Réflexions sur la place du RàPC dans trois domaines de recherche actuels

Béatrice Fuchs, Jean Lieber, Alain Mille, Amedeo Napoli

01234567890123456789012345

Réflexions sur la place du RàPC dans trois domaines de recherche actuels

Béatrice Fuchs[1], Jean Lieber[2], Alain Mille[1], Amedeo Napoli[2]

[1] *LIRIS, Bâtiment Nautibus, Université Claude Bernard, 69622 Villeurbanne Cedex*
{bfuchs,amille@liris.cnrs.fr}

[2] *LORIA, BP 239, 54506 Vandoeuvre lès Nancy,*
{lieber,napoli@loria.fr}

RÉSUMÉ : Dans cet article, nous nous intéressons à la place du raisonnement à partir de cas par rapport à trois domaines de recherche où les enjeux sont actuellement considérables : la gestion des connaissances (et l'aide à la décision), le Web sémantique (avec entre autres et la recherche d'information et le traitement de requêtes complexes), et enfin l'extraction de connaissances dans des bases de données.

MOTS-CLÉS : raisonnement à partir de cas, Web sémantique, gestion des connaissances et mémoire d'entreprise, extraction de connaissances dans des bases de données.

1. Introduction

Dans cet article, nous nous intéressons à la place du raisonnement à partir de cas par rapport à trois domaines de recherche où les enjeux sont actuellement considérables : la gestion des connaissances (et l'aide à la décision), le Web sémantique (avec entre autres et la recherche d'information et le traitement de requêtes complexes), et enfin l'extraction de connaissances dans des bases de données.

Tout d'abord, nous replaçons le cadre de la représentation des connaissances, ici à base de concepts (SRBC), qui est commun aux trois domaines et qui concerne essentiellement les logiques de descriptions et les systèmes de représentation de connaissances par objets (RCO) [10, 17]. La fonction d'un SRBC est de stocker et d'organiser les connaissances autour de la notion de concept et de fournir des services inférentiels destinés à compléter l'information disponible et à faire émerger des informations implicites. Un SRBC s'appuie sur une hiérarchie de concepts liés entre eux par une relation de subsomption. Un concept a une identité et un état, quelquefois un comportement comme dans les systèmes de RCO [11]. Un concept possède une intension, qui se définit par l'ensemble des propriétés caractéristiques, ou attributs, du concept et une extension, qui regroupe l'ensemble des instances recouvertes par le concept. L'intension peut s'appréhender comme l'ensemble des conditions nécessaires et suffisantes devant être vérifiées par un objet pour être instance du concept. Les intensions et les extensions sont emboîtées en sens inverse l'une de l'autre : plus un concept est général plus il recouvre d'individus et moins il recouvre de propriétés, et réciproquement.

La hiérarchie des concepts est exploitée pour résoudre des problèmes, par l'intermédiaire de mécanismes de raisonnement comme la classification de concepts ou la classification d'instances. La classification de concepts consiste à placer un nouveau concept dans la hiérarchie, tandis que la classification d'instances cherche à déterminer les concepts dont un individu donné peut être instance. La classification s'appuie sur le test de subsomption qui consiste à vérifier qu'un concept donné est plus général qu'un autre concept. Les classifications de concepts et d'instances sont à la base du raisonnement par classification.

Les logiques de descriptions proprement dites s'appuient sur les notions de "concept" qui représentent des classes d'individus, de rôles qui représentent des relations (binaires) entre concepts et d'individus, qui correspondent aux instances des concepts [3]. La description des concepts s'appuie sur une syntaxe qui autorise l'emploi d'un certain nombre de constructeurs, à laquelle est associée une sémantique. Les concepts, primitifs et définis, sont organisés en une hiérarchie par l'intermédiaire de la relation de subsomption : un concept `C` subsumé par un concept `D` si `C` est moins général que `D` pour toute interprétation (au sens où l'extension de `C` est inclue dans l'extension de `D`).

Dans ce milieu hiérarchique, le raisonnement à partir de cas (RÀPC) peut se voir comme une extension naturelle du raisonnement par classification. Le RÀPC 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` (mais aussi adaptable pour garantir qu'une solution de `cible` peut effectivement être construite [13, 16]). 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.

