

Rapport INRIA 1994 — Programme 3
Représentation des connaissances

Projet REPCO

3 mai 1995

Projet REPCO

Représentation des connaissances

Localisation : *Rennes*

Mots-clés : analyse de données (1), apprentissage automatique (1), classification automatique (1), langage naturel (1), logique non classique (1), REPCO (1), tuteur intelligent (1).

Repeco est un projet commun Inria/CNRS (URA 227).

1 Composition de l'équipe

Responsable scientifique

Philippe Besnard, CR CNRS

Secrétaire

Evelyne Livache, SAR Inria

Personnel Inria

Jacques Nicolas, CR

Yves Moinard, CR

René Quiniou, CR

Personnel Ura 227

Catherine Belleannée, maître de conférences, université de
Rennes 1

Marie-Odile Cordier, professeur, université de Rennes 1

Israël-César Lerman, professeur, université de Rennes 1

Véronique Masson, maître de conférences, université de Rennes 1

Laurent Miclet, professeur associé à l'Enssat
Mohammed Ouali Allah, maître de conférences, université de
Rennes 1, IUT de Lannion
Dominique Py, maître de conférences, université de Rennes 1
Sophie Robin, maître de conférences, université de Rennes 1
Pascale Sébillot, maître de conférences, université de Rennes 1
Basavenepa Tallur, maître de conférences, université de
Rennes 1
Raoul Vorc'h, maître de conférences, université de Rennes 1

Chercheurs doctorants

Jean-Maurice Blin, allocataire MESR depuis le 1^{er} octobre 1994
Joaquim Da Costa, assistant à l'université de Porto
Frédéric Danna, allocataire Inria
Cécile Fabre, allocataire Inria
Jean-Yves Giordano, allocataire MESR
Robin Gras, bourse université depuis le 1^{er} décembre 1994
Stéphane Le Peutrec, allocataire MESR
Laurent Miralabe, allocataire MESR jusqu'au 31 octobre 1994
Roger Ngouenet, bourse CIES
Laurence Rozé, bourse CNRS
Olivier Siegel, allocataire MESR
Emmanuel Siou, bourse Inria Région
Sylvie Thiébaux, bourse BDI-Région

Chercheurs invités

Torsten Schaub, bourse HCM depuis le 1^{er} juin 1993

Collaborateurs extérieurs

François Carpentier, université de Bretagne Occidentale, Brest
Henri Leredde, maître de conférences, université Paris Nord
Philippe Peter, maître de conférences, Ireste, Nantes
Raymond Rolland, maître de conférences, université de Rennes 1

2 Présentation des objectifs

Le projet traite de l'approche logique et statistique des bases de connaissances, composante caractéristique des systèmes d'intelligence artificielle.

A partir du principe d'une représentation des connaissances par la logique, le projet s'attache à la conception et à la mise en œuvre de nouveaux schémas d'inférence, tant au travers de la logique classique (dans le cadre de l'apprentissage symbolique automatique et de la compréhension du langage naturel) qu'au travers de logiques non classiques (dans le cadre de la formalisation de divers types de raisonnements – notamment la recherche de diagnostics).

Cette activité à caractère davantage théorique sert de support à plusieurs applications du domaine de l'enseignement intelligemment assisté par ordinateur. Dans ce contexte, la modélisation de l'usager est naturellement un thème à la croisée des recherches menées dans le projet.

D'autre part, le projet a une activité de recherche en analyse des données dans une perspective de structuration de connaissances. Aussi, l'interaction et la complémentarité de l'apprentissage symbolique et de la classification déterminent un autre point de rencontre pour les études effectuées dans le projet. Cela est concrétisé dans une application portant sur les séquences génétiques.

3 Actions de recherche

3.1 Raisonnements et logiques non classiques

Participants: Philippe Besnard, Marie-Odile Cordier, Yves Moinard, Raymond Rolland, Torsten Schaub, Olivier Siegel

Le raisonnement sur des bases de connaissances un tant soit peu complexes est mal modélisé par la logique classique. Il repose toutefois sur des bases rationnelles et assez générales pour justifier l'élaboration d'un cadre formel complètement indépendant du domaine particulier concerné. Ces considérations ont abouti à l'introduction de logiques non classiques, qui restent très proches dans leur esprit et leur nature de la logique classique, mais qui s'en distinguent par le rejet de certaines propriétés, et donc aussi par l'accent mis sur d'autres propriétés qui étaient annexes dans la logique classique. Deux propriétés souvent rejetées sont la monotonie (en logique classique, si l'ensemble des données croît, l'ensemble des déductions possibles ne peut que croître) et la règle du "ex falso quod libet" (si les données contiennent une contradiction, on peut en déduire n'importe quoi).

Dans cette perspective, nous étudions quelques logiques non classiques parmi les mieux appropriées.

Au sein du groupe Léa Sombé, nous avons achevé une analyse générale des solutions actuellement proposées pour la révision d'une base de connaissances [1, 2, 8]. Dans la même ligne générale, nous avons poursuivi l'étude de l'opération de "mise à jour" au sein du projet inter-PRC "Gestion de l'incertain et de l'évolutif dans une base de connaissances" (voir la section suivante). Nous avons aussi fait un recensement et une comparaison des modèles logiques possibles pour le raisonnement à partir d'informations contradictoires. Nous avons finalement tenté de relier les problématiques de raisonnement incertain et d'informations incomplètes, contradictoires, ...

Dans le but de mieux cerner les propriétés importantes des logiques non classiques nous avons complété plusieurs études. D'une part, nous avons continué l'effort de généralisation de la notion de circonscription en terme de modèles préférentiels. Les modèles préférentiels, ou plus précisément ici, les modèles minimaux, constituent une méthode simple et naturelle pour choisir certains modèles d'une théorie logique classique : on ne garde que les modèles qui sont minimaux pour une certaine relation binaire définie sur la classe de toutes les interprétations du langage. La plupart des circonscriptions définies dans la littérature peuvent s'exprimer en terme de modèles minimaux, y compris les variantes les plus récentes (circonscription limitée, et circonscription emboîtée) qui sont apparues en réponse à des problèmes fondamentaux de représentation des connaissances. Un domaine actuellement très actif consiste à donner les propriétés générales de la notion d'inférence par modèles préférentiels. Nous avons obtenu la première caractérisation complète de cette inférence dans un cadre général qui couvre réellement toutes les circonscriptions [29].

D'autre part nous avons mené une étude des relations d'inférence non monotones dans un cadre très général, y compris lorsque la logique "de base" est elle-même non classique, intuitionniste par exemple. Nous avons alors exprimé non plus simplement des propriétés pertinentes des logiques non monotones, mais plus généralement les liens existant entre ce genre de propriété et les diverses propriétés que peut avoir la logique de base. Cela devrait permettre d'aider à répondre de façon convaincante et précise à la question de savoir de quelle logique de base il convient de partir, selon la classe de problèmes que l'on veut résoudre [14].

Nous avons aussi poursuivi l'étude d'une logique non monotone particulièrement intéressante, la logique des défauts. Il s'agit d'un des formalismes les plus faciles à appréhender, et les plus féconds. Nous avons fait porter nos efforts sur l'implémentation de systèmes de raisonnement basés sur la logique des défauts. Nous essayons de tirer parti de travaux existants en déduction automatique et de les adapter à la non-monotonie [10]. Nous le faisons en tenant, en plus, compte des problèmes de priorité, notamment de spécificité. Il s'agit d'une notion indispensable dès que l'on prétend traiter rationnellement des règles avec exceptions (ce qui est le cas pour la logique des défauts), et qui pose des problèmes avec tous les formalismes non monotones actuels. Ce problème apparemment simple n'a pas encore trouvé de solution générale satisfaisante. Nous avons fourni une solution qui intègre à la logique des défauts un mécanisme de choix qui demeure calculable dans les cas simples et qui fournit une réponse adaptée dans un cadre suffisamment général.

