INRIA

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Action READ

Reconnaissance de l'Ecriture et Analyse de Documents

Nancy

THÈME 3A

Table des matières

READ est une équipe du LORIA et une action INRIA depuis le 1er janvier 1997

1 Composition de l'équipe

Responsable scientifique

Abdel Belaïd [Chargé de recherche au CNRS]

Assistante de projet

Isabelle Herlich [à temps partiel]

Ingénieur expert

François Parmentier [du 1/8/98 au 31/10/1999]

Olivier Lescop [du 15/11/1998 au 31/3/1999]

Personnel Université Nancy 2

Yolande Belaïd [Maître de conférences]

Amos David [Maître de conférences (adhésion à l'équipe depuis le 18 octobre 1999)]

Chercheur doctorant

Christophe Choisy [allocataire MESR]

Ingénieurs invités ou associés

Norbert Valverde [Ingénieur ITESOFT, du 1er février 1999 jusqu'au 1er février 2000]

Laurent Pierron [Ingénieur Loria, du 1er mars 1999 au 30 mars 1999, puis du 1 octobre 1999 au 31 décembre 2000]

Stagiaires

Samia Maddouri [Doctorante tunisienne, 15 août au 15 septembre 1999]

Emilie Pérales [Doctorante tunisienne, 15 août au 15 septembre 1999]

Sébastien Jacob [IUT informatique, 1er avril au 31 juillet 1999]

Vincent Kneip [IUT informatique, 1er avril au 15 juin 1999]

Sébastien Ferry [Maîtrise informatique, du 1er juillet 1999 au 31 août 1999]

Nicolas Galmiche [Maîtrise informatique, stagiaire payé par la société ITESOFT, depuis le 1er juillet 1999]

2 Présentation et objectifs généraux

Notre objectif est de proposer des modèles et des techniques de reconnaissance autour de la communication écrite et de la recherche d'information.

La première partie de nos travaux concerne la reconnaissance du script manuscrit. La caractéristique principale de la recherche dans ce domaine est de prendre en compte le double ancrage du sens de la communication écrite : linguistique et spatial. L'analogie est souvent faite avec la parole qui partagent la caractéristique d'être chacune une forme d'expression temporelle de la langue. Cependant, l'écrit se distingue par sa nature fondamentalement 2D, rendant moins performante une analyse linéaire du signal. Les techniques de reconnaissance explorées dans READ sont de type stochastique bidimensionnel bien connus pour leurs propriétés d'absorption du bruit et d'incorportaion du contexte lexical.

La reconnaissance de l'écriture voit aujourd'hui son domaine s'étendre à l'analyse de documents sous l'influence du progrès enregistré par l'édition électronique et les réseaux de communications en ligne. L'aspect spatial est encore plus difficile que dans le cas de l'écriture car l'information est répartie dans une structure de mise en page très complexe de laquelle il faut extraire une structure logique. Cette complexité est encore accrue quand les documents ne sont pas normalisés comme les documents commerciaux ou les formulaires. La seconde partie de notre thématique tente d'apporter une réponse à cette difficulté par la proposition de méthodes de segmentation adéquates, de formalismes génériques de modélisation des structures de documents en relation avec les normes standards dans des domaines spécifiques (bibliothèques, édition, etc.), et la proposition de technologies adaptatives pour l'analyse des documents hétérogènes.

Enfin, READ s'intéresse depuis peu à l'exploitation des documents par l'utilisateur dans un contexte de recherche d'information (RI). Notre objectif de départ a été de montrer comment les techniques de RI peuvent être employées pour contribuer à une modélisation efficace des documents. Nous travaillons sur des méthodes d'exploitation du contenu par la définition et l'intégration des méthodes d'analyse de l'information et d'un modèle de l'utilisateur.

3 Fondements scientifiques

3.1 Reconnaissance de l'écriture

Mots clés : reconnaissance de l'écriture non-contrainte, modélisation stochastique 2D, modèles et champs de Markov, réseau de neurones transparent.

Participants : Samia Maddouri, Christophe Choisy, Abdel Belaïd.