Dans ce qui suit, nous discutons des problématiques du Web sémantique, de la gestion des connaissances et de l'extraction de connaissances dans les bases de données, en essayant de montrer les apports actuels et potentiels du RÀPC dans les trois domaines.

2. Le Web sémantique

2.1. La problématique du Web sémantique

Le Web aujourd'hui est exploité par des personnes, qui en général, recherchent une information ou posent des questions via un moteur de recherche, et analysent le résultat elles-mêmes. Le Web est appelé à devenir "sémantique", au sens où il va être exploité en priorité par des machines qui vont traiter des problèmes posés par des personnes et qui vont délivrer les résultats obtenus à ces personnes [12]. Le Web sémantique va devenir ainsi un espace d'échange d'informations entre machines, permettant l'accès à de très grands volumes d'informations, et fournissant les moyens de gérer ces informations. Cet espace d'échange doit être qualitatif aussi bien que quantitatif, personnalisé et sûr. Toutefois, une machine ne peut être en mesure d'appréhender le volume des informations disponibles sur le Web et peut fournir une aide conséquente aux personnes, que si on la dote d'une certaine "intelligence". Parmi les besoins qui découlent de cette vision figurent

la nécessité de disposer de langages pour exprimer le contenu des documents, d'une sémantique associée à ces langages, et de moteurs d'inférences associés, qui s'appuient sur cette sémantique pour raisonner. Des ressources de plusieurs types sont également nécessaires, en particulier des bases de connaissances comprenant des *ontologies* [20], des bases de règles, d'individus, de bases de fonctions ou de services, des bases de données (documents), des thésaurus, etc.

L'omniprésence du Web modifie la manière d'envisager la recherche et l'échange de documents de toutes natures. Les contraintes de fonctionnement du Web privilégient la modularité et l'autonomie des documents manipulés. De ce point de vue, les technologies à base de concepts sont souvent utilisées pour répondre à ces contraintes. Qu'ils se trouvent mobilisés pour coder les documents, pour représenter leur contenu ou pour implanter des serveurs, les concepts et les objets sont présents au cœur du Web, et leur rôle est appelé à se renforcer.

Par ailleurs, la nécessité de contrôler les informations, documents ou données, par l'intermédiaire d'une sémantique, renvoie aux problématiques de représentation des connaissances en général et par objets en particulier. C'est pourquoi, alors que se déploient d'importants projets sur le Web sémantique, les concepts et les objets se font de plus en plus indispensables, à la fois en tant qu'éléments de représentation des connaissances, en tant que support de programmation et d'échange et en tant qu'unité de déploiement modulaire de services. En plus, l'utilisation de XML sous toutes ses facettes comme passerelle entre documents et objets trouve dans l'idée de Web sémantique une justification naturelle.

En tant que thème de recherche, le Web sémantique constitue un cadre fédérateur pour une variété de travaux de recherche qu'il faut combiner, parmi lesquels se trouvent la représentation et la gestion de connaissances, la formalisation du raisonnement, en particulier par classification et à partir de cas, mais aussi à base de règles, la l'extraction de connaissances, la fouille de textes, et la recherche d'informations.

2.2. La notion d'ontologie

En représentation des connaissances, le terme *ontologie* fait référence à un modèle opérationnel utilisé pour décrire un domaine particulier du monde réel [12, 20]. Dans cet ordre d'idées, les *concepts* apparaissent comme des briques de base des ontologies [14].

