et évaluation d'un graphe d'implication issu de données binaires dans le cadre de la didactique des mathématiques*. PhD thesis, Université de Rennes I, France, 1981.

[Roy et Bouyssou, 1993] B. Roy et D. Bouyssou. *Aide Multicritère à la Decision : Méthodes et Cas*. Paris : Economica, 1993.

[Shanteau, 1988] J. Shanteau. Psychological characteristics of expert decision-maker. *Acta Psychologica*, 91 :215–302, 1988.

[Shneiderman, 1996] B. Shneiderman. The eyes have it : A task by data type taxonomy in information visualization. In *IEEE Symp. on Visual Languages*, pages 336–342. IEEE Press, 1996.

[Spence, 2001] R. Spence. *Information Visualization*. Addison-Wesley, 2001.

[Srivastava *et al.*, 2000] J. Srivastava, R. Cooley, M. Deshpande, et P. Tan. Web usage mining : discovery and applications of usage patterns from web data. *SIGKDD Explorations*, 1(2) :12–23, 2000.

[Sugiyama *et al.*, 1981] K. Sugiyama, S. Tagawa, et M. Toda. Methods for visual understanding of hierarchical systems. *IEEE Trans. Syst. on Man and Cybernetics*, 11(2) :105–129, 1981.

[Tufte, 1983] E. Tufte. *The visual display of quantitative information*. Graphics Press, 1983.

[Tukey, 1977] J.W. Tukey. *Exploratory data analysis*. Addison-Wesley, 1977.

[Vincke, 1992] P. Vincke. *Multicriteria Decision-aid*. J. Wiley and Sons, 1992.

Summary

Based on hypothesis from anthropocentric systems, this paper presents an interactive environment for association rule mining : the user acts as a heuristics for driving the process in a complex search environment. To allow both an intelligible visual representation and an easy instanciation of the interactive tools, the chosen model is a layered digraph where the layers are associated with the cardinalities of the rule antecedent. This process has been implemented in a prototype software. The first obtained results on real life applications open new prospects for the behavioural analysis of a user during a mining process.