INRIA

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Action ORPAILLEUR

Représentations, raisonnements, et extraction de connaissances dans les bases de donneées

Nancy

THÈME 3A

Table des matières

1 Composition de l'équipe

Responsable scientifique

Amedeo Napoli [(CR CNRS)]

Responsables permanents

Florence Le Ber [(CR INRIA – détachement)]

Jean Lieber [(MdC, Université Henri Poincaré — Nancy 1)]

Jean-François Mari [(Professeur, Université de Nancy II)]

Yannick Toussaint [(CR INRIA)]

Assistante de projet

Jamila Merikhi [(jusqu'à septembre 2000)]

Antoinette Courrier [(Technicienne CNRS, depuis octobre 2000)]

Chercheurs doctorants

Rim Al Hulou [(doctorante, bourse co-financée Syrie – INRIA)]

Sandra Berasaluce [(doctorante avec co-encadrement, bourse MENRT)]

Fairouz Chakkour [(doctorante, bourse co-financée Syrie – INRIA)]

Hacène Cherfi [(doctorant, bourse co-financée Région – INRIA)]

Jean-Luc Metzger [(doctorant, bourse co-financée INRA – INRIA)]

Emmanuel Nauer [(doctorant–ATER, Université de Metz)]

Arnaud Simon [(doctorant–ATER, IUT de Illkirch, jusqu'à octobre 2000)]

Chercheurs post-doctorants

Benoît Bresson [(Collaboration Orpailleur-CAV Nancy)]

Rafik Taouil [(Collaboration Dyade-Bull-Orpailleur)]

Stagiaire

Sébastien Hergalant [(DEA de bioingénierie)]

2 Présentation et objectifs généraux

L'orpailleur est l'artisan qui recueille par lavage — à travers un tamis — les paillettes d'or dans les fleuves et les terres aurifères. L'or, dans le cadre de la conception de systèmes à bases de connaissances (SBC dans la suite), correspond à la connaissance. Cette connaissance est de plusieurs types et a plusieurs origines : elle peut reposer sur de l'expertise, des expériences, des explications, des stratégies et des façons de faire. Elle peut être donnée de façon explicite — par des spécialistes — ou exister de manière implicite — dans des bases de données de toutes natures. Pour être opérationnelle, cette connaissance doit être représentée et manipulée de façon adéquate par des procédures de raisonnement.

L'objectif du projet Orpailleur est de concevoir des systèmes intelligents mettant en œuvre des connaissances pour résoudre des problèmes. Ces systèmes intelligents sont multi-formes et sont appelés à fonctionner dans différents domaines d'application, aux premiers rangs desquels se trouvent l'agronomie, l'analyse de textes scientifiques et techniques, la bibliométrie, la chimie (planification de synthèses organiques), la médecine et la muséologie.

3 Fondements scientifiques

3.1 Systèmes à bases de connaissances, représentation des connaissances et raisonnements

Mots clés : systèmes à bases de connaissances, représentation des connaissances (par objets), systèmes classificatoires, logiques de descriptions, structures ordonnées, représentation de l'espace, treillis de relations spatiales, raisonnement par classification, raisonnement à partir de cas.

Participants : Rim Al Hulou, Sandra Berasaluce, Benoît Bresson, Fairouz Chakkour, Hacène Cherfi, Florence Le Ber, Jean Lieber, Jean-François Mari, Jean-Luc Metzger, Amedeo Napoli, Emmanuel Nauer, Arnaud Simon, Rafik Taouil, Yannick Toussaint.

Résumé : *Dans le cadre de la conception de* SBC*, nous nous intéressons essentiellement aux systèmes de* RCO*— représentation de connaissances par objets — aux systèmes classificatoires et aux logiques de descriptions. Ces systèmes s'appuient sur une hiérarchie de classes (ou concepts) instanciables qui sont organisées en hiérarchie(s) par l'intermédiaire d'une relation d'ordre partiel (spécialisation ou subsomption). La hiérarchie des classes peut être consultée pour résoudre des problèmes par l'intermédiaire de procédures (approche procédurale) ou de mécanismes de raisonnement comme la classification de classes ou d'instances (approche déclarative). Un ensemble d'assertions décrivant les faits dans lesquels interviennent les classes et leurs instances — instanciations de classes et instanciations de relations entre classes — complète la représentation de l'univers étudié.*

3.1.1 Systèmes classificatoires et raisonnements

Appréhender un système de RCO comme un système logique a donné naissance à la théorie des systèmes classificatoires, qui s'appuie sur les développements théoriques réalisés dans le cadre des logiques de descriptions. Les opérations principales qui sont à la base du raisonnement sont :

- le test de subsomption vérifie qu'une classe `C` est plus générale qu'une classe `D`,
- la classification de classes qui consiste à placer une nouvelle classe `X` dans une hiérarchie $\mathcal{H}$, la classification d'instances qui consiste à déterminer les classes dont un objet `x` donné peut être une instance (en particulier, une classe `C` n'est satisfiable que si elle peut avoir effectivement des instances),
- la recherche de propriétés qui consiste à retrouver les propriétés détenues par une classe ou une instance.

Le raisonnement par classification s'appréhende comme une procédure de déduction opérant sur une hiérarchie. Sa mise en œuvre repose sur un cycle comprenant trois étapes :

- initialisation (création d'un nouvel objet `x` à classer),
- classification (recherche de la position de `x` dans la hiérarchie),
- mise en place de `x` dans la hiérarchie et exploitation de cette mise en place (ce qui peut ramener le cycle à sa première étape).

Le RÀPC (raisonnement à partir de cas) se propose de faire correspondre à l'énoncé d'un nouveau problème `P` une solution `Sol(P)` en tirant parti d'un ensemble de cas, qui sont des problèmes déjà résolus accompagnés de leurs solutions. Un cas mémorisé, ou cas source, est la donnée d'un couple énoncé de problème – solution (`P`, `Sol(P)`) et fait partie d'une base de cas. Le processus du RÀPC se décompose en trois opérations principales : la remémoration, l'adaptation et la mémorisation. Étant donné un problème `cible` à résoudre, la remémoration consiste à retrouver dans la base de cas un énoncé de problème `source`, jugé similaire ou analogue à `cible`. Si `source` existe, sa solution `Sol(source)` est adaptée pour produire une solution `Sol(cible)` de `cible`. Une étape de mémorisation peut compléter les deux étapes précédentes.

Bibliographie

Systèmes de RCO : [2] [8] [23].
RÀPC : [13] [14] [15] [16] [28] [7] [6] [21].

3.1.2 Systèmes à bases de connaissances pour l'étude des organisations spatiales agricoles

Dans ce cadre, le travail de recherche est effectué en collaboration avec des chercheurs de l'INRA (Centres de Nancy, Mirecourt et Montpellier). Il porte sur l'étude des formes d'organisation spatiale de l'agriculture. Trois projets ont été développés ou sont en cours de développement :

- interprétation d'images satellitaires,
- simulation d'organisations spatiales,
- exploitation d'une base d'enquêtes (cartes et explications).

Le premier projet, interprétation d'images satellitaires, a fait l'objet de la thèse de L. Mangelinck en 1998. Les principaux résultats ont porté sur la représentation dans un système de RCO de structures spatiales définies comme des ensembles d'entités spatiales reliées entre elles par des relations spatiales qualitatives. Nous avons travaillé sur la représentation des relations topologiques dans les systèmes de RCO et sur leur organisation sous forme de treillis. L'année 2000 a permis de faire une synthèse de ce travail par des publications, parues ou soumises.

Le deuxième projet, simulation d'organisations spatiales, s'oriente vers la réécriture d'un modèle multi-agents d'allocation de surfaces plus général, qui permettrait d'intégrer des informations telles que celles issues de la fouille de données *Ter Uti* (représentation de la dynamique des organisations spatiales). Une collaboration en ce sens est en cours avec MAIA.

Le troisième projet a débuté en 99 dans la cadre du stage de DEA d'E. Kaboré et s'est poursuivi dans le cadre du stage de DEA de J.-L. Metzger. Nous nous sommes intéressés à la représentation sous forme de graphes de structures spatiales et au calcul de similarités entre structures spatiales. Nous avons étudié différents systèmes (RCO, logiques de descriptions) pour représenter ces structures. Une phase importante d'acquisition de connaissances et de formulation de problèmes avec les chercheurs de l'INRA a également eu lieu et a permis de finaliser un sujet de thèse pour J.-L. Metzger, sujet intitulé : "Élaboration de formalismes de représentation et de raisonnement pour les systèmes d'informations géographiques". Cette thèse est co-financée par l'INRA et l'INRIA et débute fin 2000.

Bibliographie

Représentation et manipulation de données spatiales : [31] [19] [20] [29] [18] [30] [5].

3.2 Extraction de connaissances dans les bases de données

Mots clés : extraction de connaissances dans les bases de données, méthodes symboliques pour la fouille de données, classification par treillis, recherche de motifs fréquents (dans des tableaux de données), extraction de règles, modèles de Markov cachés pour la fouille de données.

Participants : Rim Al Hulou, Sandra Berasaluce, Benoît Bresson, Fairouz Chakkour, Hacène Cherfi, Florence Le Ber, Jean Lieber, Jean-François Mari, Jean-Luc Metzger, Amedeo

Napoli, Emmanuel Nauer, Arnaud Simon, Rafik Taouil, Yannick Toussaint.

3.2.1 ECBD symbolique

L'extraction de connaissances à partir des bases de données — abrégée en ECBD — est une activité qui consiste à analyser un ensemble de données brutes de façon à en extraire des connaissances exploitables. C'est un expert du domaine relatif aux données — l'« analyste » — qui est chargé de diriger l'extraction. En fonction de ses objectifs, l'analyste va sélectionner des données et utiliser des outils de *fouille de données* pour construire des modèles expliquant les données. L'analyste peut ensuite sélectionner et exploiter les modèles qui représentent un point de vue satisfaisant. Pour mener à bien son activité, l'analyste met à contribution ses connaissances mais aussi un ensemble d'outils regroupés au sein d'un système d'ECBD. Un système d'ECBD s'articule autour de quatre composantes principales :

- les bases de données et leurs systèmes de gestion,
- un système à base de connaissances pour la gestion des connaissances et la résolution de problèmes sur le domaine relatif aux données,
- un système d'analyse de données symboliques pouvant s'appuyer sur des techniques d'induction, de classification (par treillis et par arbres de décision, etc.), éventuellement couplé à un système d'analyse de données numériques et de statistiques,
- une interface se chargeant des interactions et de la visualisation des résultats intermédiaires et finaux.