Résumé : *Nous proposons dans cette section deux techniques de reconnaissance de l'écriture manuscrite hors-ligne. Elles sont toutes deux basées sur un compromis entre une estimation locale des lettres et une vision globale du mot par vérification de la cohérence lexicale. Cette combinaison tente de trouver une solution au dilemme de Sayre : pour reconnaître, il faut segmenter et pour segmenter, il faut reconnaître !*

3.1.1 Approche bidimensionnelle markovienne

La reconnaissance de l'écriture manuscrite non contrainte soulève de nombreux problèmes en raison de la variabilité de l'écriture, accentuée par l'aspect omniscripteur. Deux approches s'opposent en reconnaissance de mots : la première a une vision globale du mot, disposant de beaucoup d'information, elle est robuste aux variations, mais cet aspect généraliste la limite à des vocabulaires distincts et réduits. La seconde permet de s'affranchir de ces limites mais nécessite une interprétation locale et une segmentation correcte du mot. Nos approches se proposent de tirer avantage des deux méthodes, réduisant la complexité de la méthode globale en l'appliquant sur des entités plus petites (lettres), l'approche analytique détermine les meilleures frontières entre les lettres, évitant ainsi une segmentation préalable.

Christophe Choisy a proposé comme estimateur global pour les lettres, une version améliorée des champs de Markov causaux développés par G. Saon dans notre équipe. Le modèle analytique se base sur des modèles de Markov (HMM)s. Ces travaux ont permis de mettre en évidence la diminution de la complexité liée à l'approche analytique. Une estimation de cette réduction fait apparaître un facteur 7 entre cette dernière et l'approche globale. Ce modèle présente une structure très intéressante, ouvrant la possibilité d'un apprentissage croisé des lettres au travers des mots entiers. L'idée de cet apprentissage croisé est d'extraire les informations relatives aux lettres dans des modèles de mots générés automatiquement, puis de combiner ces informations entre des mots de différentes orthographes. Bien que les conditions d'expérimentation ne soient pas parfaites (base de mots limitée), les résultats obtenus se placent très correctement par rapport aux travaux actuels dans la littérature. Ils ont fait l'objet d'un article ([3]).

3.1.2 Approche séquentielle neuronale

Samia Maddouri propose une approche hybride qui assure une complémentarité entre une vision globale par un réseau de neurones et une vision locale par HMM. Le système de reconnaissance se base sur un réseau de neurones à quatre couches (couche des caractéristiques, lettres, parties de mots, et couche mots) qui identifie un mot à partir des caractéristiques structurelles apparentes qui le décrivent à travers deux processus. Le processus ascendant correspondant à la phase de propagation de l'information et le processus descendant qui se traduit par rétropropagation de l'information afin de générer de nouvelles hypothèses pour un nouveau cycle de propagation/rétropropagation. Une correspondance entre les zones de lettres dans l'image d'entrée du mot manuscrit et les lettres dans chaque mot imprimé du lexique apportées par les niveaux inférieurs ainsi qu'une méthode de normalisation et de reconnaissance, est effectuée à la fin de chaque cycle. La sortie est une liste de mots imprimés candidats qui seront classés selon l'ordre décroissant de leur degré d'activation afin de décider sur le résultat de la reconnaissance. La particularité de ce système est le traitement transparent et parallèle de l'information et la propagation progressive des activations entre les niveaux adjacents des neurones. Son avantage est qu'il ne nécessite ni une phase d'apprentissage ni une large base d'exemples.

3.2 Reconnaissance de documents

Participants : Yolande Belaïd, Afef Kacem [correspondant], Norbert Valverde [Ingénieur ITESOFT], Abdel Belaïd.

Mots clés : Technologie adaptative, Traitement de formulaires, Segmentation de documents mathématiques, Authentification de documents, Bibliothèque virtuelle, Partie du dicours.

Résumé :

Dans cette partie, nous décrivons nos travaux portant sur la reconnaissance de documents. Ces travaux concernent : 1) la technologie adaptative en segmentation de documents administratifs ou commerciaux, 2) la segmentation de documents mathématiques, 3) la reconnaissance de sommaires pour la consultation à distance de la documentation; et enfin 4) la distribution du traitement de document pour le télétravail.