3.2 Prise en compte de l'incertain pour la modélisation et le suivi d'environnements évoluant dans le temps

Participants : Marie-Odile Cordier, Laurence Rozé, Sylvie Thiébaux

Les recherches effectuées peuvent se décrire selon les trois axes ci-dessous :

3.2.1 Mises à jour

L'opération consistant à mettre à jour une base de connaissances représentant l'état du monde au temps t pour prendre en compte les informations acquises sur son état au temps t' est une forme de révision appelée "mise à jour" (ou update) pour l'opposer à la "révision de croyances" (belief revision). Cette opération est fondamentale dès que l'on s'intéresse à des environnements (ou systèmes) évoluant dans le temps : elle est en particulier nécessaire pour prendre en compte les conséquences d'une action dans le cadre de la planification ou encore les observations faites régulièrement par des capteurs lors du suivi de systèmes dynamiques. Nous avons proposé avec Pierre Siégel (LIUP Marseille) d'exprimer de manière explicite la persistance des informations entre deux états par un ensemble de contraintes de transition. Les contraintes strictes (ou obligatoires) sont distinguées des contraintes préférentielles qui sont partiellement ordonnées. Nous avons montré que

l'opération de mise à jour peut alors être formalisée en utilisant la circonscription avec priorités. Ce travail étend les résultats de M. Winslett. Un travail en cours avec Jérôme Lang (Irit, Toulouse) consiste à étudier les rapports existant entre cette approche et les approches syntaxiques de maintien de la consistance dans les bases de connaissances.

3.2.2 Planification

La planification consiste en la génération automatique d'actions permettant de faire évoluer un environnement selon un objectif donné. Nous nous intéressons tout particulièrement au cas où l'environnement et les moyens d'actions sont décrits de manière incomplète, incertaine et/ou imprécise.

Le formalisme d'actions sur lequel nous nous appuyons est celui proposé par Brewka et Hertzberg et permet de prendre en compte de manière explicite l'incertitude existant sur la description de l'environnement et sur les effets des actions [44]. Les plans sont engendrés sous la forme d'automates d'états finis, ce qui permet de réagir en cours d'exécution à certains événements imprévus. Le planificateur est capable de respecter les contraintes de temps qui lui sont imposés grâce à l'utilisation d'algorithmes de type "anytime". Ceux-ci s'appuient sur une estimation de la qualité des plans calculée par des méthodes probabilistes.

La maquette de Pascale, le planificateur basé sur ces concepts, a été améliorée du point de vue de son efficacité d'exécution et a été dotée d'une interface graphique. Nous évaluons actuellement l'adéquation de Pascale à la génération réactive de plans de reprise de service, suite à un incident sur un réseau de distribution d'électricité (contrat EDF).

3.2.3 Diagnostic de systèmes évolutifs

Le diagnostic de systèmes de type "model-based" consiste à identifier l'état des composants d'un système à partir des observations et de son modèle de fonctionnement. Nous avons proposé une nouvelle vue du diagnostic model-based de systèmes évolutifs, dans laquelle le modèle de fonctionnement d'un système est décrit par l'ensemble des événements (changements internes, externes, actions de réparations et de test...) susceptibles de le faire évoluer [18]. Le diagnostic consiste alors, non pas en une caractérisation des états successifs des composants du système, mais en la détermination de la suite d'événements qui, s'ils

se sont produits, expliquent des observations successives. Nous nous sommes intéressés à l'adaptation de représentations et d'algorithmes issus du domaine de la planification à cette approche du diagnostic à base d'événements.

D'autre part, une collaboration existe avec le LAAS (Louise Travé) et l'Institut Galilée-Paris13 (Philippe Dague et François Lévy) dans le cadre d'un contrat CNET/France-Telecom. Nous travaillons plus particulièrement sur l'utilisation de techniques à base de modèles pour la supervision et le diagnostic de systèmes de transmission de données. L'analyse de ce problème a donné lieu à un premier rapport d'avancement (juillet 1994). Une modélisation du système étudié a été réalisée ; celle-ci complète une première ébauche de modèle proposée par C. Ungauer (CNET, Lannion). La validation de ce modèle est en cours.

Le travail commencé dans le cadre du projet inter PRC "Gestion de l'incertain et de l'évolutif" se poursuit. Nous nous intéressons en particulier au diagnostic de systèmes dynamiques en illustrant nos propositions sur l'exemple des taxis-robots choisi par ce groupe de travail. Un rapport est en cours de rédaction.

3.3 Traitement du langage naturel

Participants : Cécile Fabre, Pascale Sébillot

Nos travaux portent sur la détermination automatique de la sémantique des composés anglais et français. Il s'agit de prédire à partir de la forme d'un composé et de la sémantique des mots qui le composent quelle relation entretiennent entre eux les constituants. Cette interprétation automatique conduit à la production d'une représentation du sens du composé, représentation qui est intégrée à différents projets : production d'une paraphrase du composé dans un tuteur d'apprentissage de l'anglais, utilisation directe dans un projet d'indexation automatique (dans le cadre d'un contrat avec le CNET), etc. Nous avons montré que des modèles linguistiques performants sont disponibles pour l'interprétation sémantique des composés dotés d'un élément prédicatif, c'est-à-dire en général d'un verbe (*spoon feed*) ou d'un déverbal (*window-washer*, *extracteur à fumées*). Une analyse basée sur des théories linguistiques génératives et exprimées en termes de satisfaction de la structure argumentale du mot prédicatif peut être menée. Par exemple, la structure argumentale du prédicat *wash* indique que ce verbe nécessite un argu-

ment complément d'objet direct et un agent. Dans *window-washer*, la suffixation en *-er* fait jouer ce rôle agent au mot *washer* ; le mot *window* est donc obligatoirement le complément d'objet direct de *wash*. Le composé s'interprète donc comme un "laveur de carreaux". Nous nous intéressons actuellement aux composés *Nom Nom anglais* (*dog house*) *Nom à Nom français* (*couteau à fromage*) pour lesquels n'existe pas d'information prédicative apparente permettant de prévoir la relation entre les constituants. Pour déterminer automatiquement cette relation, nous étudions deux types d'analyses : l'un, représenté par les travaux de P. Downing, consiste à déterminer la relation en fonction de la classe sémantique du nom tête. L'autre vise à rétablir le prédicat sous-jacent en l'absence d'information morphologique : les travaux que T.W. Finin consacre aux composés, ainsi que ceux de J. Pustejovsky sur la sémantique lexicale en général, permettent de proposer des solutions dans ce sens. T.W. Finin définit une catégorie particulière de noms, les noms rôles (*role nominals*), qui sont associés typiquement à un verbe dont ilsinstancient un des rôles. Par exemple, *house* est associé typiquement à *live in*, *dwel* ou *occupy* et *couteau* à *couper*. Une fois cette association enregistrée dans le lexique, le mécanisme d'interprétation peut suivre les mêmes étapes que dans le cas des déverbaux (identification de la structure argumentale du verbe, vérification des contraintes de sélection des arguments, etc.) [21]. De plus, nous enrichissons notre modèle linguistique grâce au principe de co-compositionalité de J. Pustejovsky, principe qui traduit le fait qu'en sémantique lexicale, le produit de l'application d'une fonction est sensible à la fois à la fonction (ici, le prédicat impliqué par le nom tête du composé) *et* à son argument (le deuxième constituant).