Un système d'ECBD vise à traiter des bases de données volumineuses et évolutives, et il peut, pour ce faire, s'appuyer sur des connaissances du domaine lors du processus d'extraction des connaissances. L'ECBD peut être ainsi vue comme le processus alimentant un système à base de connaissances ; les connaissances extraites sont stockées dans la base pour être réutilisées dans d'autres applications, et mises à jour le cas échéant (lors de l'arrivée de nouvelles données par exemple). Pour répondre à ces deux exigences, un système d'ECBD doit assurer la communication entre le système de gestion des bases de données et le système à base de connaissances.

Les classifications par treillis et par arbres de décision font partie des techniques utilisées dans le projet Orpailleur pour la conception d'outils de fouille de données symbolique. Ces techniques relèvent de l'analyse de « tableaux de données » et sont adaptées au traitement de données représentées dans un système de RCO. D'une part, un système de RCO fournit des services inférentiels et d'accès à l'information qui sont mis à profit pour l'ECBD. D'autre part, le formalisme de représentation associé au système de RCO permet de manipuler des données « complexes », où les attributs peuvent être multivalués ou définir des relations. Les fonctionnalités de représentation et de raisonnement du système de RCO sont exploitées pour la construction d'arbres de décision et de treillis de Galois. Les structures ainsi construites sont des hiérarchies de classes et d'objets, qui relèvent du modèle de RCO. Une telle approche a pour avantage de permettre la prise en compte des connaissances disponibles et de produire des

structures directement interprétables — ces structures rendent compte de l'organisation des données et mettent en évidence des règles de classification — et réutilisables, pour mener à bien des raisonnements. Les treillis de Galois permettent de faire émerger des règles d'association (exactes ou partielles). En tenant compte des connaissances sur le domaine des données, il est alors possible d'élaguer l'ensemble des règles extraites obtenues — le processus d'extraction est de complexité exponentielle — et de ne conserver que des règles jugées intéressantes.

Un système d'ECBD s'appuyant sur le système de RCO Y3 — appelons-le ORPAILLEUR — a été réalisé dans le projet Orpailleur. Les développements logiciels ont consisté à étendre Y3 par des fonctionnalités de classification (de classes et d'instances) et de filtrage. Deux modules de classification par treillis et par arbres de décision ont été ainsi implantés, et ont été couplés à des modules de visualisation (hiérarchies d'objets et données). En particulier, le système ORPAILLEUR a été utilisé pour analyser des données médicales et des données textuelles.

Les expérimentations menées ont permis de valider (partiellement) le système mais elles ont aussi fait émerger plusieurs problèmes qui font l'objet de préoccupations actuelles. Ainsi, les règles de classification et d'association extraites doivent être validées par une mesure d'ordre statistique (mesure de probabilité, de confiance). La mesure actuellement employée n'est pas bien adaptée à de gros volumes de données et rend difficilement compte de «l'utilité» potentielle des règles extraites. Pour traiter ce problème, une collaboration avec Régis Gras (professeur émérite à l'université de Nantes) est en cours, dont l'objectif est de proposer une nouvelle mesure statistique pour la validation des règles d'association. Par ailleurs, la construction de treillis de Galois est un processus exponentiel qui nécessite un temps de calcul plutôt élevé. L'utilisation de techniques de programmation parallèle devrait réduire ce temps de calcul est donc permettre une plus grande interactivité avec l'analyste. Des travaux dans ce sens ont débuté en collaboration avec le projet HIPPO de l'université de Newcastle upon Tyne (Arnaud Simon est depuis octobre 2000 en stage post-doctoral dans cette université). Enfin, deux problèmes importants subsistent et doivent être traités dans un avenir proche : le système ORPAILLEUR repose sur le système Y3 — écrit lui-même en Lisp — ce qui lui donne une portée et une portabilité très limitées ; le système ORPAILLEUR n'est pas toujours pas en liaison (directe) avec un système de gestion de bases de données. Ainsi, une migration du système ORPAILLEUR vers un environnement JAVA est en cours d'étude et de réalisation depuis le milieu de l'année 2000. L'objectif est de disposer à moyen terme d'un système portable autorisant un couplage effectif avec les systèmes de gestion de bases de données existants.

3.2.2 ECBD et bases de données

Le projet Orpailleur est associée au projet Dyade (INRIA Rocquencourt) et Bull pour un projet qui concerne les machines parallèles NEC et l'exploitation par ces machines de bases de données volumineuses. Une action INRIA, nommée PARANA, est en cours de création sur le sujet. Actuellement, l'exploitation des bases de données est classique et s'articule essentiellement autour de fonctions d'interrogation standards. Une extension naturelle est d'étudier la mise en œuvre de techniques et d'outils permettant de pratiquer de l'extraction de connaissances dans les bases de données, parallèlement aux fonctions d'interrogation. Pour cela, il est nécessaire de bien connaître la gestion des bases de données, mais aussi les méthodes symboliques d'ECBD, comme la classification par arbres de décision et par treillis, et l'extraction de règles

d'implications partielles dans un treillis.

Dans ce cadre, Rafik Taouil bénéficie d'une bourse post-doctorale INRIA, qui a débuté au 1er octobre 2000 pour un an (prise en charge par Dyade), et a depuis intégré le projet Orpailleur. L'objectif du travail de recherches de Rafik Taouil est d'étudier les diverses extensions possibles d'un langage classique d'interrogation de bases de données dans l'optique de la fouille de données. Un langage d'interrogation peut être étendu pour lui-même, pour autoriser des filtrages plus complexes, plus précis et plus fins, mais aussi être étendu pour faire émerger des motifs fréquents et des règles d'implication partielles dans des tableaux de données volumineux. Les motifs fréquents sont des groupes de propriétés partagées par des individus, qui apparaissent avec une certaine fréquence, et les règles extraites matérialisent des implications entre motifs. Pour ce travail de recherches, qui comporte également une part importante de réalisation pratique, les techniques de recherche par niveaux de motifs fréquents et de classification par treillis vont être exploitées en priorité.

3.2.3 ECBD numérique

Une des originalités du projet Orpailleur est de réutiliser certains travaux de classification numérique en reconnaissance de la parole pour procéder à de la fouille de signaux spatiaux-temporels, plus précisément pour étudier la classification de données temporelles ou spatiales, par exemple pour traiter des données issues d'un processus industriel comme les caractéristiques de tôles laminées par un train à bande ou les successions de cultures sur une parcelle géographique donnée. Dans ces deux domaines d'application, des outils à base de modèles stochastiques — les modèles de Markov cachés d'ordre 1 ou 2 (HMM1 et HMM2) — développés initialement pour la reconnaissance de la parole et l'identification du locuteur, sont utilisés. Ces recherches en ECBD, d'une nature particulière et originale, visent à accroître le coté générique des outils de reconnaissance en investissant un domaine de recherche plutôt vierge. Elles constituent aussi un bel exemple d'inter-disciplinarité.

L'émergence des techniques stochastiques est principalement due à l'apparition de nouveaux serveurs de calcul puissants. Beaucoup d'hypothèses simplificatrices ont été posées dans les années 1980 pour implanter des algorithmes d'apprentissage et de reconnaissance ; l'utilisation de chaînes de Markov du premier ordre est la plus connue. Pour notre part, ce sont les modèles stochastiques d'ordre supérieur comme les modèles de Markov d'ordre 2 qui permettent une meilleure prise en compte des durées des suites d'états stationnaires et transitoires et qui nous intéressent.

Reconnaissance de successions culturales

Nous avons étudié les successions culturales pratiquées en Lorraine afin d'intégrer cette connaissance dans un modèle de simulation d'organisations spatiales agricoles en cours de développement à l'INRA. Pour réaliser cette étude, nous exploitons des données *Ter Uti* qui constituent un relevé de l'utilisation du territoire depuis une vingtaine d'années. Ces données sont traitées avec des algorithmes d'apprentissage développés au LORIA pour la reconnaissance de la parole. Ces algorithmes s'appuient sur les HMM1 et HMM2, qui permettent de représenter des observations spatio-temporelles, comme des successions d'états où les transitions entre états dépendants, suivant l'ordre du modèle, de l'état courant et des n états précédents.

Durant l'année 2000, nous avons évalué ces outils en relation étroite avec des experts agronomes de l'INRA. Une étudiante de DAA – diplôme d'agronomie approfondie – de l'ENSAIA, qui étudiait les pratiques culturales et leurs influences sur la qualité de l'eau dans le bassin parisien a utilisé intensivement les HMM (1 et 2) dans son travail de modélisation. À ce titre, nous sommes impliqués dans le projet PIREN-Seine avec la station INRA de Mirecourt.

A un niveau plus théorique, nous nous sommes intéressés à la classification non supervisée de territoires à l'aide de HMM sur des critères spatiaux [22]. Ce travail se poursuit, toujours dans le cadre de l'appel d'offres PSIG'2000 – programme des systèmes d'information géographique — et nous abordons maintenant le problème de la segmentation de données spatio-temporelles.

3.3 Fouille de textes et analyse de l'information scientifique et technique

Mots clés : information scientifique et technique, informatique linguistique, terminologie, interprétation, fouille de textes, synthèse de textes, classification, logiques de descriptions, treillis.

Participants : Fairouz Chakkour, Hacène Cherfi, Amedeo Napoli, Emmanuel Nauer, Yannick Toussaint.

La fouille de données textuelles consiste en l'analyse d'un volume important de documents textuels pour fournir à l'utilisateur une vision synthétique et interprétable de leur contenu. La quasi-totalité des travaux actuels en fouille de textes reposent sur des méthodes numériques de classification, statistiques ou neuronales. Nous avons adopté une méthode symbolique qui nous permet d'associer au processus de fouille des connaissances du domaine et d'envisager un traitement beaucoup plus fin de l'information dans les textes.