3.2.1 Technologie adaptative en traitement de formulaires

L'analyse de formulaires est un domaine en plein essor. De nombreuses sociétés et administrations ont à traiter rapidement le contenu de leurs formulaires tels que des bons de commande, des liasses fiscales, des questionnaires à choix multiples, etc. Quelle que soit l'application, le contenu d'un formulaire est hétérogène. Il contient du texte, des logos et des lignes graphiques empêchant la localisation directe du contenu.

Lors d'une première expérience concernant l'analyse des formulaires de la CNAM, une technique de précalssification a été proposée. Elle repose sur l'extraction des traits par la transformée de Hough et sur la classification du contenu des cellules. Les cellules sont extraites par l'examen des cycles formés par les traits et leur contenu est analysé et classé à l'aide de méthodes neuronales, de type perceptron multi-couche.

Actuellement, ce travail se poursuit en collaboration avec la société ITESOFT sur l'analyse de bordereaux destinés à la vente par correspondance. Il s'agit de concevoir un système générique de localisation de zones d'information telles que les zones de correspondance et de commandes. A cause de la forte variabilité de l'emplacement et de la forme des zones, nous avons choisi de modéliser ces zones à l'aide de graphes de contraintes entre "points fixes". Ces points fixes correspondent à des champs logiques textuels dont la forme reste immuable quelle que soit la mise en page. Les contraintes sont de type topographique. Elles serviront à limiter la recherche à l'espace des zones. La localisation des zones est opérée par constitution des graphes et recherche de consistance d'arcs limitant les candidats aux seuls champs représentatifs des points fixes de chaque zone.

3.2.2 Logique floue en analyse de documents mathématiques

Un document mathématique contient du texte et des formules mathématiques. Contrairement au texte qui a une structure linéaire, les formules obéissent à des règles de structure

spécifiques qui échappent à un lecteur optique. Afin de restituer aux formules la structure planaire, deux solutions sont souvent proposées : reconnaissance de caractères puis restructuration ou bien étiquetage puis reconnaissance. La première solution suppose que le lecteur optique réussit à segmenter les formules et est capable de fournir l'emplacement de chaque caractère. La seconde facilite le travail puisqu'elle segmente la formule en caractères avant de les présenter individuellement à l'O.C.R. Cette méthode évite les procédures de segmentation trop généralistes de l'O.C.R. Etant donné le peu de succès de la première méthode, nous avons expérimenté la seconde en opérant une segmentation adaptative du texte. L'idée est d'effectuer un étiquetage en plusieurs étapes : extraction des lignes, repérage des formules isolées, puis repérage des formules à l'intérieur du texte. La méthode choisie [?] est fondée sur l'étiquetage des symboles mathématiques et l'extension du contexte aux symboles voisins. A cause de la faiblesse des critères typographiques à ce niveau, on utilise la logique floue pour l'étiquetage. Des règles de voisinage structurel sont mises à profit pour propager la structure aux éléments proches tels les indices, les exposants, etc. et délimiter l'espace des formules. Cette extension reste partielle pour les formules isolées du texte dont la délimitation revient à une simple vérification de l'existence de quelques symboles clés. Le taux d'étiquetage primaire des composantes connexes est de l'ordre de 95.3%. Mais leur étiquetage secondaire accroît ce taux d'environ 4%. Les résultats obtenus montrent l'applicabilité de notre système puisque 95% des formules mathématiques sont bien extraites des documents imprimés de bonne qualité.

3.2.3 Reconnaissance de sommaires de revues

Ce travail s'inscrit dans le cadre du projet Calliope en collaboration entre le centre de recherche de Xerox et l'INRIA. Calliope est un projet de bibliothèque électronique qui permet à des chercheurs d'accéder depuis leur station de travail à un ensemble de périodiques scientifiques physiquement stocké sur un site distant. Calliope développe le concept de téléphotocopie, c'est-à-dire la numérisation à la demande d'articles scientifiques et leur impression à distance. La sélection des articles se fait par les tables des matières fournies sous forme électronique par une entité tierce.

Aussi importante que soit cette base de données de tables des matières, de nombreux périodiques n'y figurent pas et ne peuvent être intégrés à Calliope à moins de procéder à une resaisie manuelle toujours longue et fastidieuse.

