

Compte-rendu de la conférence

Ingénierie des Connaissances 2007

L'édition 2007 de la conférence Ingénierie des Connaissances s'est tenue dans le cadre de la plateforme biannuelle de l'AFIA (Association Française pour l'Intelligence Artificielle) du 2 au 6 juillet 2007 à Grenoble.

Sur 54 articles soumis (émanant d'équipes de recherche localisées en Belgique, Canada, France, Luxembourg et Tunisie), 27 ont été acceptés en version longue (représentant un taux de sélection de 50%) et 12 ont été acceptés en version courte. Ces travaux ont été présentés dans le cadre de 7 sessions thématiques décrites ci-après.

Lors de la session plénière de la plateforme, Jean-François Perrot, professeur émérite du LIP6, a animé une conférence invitée intitulée « Etre ou ne pas être : est-ce bien la question ? »

« La pratique de la programmation par objets conduit à "réifier" sans discrimination. De même, l'activité qui se déploie autour des terminologies et des ontologies tend à tout étiqueter par des noms. Pourtant, il est manifeste que tout ce qui entre dans nos modèles ne peut pas sans violence être traité comme un objet. Pour essayer d'expliquer cette dérive, afin de mieux la combattre, on analyse les présupposés de la tradition philosophique dont nous sommes imbus, en les confrontant à l'exigence de rigueur familière aux informaticiens. »

Session 1 « Ontologies » animée par C. Reynaud - LRI, Orsay

Quatre présentations ont été effectuées durant cette première session portant sur les ontologies. Deux problématiques ont été abordées : la construction d'une ontologie et le problème de l'alignement. La construction d'ontologies a fait l'objet de 2 exposés. L'un, concernant une ontologie générale des programmes, COPS, a consisté à décrire son contenu ainsi que la méthodologie ayant permis de l'élaborer par spécialisation de l'ontologie DOLCE et réutilisation d'ontologies de domaine, situées à un niveau d'abstraction plus élevé, sur les documents (I&DA) et les artefacts. Le second concernait une ontologie, également décrite du point de vue de son contenu, utilisée pour décrire des licences d'une manière unifiée et indépendante des systèmes de DRM propriétaires tout en étant compatible avec les langages d'expression des droits (REL) de ces systèmes. Le problème de l'alignement, quant à lui, a été étudié sur deux types de données. Une première présentation s'est intéressée à l'alignement de hiérarchies de classification utilisées pour la recherche de documents. Une méthode a été proposée, capable de trouver des relations d'équivalence ou de subsomption entre les concepts de deux hiérarchies. Celle-ci est basée sur le paradigme des règles d'association pour lesquelles des mesures d'intérêt existent. L'étude présentée a consisté à évaluer les

résultats obtenus par la méthode proposée selon ces différentes mesures d'intérêt. La seconde présentation s'est intéressée à la mise en correspondances d'éléments de données (EDs) extraits de 11 sources biomédicales. L'originalité de la méthode d'alignement proposée repose sur l'utilisation conjointe de techniques au niveau schema et au niveau instances, les techniques au niveau schema étant basées d'une part sur l'utilisation d'UMLS, une ressource terminologique biomédicale de référence, et sur WordNet.

Session 2 « Web Sémantique et Recherche d'Information » animée par J. Charlet - AP-HP et INSERM, Paris

Michel Buffa, Guillaume Ereteo et Fabien Gandon nous ont présenté un travail intitulé « Wiki et Web Sémantique ». L'exposé de Michel Buffa présentait le principe de fonctionnement des wikis, leurs avantages ainsi que leurs faiblesses en termes de rigidité de l'interaction. Ensuite, SweetWiki était présenté, un wiki qui cherche à réconcilier les 2 tendances du Web, le Web 2.0 et le Web sémantique. La modélisation y est assez développée puisque les auteurs construisent une ontologie du wiki lui-même. Par ailleurs, un certain nombre de mécanismes permettent d'importer des ontologies pour tagger les éléments du wiki. Enfin, d'autres mécanismes, de requête avec le moteur CORESE ou d'organisations des folksonomies par les utilisateurs proposent une évolution du concept de wiki. La discussion a porté, logiquement au sujet des wikis, sur la façon dont cet enrichissement favorisait ou pas l'interaction avec les utilisateurs avec des propositions d'évolution puis de fixations des folksonomies.