3.4 Apprentissage symbolique automatique

Participants : Catherine Belleannée, Jacques Nicolas, Laurent Miclet, Raoul Vorc'h, Jean-Yves Giordano, Robin Gras, Laurent Miralabe

Le but de l'apprentissage automatique supervisé est de construire un ensemble de fonctions de discrimination. Ses entrées sont un ensemble de classes à discriminer dont on connaît la description de certains éléments (ensemble d'instances d'apprentissage) et un langage permettant de générer toutes les fonctions de discrimination possibles.

Les problèmes que nous étudions sont :

- trouver des couplages intéressants de méthodes issues de l'apprentissage et de l'analyse de données (approche “numérique symbolique”), ces deux disciplines partageant des préoccupations communes importantes ;
- rechercher *l'ensemble* des solutions d'un problème d'apprentissage, par opposition à une approche classique où l'on recherche *une* meilleure solution.

Nous décrivons plus précisément les actions portant sur le couplage numérique-symbolique et l'inférence grammaticale (i.e. lorsque les fonctions de discrimination sont des grammaires). Les aspects relevant plus particulièrement de la classification sont reportés dans la partie «analyse des données». Ce dernier travail a débuté dans le cadre d'actions sur l'analyse de séquences dans les génomes (projet GREG). L'analyse de séquences biologiques est la principale application que nous développons. Notre but à terme est de fournir au biologiste un outil puissant d'analyse de grandes séquences lui permettant d'être assisté dans l'exploration de celles-ci. Un outil de visualisation basé sur la manipulation de l'arbre des sous-mots d'une séquence a été développé cette année comme premier support de ces recherches. L'algorithme de base utilisé est celui de Chen et Seiferas. Des séquences de taille supérieure à 100 000 ont pu être traitées ainsi.

3.4.1 Couplages symboliques-numériques

Dans le cadre du projet inter PRC “méthodes symboliques-numériques de discrimination”, le premier point a fait l'objet d'une étude synthétique sur les méthodes existantes et l'apport respectif des opérations numériques/symboliques dans les systèmes hybrides de discrimination. Les deux conclusions principales de ce travail sont l'intérêt d'appliquer des opérations symboliques pour structurer l'espace de solutions, afin d'augmenter le pouvoir d'exploration des méthodes et l'intérêt d'appliquer des opérations numériques pour diriger la recherche dans cet espace, afin d'augmenter la sélectivité des méthodes.

Notre méthode consiste à utiliser les techniques d'analyse des données pour reformuler et simplifier le travail combinatoire des algorithmes d'apprentissage. En particulier, la classification permet de sélectionner et d'ordonner la présentation des instances d'apprentissage.

Cette étude a été validée sur un problème artificiel de reconnaissance de formes d'ondes initialement proposé par Breiman [34]. Conformément à nos buts, nous avons conservé la méthode d'apprentissage de l'espace des versions définie par Mitchell, où l'on recherche les antichaînes extrémales d'un ordre partiel. Ceci nous a permis de raffiner nos outils de gestion d'espaces de versions et de montrer la faisabilité de la production de solutions en un temps de l'ordre de la seconde pour un langage de vecteurs d'attributs binaires de taille supérieure à la centaine, avec des ensembles d'apprentissage de plusieurs centaines d'éléments (ce problème était jusqu'à présent considéré comme exponentiel en fonction du nombre d'attributs et non traitable pour un nombre d'attributs dépassant la dizaine).

3.4.2 Inférence grammaticale

Nous avons également étudié dans le cadre de l'espace des versions le problème de l'inférence grammaticale (les exemples et contre-exemples sont alors des mots et les classes des grammaires ou des automates de langages à déterminer). Étant donnée la complexité du problème, notre action s'est cependant focalisée sur la seule production de l'ensemble des solutions les plus générales (G-set ou BS, Boarder Set).

Quelle que soit l'application, les deux principales questions à résoudre dans le cadre de l'espace des versions concernent 1) la spécification d'un ordre de généralité et la comparaison effective de deux hypothèses selon cet ordre et 2) la généralisation ou la spécialisation d'une hypothèse pour prendre en compte le nouvel exemple ou contre-exemple.

Inférence de grammaires régulières à partir d'exemples et de contre-exemples.

Participants : Laurent Miclet

L'espace de recherche en inférence de grammaires régulières est un treillis construit à partir d'un automate canonique reconnaissant uniquement les exemples de mots du langage à inférer (automate le plus spécifique). Les éléments du treillis sont dérivés de cet automate par une fonction correspondant à la fusion des états de l'automate. Il existe deux automates canoniques principaux pour un même problème : l'automate canonique maximal et l'accepteur de l'arbre des préfixes. Sur le plan

théorique, notre travail a consisté à éclaircir les liens existant entre les deux treillis [28]. Un certain nombre de résultats ont été obtenus, en particulier la preuve que le second treillis est inclus strictement dans le premier et diverses propriétés sur l'ensemble BS des généralisations maximales.

Sur le plan des résultats pratiques, un calcul de BS a été proposé, travaillant essentiellement sur des structures de partition, permettant de produire des BS de plusieurs dizaines de milliers d'éléments. Ceci permet d'observer la répartition des éléments du BS dans le treillis. Nous avons également travaillé sur une version semi-itérative (sur les exemples positifs) de l'algorithme [27].

Du point de vue validation, un protocole expérimental a été mis au point en collaboration avec plusieurs laboratoires (CNET Lannion et Université de Valence) qui permettra de comparer plusieurs algorithmes sur une base commune.

Inférence de grammaires non régulières

Participants : Jean-Yves Giordano, Laurent Miralabe, Jacques Nicolas

L'exploration d'un ensemble de grammaires partiellement ordonné est un procédé classique en inférence grammaticale. L'éclairage apporté sur ces méthodes par l'approche espace des versions nous a amené à explorer deux voies possibles pour l'inférence de grammaires algébriques.

En premier lieu, nous examinons les possibilités offertes par une relation d'ordre peu utilisée en inférence grammaticale : l'inclusion structurelle (on compare les arbres de dérivation). Cette relation est calculable en temps polynômial (travaux de Hunt et Rosenkrantz) sur une forme normale : les grammaires inversibles, qui constituent alors l'ensemble des hypothèses. Afin de réduire la taille de l'espace de recherche, une contrainte est mise sur le nombre de non terminaux utilisés. Les opérateurs de spécialisation sont au nombre de deux : suppression d'une règle, et substitution sur la partie gauche d'une règle. Bien que l'exploration de l'espace reste incomplète, le nombre de grammaires produites est important, et des heuristiques doivent être mises en œuvre pour éviter l'explosion combinatoire [22].

En parallèle, nous avons développé un algorithme utilisant une autre relation d'ordre, la couverture de Reynolds, ainsi que l'opérateur de spécialisation associé. Il s'agit d'une méthode descendante, produisant les solutions par fission des non-terminaux d'une grammaire initialement

universelle. Pour les langages réguliers, les résultats obtenus sont comparables à ceux des méthodes existantes, tout en permettant d'espérer la production plus rapide des solutions que dans la méthode ascendante. Nous procédons à la phase de test pour les grammaires algébriques.