Nous avons proposé [7] un outil automatique adapté à la reconnaissance de ces tables des matières. Le travail consiste à numériser les tables de matières et à reconnaître automatiquement leurs articles. La méthode de reconnaissance des articles s'appuie essentiellement sur une procédude de marquage linguistique et d'étiquetage des champs auteurs et titres. Nous avons une procédure de marquage du type partie du discours. Elle consiste à affecter à chaque mot une étiquette correspondant à sa catégorie grammaticale, s'il s'agit d'un nom commun, et son type : prénom, initiale ou nom s'il s'agit d'un nom propre. Ensuite, cet étiquetage est étendu aux mots voisins dans le but de corriger les mots non étiquetés (correspondant à des mots erronés, n'ayant pas été correctement reconnus par OCR). Des règles contextuelles sont appliquées de manière récursive.

Le prototype logiciel développé fonctionne de manière très satisfaisante sur des sommaires de formats très différents. Les performances s'élèvent à 98% de bonne reconnaissance pour des

sommaires textuels appartenant à des revues scientifiques.

3.2.4 Distribution du traitement de document pour le télétravail

La réalisation de plate-formes d'analyse de documents en milieu industriel nécessite de prendre en compte plusieurs contraintes relatives d'une part au volume des données (des milliers de documents sont traités par jour), d'autre part à la diversité et à la complexité du traitement des images (compression, stockage, conversion, extraction d'information, ...), et enfin à l'efficacité et à la rapidité des traitements (distribution des tâches, parallélisation, optimisation, ...). Ces tâches deviennent vite insurmontables si l'infrastructure locale ne permet pas une gestion efficace des ressources et des outils.

Dans le cadre d'une collaboration avec le CETIR (Centre Européen des Technologies de l'Information en milieu Rural), nous avons entrepris des recherches sur la distribution du traitement de documents en mettant à profit les possibilités de l'Internet pour la communication et l'échange de documents. Le document n'est plus considéré comme une entité indivisible traitée sur un seul site, mais comme un ensemble d'objets qui peuvent être répartis sur différents sites. Cette vue du traitement du document a nécessité une conception objet des documents et une architecture supportant le traitement réparti.

Les éléments de réflexion ont porté d'une part sur les aspects normalisation des données et des protocoles de transports et d'autre part sur la conceptualisation de la gestion des tâches. La normalisation des documents se fonde sur les standards et recommandations reconnus au niveau international, comme XML pour W3C. De la même manière, les protocoles de transport bénéficient naturellement des infrastructures réseaux existantes, comme les serveurs et protocoles HTTP. Une partie de l'effort a été mise sur la description et la représentation des messages (documents et accompagnements), et sur les modalités d'interaction entre les différents acteurs ou parties du réseau. Par ailleurs, pour permettre l'ouverture du système sur de nouvelles ressources (nouvelle application, par exemple), une conceptualisation des types de données et des outils a été rendue nécessaire. Une association intelligente des deux par l'intermédiaire de graphes conceptuels et de définitions de propriétés a été choisie. Cette spécification s'aligne sur le modèle de la recommandation RDF (Resource Description Framework de W3C). L'aspect distribué d'une telle plate-forme qui sera étudié ultérieurement nécessitera également une standardisation de sa mise en œuvre. On pourra prendre pour modèle CORBA qui définit une architecture pour applications distribuées orientées objet. L'application visée concerne le vidéocodage pour la correction à distance de vignettes de montants de bons de commande.

3.3 Recherche d'information

Participants : Amos David, David Bueno [correspondant], Vincen Kneip, Abdel Belaïd.

Mots clés : Recherche d'information, Recherche coopérative d'information, Modélisation de l'utilisateur, Personnalisation des réponses.

Résumé : *Cette section porte d'une part sur l'emploi des techniques de recherche d'information dans un processus de reconnaissance de documents et d'autre*

part sur l'étude de l'apport de la reconnaissance aux systèmes de recherche d'information. En l'occurrence, il s'agit d'utiliser des requêtes pour assister à l'identification des structures génériques d'une classe de documents.