Pratiquement, une ontologie $\mathcal{O}$ se présente comme un système formel constitué d'un ensemble de concepts et d'un ensemble de relations binaires spécifiées par des couples de concepts (D,R) de *domaines* et de *co-domaines*, d'un ensemble A d'axiomes, d'une relation de *subsomption* (ou spécialisation) notée $\sqsubseteq$, qui est généralement réflexive, antisymétrique et transitive, qui permet d'organiser les concepts et les relations en une hiérarchie, et qui autorise les inférences. En particulier, $\mathtt{C}_1 \sqsubseteq \mathtt{C}_2$ signifie que $\mathtt{C}_1$ est un sous-concept de $\mathtt{C}_2$: l'extension de $\mathtt{C}_1$ est contenue dans celle de $\mathtt{C}_2$ tandis que l'intension de $\mathtt{C}_1$ contient celle de $\mathtt{C}_2$. Il est possible alors d'inférer qu'un individu est instance d'un certain concept ou qu'un concept partage certaines propriétés avec un autre concept. Suivant ce schéma, une base de connaissances s'appuie sur un couple constitué d'une ontologie et d'une base d'assertions ou de faits (dans lesquels interviennent les individus). Ainsi, une base de connaissances contient des unités de formes et de niveaux d'abstraction différents, comme par exemple des concepts, des instances, mais aussi des règles manipulant des faits, des stratégies (*heuristiques*) exprimant la façon de se servir des connaissances de la base.

Les ontologies ont une place d'importance croissante dans des domaines comme la gestion des connaissances, les systèmes coopératifs, l'intégration et la recherche d'information, le commerce électronique et bien sûr le Web sémantique. Les ontologies sont appelées à jouer là un rôle clé en établissant une terminologie commune entre les agents — logiciels et humains — qui peuvent ainsi partager la même sémantique sur les concepts et les relations manipulés. Les ontologies sont également au cœur de la gestion des connaissances, où, à l'image de la gestion de bases de données, ce ne sont plus simplement des données (syntaxiques) mais des connaissances (munies d'une syntaxe et d'une sémantique) qui sont considérées et manipulées.

Toutefois, les objets du quotidien obéissent rarement à des lois rigoureuses, comme les êtres humains et au contraire des objets mathématiques. Intégrer les différentes natures des objets du quotidien dans un formalisme de représentation pose des problèmes difficiles à résoudre comme la représentation de modalités (statut et degré de vérité des informations), la représentation de connaissances typiques et exceptionnelles, la représentation de connaissances incomplètes, évolutives, interdépendantes, etc.

Dans cet ordre d'idées, la RÀPC a un rôle à jouer qui est celui de retrouver des objets présentant une certaine similarité pour résoudre un problème de recherche donné, recherche d'information, de documents sur la base de leur contenu [2].

2.3. L'accès à l'information sur le Web par le contenu

Le besoin en information est primordial dans des domaines comme la veille technologie (et donc la gestion des connaissances et la mémoire d'entreprise), tandis que de plus en plus données sont disponibles. La quantité évolue sans que la qualité suive, ni que les systèmes capables de prendre en charge ces données évoluent de façon à pouvoir les appréhender à leur juste mesure, que ce soit au niveau de la gestion par le contenu comme au niveau de l'extraction de connaissances à partir de ces données. Ainsi, cette quantité croissante de données nécessite de mettre en œuvre des moyens particuliers pour les exploiter, de favoriser l'accès à l'information dans des fonds volumineux et hétérogènes tel qu'il en existe sur le Web. Un objectif est de fournir aux "clients et consommateurs d'information" un environnement dans lequel ils puissent exploiter les données de leur domaine pour leurs besoins.

Les moteurs de recherche sont souvent débordés par l'explosion du Web et ne répondent pas toujours correctement aux tâches de recherche d'information. Une préoccupation est de favoriser un accès intelligent aux données du Web en exploitant des connaissances relatives au domaine des données traitées, en s'appuyant sur une ontologie du domaine par exemple. Dans ce cadre, une approche générale consiste à coupler ontologie (et donc connaissances, éventuellement extraites des documents étudiées) et recherche d'information. L'idée ici est que la fouille de données et la recherche d'information sont deux approches complémentaires pour appréhender des données structurées ou non : la fouille de données permet de guider la recherche d'information à partir des connaissances extraites des données, et, inversement, la recherche d'information permet de guider la fouille de données par l'exploitation des connaissances issues de la fouille de données elle-même.

La représentation du contenu d'un document permet de manipuler ce document pour faire de la recherche par spécialisation, par similitude, par analogie, etc. Ce type de manipulation peut être pris en charge par un formalisme de RÀPC, que ce soit pour la recherche de documents similaires [2], la recherche de documents adaptables, le traitement de requêtes

analogues sur la base de chemins de similarité.