Nos travaux se structurent suivant deux directions complémentaires : une approche informatique linguistique du texte dont le but est d'extraire du texte des structures conceptuelles. Bien que cette problématique ait fait l'objet de nombreuses publications en linguistique informatique, des questions spécifiques à la fouille se posent. La seconde orientation est une approche basée sur les connaissances et le raisonnement pour réaliser la tâche de fouille de données proprement dite sur les structures conceptuelles indexant les textes.

La linguistique de corpus pour l'Information Scientifique et Technique

Notre approche de la linguistique est guidée par les travaux d'analyse des textes ou de raisonnements que nous effectuons en aval. Dans ce contexte, nous ne recherchons pas à expliquer des phénomènes de langue ou à les caractériser de façon plus ou moins universelle. Nous recherchons au contraire à exploiter les caractéristiques communes aux différents textes scientifiques ou techniques mais également les spécificités de certains types de textes pour en extraire des informations plus pertinentes.

Indexation conceptuelle des textes

Les objectifs que nous nous sommes fixés sont d'indexer des documents textuels par des structures conceptuelles qui nous permettent de dépasser la simple indexation au niveau du terme que nous utilisons jusqu'à présent. Ces structures conceptuelles sont des objets que nous

structurons au sein d'une base de connaissances sur lesquelles il devient alors possible de faire des calculs de spécialisation ou de généralisation.

L'indexation conceptuelle fait appel à deux autres domaines de recherche. D'une part, les travaux sur l'analyse syntaxique robuste qui permet de découper une phrase en syntagmes nominaux et verbaux. À partir de ce découpage se pose alors un premier problème, encore mal résolu au niveau linguistique, d'identification de la dépendance entre ces syntagmes. Le second problème est la normalisation conceptuelle : de nombreuses formes linguistiques différentes expriment en réalité, un seul et même concept que nous devons identifier pour éviter une trop grande dispersion des données collectées à partir des textes. De ce point de vue, cette indexation se rapproche des travaux sur l'extraction d'information avec un besoin de structuration des données plus fort, lié à l'utilisation d'outils de fouille de données par la suite.

Les travaux que nous menons sur l'indexation se basent sur le raisonnement à partir de cas [14, 15]. Cette approche permet d'exploiter la régularité de certaines constructions syntaxiques et de rapprocher certaines de ces structures en effectuant des transformations pour créer la représentation conceptuelle.

Éléments linguistiques pour l'analyse de références bibliographiques

À la suite des travaux sur les tables des matières [12], le projet *Citations* est mené en coopération avec l'INIST et l'équipe READ. Il vise à identifier dans les bibliographies d'articles les différents champs présents pour mettre en œuvre, par la suite, des algorithmes de fouille de données. Une première étape de reconnaissance de caractères a déjà été réalisée. La diversité des formats rencontrés dans les différentes revues mettent en échec un certain nombre de règles exploitant les séparateurs de champs (la virgule, le point, les deux points) ou des marqueurs (actes, journal...). Nous avons donc calqué notre approche sur les étiqueteurs morpho-syntaxiques qui associent à chaque mot d'une phrase une catégorie syntaxique. Cette notion de catégorie syntaxique étant trop spécifique par rapport à notre besoin, nous définissons actuellement un nouvel ensemble d'étiquettes et étudions actuellement l'apport de ces informations pour la définition de nouvelles règles de découpage.

Fouille de textes basée sur des méthodes symboliques

Nous avons entrepris des travaux dans deux directions. La première se base sur des méthodes symboliques de fouille de données. Une première thèse sur l'utilisation des logiques de descriptions a été soutenue en 1999 (thèse de Nicolas Capponi [1]), et nous souhaitons compléter ces travaux en utilisant plus spécifiquement les treillis. La seconde direction vise à étudier la différence et la complémentarité des méthodes symboliques et numériques, ces dernières étant plus classiquement utilisées en fouille de textes ou en recherche d'information.

Travaux sur les treillis

À la suite de la thèse d'Arnaud Simon [32], une thèse est en cours sur l'intérêt des treillis pour faire de la fouille sur des données issues de textes [27, 26]. Nous avons mené plusieurs expérimentations s'appuyant sur une indexation des textes au niveau du terme. Cette analyse produit un ensemble de classes structurées en treillis et un ensemble de règles d'association. Les travaux que nous menons sur l'indexation conceptuelle doivent nous permettre de caractériser

plus finement les textes. Les règles d'association ainsi extraites devraient nous permettre une meilleure analyse d'un domaine. Les principales questions que l'on souhaite aborder actuellement sont les suivantes :

- la prise en compte de la structure hiérarchique des index conceptuels dans la méthode de classification hiérarchique,
- l'extraction de règles d'association exploitant la hiérarchisation des index conceptuels,
- le suivi d'une classification lors de l'ajout de nouvelles données,
- la lisibilité et l'évaluation des résultats : cette phase se fera par l'intervention d'experts de l'INIST.

Complémentarité entre méthodes symboliques et méthodes numériques

Nous avons commencé cette année à étudier la complémentarité des méthodes symboliques avec les méthodes numériques telles que les cartes de Kohonen [17]. Nous nous sommes intéressés plus spécifiquement aux points suivants :

- Les méthodes symboliques ne permettent pas facilement de prendre en compte la pondération des termes (ou plus généralement des propriétés) telle que cela est pratiqué en recherche d'information. La mise en correspondance d'une classification symbolique et d'une classification numérique fait donc appel à des choix de projection d'une classe numérique sur une classe symbolique dont les conséquences sont importantes et que nous souhaitons pouvoir cerner.
- La méthode de classification de Kohonen permet d'activer des mécanismes de propagation de propriétés d'une carte à une autre suivant les points de vue observés. Il est cependant difficile de caractériser, par une règle basée sur des propriétés les phénomènes observés. Les treillis, quant à eux, permettent l'extraction de règles d'association. Nous cherchons donc à caractériser les liens entre ces deux approches.

Bibliographie globale sur l'ECBD

ECBD symbolique : [11] [25] [32] [9].
ECBD numérique : [22].
Fouille de textes : [14, 15] [12] [1], [27, 26] [17].

3.4 Extraction de connaissances et aide à l'interrogation de bases de données chimiques

Mots clés : systèmes de représentation de connaissances par objets, extraction de connaissances dans des bases de données chimiques, classification, recherche d'information,

interrogation et navigation dans des bases de données, perception, synthèse en chimie organique.

Participants : Sandra Berasaluce, Claude Laurenço [CCIPE et LIRMM Montpellier], Jean Lieber, Amedeo Napoli.

Le travail de thèse de Sandra Berasaluce est co-dirigé par Claude Laurenço (CCIPE et LIRMM de Montpellier) et Amedeo Napoli. Il porte sur l'extraction de connaissances et l'aide à l'interrogation et à la navigation dans des bases de données de chimie organique. L'extraction de connaissances dans des bases de données est une question actuelle en chimie organique. Il existe de très grandes bases de données publiques qui portent sur plus de 18 millions de substances décrites — avec leurs propriétés chimiques, physiques et biologiques — et sur plus de 10 millions de réactions. L'interrogation de ces bases de données est le plus souvent difficile et frustre, car elle est conçue pour répondre à des besoins de documentation plus que pour aider à la résolution de problèmes, et elle se fait avec des moyens plutôt limités (sans tenir compte de connaissances du domaine par exemple).

Ainsi, il est possible actuellement de retrouver dans les bases de données de réactions disponibles la plupart des réactions qui forment un composé donné. Toutefois, il est très difficile, voire impossible, d'obtenir des informations sur des méthodes générales de synthèse relatives à une telle construction, car les bases sont généralement des collections de réactions particulières indépendantes les unes des autres. Ainsi, une requête sur la formation des cycles à 7 chaînons fournira plusieurs milliers de réponses : ce volume est trop important pour pouvoir mener une analyse sur ces réponses et faire émerger des méthodes générales de synthèse. L'idée est donc d'utiliser des techniques d'ECBD — comme la classification conceptuelle, la classification par treillis et par arbres de décision — pour découvrir des régularités dans les données, par exemple faire émerger des schémas réactionnels génériques à partir de descriptions de réactions spécifiques.

À l'aide du logiciel RÉSYN-ASSISTANT, dont l'objectif principal est la perception de molécules en chimie organique avec extraction de blocs composants, Sandra Berasaluce fait émerger une représentation des réactions en termes de blocs (composant la réaction) et des changements observés entre les blocs. Les intérêts de cette perception originale des réactions sont nombreux : il est alors possible d'envisager le développement d'un nouveau mode d'interrogation des bases de données de réactions, et l'utilisation d'une base de données de méthodes de synthèse en parallèle avec des bases de données réactionnelles (collection d'exemples particuliers). Ce travail de recherche est polyvalent et revêt un ensemble d'intérêts théoriques et pratiques en informatique et en chimie. Parmi ces intérêts, il faut mentionner pour l'informatique : la gestion de bases de données (traitement de requêtes, indexation), l'extraction de connaissances dans les bases de données, la représentation de connaissances et le raisonnement ; pour la chimie : la modélisation de réactions et leurs représentations sous différents points de vue.

3.5 Systèmes intelligents de traitement de l'information

Mots clés : systèmes d'information intégrés, données semi-structurées, navigation intelligente sur le Web, grandes bases de connaissances, grandes bases de données, bases de

données à objets, bases de données distribuées.

Participants : Rim Al Hulou, Hacène Cherfi, Amedeo Napoli, Emmanuel Nauer, Yannick Toussaint.

Accès à l'information sur le Web

La maîtrise de l'accès à l'information dans un fonds volumineux et hétérogène tel que le Web représente un enjeu majeur pour les consommateurs d'information (chercheurs, entreprises, etc.). Les moteurs de recherche sont débordés par l'explosion du Web et ne répondent plus aux tâches de recherche d'information. Le but de ce travail de recherche est d'étudier la combinaison de la technologie des systèmes à bases de connaissances et de la technologie du Web ; ceci afin de fournir une aide efficace lors de la consultation du Web. Notre approche consiste à utiliser des données structurées d'un domaine, comme les références bibliographiques, les thésaurus, etc., pour faire émerger des connaissances sur ce domaine, comme des réseaux d'auteurs, le vocabulaire employé par tel ou tel auteur, etc. Ces connaissances sont alors exploitées pour la recherche ou le filtrage d'information sur le Web. L'accès aux données du Web est réalisé via les moteurs de recherche classiques comme AltaVista, Excite, ou encore Google, qui sont utilisés comme des outils distants. Les connaissances émergeantes devant servir à guider l'utilisateur (i) en amont des moteurs de recherche pour la détermination de requêtes, pour l'expression du besoin, etc. et (ii) en aval pour l'évaluation et une présentation plus intelligible des documents.