3.3.1 Modélisation générique des structures de documents

Dans ce travail, nous avons posé le problème de la modélisation automatique de documents par la combinaison d'outils de recherche d'information (RI) et des méthodes classiques utilisées en apprentissage. Aussi, notre objectif de départ a été de montrer comment les techniques utilisées en RI peuvent être utilisées pour améliorer les performances des processus de reconnaissance.

Notre recherche a été centrée sur les types de requêtes et les résultats qui peuvent être obtenus à partir d'un système de RI [2, 4]. Notre approche a été de montrer comment représenter la connaissance sur les structures de document en forme d'une base de données qui nous permet d'appliquer ensuite des fonctions de RI pour accéder à la connaissance sur les structures de documents. Deux catégories de structures ont été considérées : la première concerne l'organisation de la structure souvent disponible sous forme normalisée, comme le format BibTeX pour les références bibliographiques ou SGML pour les documents structurés. La seconde catégorie concerne la mise en page qui est généralement associée à la structure physique.

3.3.2 Personnalisation des réponses et recherche coopérative dans les SRI

La pertinence d'une réponse fournie par un système de recherche d'information (SRI) concerne l'exactitude de la réponse par rapport à la requête de l'utilisateur et son adéquation par rapport au niveau de connaissance et aux préférences de l'utilisateur. Le premier type de pertinence reçoit un intérêt particulier depuis les premiers SRI. Le deuxième type de pertinence est beaucoup plus difficile à traiter car il s'agit d'adapter le système aux spécificités de chaque utilisateur.

Pour personnaliser les réponses d'un SRI, nous avons choisi de modéliser l'utilisateur. Ce choix consiste à représenter explicitement chaque utilisateur dans le système selon ses comportements et les connaissances qu'il emploie pour ses recherches. Chaque activité de l'utilisateur est représentée sous forme de document. Les activités portent sur : les objectifs (et des sous-objectifs éventuellement) formulés par l'utilisateur, les types de techniques de recherche qu'il emploie (par exemple, observation, demande de renseignement ou symbolisation et raisonnement, et spécification explicite de préférence) et les termes employés. Cette approche nous permet d'analyser les informations sur un utilisateur par des techniques issues de l'analyse de données. Nous obtenons ainsi une description synthétique du comportement et du niveau de connaissance d'un utilisateur.

En SRI, la personnalisation des réponses est jusqu'à présent limitée au traitement automatique du système : nous avons élargi notre champ d'étude à l'exploitatioon des compétences humaines dans un processus de recherche d'information. Pour ces études, nous avons défini une architecture pour un système de recherche coopérative d'information (SRCI). Ce type de système permet trois modes de recherche d'information: l'autonomie, l'observation et la coopération. En mode coopération par exemple, un expert peut, à distance, intervenir direc-

tement dans le processus de recherche d'un utilisateur pour l'assister dans son processus de recherche et obtenir la description synthétique du comportement et du niveau de connaissance de l'utilisateur. En mode observation, un utilisateur ne peut qu'observer les activités d'un autre utilisateur sans pouvoir intervenir dans le processus de recherche d'information. En mode autonomie, l'utilisateur travail seul sans aucune collaboration avec d'autres utilisateurs.

Les modes d'observation et de coopération permettent donc d'exploiter les compétences d'un autre utilisateur pendant un processus de recherche d'information.

Les modes coopératifs ainsi que l'architecture pour les réaliser sont applicables dans bien d'autres domaines que le SRI. Par exemple, pour le télétravail, la formation à distance, etc.

Nos propositions sont expérimentées dans le prototype STREEMS développé dans le cadre du projet européen LEONARDO pour gérer des informations sur les arbres autorisés pour le reboisement par l'union européenne et dans le prototype METIORE qui permet des recherches bibliographiques sur les publications enregistrées au centre de documentation du LORIA.

Ces travaux [5, 6, 1] sont poursuivis en collaboration avec l'université de Malaga, Espagne, où Amos DAVID est co-directeur de la thèse de David BUENO.

4 Logiciels

4.1 Technologie adaptative

Participants : Yolande Belaïd, Norbert Valverde, Emilie Pérales, Nicolas Galmiche, Abdel Belaïd.