3. La gestion des connaissances

L'utilisation des technologies du Web permet de partager des documents et des connaissances. Les documents numériques et numérisés peuvent être rendus accessibles de manière standard et transparente auprès des utilisateurs concernés. Une ambition, à terme, est de réaliser de véritables serveurs de connaissances, permettant la recherche et la manipulation de ressources. Cependant, la mise en place de portails et l'organisation de sites Web se révèlent des tâches coûteuse : la recherche de documents pertinents et l'interrogation d'un site sur la base de son contenu sont devenus des nécessités mais sont souvent peu efficaces : les formalismes de représentation des connaissances et de raisonnement sont les formalismes adéquats pour représenter le contenu de ces éléments et peuvent dans une certaine mesure aider à résoudre une partie de problème de recherche et de gestion qui se posent.

3.1. La notion de serveur de connaissances

Ainsi, la *gestion de connaissances* s'articule autour des notions d'acquisition, de diffusion, d'évaluation, d'évolution et de maintenance des connaissances [9]. À l'heure actuelle, la conception des systèmes de gestion des connaissances (et de mémoire d'entreprise) nécessite :

- d'exploiter des bases de connaissances et des ontologies,
- d'exploiter des bases de données de natures différentes et de volumes importants comme le Web par exemple,
- d'assurer la navigation dans informations et la recherche des informations par le contenu,
- de traiter des problèmes complexes comme l'intégration et la fouille de données hétérogènes.

Dans ce cadre, le langage XML est bien adapté à la description de documents textuels — c'est une de ses raisons d'être — mais la résolution de problèmes nécessitant des raisonnements et de la recherche d'information par le contenu doit faire appel aux formalismes de représentation des connaissances et de raisonnement. XML doit alors jouer le rôle de passerelle entre l'univers des données et celui des connaissances [1]. De cette façon, les travaux sur la gestion des connaissances sont en relation directe avec les travaux sur le Web sémantique. Il est clair qu'aujourd'hui, un système de gestion des connaissances et de gestion d'une mémoire d'entreprise doit s'appuyer sur les principes du Web sémantique pour assurer une parfaite adéquation entre gestion des connaissances et gestion des données, et en particulier des données du Web.

Ainsi, l'architecture d'un serveur de connaissances pour la gestion des connaissances et la gestion d'une mémoire d'entreprise doit prendre en compte les informations et les connaissances propres à une entreprise, mais aussi les informations disponibles sur le Web, et mettre en œuvre des mécanismes de représentation et de raisonnement. Ici, il faut faire émerger la notion plus spécifique de *serveur de connaissances multidimensionnel*, qui peut être vu comme un un "système d'information intelligent" capable de gérer un

référentiel multidimensionnel de connaissances pour une organisation ou une entreprise. Autour du référentiel gravitent les éléments d'information circulant dans l'organisation ou l'entreprise comme des données, des connaissances, et des informations de toutes natures : messages, notes, notices, documents, modes d'emploi, etc. Les connaissances et les informations associées sont considérées comme des ressources devant être exploitées, enrichies et étendues. Les techniques de conception mises en œuvre relèvent à la fois de la technologie des systèmes de connaissances pour la résolution de problèmes complexes et l'aide à la décision, et de la technologie des systèmes d'information pour gérer conjointement des informations et des connaissances, les stocker sous une forme persistante, les retrouver, les interroger en fonction de leur contenu, les visualiser et les réutiliser sous différents points de vues.

3.2. Le raisonnement à partir de cas et l'aide à la décision

Dans le cadre de la gestion des connaissances et de la mémoire d'entreprise, l'aide à la décision sur la base de critères qualitatifs peut s'appuyer sur le raisonnement à partir de cas. En particulier, des connaissances et méta-connaissances d'ordre stratégique et tactique, des historiques et des connaissances temporelles peuvent être appréhendés et exploités par le RÀPC. D'une part, il faut distinguer la "mémoire", avec la prise en compte d'expériences, de plans de résolution ou de maintenance, d'historiques et de comportements standards ou exceptionnels ... D'autre part, il faut distinguer la résolution de problèmes actuels sur la base de problèmes déjà résolus, de la mise en œuvre de stratégies de résolution ou de stratégies d'évolution. Dans tous les cas, le RÀPC peut jouer un rôle de premier plan au niveau d de la prise de décision, que ce soit du côté des décideurs, des concepteurs, des techniciens et des utilisateurs.