Un enjeu est la production de connaissances — dites aussi *informations élaborées* dans le milieu des systèmes d'information — qui donnent une idée synthétique d'un ensemble de données et qui puissent être exploitées dans un raisonnement. Concrètement, cette production de connaissances peut se voir comme un processus d'ECBD qui agit sur des informations scientifiques et techniques (IST). Elle peut consister à chercher à dégager les principaux thèmes de recherche sous-jacents à un corpus de références bibliographiques, ou encore les collaborations entre auteurs, l'émergence d'une technique bien particulière, etc. Nous touchons en cela au domaine de la bibliométrie qui fixe les bases d'exploitation de l'IST. Là aussi, une normalisation minimale des données à exploiter est indispensable pour éviter des biais statistiques.

Intégration et traitement de l'information : les données semi-structurées et XML

L'intégration et le traitement d'informations provenant de sources variées et hétérogènes sont des questions préalables à la construction de systèmes de fouille de données sur le Web (en particulier). À partir de ces données hétérogènes, nous souhaitons modéliser le domaine — déterminer les concepts du domaine, les liens entre les concepts et le vocabulaire attenant aux concepts — en utilisant un système de représentation capable d'effectuer des raisonnements sur ces données. De ce fait, il est indispensable de maîtriser l'hétérogénéité du vocabulaire utilisé et de sa portée sémantique. La construction d'un vocabulaire de base est donc un point essentiel dans notre approche. Le fait que les données soient hétérogènes et plutôt non régulières nécessite aussi de s'intéresser au traitement de *données semi-structurées* (DSS) et à l'intégration d'informations provenant de sources différentes.

Les données semi-structurées sont des données hétérogènes, non régulières, sans format fixe

bien déterminé qui décrive leur structure et leur organisation. De telles caractéristiques rendent difficile voire impossible la manipulation de telles données par des systèmes de gestion de bases de données (SGBD) classiques sans autre extension ou modification.

Une des options prises dans le groupe Orpailleur pour prendre en compte et manipuler des données semi-structurées textuelles consiste tout d'abord à les décrire avec le langage de description de données textuelles XML [10]. Les caractéristiques et les fonctionnalités de XML le rendent particulièrement bien adapté à la description de DSS. La mise en œuvre de raisonnements sur de telles données — par exemple pour des besoins de résolution de problèmes ou d'ECBD — est ensuite dévolue à un système de RCO, où émerge la notion de classe *polythétique*, classe qui peut se définir à la fois par des disjonctions et des conjonctions d'attributs, par opposition aux conventionnelles conjonctions d'attributs. Les données semi-structurées sont alors transformées en objets semi-structurés et sont manipulées par des processus de classification adaptés (prise en compte des disjonctions par exemple [10]).

Expérimentations et réalisations : vers la veille technologique

Les travaux menés jusqu'à présent concernent l'intégration des données (locales ou distantes, multi-sources et de natures différentes) et la construction d'informations élaborées à partir de ces données. Un système d'investigation sur le Web a été mis en place, qui permet un accès indifférencié aux données locales ou distantes, ainsi que des croisements entre ces données. Concrètement, nous avons développé un ensemble de modules permettant (i) de réaliser des opérations fondamentales d'intégration de données telles la normalisation des données manipulées, la convergence du vocabulaire, la suppression des doublons dans les données, etc. ; (ii) d'accéder à l'ensemble des données par une interface Web (génération dynamique de pages HTML).

D'un point de vue pratique, une collaboration avec des chercheurs de l'INRS (Institut National de Recherche et Sécurité) a orienté notre contexte d'application vers le domaine médical. Ce domaine constitue un terrain particulièrement propice aux expérimentations, du fait qu'il s'avère riche en fonds structurés (bases de données, thésaurus, etc.) et en données en ligne. L'utilisation et le croisement de données structurées et hétérogènes pour la construction d'un système intelligent de recherche d'informations ont permis :

- La structuration du domaine pour un accès hiérarchisé à l'information : des accès thématiques sont construits automatiquement par des méthodes de classification (à partir de descripteurs de références bibliographiques par exemple).

- La traduction de termes pour un accès multilingue : une traduction automatique du vocabulaire du domaine peut être effectuée via un thésaurus ou encore par des corrélations de descripteurs au travers de références bibliographiques multilingues.

- La génération d'un environnement d'investigation spécialisé (et intégré) sur le Web permettant à l'utilisateur d'être assisté dans l'étape consistant à définir le vocabulaire de la requête à soumettre à un moteur de recherche (pour une recherche d'information sur Internet), ou encore d'obtenir des compléments d'informations (références bibliographiques locales) sur les documents du Web retrouvés.

– Le filtrage d'information sur Internet : à partir des critères sélectionnés par l'utilisateur, une requête est générée automatiquement et est soumise aux moteurs de recherche. Cette requête est précisée par l'ajout d'un contexte de recherche (vocabulaire proche) aux critères sélectionnés.

Il s'agit maintenant de déterminer aussi quelles connaissances vont permettre l'émergence de documents pertinents. Ces connaissances vont servir à favoriser la recherche d'information sur le Web (formulation automatique de requêtes). Elles devront également permettre d'analyser (valider, rejeter, juger, etc.) et de classer les documents, proposés en réponse par les moteurs de recherche. Dans ce but, la classification et le raisonnement ont un rôle essentiel à jouer. Il devient alors nécessaire de prendre en compte ces documents (textuels) hétérogènes, de les coder dans un formalisme de représentation pour être en mesure d'effectuer des raisonnements : par exemple, traiter des requêtes analogues, reconnaître qu'une requête est plus générale qu'une autre, classifier des requêtes, etc. Les résultats de ces travaux peuvent être étendus à la gestion de grandes bases de connaissances et de grandes bases de données.

Ces travaux prennent actuellement part, dans le cadre de l'action Escrire, à des collaborations avec les projets ACACIA et SHERPA de l'INRIA. Le but de cette action est d'étudier différents types de représentation des connaissances pour la gestion de documents scientifiques et techniques.

Bibliographie : [10] [23] [4] [3] [24].

3.6 Méthodes classificatoires pour l'étude du génome

Mots clés : génome, classification, reconstruction d'arbres phylogénétiques, modèles de Markov.

Participants : Bertrand Aigle [Laboratoire de Génétique et Microbiologie, UHP Nancy 1], Bernard Decaris [Laboratoire de Génétique et Microbiologie, UHP Nancy 1], Sébastien Hergalant, Florence Le Ber, Pierre Leblond [Laboratoire de Génétique et Microbiologie, UHP Nancy 1], Jean-François Mari, Amedeo Napoli.

Certains membres du projet Orpailleur commencent à s'intéresser de près à l'étude du génome et à l'application de méthodes propres à l'ECBD pour ce faire. Comme pour l'ECBD, deux approches peuvent être considérées, l'une plutôt numérique et l'autre plutôt symbolique.

Dans le cadre symbolique, une première étude a été réalisée par Sébastien Hergalant lors de son stage DESS double compétence. Le sujet du stage consistait à étudier les liens potentiels entre la classification en phylogénétique — plus précisément la reconstruction d'arbres phylogénétiques — et la classification dans les systèmes classificatoire. Pour cela, il a fallu se familiariser avec les logiciels de classification phylogénétique et de reconstruction d'arbres phylogénétiques, les recenser et en tester quelques uns. Ensuite, une double étude a été menée : (i) comment l'adaptation est-elle perçue conjointement en reconstruction phylogénétique et dans la phase d'adaptation du RÀPC (bien connue de certains membres du projet Orpailleur [28] [16]), (ii) par suite, comment la méthodologie exploitée en reconstruction phylogénétique

peut-elle être transférée en RÀPC, et bien sûr réciproquement. Les résultats de cette première étude préliminaire (voir ci-dessous) sont encourageants et ouvre la voie à de nombreux développements.

Un autre projet de recherche sur le génome est en cours avec le Laboratoire de Génétique et Microbiologie (UA INRA 952, Université Henri Poincaré, Nancy 1) sur le transfert de matériel génétique entre bactéries. Au cours de cette première année, le projet s'est concrétisé par des avancées à la fois sur les plans biologique et informatique. Du point de vue biologique, une banque de 10 cosmides chevauchants recouvrant la répétition terminale inversée (TIR) de l'ADN chromosomique de Streptomyces ambofaciens ainsi qu'une partie adjacente à cette TIR — soit un total d'environ 300 kilobases — a été fournie au Génoscope (CNS, Evry) pour le séquençage de cette région. Les premières séquences commencent à être disponibles.

Du point de vue informatique, nous avons testé et comparé différents logiciels de recherche d'ORF (séquence codante qui définit à l'aide des bases `A`, `C`, `G` et `T` une protéine), tels que Artémis, GenMark et Glimmer2 à l'aide des séquences déjà annotées et disponibles du génome de Streptomyces coelicolor (`http://www.sanger.ac.uk`), un proche parent de S. ambofaciens. Nous envisageons également de tester le programme TBparse et de comparer l'ensemble de ces outils avec ceux développés au LORIA. La plupart des logiciels sont basés sur des modèles de Markov cachés (HMM). L'un des ordinateurs du laboratoire de Génétique et Microbiologie a été spécialement dédié à ce projet et affecté pour l'installation de ces logiciels afin d'entreprendre l'analyse des séquences obtenues.

Dans le cadre de l'étude proprement dite des transferts horizontaux dans la région terminale du chromosome de S. ambofaciens, nous nous intéressons à l'utilisation de modèles de Markov d'ordre supérieur. Ces modèles développés par l'équipe Orpailleur sont actuellement évalués à l'aide des séquences disponibles du chromosome de S. coelicolor.