3.4.3 Symétries et tableaux analytiques

Participants : Catherine Belleannée, Raoul Vorc'h

La prise en compte de symétries dans l'énoncé d'un problème permet d'effectuer un élagage important dans l'arbre de recherche des solutions. Les méthodes proposées jusqu'à présent s'appliquent à des énoncés sous forme clausale. Nous cherchons à étendre une telle démarche à des formules quelconques. Après avoir caractérisé les types possibles d'élagage dans le cadre des preuves par tableaux analytiques, nous avons cette année travaillé plus particulièrement sur le problème de la détection de symétries pour des formules quelconques.

Nous avons tout d'abord élaboré une technique de comparaison efficace des formules modulo la commutativité, en définissant une forme canonique adaptée pour les formules. Deux formules sont syntaxiquement égales modulo la commutativité si et seulement si elles ont la même forme canonique.

D'autre part, nous avons montré que le problème de détection de symétries dans un ensemble de formules quelconques pouvait se ramener à un problème de détection de symétries dans un ensemble de listes d'atomes (ce qui permet en quelque sorte de se ramener au cas clausal). Cet ensemble de listes, construit en tenant compte de la structure des formules, est partitionné de façon à minimiser l'espace de recherche pour l'algorithme de détection.

Ainsi, on peut scinder le problème de détection de symétrie en deux questions indépendantes : d'une part construire un ensemble de listes d'atomes pertinents, d'autre part élaborer un algorithme de détection de symétries dans un ensemble de listes d'atomes.

Des résultats satisfaisants ont été obtenus pour la détection de symétries dans des théories classiques (problèmes des reines, des pigeons ...) [12], en appliquant l'algorithme de Sais et Benhamou (conçu initialement pour le traitement de clauses).

S'inspirant des approches incomplètes qui connaissent actuellement un vif succès pour le problème SAT (et plus généralement pour les problèmes NP-complets) nous travaillons actuellement à la conception d'un algorithme de détection de symétries procédant par réparation.

3.5 Enseignement Intelligemment Assisté par Ordinateur

Participants : François Carpentier, Frédéric Danna, Véronique Masson, Dominique Py, René Quiniou, Pascale Sébillot, Emmanuel Siou

L'EIAO a pour objectif la transmission des savoirs. Ceci nécessite trois types de connaissances : des connaissances sur le domaine enseigné, des connaissances sur l'élève, et des connaissances pédagogiques. Les connaissances sur l'élève sont habituellement regroupées dans un "modèle de l'élève". Les techniques de constitution et de mise à jour des modèles de l'élève peuvent être considérées comme un cas particulier des techniques employées pour les "modèles de l'utilisateur".

Nos travaux sur la modélisation de l'utilisateur concernent trois domaines :

- le diagnostic des déficits du langage. L'utilisateur est un adulte affecté de troubles du langage.
- les composés anglais. L'utilisateur est un adolescent ou un adulte qui apprend les mécanismes de composition en anglais.
- la démonstration en géométrie. L'utilisateur est un collégien qui apprend à construire des preuves.

3.5.1 Sarah : modélisation de l'utilisateur à l'aide de techniques d'apprentissage automatique

Participants : Véronique Masson, René Quiniou, Emmanuel Siou

Nous étudions l'apport de l'apprentissage automatique à la modélisation de l'utilisateur ou l'apprenant. Nous examinons particulièrement les techniques de la programmation logique inductive qui consiste à apprendre un prédicat logique (ou par extension, un programme logique) à partir de faits satisfaits ou niés par ce prédicat. Il est particulièrement intéressant de synthétiser le modèle de l'utilisateur sous forme d'un programme logique car ce modèle permet, d'une part, de simuler pratiquement directement le comportement de cet utilisateur et, d'autre part, de découvrir

les erreurs de ce programme par rapport à une spécification donnée, ce qui correspond, en quelque sorte, à effectuer le *diagnostic cognitif* de cet usager. Les erreurs découvertes de cette manière correspondent aux carences de l'usager, qui peuvent être par exemple, un manque de connaissance, une connaissance erronée, une mauvaise interprétation d'une connaissance correcte ou encore une généralisation abusive d'une connaissance correcte.

Les caractéristiques de l'application visées (Sarah, Système d'Aide à la Rééducation des ApHasiques, cf. ci-dessous) sont sévères : pas de contre-exemples et nécessité d'un apprentissage incrémental. Nous avons développé une procédure de production de clauses qui respecte ces contraintes. La procédure suit une stratégie de recherche bi-directionnelle et tire parti des déclarations de type des arguments des prédicats utilisables pour construire la définition du prédicat à apprendre [40]. Cette stratégie permet de repousser l'explosion combinatoire à un niveau généralement acceptable (clauses comportant jusqu'à six littéraux au lieu de clauses limitées à trois littéraux pour la stratégie uni-directionnelle).

Nous avons également commencé à étudier l'*invention de prédicat* dans ce même cadre. L'idée est de ne pas se limiter à l'ensemble des prédicats constituant la connaissance du domaine pour construire la définition du prédicat à apprendre mais de pouvoir *inventer* un prédicat intermédiaire qui représente un concept intéressant dans le cadre de l'application. La définition de ce prédicat intermédiaire est également construite au cours de l'apprentissage du prédicat principal. Nous nous sommes limités pour l'instant à l'invention de prédicats non récursifs.

Ces idées sont appliquées au système Sarah dont le but est de diagnostiquer les troubles et déficits des patients aphasiques au moyen d'exercices pertinents, générés de façon automatique, puis d'aider le patient à compenser ses déficits en lui faisant prendre conscience de ses erreurs [7]. Ce travail est effectué en étroite collaboration avec H. Guyard de l'UER Langage, université de Rennes 2.

3.5.2 Projet Mentoniez

Participants : Dominique Py, François Carpentier

Le projet Mentoniez vise à l'élaboration d'outils pour l'aide à la résolution de problèmes, sur le domaine de la géométrie euclidienne plane, en classe de quatrième et troisième. Il est développé en collaboration avec le

groupe de recherches CNRS “Didactique et acquisition des connaissances” depuis 1985.

Nous avons étudié un cas particulier d’activité géométrique, qui consiste pour l’élève à retrouver l’élément manquant dans un pas de démonstration. Des tests réalisés en collège ont permis de recueillir un vaste échantillon de réponses. L’analyse de ces réponses a mis en évidence le fait que l’activité de l’élève peut être modélisée comme un raisonnement analogique, soumis à des contraintes syntaxiques et sémantiques. Un programme de simulation du comportement des élèves a été réalisé sur ce principe. Il exploite une base de connaissances comprenant des schémas de théorèmes et des règles d’instanciation, qui sont combinées afin de produire un ensemble de réponses (correctes et incorrectes) à la question posée. La plausibilité des mécanismes utilisés est attestée par le fait que, pour une même question, les réponses fournies par le programme sont très proches des réponses obtenues auprès des élèves [36].

Nous travaillons actuellement à l’intégration de ce module dans un système de diagnostic. Il s’agit, étant donné un couple question-réponse, de déterminer le raisonnement qu’a effectué l’élève pour produire cette réponse. Le diagnostic est ici formalisé comme un processus de construction d’hypothèses, qui cherche à générer une explication au phénomène observé.