Nicolas Guelfi, Cédric Pruski et Chantal Reynaud nous ont présenté un travail intitulé « Les ontologies pour la recherche ciblée d'information sur le Web : une utilisation et extension d'OWL pour l'expansion de requêtes ». Dans ce travail, les auteurs étendent OWL pour pouvoir exprimer des notions de méronymie ou d'antonymie dans les requêtes. Des règles d'expansion de requêtes fondées sur des études permettent de régler cette expansion à travers l'ordre de déclenchement des mécanismes, en fonction des résultats communément admis par les utilisateurs. Les premiers résultats expérimentaux justifient l'intérêt de cette approche.

Laurent Mazuel, Nicolas Sabouret nous ont présenté un travail intitulé « Degré de relation sémantique dans une ontologie pour la commande en langue naturelle ». Dans ce travail, les auteurs proposent une mesure de degré de relation sémantique entre deux concepts d'une ontologie pour faciliter la phase dite « d'interprétation sémantique » dans le cadre de la communication homme-machine en langage naturel. Cet appariement entre concepts sert à calculer la proximité entre les événements d'un agent et la commande de l'utilisateur. Cet article où les ontologies sont mises en œuvre par d'autres domaines de recherche que l'IC ouvre des voies de recherches intéressantes.

Session 3 « Web Sémantique et Recherche d'Information » animée par F. Gandon - INRIA, Sophia-Antipolis

Annotation sémantique par classification, Yannick Toussaint, Sylvain Tenier, LORIA / INIST – Nancy

Sylvain présente une méthode d'annotation prenant en compte la dépendance entre concepts i.e. le fait qu'un élément dans une page doit être annoté par un concept parce que certains autres éléments de cette même page sont annotés par d'autres concepts. La représentation en Logiques de Descriptions de la structure des documents, de l'ontologie et de l'annotation lui permet de reposer sur les mécanismes de classification fournis par ces logiques en introduisant la notion de classe d'annotation et en définissant le processus d'annotation comme une opération d'instanciation.

Extraction Contextuelle de Concepts Ontologiques pour le Web Sémantique, Lobna Karoui, Marie-Aude Aufaure, Nacera Bennacer, Supelec - Gif sur Yvette

Lobna présente une méthode d'extraction des concepts à partir de documents HTML reposant sur un algorithme de clustering hiérarchique non supervisé qui utilise de façon incrémentale l'algorithme de partitionnement Kmeans et est guidé par un contexte structurel. Elle exploite la structure HTML ainsi que la position du mot afin d'optimiser la pondération de chaque terme ainsi que la sélection du co-occurent le plus proche sémantiquement. L'algorithme incrémental propose un raffinement successif entièrement automatique ou interactif.

Traitement d'attributs inter-dépendants pour la recherche d'information par treillis, Nizar Messai, Marie-Dominique Devignes, Amedeo Napoli, Malika Smaïl-Tabbone, LORIA – Nancy

Nizar propose d'étendre l'analyse de concepts formels pour gérer les contextes formels où il existe des relations entre les attributs et par la suite d'appliquer cette extension à la recherche d'information par treillis. Les relations entre attributs permettent de distinguer des attributs plus importants traduisant les préférences et priorités d'un utilisateur. L'extension permet une navigation guidée par des connaissances d'un treillis de concepts ou une spécification de la façon dont les mots clés d'une requête doivent être considérés et interprétés afin d'obtenir la réponse la plus satisfaisante.

Session 4 « Textes et Ontologies » animée par N. Aussenac-Gilles - IRIT, Toulouse

Cette session a regroupé 4 exposés qui témoignent de la diversité des travaux actuels sur la construction d'ontologies à partir de textes, la place significative des approches statistiques et le souci constant de bien peser la place de l'information tirée des textes par rapport à la connaissance que l'on veut représenter au sein d'une ontologie.

Le premier exposé "Construction d'une ontologie à partir d'un corpus de textes avec l'ACF" présenté par Rokia Bendaoud (et écrit avec Mohamed Rouane Hacene, Yannick Toussaint, Bertrand Delecroix, Amedeo Napoli du LORIA - Nancy), est un exemple très pertinent de démarche scientifique d'une part et de la combinaison

d'analyses linguistiques et statistiques pour la construction d'ontologies d'autre part. En effet, il s'agit d'une adaptation au français et au domaine de la génomique d'une approche proposée par P. Cimiano pour l'anglais. Cette approche s'appuie sur des textes analysés syntaxiquement et le réseau de termes qui en est tiré pour conduire une analyse de concepts formels (ACF). L'exposé souligne la nécessité d'adapter les outils d'analyse syntaxique et d'extraction de termes aux particularités du corpus pour obtenir des résultats pertinents.