Des moyens humains ont été spécifiquement dédiés au développement du projet : l'équipe du laboratoire de Génétique et Microbiologie a été renforcée cette année par l'arrivée de Bertrand Aigle, nommé maître de conférences en septembre 2000 dans le laboratoire de Génétique et Microbiologie, et l'arrivée de Sébastien Hergalant — précédemment en stage DESS chez Orpailleur — étudiant en DEA de bioingénierie. Les activités de recherche de Bertrand Aigle sont principalement basées sur le projet bioinformatique de Le stage pratique de DEA de Sébastien Hergalant, en co-tutelle entre Orpailleur et le Laboratoire de Génétique et Microbiologie, concernera essentiellement l'utilisation des modèles de Markov d'ordre supérieur dans le cadre de la recherche des transferts horizontaux.

Bibliographie : S. Hergalant, Classification et reconstruction phylogénétique, Rapport de stage de DESS IDC2, LORIA, Nancy, 2000.

4 Logiciels

4.1 Les logiciels RÉSYN et RÉSYN-ASSISTANT

Participants : Sandra Berasaluce, Claude Laurenço [CCIPE et LIRMM Montpellier], Jean Lieber [correspondant], Amedeo Napoli.

Résumé : *À l'origine, le système* RÉSYN *a été développé en Y3 dans le cadre du* GDR CNRS *1093 « Traitement Informatique de la Connaissance en Chimie Organique », sous la direction de Joël Quinqueton au* LIRMM *à Montpellier. Le système* RÉSYN *a pour objet la planification de synthèses en chimie organique. Une extension de* RÉSYN*, appelée* RÉSYN/RÀPC *a été développée par Jean Lieber, pour intégrer le raisonnement à partir de cas (*RÀPC*) dans* RÉSYN *et ainsi compléter le seul raisonnement par classification utilisé dans* RÉSYN*. Actuellement, c'est le prototype* RÉSYN-ASSISTANT *qui a pris la relève : le système est écrit en Java (pour des problèmes de portabilité) et reprend une bonne partie de développements effectués sur* RÉSYN*. dans le cadre du* GDR CNRS *1093. L'objectif de* RÉSYN-ASSISTANT *est de proposer une aide à la compréhension des problèmes de synthèse organique. Pour cela, des outils permettant de percevoir des molécules ont été développés, ce qui a conduit à une représentation du problème de la synthèse organique avec une représentation abstraite par blocs. Jusqu'à présent,* RÉSYN-ASSISTANT *était plutôt dédié à la perception des molécules, mais les développements actuels l'oriente vers l'extraction de connaissances dans des bases de données de réactions. Ainsi, les données sur les réactions peuvent être pour* RÉSYN-ASSISTANT *soit saisies manuellement (exploitation de la connaissance des experts), soit extraites automatiquement à partir de fichiers exportés depuis diverses bases de données commerciales de réactions. Cette deuxième possibilité n'était pas présente dans* RÉSYN *et la première était beaucoup plus fastidieuse qu'elle ne l'est désormais dans* RÉSYN-ASSISTANT.

4.2 PALÉTUVIER et CASIMIR/PROTOCOLAIRE : représentation hiérarchique de connaissances

Participants : Benoît Bresson [correspondant], Jean Lieber, Amedeo Napoli.

Résumé :

PALÉTUVIER *est un outil développé en* JAVA *pour la gestion de hiérarchies. Cet outil permet de créer et de mettre à jour une hiérarchie de concepts dans le cadre d'une application donnée, à partir de la description des concepts de cette application et de la relation d'ordre qui lie ces concepts. Des fonctionnalités de création de concepts primitifs ont été développés. Un outil convivial pour visualiser des hiérarchies a été mis au point.*

CASIMIR/PROTOCOLAIRE *est un système d'aide au traitement du cancer du sein qui s'appuie sur* PALÉTUVIER*. Ce système repose sur une représentation hiérarchique d'un protocole de traitement de ce type de cancers. Un concept de cette*

hiérarchie représente une catégorie de tumeurs. À certains de ces concepts sont associés des traitements. Classer une tumeur donnée dans cette hiérarchie permet ainsi d'indiquer les traitements de cette tumeur qui sont donnés par le protocole. La base de connaissances représentant le protocole est stockée dans un format XML. *Une phase de validation de cette base et d'amélioration de l'ergonomie de ce logiciel est en cours. Des versions de* CASIMIR/PROTOCOLAIRE *destinées à un fonctionnement en réseau ont été implantées (versions* CORBA, HTTP *et* applet*). Enfin, un logiciel destiné à l'aide au diagnostic et au traitement pour le cancer de la prostate est également développé, en s'appuyant sur l'architecture de* CASIMIR/PROTOCOLAIRE.

4.3 Un logiciel pour l'interprétation des paysages

Participants : Florence Le Ber [correspondant], Amedeo Napoli.

Résumé :

Un système de reconnaissance de modèles d'organisations territoriales agricoles à partir d'images satellitaires a été réalisé en Y3 (pas de développements nouveaux depuis 1998). Ce système est destiné à aider les agronomes à interpréter les images dans un but de diagnostic et de prévision de l'évolution des territoires. La reconnaissance de modèles s'exprime comme une classification de structures, où les structures sont des ensembles d'objets reliés entre eux. Le système produit une reconnaissance cartographiée, c'est-à-dire qu'il produit une image finale où sont représentées par une même couleur les parties de l'image initiale associées à un même modèle.

Parallèlement, ont été développés des logiciels de simulation : à partir des données d'un territoire et d'un système de production agricole, il s'agit d'organiser l'occupation de l'espace comme pourrait le faire un agriculteur et de produire des cartes possibles d'occupation du sol. Trois modèles ont été implantés : un modèle à base de règles, un modèle multi-agents et un modèle de recuit simulé. Ces trois systèmes sont utilisables pour des objectifs distincts.

4.4 Logiciels pour l'ECBD symbolique

Participants : Amedeo Napoli, Arnaud Simon [correspondant].

Résumé : *Le système d'*ECBD ORPAILLEUR *a été développé à partir d'une application médicale, menée conjointement avec le Registre Lorrain du Cancer de l'Enfant (*RLCE, *Hôpital d'Enfants de Nancy-Brabois). Ce système comprend principalement deux modules de fouille et deux modules de visualisation :*

- *Le module de classification par arbres de décision repose sur l'algorithme Al-FReDO, pour « Algorithme de Fouille dans une Représentation des Données par Objets », qui utilise les techniques classiques de construction d'arbres de décision, ainsi que les principes de l'apprentissage par généralisation, pour traiter des données représentées dans un système de* RCO.

- *le module de classification par treillis repose sur un algorithme incrémental de construction de treillis de Galois, avec exploitation de connaissances du domaine. Des règles expliquant les données peuvent être extraites du treillis.*
  *Un module complémentaire qui s'appuie sur la théorie des « ensembles approximatifs » (*rough sets*) permet de prendre en compte le degré de confiance associé aux règles extraites.*
- *Un module de visualisation permet de visualiser l'organisation des données ainsi que les résultats des différents algorithmes de fouille.*
- *Un module de cartographie adaptable à tout type de cartes est appliqué pour visualiser un point de vue géographique sur les données. Ce module, conçu pour traiter le caractère essentiellement géographique de certaines données du* RLCE*, est utilisé pour mettre en évidence la répartition géographique de facteurs d'étude liés aux données. La répartition obtenue est ensuite comparée à des cartes de répartition de référence pour faire apparaître d'éventuelles corrélations.*

4.5 Logiciel pour l'ECBD numérique

Participants : Florence Le Ber, Jean-François Mari [correspondant].

Résumé : *Dans le cadre de la fouille de données pour les systèmes d'information géographique, nous avons développé des méthodes d'apprentissage et d'inférence fondées sur une modélisation à l'aide des chaînes de Markov d'ordre 2. Les données, fournies par la Direction de la Recherche Agricole et Forestière (*DRAF*) lorraine, ont été traitées à l'aide d'outils dédiés à l'origine à la reconnaissance de la parole.*

*En collaboration avec des experts agronomes de l'*INRA *(station* SAD *de Mirecourt), nous avons écrit un logiciel permettant la construction de modèles de Markov d'ordre quelconque, et leurs apprentissages sur des données temporelles et spatiales. Nous avons particulièrement mis l'accent sur les outils de visualisation qui permettent aux experts agronomes d'évaluer les résultats de la modélisation et de s'approprier la connaissance mise en lumière.*

4.6 Les logiciels pour la fouille de textes

Participants : Fairouz Chakkour, Hacène Cherfi, Arnaud Simon, Yannick Toussaint [correspondant].

Résumé : *En dehors du système* ORPAILLEUR *décrit précédemment, les ressources, outils et environnements utilisés dans le cadre de la fouille de textes sont les suivants :*

- *Étiqueteur de Brill : l'étiqueteur de Brill attribue aux mots d'un texte une fonction grammaticale. Cet outil, initialement prévu pour travailler sur l'anglais a été adapté au français et au traitement de thésaurus par l'*INALF *et*

les membres d'Orpailleur. Il met en œuvre des techniques d'apprentissage statistiques et probabilistes pour construire des règles lexicales et contextuelles utilisées ensuite pour l'étiquetage.

- *Lemmatiseur du français : le lemmatiseur du français produit le lemme d'une forme fléchie (développé en collaboration avec Fiammetta Namer, Université de Nancy 2).*
- *Classification statistique : issu des recherches menées à l'*INIST, SDOC *est un outil de classification statistique s'appuyant sur la méthode des mots associés et utilisant l'*indice d'équivalence.
- CLASSIC *: une logique de descriptions qui est utilisée dans le cadre de la généralisation inductive.*
- *Des corpus : le projet* ILIAD *nous a amené à constituer un corpus de textes. Ce sont environ 10 000 résumés d'articles scientifiques en provenance de la base Pascal de l'*INIST *qui ont ainsi été collectés.*

4.7 Les logiciels pour le traitement de l'information

Participants : Rim Al Hulou, Amedeo Napoli, Emmanuel Nauer [correspondant].