3.5.3 Compounds : un système d’enseignement des composés anglais

Participants : Frédéric Danna, Pascale Sébillot

Le phénomène de composition est très productif en Anglais ; or, des études menées auprès d’étudiants francophones de la langue anglaise ont montré les énormes difficultés de ces derniers à maîtriser ce phénomène. L’élaboration du système Compounds se fait en collaboration avec P. Boucher, professeur de Linguistique anglaise à l’université de Nantes.

Le module expert de Compounds est une intégration des résultats de nos travaux sur la sémantique des composés anglais à partir de théories linguistiques génératives. Il comprend un ensemble de règles de “bonne construction” des composés permettant de produire le composé correspondant à une définition et, réciproquement, de trouver la

définition correspondant à un composé donné. Nous travaillons actuellement à l'élaboration du modèle de l'élève au sein de ce tuteur. Outre les types d'erreurs traditionnels, tels que la méconnaissance d'un phénomène, l'utilisation de règles incorrectes ou l'utilisation incorrecte de règles correctes, des problèmes nouveaux se greffent dans le cadre de l'EIAO des langues, tels que la nécessité de modéliser l'influence de la langue maternelle. Une étude de tests réalisés auprès d'étudiants nous a permis de classifier les erreurs commises en fonction des notions fondamentales de la composition anglaise : rôle de la tête, connaissance de la structure argumentale des prédicats, rôle des affixes, etc. À partir des réponses fournies par chaque élève à un exercice, nous produisons une formule regroupant les connaissances correctes et erronées de ces principes. Nous travaillons d'une part à la synthèse de ces formules dans le but de constituer le modèle et d'autre part à la modélisation en logique de l'évolution de ce modèle.

3.6 Gide

Participants : Sophie Robin, Stéphane Le Peutrec

Dans le cadre du projet Gide qui visait la conception d'un système destiné à assister un médecin épiléptologue dans la manipulation des dossiers médicaux de ses patients, nous avons mis en évidence le besoin d'intégrer un certain nombre de schémas de raisonnement que nous avons regroupés sous le terme d'abstraction, ignorant ainsi la distinction généralement faite entre généralisation et abstraction.

Ce système, qui doit notamment être capable de restituer au médecin les informations significatives d'un dossier dans un contexte d'utilisation donné (début de consultation, prescription thérapeutique, ...), doit donc disposer d'une faculté d'abstraction qui lui permette, d'une part, d'ignorer certaines informations en marge des préoccupations de l'utilisateur, d'autre part, de condenser les informations significatives avec un niveau de détail approprié à la situation. Cette première étude a permis de dégager un certain nombre de mécanismes suffisamment généraux pour sortir du cadre de Gide, nous conduisant à débiter une étude plus générale sur l'abstraction.

Les travaux menés sur l'abstraction dans le cadre de la démonstration de théorèmes notamment, s'attaquent de front à deux difficultés majeures qui sont d'une part, assurer que les déductions obtenues dans la base

abstraite restent valides pour la connaissance initiale, d'autre part, déterminer les abstractions les plus judicieuses pour un problème donné. Notre démarche s'inscrit davantage dans ce second aspect des choses.

Si on se tourne vers les sciences cognitives, tous les travaux soulignent l'aspect fortement contextuel de l'abstraction et le lien existant entre l'abstraction et un modèle de structuration de la connaissance en niveaux. Ceci nous conduit naturellement vers les représentations centrées objet et plus particulièrement vers des travaux sur les points de vue et les perspectives.

3.7 Analyse des données

Nous travaillons sur les points suivants :

- classification de séquences génétiques ;
- arbres de décision ;
- rôle des algorithmes génétiques dans la représentation euclidienne d'une matrice de dissimilarité AVL (analyse de la vraisemblance des liens) ;
- analyse ensembliste et statistique du problème de la satisfiabilité ;
- contribution à Modulad

3.7.1 Classification de séquences génétiques

Participants : Israël César Lerman, Jacques Nicolas, Philippe Peter, Basavenepa Tallur

Nos travaux se sont activement poursuivis sur la classification de séquences protéiques qu'elles soient préalablement alignées ou non. Dans le second cas, chaque séquence est formellement un mot empruntant ses lettres dans un alphabet de taille 20, représentant les vingt acides aminés. Alors que dans le premier cas, où les mots de même longueur sont calés les uns par rapport aux autres, l'alphabet comprend une vingt-et-unième lettre, représentant une délétion. L'avancée la plus significative concerne la conception d'une matrice de similarité de type AVL (dans sa version "quantité d'information") sur l'alphabet étendu. Cela a nécessité une fine analyse comparée des méthodes de construction de deux familles de matrices d'association entre acides aminés. La première est due à Dayhoff et se trouve très utilisée dans le milieu bioinformatique ;

et la seconde, récente, est due à Hénikoff (1992). En partant de familles préalignées à partir de considérations définies par la structure protéique, d'excellents résultats ont été obtenus. Très précisément, il s'agit du cas du Cytochrome C ; et de celui, beaucoup plus difficile, des synthétases t-RNa. Pour la classification adoptée (AVL (0.5)), nous représentons un même site, par un attribut "graphe valué complet".

Pour la comparaison de séquences non alignées et de longueurs différentes, on considère une approche de type "fenêtre significative". On prend une longueur fixe pour cette fenêtre qui glissera le long des deux séquences à comparer, pour sélectionner au moyen d'une recherche en "faisceau" l'ensemble des couples de fenêtres significatives par rapport à une hypothèse d'indépendance. Ces derniers vont pourvoir la similarité entre séquences, à partir de la matrice de Dayhoff. La classification AVL de deux familles (cytochromes et globines) a, dans chacun des cas, été très proche de la phylogénie connue de l'évolution. Des résultats, également tous en accord avec les connaissances biologiques, ont été obtenus sur une famille de synthétases.

Soulignons que l'ensemble de ces travaux est mené en étroite collaboration avec J. Lebbe du projet Clorec de l'Inria et avec J.L. Risler du centre de génétique moléculaire de Gif sur Yvette.

3.7.2 Arbres de décision

Participants : Israël-César Lerman, Joaquim Da Costa

La technique des arbres de décision cristallise aussi bien les efforts méthodologiques en analyse des données qu'en apprentissage. Leur construction est usuellement basée sur un coefficient d'association entre attributs qualitatifs (exemple : "entropié", X2, Gini). Les coefficients pratiquement utilisés ne tiennent aucun compte d'une sémantique qui peut exister de façon sous jacente à l'ensemble des valeurs (on dit encore *modalités* ou *catégories*) des variables à comparer. Mais c'est le contraire de ce que nous faisons pour la classification de variables qualitatives. Et, il s'agit de transférer cette prise en compte pour la formation de l'arbre de décision. Cette approche nous permet d'introduire dans le domaine, la nouvelle famille, issue d'AVL, de coefficients d'association entre variables relationnelles. De façon conjointe, nous nous intéressons à la réduction de la complexité de recherche dans un arbre de décision.

Le programme Arcade (arbre de classification et de décision) mis au point, mais qui reste encore d'un niveau expérimental, reprend les différents aspects du programme bien connu Cart ; mais, en y introduisant nos préoccupations.

Sur des données très difficiles, l'expérimentation (très avancée) doit bientôt aboutir.

3.7.3 Les algorithmes génétiques dans la représentation euclidienne d'une matrice de dissimilarité

Participants : Israël-César Lerman, Roger Ngouenet

L'approche dite MDS (*multi dimensional scaling*) se propose, à partir d'une matrice de dissimilarités sur un ensemble E d'unités de données (objets, concepts ou variables de description), d'en fournir une représentation euclidienne dans un espace géométrique de faible dimension (2 ou 3). Dans ces conditions, les distances inter-points doivent refléter au mieux les dissimilarités entre les unités de données. Dans notre cas, la notion de dissimilarité sera directement associée à la quantité d'information que suppose l'indice probabiliste pris dans la famille des coefficients AVL.