Le deuxième exposé « Construction d'une ontologie du domaine HSE » par Sylvie Després présentait un travail plus classique de construction d'ontologie à partir de textes dans le domaine du droit, réalisé en collaboration par le LARIA (Frédéric Fürst) et le LIPN (S. Després avec Sylvie Szulman). La particularité de ce travail est le volume important du corpus de textes et l'étude de la complémentarité des connaissances tirées des textes et de la réutilisation d'un noyau d'ontologie propre au droit.

Lors de l'exposé suivant, "Vers une classification des similarités basées sur le contenu informationnel des concepts d'une hiérarchie de subsomption", E. Blanchard du LINA a présenté un travail théorique conséquent et très utile à la communauté scientifique, réalisé avec Mounira Harzallah et Pascale Kuntz-Cosperec. Les mesures de similarité en tre concepts étudiées sont tirées des domaines de la classification et de l'ingénierie des connaissances. Les auteurs montrent que l'estimateur de contenu informationnel introduit par Resnik permet de juger de l'adéquation de ces mesures à un besoin et de comparer les propriétés de ces mesures.

Pour terminer la session, avec "Dire n'est pas concevoir", C. Roche (Equipe Condillac - Université de Savoie) a présenté un exposé qui se voulait provocateur et qui rappelait des principes certes pertinents, mais pas aussi oubliés que le prétendait l'orateur. Parmi ceux-ci, C. Roche a insisté sur la nécessité de bien distinguer l'étude du comportement des termes en corpus de la définition des concepts (le corpus, avec ses limites, ne pouvant fournir tous les éléments nécessaires au choix des concepts), sur la nécessité de poser un point de vue lors de la construction d'une ontologie, et de s'interroger sur l'explicitation des principes d'organisation des concepts. C. Roche défend une approche aristotélicienne basée sur la différenciation de tout concept de son concept père par une propriété, et insiste sur l'importance de la formalisation comme achèvement de l'ontologie. Un débat animé s'est engagé avec la salle sur l'analyse faite des approches actuelles de construction d'ontologies à partir de textes.

Session 5 « Applications de l'IC » animée par S. Despres - CRIP5, Paris

La session intitulée « Applications de l'IC » qui s'est déroulée vendredi matin a donné lieu à quatre présentations qui ont suscité de nombreuses questions et donné lieu à un débat riche et animé. Si les trois dernières communications relevaient du domaine médical, la première était consacrée à un système question/réponse évalué dans le domaine du tourisme.

La première présentation « Intégration d'une ontologie de domaine dans un système question/réponse » a été faite par Farah Benamara de l'IRIT à Toulouse. Le travail présenté traitait de la relaxation des concepts à l'origine des échecs d'un de question/réponse dans un domaine fermé en s'appuyant sur une ontologie. Les concepts de l'ontologie sont par des propriétés structurelles (essentiels) et fonctionnelles (contingentes) avec une pondération qui permet de moduler leur importance dans la définition du concept. Les mesures de similarité prennent en compte ces différents aspects du modèle. Les questions ont porté sur la modélisation des concepts, les mesures de similarité utilisées et sur la façon dont les pondérations sont définies. Les réponses apportées ont précisé la façon dont les concepts sont élaborés et ont explicité les mesures de similarité construites. Les perspectives du travail ont également été abordées. Il s'agira en particulier de confronter les résultats obtenus avec les utilisateurs et d'enrichir l'ontologie exploitée.

Les trois dernières présentations relevaient du domaine médical ce qui a donné une certaine unité à la fin de la session même si les sujets traités étaient assez différents.

La première d'entre elles « Une modélisation qualitative de la plausibilité du diagnostic de pneumopathie médicamenteuse » a été présentée par Jacques Bouaud de l'Assistance Publique des Hôpitaux de Paris. Le sujet exposé concernait la modélisation de connaissances plausibles pour réaliser un système d'aide à la décision dans le domaine de la pneumopathie médicamenteuse. Après avoir rappelé les différents modèles de l'incertain existants et exposé les spécificités des problèmes propres au diagnostic de la PM, un modèle empirique a été présenté de manière claire et didactique. Les questions ont porté essentiellement sur la difficulté à obtenir des valeurs auprès des experts et sur la validation du système par les médecins.