Mots clés : Technologie adaptative, Signature de région, Propagation de contraintes.

Un prototype logiciel a été développé pour la société ITESOFT permettant la segmentation automatique de formulaires de type bon de commande. Ce prototype a été mis au point à partir des formulaires de la société QUELLE, spécialisée en vente par correspondance.

4.2 TDM

Participants : Abdel Belaïd, François Parmentier, Olivier Lescop, Laurent Pierron.

Mots clés : Partie du discours, Forme canonique, Reconnaissance de caractères.

Nous avons dévéloppé le logiciel TDM permettant la reconnaissance automatique d'articles de revues placés dans leurs sommaires. Les résultats servent à alimenter le serveur de tables de matières de Calliope.

4.3 STREEMS

Participants : Amos David, David Bueno.

Mots clés : Système de Recherche Coopérative d'Information, personnalisation de réponses, modélisation de l'utilisateur.

STREEMS est un prototype développé dans le cadre du projet européen LEONARDO. Il

permet de gérer des informations botaniques et multimédia sur les arbres autorisés par l'union européenne pour le reboisement des forêts.

4.4 Metiore

Participants : Amos David, David Bueno.

Mots clés : Système de Recherche Coopérative d'Information, personnalisation de réponses, modélisation de l'utilisateur.

METIORE est un SRI qui permet d'effectuer des recherche sur des publications. La base est constituée des publications du LORIA. Les trois opérations de recherche (recherche multi-critères ; analyse croisée ; analyse croisée avec contraintes) son implémentées dans le prototype.

5 Contrats industriels (nationaux, européens et internationaux)

5.1 Technologie adaptative

Participants : Norbert Valverde, Yolande Belaïd, Nicolas Galmiche, Emilie Pérales.

Mots clés : Bon de commande, Vente par correspondance, Analyse de formulaires.

C'est un projet de collaboration avec la société ITESOFT et la société QUELLE pour la réalisation d'un logiciel générique d'analyse de bordereaux de vente par correspondance (VPC). Ce logiciel sera introduit dans la plate-forme actuelle FORMSCAN de la société ITESOFT. Une version de ce logiciel est en cours de validation industrielle et sera installé dans la société QUELLE.

5.2 Reconnaissance de tables de matières

Participants : Laurent Pierron, Abdel Belaïd, François Parmentier, Olivier Lescop.

Mots clés : Tables de matières, Marquage morphologique, Partie di discours.

Ce logiciel a fait l'objet d'un premier contrat avec le centre de recherche de Rank Xerox dans le cadre du projet Calliope. Il fait aujourd'hui l'objet d'une convention de collaboration STIC (bibliothèque virtuelle) entre l'INRIA et l'Université de Tunis II. Ce logiciel sera intégré dans le projet national tunisien de connexion des bibliothèques universitaires ainsi que dans un autre projet sur l'arabisation.

5.3 Navigation du Nord-Est, Nancy

Participant : Amos David.

Mots clés : Base de données, Indexation d'images, Cartes géographiques.

Nous sommes en collaboration avec Navigation du Nord-Est pour la réalisation d'une base de données multimédia portant sur des cartes géographiques et historiques d'inondation. L'ob-

jectif scientifique est de proposer une méthodologie pour l'indexation des images et les cartes, ainsi que la proposition des méthodes d'accès aux images.

5.4 Rétroconversion de documents

Participants : Abdel Belaïd, Laurent Pierron.

Mots clés : Saisie automatisée, Rétroconversion de structures.

C'est un projet de collaboration avec la société Berger-Levrault qui cherche à se munir d'un système de saisie automatique de documents imprimés. L'objet actuel du projet est de réaliser une étude préalable permettant de mesurer les efforts humains et matériels à fournir pour la réalisation d'une telle plate-forme. Le travail de recherche proprement dit concernera la reconnaissance multifonte, la segmentation physique et la reconnaissance de la structure logique des documents.

6 Actions régionales, nationales et internationales

6.1 Actions nationales

L'équipe participe aux activités de l'association GRCE (Groupement de Recherche en Communication Ecrite) qui réunit les spécialistes francophones dans ce domaine. Ce groupe a pour but de permettre des échanges scientifiques entre les chercheurs de différents laboratoires et les industriels. L'équipe participe également aux actions du GDR-PRC dans le groupe 5.2. sur l'écrit et le document.