Résumé : *Deux systèmes principaux de traitement de l'information sont actuellement en cours de développement. Un système générique de traitement d'informations et de données brutes — en fait une boîte à outils composée d'un ensemble de modules — est actuellement en cours de développement. Le système baptisé « Moteur de Recherche, Filtrage et Classification d'Informations sur Internet », dont la finalité est l'aide à la navigation et à la recherche d'information sur le Web, repose sur un choix particulier d'assemblage de modules. Les modules proviennent de différents horizons. La boîte à outils* DILIB*, qui est une plate-forme dédiée au traitement de l'information reposant sur le format* SGML*, a fourni un certain nombre de modules. D'autres modules nécessaires à des traitements spécifiques ont été développés de façon* ad hoc *: un module de mise en corrélation de descripteurs de langues différentes dans des notices multilingues, un module de classification par treillis de documents suivant un treillis de concepts, un module de normalisation des auteurs, et, actuellement en cours de développement, un module de normalisation des descripteurs dans un contexte multi-bases. D'autres modules encore proviennent du réseau — lemmatiseur, grapheur — ou sont directement utilisables sur le réseau (moteurs de recherche, service de traduction, etc.).*

Un système dont la finalité est la prise en compte et la manipulation de données semi-structurées est développé dans le cadre de l'intégration de bases de données et la résolution de problèmes dans le domaine des données. Les données sont essentiellement des documents textuels, décrits en XML*. Dans un tel cadre, le langage* XML *sert de support à la description des documents tandis que la logique de descriptions* CICLOP *(et parallèlement le langage* OIL*) permet de mettre en œuvre des raisonnements par classification et d'exploiter des connaissances du domaine, pour l'aide*

à la navigation et recherche d'informations dans des bases de données hétérogènes, la classification de requêtes et le traitement de requêtes analogues.

5 Actions régionales, nationales et internationales

5.1 Deux actions locales

5.1.1 Collaboration URI et Orpailleur

Participants : Rim Al Hulou, Dominique Besagni [INIST], Benoît Bresson, Fairouz Chakkour, Hacène Cherfi, Claire François [INIST], Luc Grivel [INIST], Florence Le Ber, Jean Lieber, Jean-François Mari, Bernard Maudinas [INIST], Amedeo Napoli, Emmanuel Nauer, Xavier Polanco [INIST], Ivana Roche [INIST], Jean Royauté [INIST], Arnaud Simon, Rafik Taouil, Yannick Toussaint.

La collaboration entre l'équipe URI (Unité de recherches et d'innovation) de l'INIST et le groupe Orpailleur cherche à mettre à profit la spécificité et les contextes propres aux deux équipes pour faire avancer les recherches et le développement de logiciels dans le cadre de l'analyse de l'information scientifique et technique. Les finalités et la valorisation de la collaboration portent essentiellement sur la rédaction commune d'articles scientifiques et la mise en œuvre opérationnelle des recherches dans le contexte de l'INIST. Des contacts permanents existent entre les deux équipes, globalement et individuellement. Parmi les thèmes principaux qui intéressent cette collaboration se trouvent l'ECBD et plus particulièrement la fouille de textes. Plus précisément, des travaux sont en cours de développement sur un certain nombre de points dont :

- L'étude des stratégies d'interrogation de grandes bases de données textuelles et l'élaboration d'une typologie de requêtes.
- La prise en compte de données semi-structurées provenant de bases de données textuelles hétérogènes.
- L'étude et la mise en œuvre d'une méthodologie pour la fouille de textes, avec l'extraction et l'analyse de structures prédicatives et l'utilisation du système NEURODOC pour l'ECBD.
- L'étude de XML comme une plate-forme intermédiaire pour la description de documents textuels (scientifiques et techniques), en vue d'une manipulation intelligente de ces documents dans l'environnement d'un système de RCO.

Par ailleurs, un autre projet commun vient d'être initié, concernant la réalisation d'une plate-forme expérimentale d'analyse de l'information textuelle. L'objectif est de pouvoir inclure dans cette plate-forme les différents outils de traitement des textes et de classification.

5.1.2 La collaboration READ et Orpailleur

Participants : Fairouz Chakkour, Hacène Cherfi, Yannick Toussaint.

Le projet *Citations* vise à traiter automatiquement les références bibliographiques des articles scientifiques qui ont été numérisées. L'objectif est donc d'utiliser conjointement des méthodes linguistiques simples et des règles d'agglomération pour aider à la segmentation des références et retrouver les champs bibliographiques de chacune des entrées.

5.2 Actions nationales

5.2.1 GDR TICCO 1093 CNRS

Participants : Sandra Berasaluce, Jean Lieber, Amedeo Napoli.

Le GDR CNRS 1093 TICCO — *Traitement informatique de la connaissance en chimie organique* — réunit des chercheurs en chimie organique du CCIPE à Montpellier, des chercheurs en informatique du LIRMM (Montpellier) et du LORIA et des chercheurs de l'industrie pharmaceutique (Sanofi-chimie, Roussel-Uclaf, et Institut de Recherches Servier entre autres). L'objectif du GDR est l'étude et la mise en œuvre de systèmes d'aide à la planification de synthèses de molécules avec comme base informatique les systèmes de RCO, le raisonnement par classification et le RÀPC. Ce travail nécessite des recherches sur une représentation des objets de la chimie organique, une représentation des plans de synthèses de molécules, une modélisation des raisonnements élémentaires et des stratégies de synthèse employés par les chimistes pour résoudre un problème de synthèse.

Le travail de thèse de Sandra Berasaluce, co-encadré par Claude Laurenço (CCIPE et LIRMM Montpellier), entre dans le cadre du GDR TICCO. À ce titre, Sandra Berasaluce peut bénéficier de la double expertise chimie et informatique, et le GDR TICCO offre un environnement idéal pour ce travail de recherches bidisciplinaire.

5.2.2 Projet Casimir

Participants : Benoît Bresson, Jean Lieber, Amedeo Napoli.

Le projet Casimir (Conception continue d'un savoir casuel) vise à élaborer un système qui fournisse une aide à la décision thérapeutique pour la prise en charge de malades souffrant d'un cancer du sein, ainsi qu'une aide au suivi de l'évolution des règles d'actions prises pour soigner les malades. Ce projet s'articule autour de deux champs de recherches d'actualité : le raisonnement à partir de cas et la mémoire organisationnelle. La conception de ce type de mémoire est vue comme une activité de conception portant sur le savoir mis en œuvre, donc ici les protocoles de traitement. Cela suppose que le savoir préexiste et qu'il est conservé dans une mémoire. Un des objectifs premiers du projet Casimir est de collecter ce savoir puis de le représenter sous une forme informatique réutilisable (dans un système à base de connaissances par exemple). L'objectif de la construction d'une mémoire organisationnelle n'est pas seulement de collecter et d'expliciter les savoirs, mais aussi d'élaborer à partir de ces savoirs une réflexion sur l'activité fonctionnelle liée à ces savoirs, pour les analyser et les faire évoluer.

Pour l'instant, deux phases de ce travail peuvent être considérées comme achevées, et une troisième est en cours de réalisation. La première est une étude théorique sur l'apprentissage à partir d'échecs, pour engendrer des explications devant servir dans les mises à jour des règles d'actions. Le protocole de traitement évolue : son utilisation à un instant donné peut conduire à des décisions erronées, car obsolètes, ou encore à des impasses, avec obligation d'adapter le protocole, compte tenu de l'état actuel des connaissances en cancérologie du sein. L'apprentissage à partir d'échecs, à travers une analyse de la décision erronée, permet de faire évoluer le protocole.

La deuxième phase de ce travail est applicative : pour pouvoir raisonner avec le protocole et le faire évoluer, il faut le connaître et le représenter informatiquement. C'est l'application CASIMIR/PROTOCOLAIRE, décrite dans ce document, qui permet de le faire. Actuellement, CASIMIR/PROTOCOLAIRE est sur le point d'entrer dans une phase opérationnelle et devrait pouvoir être utilisé par les oncologues de la région Lorraine.

La troisième phase est l'étude de l'adaptation du protocole de traitement du cancer du sein aux situations dans lesquelles l'utilisation « littérale» de cette base de connaissances n'est pas satisfaisante. L'objectif est l'implantation d'un système de raisonnement à partir de cas — CASIMIR/NON PROTOCOLAIRE — permettant de réaliser ces adaptations. Actuellement, l'acquisition et la modélisation des connaissances d'adaptation nécessaires à un tel système est en cours.

Deux présentations (en français et en anglais) de CASIMIR/PROTOCOLAIRE (tel qu'il existe) et de CASIMIR/NON PROTOCOLAIRE (tel qu'il est envisagé) sont parues dans [13] et [6].

5.2.3 Collaboration sur le thème du RàPC (Universités de Lyon 1 et Lyon 3)

Participants : Béatrice Fuchs [IAE-MODEME, Université de Lyon 3], Jean Lieber, Alain Mille [LISI, Université de Lyon 1], Amedeo Napoli.

Dans le cadre du RàPC, l'étape d'adaptation joue un rôle central. C'est malheureusement une étape très peu modélisée dans la littérature. Des modèles ont été proposés parallèlement dans l'équipe Orpailleur et dans l'équipe de recherches dirigée par Alain Mille (professeur des Universités depuis cette année, au LISI à l'Université Claude Bernard Lyon 1). Une collaboration s'est engagée avec Alain Mille et Béatrice Fuchs (maître de conférences à l'Université de Lyon 3, équipe IAE-Modeme), pour confronter ces modèles de l'adaptation et les enrichir par nos expériences respectives. Ce travail a conduit à un premier modèle qui s'appuie sur deux idées principales. La première est le fait de considérer un cas comme un chemin dans un espace de recherches, ce qui permet de bénéficier des recherches en planification à partir de cas. La seconde est de décomposer la relation entre le problème à résoudre et le problème dont on connaît une solution, de façon à décomposer la tâche complexe de l'adaptation en sous-tâches plus simples.

Dans la continuité de ces recherches, un algorithme d'adaptation générique a été proposé dans [16]. Il s'appuie sur les notions d'appariement entre problèmes et de dépendance entre un problème et la solution qui lui est associée. Bien que cet algorithme repose sur une représentation très simple des problèmes et des solutions (par des n-uplets de réels), il peut se généraliser à des représentations plus complexes.