Jean Charlet de l'INSERM a ensuite exposé les travaux relatifs à la « Construction d'une ontologie pour la prise en charge de l'hypertension artérielle ». L'ontologie présentée a été construite à partir de conceptualisations théoriques et de traces textuelles liées à la pratique clinique. La SNOMED est utilisée pour lier manuellement les concepts du modèle aux termes qu'elle contient. Les outils SYNTAX et UPERY sont utilisés et les principes différentiels de construction d'une ontologie sont mis en œuvre via l'éditeur DOE.

Les questions ont porté sur la réutilisation de ressources existantes dans le domaine de la médecine et sur les niveaux d'abstraction des ontologies. La réponse est claire pour Jean Charlet qui estime que l'inadéquation des ontologies générales et l'absence d'une ontologie générique du domaine a conduit au développement d'une ontologie spécifique.

Enfin Anis Ellini membre du LIM & BIO et du département de santé public de l'hôpital Avicennes a relaté « Une expérience originale de test automatisé d'un système d'aide à la décision clinique basé sur les guides de bonnes pratiques ». Cette méthode de test a été élaborée à partir de spécifications papier et a permis la modélisation d'algorithmes cliniques sous forme d'organigrammes de séquences d'exécution. Cette approche est innovante puisqu'elle permet de visualiser les interactions mutuelles.

Les questions ont porté sur les spécifications du projet et sur la manière dont le système pourrait être perçu par les équipes de médecins. L'utilisation d'un tel

système constitue-t-elle un surcroît de travail pour les utilisateurs. En raison du manque d'expérimentation pratique les réponses sont restées prospectives mais optimistes. D'autres domaines d'application sont envisagés.

Session 6 « Epistémologie et coopération » animée par M. Lewkowicz - Tech-CICO, Troyes

Cette session était composée de 4 présentations traitant du partage de connaissances au sein de collectifs humains. La première, par Charles Delalonde, de l'équipe Tech-CICO de l'UTT a permis de présenter le modèle et l'outil DemonD, développés dans le cadre d'une thèse CIFRE à Orange R&D. DemonD est dédié à la recherche de connaissances entre équipes distribuées géographiquement. Sur la base du constat selon lequel on ne peut prescrire ni le social ni l'usage des outils au sein de communautés, et en mobilisant la théorie de l'acteur réseau, la proposition consiste, face à une requête, à proposer à la fois des ressources documentaires et des personnes expertes du sujet. Un espace collaboratif est généré automatiquement afin de permettre à la personne à la recherche d'informations et aux experts identifiés d'interagir afin de répondre à la question initiale. La seconde présentation, par Nicolas Férey du LIMSI, a montré comment un outil de visualisation en 3D couplé à une plate-forme de réalité virtuelle peut permettre à des acteurs de discuter de classifications afin de construire des vocabulaires spécialisés d'un domaine. Ce travail pluridisciplinaire est pour l'instant focalisé sur les mécanismes de visualisation des classifications et les mesures de distance entre les classes proposées par des experts différents. Il est prévu dans un second temps d'observer et d'analyser les interactions au sein de l'environnement de réalité virtuelle. La troisième présentation, par Daniel Memmi du département d'Informatique de l'Université du Québec à Montréal, a porté sur la recherche sociale et l'importance de l'expertise humaine dans la recherche d'information. La proposition consiste à modéliser la recherche d'expertise et à utiliser les concepts (représentation et algorithmes) des réseaux sociaux. L'idée est d'utiliser des graphes pour représenter la structure sociale au sein d'un groupe d'acteurs et d'associer de l'information aux nœuds de ces graphes. La quatrième proposition, par Manuel Zacklad, porte sur le Web Socio-Sémantique (W2S) et la Recherche Ouverte d'Information (ROI). En reprenant la définition de l'IC par Bruno Bachimont (IC = ingénierie des supports numériques de connaissance), Manuel Zacklad propose d'attribuer deux rôles à l'IC : d'une part pour concevoir des systèmes d'organisation des connaissances, et d'autre part pour faciliter l'orientation au sein de ces systèmes (c'est ce qu'il appelle la ROI). Il propose également de distinguer trois paradigmes d'accès à l'information : (1) par des *requêtes* sur des attributs standardisés, (2) par une *navigation* utilisant des attributs heuristiques, (3) par une *fouille* dans des ressources documentaires. Le W2S, abordant le Web comme étant un instrument de gestion documentaire, propose des techniques relevant du paradigme de la navigation (2). Dans la suite de la présentation HyperTopic est proposé comme une représentation pivot entre les trois paradigmes. En effet, HyperTopic définit des items par les attributs qui les décrivent (paradigme de requête), les thèmes qu'ils abordent (paradigme de navigation), et les ressources auxquelles ils sont liés (paradigme de la fouille). Les quatre présentations de cette session moins focalisée sur une représentation formelle des connaissances, et d'avantage ouverte sur des problématiques d'interactions entre acteurs humains, de partage et de recherche de connaissances,