Les problèmes de différentiabilité posés par les algorithmes utilisés dans la pratique (gradient et sous gradient) et qui, de toute façon ne réalisent que des optima locaux, nous ont conduit à introduire une nouvelle classe d'algorithmes stochastiques : les algorithmes génétiques. Des résultats très probants ont été obtenus sur des données de taille modérée [33]. Cependant le grand problème de ces algorithmes, est l'efficacité informatique. Ainsi, la direction actuelle des travaux portent sur la parallélisation des algorithmes séquentiels qui ont été développés et leur implantation sur la machine Paragon. Enfin, notre préoccupation porte également sur la convergence de ce type d'algorithmes, dans le cadre de notre problématique.

3.7.4 Analyse ensembliste et statistique du problème de la satisfiabilité

Participants : Israël-César Lerman

Le problème de la satisfiabilité d'une conjonction de clauses - dit problème SAT - est une expression très représentative de la classe des

problèmes NP-complets. Nous avons proposé une nouvelle représentation géométrico-logique du problème qui permet une vision synthétique des algorithmes déjà proposés. D'une part Cela peut permettre, d'une part, des améliorations notables et d'autre part, de proposer de nouvelles approches. Une première expérience, à la faveur d'un stage de fin de maîtrise, nous a permis de nous en rendre compte, à partir d'un algorithme de résolution complète. Le système de clauses est engendré de façon aléatoire, conformément à une hypothèse d'indépendance statistique. Le nombre de variables pris est de l'ordre de cent à deux cents et le nombre de clauses par rapport au nombre de variables, respecte la proportion considérée difficile de 4,25. La comparaison a été effectuée relativement à l'algorithme stochastique de résolution incomplète G-SAT, considéré dans sa première version (IJCAI 1991). Le comportement de ce dernier s'est avéré très difficile à évaluer par rapport à notre représentation.

3.7.5 Contribution à Modulad

Participants : Henri Leredde, Mohammed Ouali Allah, Philippe Peter

La contribution du groupe s'est maintenue vive dans le cadre du club Modulad. Un des moteurs principaux de cette activité est le programme Chavl (classification hiérarchique par analyse de la vraisemblance des liens). Une large famille de structures du tableau de données se trouve concernée. Il peut s'agir de la classification de l'ensemble des lignes, étiqueté par un ensemble d'objets ou de concepts. Il peut s'agir de la classification de l'ensemble des colonnes étiqueté par un ensemble d'attributs ; l'ensemble des valeurs d'un même attribut étant muni d'une sémantique et donc d'une structure plus ou moins complexe.

4 Actions industrielles

M.-O. Cordier participe au suivi du projet Sachem : «Système d'aide à la conduite de hauts-fourneaux» par Sollac.

Contrat EDF sur la "Génération réactive de plans de reprise de service sur les réseaux de distribution EDF" (M.-O. Cordier).

Convention CNET en coopération avec le LAAS et le LIPN sur le thème «Modélisation, diagnostic et supervision de réseaux de télécommunication» par M.-O. Cordier.

Convention CNET sur le thème «Calculabilité de la sémantique des composés anglais et français ; adaptation à l'indexation automatique» participation de P. Sébillot.

5 Actions nationales et internationales

Participation au projet Esprit Basic Research Actions Drums-2 (*Defeasible Reasoning and Uncertainty Management Systems*) consacré à la conception d'un système d'inférence développant des raisonnements tant révisables qu'approchés. Les partenaires sont les universités d'Amsterdam, d'Anvers, de Bruxelles, de Grenade, Imperial College de Londres, Queen Mary College de Londres, DFKI Sarrebruck, CEAB Blanes Espagne, LIUP Marseille, LRI Orsay, LSI Toulouse.

Participation au programme européen Biomed 1, action Escape (*European Standardized Computerized Assessment Procedure for the Evaluation and Rehabilitation of Brain Damaged Patients*).

Participation à 3 projets inter-PRC : «Classes polynomiales», «Méthodes numériques-symboliques» et «Gestion de l'incertain et l'évolutif dans une base de connaissances».

Participation au PRC-Greco «Intelligence artificielle».

Participation au GRECO «Didactique et acquisition des connaissances scientifiques».

Participation à l'action du GIP Greg «Informatique et génomes».

Ph. Besnard est membre du comité de rédaction de RIA, TSI et *Applied Non-Classical Logics*.

Ph. Besnard a fait partie du comité de programme de KR-94.

M.-O. Cordier est membre du bureau du PRC-IA, rédactrice en chef de RIA (*Revue d'intelligence artificielle*) et membre du comité de rédaction de AAI (*Journal of Applied Artificial Intelligence*), vice-présidente du congrès RFIA-94.

I.-C. Lerman est membre du conseil d'administration de la société Francophone de classification et du conseil de la Fédération internationale des sociétés de classification, éditeur associé de la revue *RO-Operations Research*, membre des comités de rédaction des revues suivantes : *Applied Stochastic Models and Data Analysis* (John Wiley & Sons, editors) ; *Mathématique, informatique & sciences humaines* (édité par le centre

d'Analyse et de Mathématique sociales) ; *Automatique Productique Informatique Industrielle APII* (Afcet, éditions Hermes).

Participation de I.-C. Lerman aux comités scientifique des manifestations suivantes : *Secondes rencontres de la Société Francophone de Classification*, Tours 12-13 septembre 1994, *II Jornadas de Classificação e Análise de Dados - JOCLAD 94*, Lisbonne 12-14 décembre 1994, *Méthodes d'analyses statistiques multidimensionnelles en didactique des mathématiques*, Caen 27-29 janvier 1995, *Conférence internationale sur l'analyse de données ordinales et symboliques*, Paris 20-23 juin 1995.

Participation au club Modulad (H. Leredde, I.-C. Lerman et Ph. Peter).

6 Diffusion des résultats

6.1 Colloques et congrès

Pour les conférences avec actes se reporter à la bibliographie du projet.

Ph. Besnard a été invité au *Workshop on Mathematical and Statistical Methods in Artificial Intelligence*, Udine, Italie, septembre 1994.

6.2 Séminaires, écoles et tutoriels

Ph. Besnard et T. Schaub ont donné un cours sur le raisonnement non-monotone à *6th European Summer School on Logic, Language, and Information*, Copenhague, août 1994.

D. Py : “Mentoniezh, un logiciel d'aide à la démonstration en géométrie” (journées nationales de l'APMEP, Futuroscope de Poitiers, octobre 1993), “La prise en compte de l'élève dans les tuteurs intelligents” (séminaire Irmarr, Rennes, 12 novembre 1993), “La prise en compte de l'élève dans un tutoriel de géométrie” (séminaire National de Didactique, Paris, 15 janvier 1994) et “Les erreurs en géométrie, une étude de cas” (séminaire DidaTech, Imag, Grenoble, 13 avril 1994).

6.3 Actions d'enseignement

Option de la 3^e année d'informatique, Enssat Lannion : *théorie de l'évidence* (Ph. Besnard).

Option du DEA d'informatique : *logiques d'ordres supérieurs : types et automates* (Ph. Besnard).

Option du DEA d'informatique, du DESS-ISA et 5^e année Insa : *représentation des connaissances* (Y. Moinard, R. Quiniou).