5.2.4 Collaboration avec le Musée de La Villette

Participants : Jean-Charles Lamirel [CORTEX], Arnaud Simon, Yannick Toussaint.

La Cité des Sciences et de l'Industrie de La Villette possède des collections muséologiques très riches d'objets relatifs à l'histoire des sciences et de l'industrie. Les objets des collections sont utilisés pour l'organisation d'expositions par la Cité des Sciences mais aussi par d'autres musées nationaux dans le cadre d'expositions temporaires. Une partie de ces objets est inventoriée dans une base de données relationnelle dans laquelle les possibilités d'accès aux objets se limitent à leurs numéros d'ordre. Les informations stockées dans la base concernent uniquement le suivi des restaurations et des prêts.

Le projet de collaboration avec le Musée de La Villette repose sur deux constats. D'une part les responsables d'expositions aimeraient avoir accès à l'information des collections. D'autre part les collections sont en réalité des «mines» de connaissances relatives à l'histoire des sciences et de l'industrie. Ainsi, l'objectif de ce projet est de compléter la base de données existante par les descriptions détaillées des objets, et de représenter les connaissances du conservateur, expert en histoire des sciences, grâce à un système de RCO. Le couplage entre la base de données et le système de RCO devrait assurer une meilleure gestion des objets de la base — gestion des données et des connaissances — en favorisant le filtrage et la classification par points de vue des objets, et donner ainsi une meilleure appréhension de la base et de son contenu. De plus, le couplage permet d'exploiter le système d'ECBD ORPAILLEUR pour découvrir et expliquer des corrélations entre les objets, mieux comprendre l'évolution des objets au cours du temps, et faire émerger de nouveaux thèmes d'exposition.

5.2.5 Fouille de données et systèmes d'information géographique — Collaboration avec l'INRA

Participants : Florence Le Ber, Jean-François Mari.

Une application des modèles de Markov d'ordre 1 et 2 a été mise en œuvre pour la reconnaissance de successions culturales. Cette étude se propose d'étudier les successions culturales pratiquées en Lorraine depuis une dizaine d'années, afin d'intégrer cette connaissance dans un modèle d'organisation spatiale de territoires agricoles en cours de développement à l'INRA. Ce projet, après avoir été retenu en 1998 dans l'appel d'offres «Programme de systèmes d'information géographique» initié par le CNRS et l'IGN dans le cadre du GDR Cassini, s'est vu prolonger en 1999 dans le second appel d'offres du même programme.

5.2.6 ARC A3-ILEC de l'AUF (AUPELF-UREF)

Participants : Fairouz Chakkour, Hacène Cherfi, Yannick Toussaint.

Nous participons aux travaux de l'action de recherche concertée ARC A3 sur la construction automatique de terminologies et de relations sémantiques entre termes à partir de corpus. Notre objectif est de participer à l'évaluation des outils qui ont été développés dans le cadre du projet ILIAD, et à la comparaison avec d'autres approches. Une première phase de l'ARC A3 a consisté

à évaluer des extracteurs terminologiques. La seconde phase, qui est toujours en cours, concerne plus particulièrement les comparaisons avec des approches similaires.

5.2.7 ARC INRIA Escrire

Participants : Rim Al Hulou, Hacène Cherfi, Olivier Corby [ACACIA SOPHIA ANTIPOLIS], Rose Dieng [ACACIA, INRIA SOPHIA ANTIPOLIS], Jérôme Euzenat [EXMO, INRIA RHÔNE-ALPES], Amedeo Napoli, Emmanuel Nauer, Yannick Toussaint.

Cet ARC INRIA se fait en collaboration avec le projet ACACIA (Rose Dieng, INRIA SOPHIA-ANTIPOLIS) et le projet EXMO (Jérôme Euzenat, INRIA RHÔNE-ALPES).

Un intranet, et plus généralement, l'utilisation des technologies de l'Internet, sont des opportunités pour les entreprises d'accéder et de partager des connaissances souvent difficilement accessibles sous forme documentaire. Les documents numériques et numérisés peuvent être rendus accessibles de manière standard et transparente auprès de tous les utilisateurs concernés. L'ambition, à terme, est de réaliser de véritables serveurs de connaissances permettant la recherche et la manipulation des ressources de l'entreprise. Cependant, les limites de cette approche apparaissent rapidement : l'organisation des sites se révèle une tâche coûteuse et la recherche plein texte peu efficace.

La recherche et l'interrogation d'un site en s'appuyant sur le contenu des documents devient dès lors une nécessité, et les formalismes de représentation de connaissances sont de bons candidats pour représenter ce contenu. La représentation du contenu peut permettre de manipuler ce contenu pour faire de la recherche par analogie, par spécialisation, par similitude, etc. Il existe différents formalismes de représentation des connaissances, mais aucune étude importante et poussée mettant en jeu la représentation et la manipulation de documents textuels n'a encore eu lieu, pour permettre la comparaison de leurs qualités respectives selon un tel point de vue.

Le but de l'ARC Escrire consiste donc à comparer trois types de représentations de connaissances — graphes conceptuels (GC), représentations de connaissances par objets (RCO) et logiques de descriptions (LD) — du point de vue de la représentation du contenu de documents et de leur manipulation. Pour cela, l'action s'appuie sur les compétences dans chacune des représentations des projets ACACIA (GC), EXMO (RCO) et Orpailleur (LD) respectivement. L'objectif de l'action consiste à comparer les apports de chacun des types de représentation pour la représentation du contenu dans les serveurs de connaissances.

La mise à l'épreuve de ces différents formalismes pour le traitement d'un jeu de documents, en l'occurrence des documents textuels sur le génome, nécessite de mener une réflexion méthodologique sur le passage des textes à leur représentation formelle (de façon suffisamment indépendante des formalismes employés) en lien avec le type d'accès que l'on veut avoir sur ces documents. Cette représentation formelle est définie conjointement et introduite dans un format XML. Un ensemble de requêtes définies de manière coordonnée doit être évaluée dans chacun des contextes.

À l'issue de ce travail, les différents formalismes seront comparés entre eux (mais aussi à la recherche plein-texte) selon le protocole prédéfini. Celui-ci devra apprécier des critères tant qualitatifs (expressivité des requêtes, accessibilité/lisibilité des informations, etc.) que

quantitatifs (temps de réponse à une requête, taux de précision/rappel des réponses, etc.). Cette évaluation proposera une grille d'analyse des avantages et inconvénients d'un langage de représentation formel vis-à-vis de la recherche d'informations sur le Web, et tentera de déterminer les contextes favorables à l'exploitation de chacune de ces représentations.

5.3 Actions internationales

5.3.1 Action Intégrée ECOS-CONICYT avec le Chili

Participants : Xavier Polanco [INIST], Jean Royauté [INIST], Yannick Toussaint.

En association avec l'équipe URI de l'INIST, nous avons proposé, en 1998, puis mis en place une Action Intégrée (PAI) avec deux universités chiliennes, la *Universidad de Concepción* (contacts : J. Atkinson et A. Ferreira) et la *Universidad de Chile* (contact : A. Bassi). La première année a été consacrée à la constitution de ressources sur la langue espagnole : corpus de textes et lexiques. En 1999, la deuxième année de coopération nous a permis d'adapter à l'espagnol les modules linguistiques d'étiquetage et de lemmatisation de la plate-forme ILC, plate-forme d'analyse de l'information basée sur l'extraction de termes à partir de textes et sur la classification développée par les partenaires français en 1997–1998. La troisième année a été consacrée à l'intégration des outils et un travail de recherche sur la prise en compte des aspects sémantiques de la langue a été initié.

5.3.2 Action intégrée Balaton

Participants : Katalin Bognar [Université Kossuth Lajos, Debrecen, Hongrie], Florence Le Ber, Amedeo Napoli, Emmanuel Nauer.

Un système à objets pour la représentation et la manipulation de structures

Ce projet — PAI BALATON — se fait en collaboration avec l'Université Kossuth Lajos à Debrecen (Hongrie), où notre contact est Katalin Bognar, enseignant-chercheur à l'institut de mathématiques et d'informatique.

Les objectifs scientifiques de ce projet sont de concevoir un système de RCO adapté à la représentation et à la manipulation de structures. Un tel système peut être utilisé pour la représentation de structures spatiales, textuelles ou moléculaires par exemple. Les structures sont vues comme des objets composites, dont les composants sont liés entre eux par des relations vérifiant certaines contraintes. À l'heure actuelle, il n'existe pas de langage de référence pour la représentation et la manipulation de structures. En partant de notre expérience sur les systèmes de RCO et la représentation et la manipulation de structures spatiales et moléculaires, notre but est de concevoir un système générique de représentation et de manipulation de structures. Une structure est considérée comme un graphe étiqueté dont les sommets et les arêtes sont représentées par des classes (munies d'attributs et de méthodes) dans l'univers d'une RCO. À chaque classe est associé un ensemble de relations, qui modélisent des contraintes inter-attributs. La prise en compte et la mise en œuvre d'un tel ensemble de relations est une extension du modèle classique des représentations de connaissances par objets. La manipulation de telles structures peut se faire en utilisant le raisonnement par classification et un certain nombre de variantes de ce mode de raisonnement, qui doivent être étudiées en détail dans

le cadre de cette action intégrée Balaton. Une première publication rassemble une partie des résultats obtenus au cours de cette collaboration [23].

6 Diffusion de résultats

6.1 Animation de la Communauté scientifique

- Actions internationales avec le Chili et la Hongrie.
- Participation à des groupes de travail nationaux (GDR).
- Participation à des comités de lecture de revues, à l'organisation de numéros spéciaux de revues et à l'édition d'ouvrages de recherche.
- Organisation de colloques et participation à des comités de programme.

Une journée sur l'adaptation s'est déroulée le 15 mai 2000 dans les locaux de l'INRA de Champenoux et a réuni des chercheurs en intelligence artificielle et en reconnaissance de formes. Elle a fait suite aux journées sur la classification et sur l'apprentissage qui s'étaient déroulées dans des contextes similaires.