6.2 Actions internationales

Nous collaborons de manière très active avec plusieurs pays. Notamment avec la Tunisie, où Abdel Belaïd entretient des relations de collaborations scientifiques avec les organismes de recherche tunisiens (ENSI et ENIT) où Abdel BelaId co-encadre deux thèses. Cette collaboration entre dans le cadre d'une convention cadre signée entre l'Université de Tunis et l'INRIA. Nous collaborons de manière soutenue avec l'équipe du Prof. Nabeel Murshed de l'Université Tuiuti do Parana, à Curitiba (Brésil). L'objet de la collaboration porte sur les aspects de bibliothèque virtuelle. Enfin, Nous avons participé aux projets européens mentionnés ci-dessus avec l'université de Malaga. Une concrétisation de cette collaboration est une thèse de doctorat dont Amos DAVID est co-directeur.

7 Diffusion de résultats

7.1 Animation de la Communauté scientifique

Abdel Belaïd est membre de l'IAPR. Il fait partie du comité de rédaction de plusieurs revues : Document numérique, Traitement du Signal, appartient aux comités de programme et

est relecteur de plusieurs Conférences internationales : DAUD'99 (Florence, Italie), DLIA'99 (Bangalore, Inde), ICDAR'99 (Bangalore, Inde).

Amos David est membre du comité scientifique de ISKO-France (International Society for Knowledge Organization).

Sur le plan de l'encadrement scientifique, Abdel Belaïd a encadré un DEA et trois thèses dont une a été soutenue en février 1999.

7.2 Enseignement

Yolande Belaïd est responsable de deux modules d'enseignement à l'IUT Charlemagne : Structures de données et Algorithmique avancée. Elle est auteur d'un polycopié de cours pour un troisième module sur l'initiation à l'algorithmique et à la programmation. Dans ce même institut, Abdel Belaïd assure des travaux dirigés en algorithmique et en programmation en C et en Java.

Amos David est Directeur des études du DESS IST commun aux trois universités de Nancy. Il assure les cours suivants : Architectures et évaluation des logiciels documentaires, modélisation de données, conception et developement de systèmes de base de données (relationnelle et objets), analyse de l'information

Abdel Belaïd a assuré un cours sur les modèles perceptifs en reconnaissance des formes. Ce cours est commun au DEA d'informatique et à l'ESIAL.

8 Bibliographie

Articles et chapitres de livre

[1] A. DAVID, D. BUENO, « User modeling and cooperative information retrieval in information retrieval systems », *International journal of Knowledge Organization 26*, 1, 1999, p. 30 – 45.

Communications à des congrès, colloques, etc.

[2] A. BELAID, A. DAVID, « The use of Information Retrieval Tools in Automatic Document Modeling and Recognition », *in : Tenth International Workshop on Database and Expert Systems Applications (DAUD'99), Florence, Italia*, p. 522–526, septembre 1999.

[3] C. CHOISY, « Utilisation de champs de Markov en approches globale et analytique pour la reconnaissance de l'ecriture manuscrite », *in : Journees Jeunes Chercheurs, GRCE, Tours*, GRCE, septembre 1999.

[4] A. DAVID, A. BELAID, « Information Retrieval Systems in Document Analysis and Recognition », *in : Document Layout Interpretation and its Applications (DLIA'99), Bangalore, India*, septembre 1999.

[5] A. DAVID, D. BUENO, « Personalisation of information based on the concept of relevance using a user model », *in : IV ISKO-SPAIN, Granada, Spain*, ISKO, ARMILLA, avril 1999.

[6] A. DAVID, D. BUENO, « Towards cooperative information retrieval system with user modelling », *in : 5th International conference on Information Systems Analysis and Synthesis - ISAS'99, Orlando, USA*, juillet 1999.

Rapports de recherche et publications internes

[7] A. Belaid, L. Pierron, « Reconnaissance de tables de matières », *Rapport de fin de contrat*, Xerox, avril 1999.