Option du DEA d'informatique : *intelligence artificielle et diagnostic* (M.-O. Cordier, S. Robin).

Option du DEA d'informatique, du DESS-ISA et 5^e année Insa : *analyse des données et apprentissage* (I.-C. Lerman, J. Nicolas).

Cours DIIC 3 Images numériques : *approche statistique de la reconnaissance des formes*, (I.-C. Lerman)

Cours en DESS Mathématiques appliquées et au magistère Mathématiques et Informatique de l'université de Rennes 1 : *analyse des données* (I.C. Lerman).

Cours DESS images numériques : *intelligence artificielle* (R. Quiniou).

7 Publications

Articles et chapitres de livre

- [1] P. BESNARD, E. LAENENS, «A Knowledge Representation Perspective: Logics for Paraconsistent Reasoning», *International Journal of Intelligence Systems. Special Issue: Revision and Updating in Knowledge Databases. 9*, 1, 1994, p. 153–168.
- [2] M.-O. CORDIER, P. SIEGEL, «A Temporal Revision Model for Reasoning About World Change», *International Journal of Intelligence Systems. Special Issue: Revision and Updating in Knowledge Databases. 9*, 1, 1994, p. 131–142.
- [3] F. DANNA, P. SÉBILLOT, P. BOUCHER, «COMPOUNDS: an Intelligent Tutoring System for Learning to Use Compounds in English», *Computer Assisted Language Learning 6*, 3, décembre 1993, p. 249–272.
- [4] J. DELGRANDE, W. JACKSON, T. SCHAUB, «Alternative Approaches to Default Logic», *Artificial Intelligence Journal 70*, 1-2, octobre 1994, p. 167–237.
- [5] J. HERTZBERG, S. THIÉBAUX, «Turning an action formalism into a planner—a case study», *Journal of Logic and Computation (special issue on Actions and Processes) 4*, 5, octobre 1994, version étendue de [23].
- [6] I.-C. LERMAN, «Coefficients d'association et/ou critères de comparaison entre deux structures combinatoires», *Revue Francophone de Classification*, juillet 1994, p. 8–14.

- [7] V. MASSON, R. QUINIOU, E. SIOU, H. GUYARD, «Expert knowledge for aphasia assessment and rehabilitation», in : *Applied Neuropsychology*, B. Wilson (éd.), à paraître 1994.
- [8] Y. MOINARD, «Notes on Change and Nonmonotony», *International Journal of Intelligence Systems. Special Issue: Revision and Updating in Knowledge Databases.* 9, 1, 1994, p. 101–118.
- [9] D. PY, «Reconnaissance de plan pour la modélisation de l'élève : le projet Mentoniez», in : *Didactique et Intelligence Artificielle*, N. Balacheff et M. Vivet (éd.), La Pensée Sauvage, Grenoble, 1994.
- [10] T. SCHAUB, «A New Methodology for Query-Answering in Default Logics via Structure-Oriented Theorem Proving», *Journal of Automated Reasoning*, à paraître 1995.
- [11] P. SÉBILLOT, P. BOUCHER, «Interprétation et génération automatique de noms composés anglais à l'aide de formes logiques», *Traitement Automatique des Langues, numéro spécial traitement de la composition nominale 34*, 2, 1993, p. 89–104.

Communications à des congrès, colloques, etc.

- [12] C. BELLEANNÉE, R. VORC'H, «A linear proof of the pigeon formulae using symmetry», in : *Proc. of the 3rd Workshop on Theorem-proving with Analytic Tableaux and Related Methods*, Abingdon, GB, mai 1994.
- [13] P. BESNARD, J. LANG, «Possibility and Necessity Functions over Non-Classical Logics», in : *Proc. of the 10th Conference on Uncertainty in Artificial Intelligence*, p. 69–76, Seattle WA, juillet 1994.
- [14] P. BESNARD, Y. MOINARD, «Implication connectives for logic with right weakening», in : *IPMU'94, Proc. of the 5th Int. Conf. on Information Processing and Management of Uncertainty in Knowledge-Based Systems*, p. 694–699, Paris, 1994.
- [15] P. BESNARD, Y.-H. TAN, «Non-Monotonic Reasoning with Contextual Logic», in : *Dutch-German Workshop on Non-Monotonic Reasoning*, Aix-la-Chapelle, décembre 1993.
- [16] S. BRÜNING, T. SCHAUB, «Using Classical Theorem-Proving Methods for Approximate Reasoning», in : *IPMU'94, Proc. of the 5th Int. Conference on Information Processing and Management of Uncertainty*, Paris, juillet 1994.
- [17] M.-O. CORDIER, P. SIÉGEL, «Prioritized transitions for updates», in : *18th German Conference On Artificial Intelligence (KI'94)*, septembre 1994.

- [18] M.-O. CORDIER, S. THIÉBAUX, «Event-based diagnosis for evolutive systems», in : *5th International Workshop on Principles of Diagnosis (DX'94)*, octobre 1994.
- [19] J. DELGRANDE, T. SCHAUB, «A General Approach to Specificity in Default Reasoning», in : *Proc. of the 4th Conference on the Principles of Knowledge Representation and Reasoning*, p. 146–157, Bonn, mai 1994.
- [20] J. DELGRANDE, T. SCHAUB, «Incorporating Specificity into Circumscriptive Theories», in : *Proc. of the 8th German Conference on Artificial Intelligence*, 1994.
- [21] C. FABRE, P. SÉBILLOT, «Interprétation sémantique des composés nominaux français», in : *Workshop on Compound Nouns : Multilingual Aspects of Nominal Composition*, Genève, décembre 1994.
- [22] J.-Y. GIORDANO, «Inference of Context-free Grammars by Enumeration : Structural Containment as an Ordering Bias», in : *2nd International Colloquium on Grammatical Inference*, 1994.
- [23] J. HERTZBERG, S. THIÉBAUX, «Turning an action formalism into a planner—essentials of a case study», in : *Proc. of the 8th International Symposium On Methodologies for Intelligent Systems (ISMIS-94)*, LNAI, Springer Verlag, octobre 1994.
- [24] I.-C. LERMAN, R. NGOUENET, «Un rôle pour les algorithmes génétiques dans la représentation euclidienne d'une matrice de dissimilarité», in : *XXVIèmes journées francophones de statistiques*, p. 426–429, Neufchâtel, Suisse, mai 1994.
- [25] I.-C. LERMAN, P. PETER, «Une matrice AVL pour la comparaison par paires des acides aminés», in : *Secondes journées de la société francophone de classification*, Tours, septembre 1994.
- [26] I.-C. LERMAN, «Coefficients d'association et/ou critères de comparaison entre deux structures combinatoires», in : *Journées inter-PRC Méthodes de discrimination symboliques-numériques*, Saint Malo, novembre 1993.
- [27] L. MICLET, C. DE GENTILE, «Inférence régulière à partir d'exemples et de contre-exemples : deux algorithmes optimaux (BIG et RIG) et une version heuristique (BRIG)», in : *JFA'94*, Strasbourg, mars 1994.
- [28] L. MICLET, P. DUPONT, E. VIAL, «What is the search space of the regular inference? », in : *ICGI'94*, Alicante, septembre 1994.
- [29] Y. MOINARD, R. ROLLAND, «Around a Powerful Property of Circumscriptions», in : *JELIA '94*, in *LNCS No 838*, Springer-Verlag, p. 34–49, York, GB, septembre 1994.
- [30] Y. MOINARD, R. ROLLAND, «Preferential entailments for circumscriptions», in : *KR'94*, Morgan Kaufmann, p. 461–472, Bonn, mai 1994.