ont donné lieu à plusieurs questions de fond permettant un débat avec la salle très riche.

Session 7 « Web Sémantique et Recherche d'Information - Annotations et SBC» animée par Y Prie LIRIS, Lyon

Cette session était consacrée aux systèmes d'information audiovisuelle (image et vidéo) à base de connaissances et a regroupé quatre présentations d'articles. A noter que tous les articles s'appuient sur des prototypes réels, dans la plupart des cas sur des outils open source.

Le premier article intitulé "Lecture active de documents audiovisuels : organisation de connaissances personnelles par la structuration d'annotations" (Bertrand Richard*, Yannick Prié, Sylvie Calabretto) a été présenté par Bertrand Richard. Les auteurs s'intéressent à la question des "connaissances personnelles" dans le cadre de la lecture active audiovisuelle, par exemple en vue de la création de critiques filmiques. Cette activité est celle d'un travailleur intellectuel qui inscrit ses connaissances dans le document au moyen d'annotations, mais également les organise, soit par le moyen de structures abstraites, soit par leur utilisation même dans de nouveaux documents. Un premier modèle est présenté qui permet de considérer les différentes structures d'expression des connaissances et leur transformation en cours de lecture active.

Cécile Roisin a présenté le second article, intitulé "Un système expert d'aide à la classification taxonomique de classes de descripteurs" (Marc Caillet). Le travail s'inscrit dans le cadre des travaux de valorisation du patrimoine menés à l'INA et concerne le langage FDL (Feria Description Language). Ce langage permet notamment de décrire des contraintes d'organisation de descripteurs, par exemple de spécifier l'organisation d'une pièce de théâtre en actes, scènes, en y associant des contraintes d'agrégation temporelles entre classes de descripteurs. Une classification des différentes "formes" de documents audiovisuels est ensuite proposée qui permet de les organiser en une hiérarchie de subsomption. Un algorithme permettant de réaliser cette classification est proposé, qui fonctionne pour l'instant uniquement sur les classes et les relations temporelles, mais ajoutera à terme la prise en compte de relations entre descripteurs.

Francky Trichet a présenté le troisième article intitulé "Une plateforme de développement d'espaces web sémantiques communautaires dédiés au partage de ressources multimédia" qu'il a co-écrit avec Yohan Nizon. L'article décrit la plateforme OSIRIS dédiée à l'annotation de ressources (images, sons, vidéos et documents bureautiques) au sein de communautés Web2.0 organisées autour de documents, de thesaurii et d'ontologies. Le principe est de donner à l'utilisateur un mode d'annotation simple en triplet (sujet, verbe, objet), lesquels expriment des connaissances formelles liées à des ontologies denses. L'annotation et la recherche peuvent prendre appui sur des thesaurii associés aux ontologies, qui permettent de lier mots de la langue et concepts ontologiques. Les ontologies permettent une annotation automatique des documents par déduction de triplets non explicitement proposés par les annotateurs.

Nicolas Moreau a terminé la session avec l'article "Annotation formelle graphique de documents multimédia" co-écrit avec Michel Leclère, Michel Chein et Alain Gutierrez. Ce travail prend appui sur le projet SAPHIR de gestion de documents audiovisuels décrits avec des graphes conceptuels, et traite plus particulièrement de la problématique de l'annotation de documents avec un formalisme graphique par des annotateurs non experts. Les auteurs proposent de considérer différentes dimensions dans l'annotation, au moyen d'ontologies modulaires (sous-parties d'une ontologie pertinente pour une certaine tâche). Ils s'attachent également aux notions de "graphes patrons", qui permettent de proposer des descriptions générales à instancier, ainsi que de "graphes prototypiques" exprimant des mises en contextes de noeuds, ainsi que l'atelier d'annotation CoGUI et de la méthodologie associée.