Cette journée a été organisée sur le constat suivant : le terme «adaptation» est utilisé dans des domaines variés : en raisonnement à partir de cas, dans les réseaux neuromimétiques, dans les modèles statistiques, dans les systèmes multi-agents, en biologie (reconstruction phylogénétique), etc. Il a donc paru intéressant de réunir, lors d'une même journée, différentes personnes utilisant ce terme dans leurs recherches. Le but était de passer en revue des concepts et méthodes rattachés à ce terme, afin de tenter d'en dégager des points communs et des différences. Cette journée de prospection a montré que la notion d'adaptation, comme thème de recherches, est en plein essor et ouvre des perspectives prometteuses. De plus, elle peut susciter des liens entre différents domaines de recherches. Les actes de cette journée réunissent des textes relatifs aux présentations [28].

6.2 Enseignement

- Enseignements et organisation scientifique de cours (en France et à l'étranger).
- Proposition en lien avec l'Université de Nancy 2 d'un nouveau DESS TEXTE pour «Traitements informatiques pour l'EXploitation de l'information dans les TExtes». La demande d'habilitation du DESS est en cours d'étude au Ministère.
- Encadrements de thèses, DEA, stages de DESS, étudiants de l'université, des écoles d'ingénieurs et de l'IUT.
- Participation à des jurys de thèse et de HDR.

7 Bibliographie

Articles et chapitres de livre

[1] N. Capponi, Y. Toussaint, «Interprétation de classes de termes par généralisation de structures prédicat-arguments», *in : Ingénierie des connaissances, évolutions récentes et nouveaux défis*, G. K. e. D. B. J. Charlet, M.Zacklad (éditeur), Eyrolles, janvier 2000, p. 337–357.

[2] M. Huchard, R. Godin, A. Napoli, «Objects and Classification: A natural convergence», *in : Workshop Reader, European Conference on Object-Oriented Programming, Sophia-Antipolis*, J. Malenfant et S. Moisan (éditeurs), *Lecture Notes in Artificial Intelligence 1964*, Springer, Berlin, 2000.

[3] S. Jolibois, E. Nauer, D. Chouanière, J. Ducloy, F. Grandjean, M. Mouzé-Amady, «Adaptation des normes et formats documentaires à la gestion informatisée de corpus bibliographiques», *Bulletin des Bibliothèques de France 1*, 2000.

[4] S. Jolibois, E. Nauer, D. Chouanière, J. Ducloy, F. Grandjean, M. Mouzé-Amady, «L'Unified Medical Language System (UMLS) : une base de connaissances multilingue dans le domaine biomédical», *Documentaliste - Sciences de l'information 2*, juin 2000.

[5] F. Le Ber, L. Mangelinck, A. Napoli, «Représentation de relations et classification de structures spatiales», *Revue d'intelligence artificielle 13*, 2, décembre 1999, p. 441–467.

[6] J. Lieber, B. Bresson, «Case-Based Reasoning for Breast Cancer Treatment Decision Helping», *in : Advances in Case-Based Reasoning — Proceedings of the fifth European Workshop on Case-Based Reasoning (EWCBR-2k)*, E. Blanzieri et L. Portinale (éditeurs), *Lecture Notes in Artificial Intelligence 1898*, Springer, 2000, p. 173–185.

[7] J. Lieber, A. Napoli, «Planification à partir de cas et classification», *in : Ingénierie des connaissances – Évolutions récentes et nouveaux défis*, J. Charlet, M. Zacklad, G. Kassel, et D. Bourigault (éditeurs), *collection technique et scientifique des télécommunications*, Eyrolles, 2000, p. 357–369, Le livre dans lequel est paru cet article est un recueil d'articles paru dans les actes des conférences IC (ingénierie des connaissances). En particulier, cet article est également paru dans les actes de la conférence IC-97.

[8] A. Napoli, J. Euzenat, R. Ducournau, «Les représentations des connaissances par objets», *Technique et science informatiques 19*, 1-2-3, 2000, p. 387–394.

[9] A. Simon, A. Napoli, «Un algorithme de fouille dans une représentation des données par objets : une application au domaine médical», *in : Ingénierie des connaissances — Évolutions récentes et nouveaux défis*, J. Charlet, M. Zacklad, G. Kassel, et D. Bourigault (éditeurs), Eyrolles, Paris, 2000, p. 195–207.

Communications à des congrès, colloques, etc.

[10] R. Al Hulou, A. Napoli, E. Nauer, «XML : un formalisme de représentation intermédiaire entre données semi-structurées et représentations par objets», *in : Langages et Modèles à Objets (LMO'00), Montréal*, C. Dony, H. Sahraoui (éditeurs), Hermès, p. 75–90, Paris, 2000.

[11] Y. Bastide, R. Taouil, N. Pasquier, G. Stumme, L. Lakhal, «Levelwise search of frequent patterns with counting inference», *in : 16èmes Journées Bases de Données Avancées (BDA'00), Blois*, A. Doucet (éditeur), p. 307–322, 2000.

[12] A. Belaïd, Y. Toussaint, « Une méthode d'étiquetage morpho-syntaxique pour la reconnaissance de tables de matières», *in : Colloque International Francophone sur l'Ecrit et le Document - CIFED'00*, juillet 2000.

[13] B. Bresson, J. Lieber, «Raisonnement à partir de cas pour l'aide au traitement du cancer du sein», *in : Actes des journées ingénierie des connaissances*, p. 189–196, 2000.

[14] F. Chakkour, A. Napoli, Y. Toussaint, «Le raisonnement à partir de cas pour l'identification de rôles sémantiques dans des énoncés en langue naturelle», *in : Actes du séminaire RàPC-2000, Toulouse*, Rapport IRIT/00-11-R, IRIT, Toulouse, p. 1–6, 2000.

[15] F. Chakkour, « Le raisonnement à partir de cas pour l'analyse conceptuelle des énoncés en langue naturelle», *in : Cinquièmes rencontres nationales des Jeunes Chercheurs en Intelligence Artificielle (RJCIA-2000)*, M. Ayel, J.-M. Fouet (éditeurs), p. 39–53, Universite Claude Bernard - Lyon 1, 2000.

[16] B. Fuchs, J. Lieber, A. Mille, A. Napoli, « An Algorithm for Adaptation in Case-Based Reasoning», *in : Proceedings of the 14th European Conference on Artificial Intelligence (ECAI-2000), Berlin, Germany*, p. 45–49, 2000.

[17] J.-C. Lamirel, Y. Toussaint, «Combining symbolic and numeric techniques for DL contents classification and analysis», *in : First DELOS Workshop on Information seeking, searching and querying in Digital Librairies*, décembre 2000, `http://www.loria.fr/publications/2000/A00-R-410/A00-R-410.ps`.

[18] F. Le Ber, C. Brassac, « Objets graphiques et cognition située et distribuée. Un exemple en acquisition de connaissances», *in : Représentations graphiques dans les systèmes complexes naturels et artificiels. Journées de Rochebrune*, ENST, janvier 2000.

[19] F. Le Ber, L. Mangelinck, A. Napoli, «Un système de reconnaissance d'organisations spatiales agricoles sur images satellitaires», *in : Reconnaissance des Formes et Intelligence Artificielle - RFIA'2000, Paris*, R. Deriche, M.-C. Rousset (éditeurs), *I*, AFRIF-AFIA, p. 119–128, février 2000.

[20] F. Le Ber, A. Napoli, « Representation of spatial relations and structures in object-based knowledge representation systems», *in : ECAI Workshop on Current Issues in Spatio-Temporal Reasoning, Berlin*, août 2000.

[21] J. Lieber, « Composition et décomposition de l'adaptation dans le cadre du raisonnement à partir de cas», *in : Acte de la journée sur l'adaptation, INRA de Champenoux*, F. L. Ber, J. Lieber (éditeurs), p. 31–36, 2000.

[22] J.-F. Mari, F. L. Ber, M. Benoît, « Fouille de données agricoles par Modèles de Markov cachés», *in : Journées francophones d'ingénierie des connaissances, IC2000, Toulouse, France*, p. 197–205, 2000.

[23] A. Napoli, F. Le Ber, K. Bognar, «Une proposition pour la représentation de structures dans un système de représentation de connaissances par objets», *in : Actes de IC'2000 – Ingénierie des Connaissances, Toulouse*, P. Tchounikine, N. Aussenac-Gilles (éditeurs), IRIT – Université Paul Sabatier, p. 145–152, 2000.

[24] E. Nauer, R. Al-Hulou, A. Napoli, « What can XML do for information retrieval on the WEB? », *in : Proceedings of the 7th International Workshop on Knowledge Representation meets DataBases (KRDB-2000), ECAI-2000, Berlin*, M. Bouzeghoub, M. Klush, W. Nutt, U. Sattler (éditeurs), p. 101–114, 2000.

[25] A. Simon, Y. Toussaint, «Building and interpreting term dependencies using association rules extracted from Galois lattices», *in : RIAO'2000*, 2000.

[26] Y. Toussaint, A. Simon, H. Cherfi, «Apport de la fouille de données textuelles pour l'analyse de l'information», *in : Ingénierie des connaissances - IC'2000, Toulouse, France*, mai 2000.

[27] Y. Toussaint, A. Simon, « Building and interpreting term dependencies using association rules extracted from Galois Lattices », *in : Recherche d'Informations Assistée par Ordinateur - Content-Based Multimedia Information Access, Paris*, avril 2000.

Rapports de recherche et publications internes

[28] F. Le Ber, J. Lieber, «Actes de la journée sur l'adaptation», *Rapport de recherche*, novembre 2000.

[29] F. Le Ber, L. Mangelinck, A. Napoli, « Design and comparison of lattices of topological relations : Application to satellite image understanding», *Rapport de recherche*, octobre 2000.

[30] F. Le Ber, L. Mangelinck, A. Napoli, «Treillis de relations topologiques et reconnaissance de structures spatiales sur images satellitaires», *Rapport de recherche*, novembre 2000.

[31] J.-L. Metzger, «Acquisition, modélisation et représentation de connaissances pour le raisonnement à partir de cas. Application à l'étude d'organisations spatiales agricoles», *Rapport de DEA n° A00-R-121*, 2000.

Divers

[32] A. Simon, « Outils classificatoires par objets pour l'extraction de connaissances dans des bases de données», Thèse de l'université Henri Poincaré Nancy 1, 2000.