- [31] Y. MOINARD, «Reasoning by Cases without Contraposition in Default Logic», *in: ECAI*, A. Cohn (éd.), Wiley, p. 381–385, Amsterdam, août 1994.
- [32] G. NEUGEBAUER, T. SCHAUB, «A Pool-Based Connection Calculus», *in: Proc. of the 3rd Turkish Symposium on Artificial Intelligence and Neural Networks*, 1994.
- [33] R. NGOUENET, «Représentation des objets symboliques par la MDS», *in: IVèmes Journées symbolique-numériques*, Paris, mars 1994.
- [34] J. NICOLAS, L. MIRALABÉ, B. TALLUR, «Classifying instances before generalizing», *in: Workshop Machine Learning and Statistics, ECML'94*, Catania, 1994.
- [35] J. NICOLAS, B. TALLUR, «Classification des séquences biologiques non alignées», *in: Secondes journées de la société francophone de classification*, Tours, septembre 1994.
- [36] D. PY, «Context-based diagnosis of students' errors», *in: Actes de CALISCE'94*, Paris, août 1994.
- [37] A. ROTHSCHILD, T. SCHAUB, «A Method for Query-Answering in Default Logic», *in: Proc. of the 7th Symposium on Artificial Intelligence*, Mexico, octobre 1994.
- [38] T. SCHAUB, «Computing Queries from Prioritized Default Theories», *in: Proc. of the 8th Int. Symposium on Methodologies for Intelligent Systems, LNAI*, Springer Verlag, Charlotte NC, octobre 1994.
- [39] P. SÉBILLOT, «Sémantique des composés anglais : approche générale, limites et applications», *in: ILN'93 Informatique et Langue Naturelle*, Nantes, décembre 1993.
- [40] E. SIOU, «A Bidirectional Search for Clauses», *in: Proc. of the 4th International Workshop on Inductive Logic Programming*, p. 365–376, Bad Honef, Allemagne, septembre 1994.
- [41] E. SIOU, «Diagnostic cognitif par programmation logique inductive», *in: Secondes rencontres nationales des jeunes chercheurs en intelligence artificielle*, Marseille, septembre 1994.
- [42] E. SIOU, «Inductive Logic Programming for Cognitive Diagnosis», *in: Proc. of the 4th International Conference on User Modeling*, Hyannis, Cape Cod, USA, août 1994.
- [43] E. SIOU, «Sarah ou l'intégration de techniques d'intelligence artificielle pour la rééducation de l'aphasie», *in: 9ème Congrès Reconnaissance des Formes et Intelligence Artificielle*, p. 89–100, Paris, janvier 1994.
- [44] S. THIÉBAUX, «Decision-theoretic planning with a possible models action formalism», *in: AAAI Spring Symposium on Decision-Theoretic Planning*, AAAI Press, mars 1994.

- [45] R. VORCH, «A connection-based point of view of propositional cut elimination», in: *Proc. of 3rd Workshop on Theorem-proving with Analytic Tableaux and Related Methods*, Abingdon, GB, mai 1994.

Rapports de recherche et publications internes

- [46] M.-O. CORDIER, S. THIÉBAUX, «Event-Based Diagnosis for Evolutive Systems», *publication interne n° 819*, Irisa, mai 1994, Version étendue de [18].
- [47] H. LARUELLE, P. RÉGNIER, S. THIÉBAUX, «Dossier Planification et Action», *rapport de recherche n° 94-1-R*, Irit, Toulouse, février 1994.
- [48] I.-C. LERMAN, P. PETER, J. RISLER, «Matrices AVL pour la classification et l'alignement de séquences proteiques», *publication interne n° 866*, Irisa, 1994.
- [49] J. SLANEY, S. THIÉBAUX, «Adventures in Blocks World», *Technical Report n° TR-ARP-7-94*, Australian National University, septembre 1994.

Divers

- [50] H. LARUELLE, P. RÉGNIER, S. THIÉBAUX, «Dossier Planification et Action», *Bulletin de l'Affa*, 16, pages 18-35, janvier 1994.
- [51] I.-C. LERMAN, P. PETER, H. LEREDDE, «Notice du programme CHAVL», *La Revue de Modulad*, décembre 1993, première partie.
- [52] I.-C. LERMAN, P. PETER, H. LEREDDE, «Principes et calculs de la méthode implantée dans le programme CHAVL», *La Revue de Modulad*, décembre 1993, première partie.
- [53] I.-C. LERMAN, P. PETER, H. LEREDDE, «Principes et calculs de la méthode implantée dans le programme CHAVL», *La Revue de Modulad*, décembre 1994, deuxième partie.
- [54] Y. MOINARD, «In defence of classical logic of minimal models», *Fifth International Workshop on Nonmonotonic Reasoning*, Dagstuhl, mai 1994.

8 Abstract

Project activities center around various ways in which computers handle knowledge : raw data generated knowledge (classification), structuring of knowledge (classification and machine learning), transformation and application of knowledge (natural language understanding and reasoning). Two important topics are user modelling and integrating symbolic

and numerical methods for classification and machine learning. Several applications support the project 's theoretical research : in particular, analysis of genetic sequences and rehabilitation of aphasics.

Scientific Context :

- Formalization of reasoning (including diagnostic and planning)
- Automatic natural language understanding
- Classification of structured data
- Machine learning

Developments :

- Mentoniez : tutoring system for geometry exercises
- Chavl : hierarchical classification program (Modulad library)
- Sarah : system for aphasia assessment and rehabilitation
- Pascale : system for anytime planning under uncertainty
- Avare-Avand : classification programs of attributes by associations
- Machine learning kernel (“version space” approach)

Table des matières

1	Composition de l'équipe	1
2	Présentation des objectifs	2
3	Actions de recherche	3
3.1	Raisonnements et logiques non classiques	3
3.2	Prise en compte de l'incertain pour la modélisation et le suivi d'environnements évoluant dans le temps	5
3.2.1	Mises à jour	5
3.2.2	Planification	6
3.2.3	Diagnostic de systèmes évolutifs	6
3.3	Traitement du langage naturel	7
3.4	Apprentissage symbolique automatique	8
3.4.1	Couplages symboliques-numériques	9
3.4.2	Inférence grammaticale	10
3.4.3	Symétries et tableaux analytiques	12
3.5	Enseignement Intelligemment Assisté par Ordinateur	13
3.5.1	Sarah : modélisation de l'utilisateur à l'aide de techniques d'apprentissage automatique	13
3.5.2	Projet Mentoniez	14
3.5.3	Compounds : un système d'enseignement des composés anglais	15
3.6	Gide	16
3.7	Analyse des données	17
3.7.1	Classification de séquences génétiques	17
3.7.2	Arbres de décision	18
3.7.3	Les algorithmes génétiques dans la représentation euclidienne d'une matrice de dissimilarité	19
3.7.4	Analyse ensembliste et statistique du problème de la satisfiabilité	19

Rapport d'activité INRIA 1994 — Annexe technique

3.7.5	Contribution à Modulad	20
4	Actions industrielles	20
5	Actions nationales et internationales	21
6	Diffusion des résultats	22
6.1	Colloques et congrès	22
6.2	Séminaires, écoles et tutoriels	22
6.3	Actions d'enseignement	22
7	Publications	23
8	Abstract	27