À titre d'exemple, citons un projet sur l'utilisation du RÀPC pour le raisonnement spatial qualitatif [6], qui concerne le développement d'un système de RÀPC pour l'interprétation et la comparaison de structures spatiales. Pour la mise en œuvre du système sont étudiés :

- la modélisation de structures spatiales qualitatives par des graphes,
- la représentation et la classification de graphes dans une logiques de descriptions (en l'occurrence le système RACER),
- la définition de chemins de similarité entre graphes (à la façon de [16]) modélisant des structures spatiales,
- l'adaptation des explications liées aux graphes, qui consiste pour le moment à une simple recopie des explications qui sont exploitables après remémoration.

3.3. L'ingénierie d'un système de gestion des connaissances

L'ingénierie d'un système à base de connaissances s'appuie sur un ensemble d'opérations qui se retrouvent dans la conception de tout logiciel d'envergure :

- Initialisation et modélisation : ces étapes recouvrent la spécification d'un modèle des éléments du domaine à représenter et une mise en œuvre du modèle du domaine ; des méthodologies de modélisation telles que KADS ou COMMON KADS ont été mises au point pour ces besoins [18].
- Représentation et raffinement : ces étapes recouvrent la phase de représentation proprement dite, le choix d'un langage de représentation, l'implantation du modèle et le raffinement du modèle après les premières opérations de test.

- Évaluation : cette étape suit et complète l'utilisation du système à base de connaissances dans des applications et recouvre la mise en place d'environnements logiciels adaptés aux besoins spécifiques et l'évaluation du fonctionnement du système dans la pratique.
- Maintenance : les connaissances évoluent, les modes de raisonnement aussi, ce qui fait évoluer d'autant la spécification du système, qui doit être mis à jour pour garantir la cohérence et la compatibilité des connaissances, anciennes et nouvelles.
- Diffusion : les connaissances doivent être partagées et transmises si besoin est sous une forme opérationnelle : c'est là un des fondamentaux de la conception des ontologies et donc des bases de connaissances. Le fait que les connaissances soient codées à l'aide d'un langage de représentation muni d'une syntaxe et d'une sémantique bien définies garantit que ce sont bien les mêmes éléments de connaissances qui sont envoyés et réceptionnés, au sens où ils peuvent servir à résoudre des problèmes de même nature de part et d'autre.

Du côté du monde industriel, ce sont les notions de gestion des connaissances et de mémoire d'entreprise qui émergent effectivement [15, 9, 21]. Les problèmes à résoudre consistent pour l'essentiel à gérer les connaissances liées à l'entreprise : les recenser, les mémoriser, les utiliser, les transmettre et les faire croître. Les éléments essentiels de la mémoire d'entreprise sont le savoir-faire, l'expertise, les documents scientifiques et techniques ; les différents "agents" sont ici les personnes, les connaissances, les documents et les actions (la dynamique, le flot des informations, etc.). Un des problèmes récurrents qui se pose est celui de "trouver la bonne information" : qui sait ou peut savoir où elle se trouve, comment s'en servir et sinon comment faire.

Les principales étapes de la conception d'une mémoire d'entreprise sont calquées sur le schéma ci-dessus d'ingénierie d'un système à base de connaissances : la détection des besoins et la conception du modèle, la construction de la mémoire, l'utilisation, l'évaluation et la diffusion de la mémoire. Il faut faire ressortir ici plusieurs points particuliers : l'utilisation effective de la mémoire d'entreprise par des personnes aux statuts différents et donc aux besoins différents (décideurs et techniciens par exemple), l'accès nécessaire à des systèmes et à des bases d'informations qui sont associés ou qui dépendent de la mémoire d'entreprise, le partage et la diffusion des éléments de la mémoire sur un plan interne mais aussi externe, et enfin le "retour sur expérience" qui autorise une mise au point et un raffinement progressifs de la mémoire d'entreprise. C'est précisément à ce niveau là que le RÀPC peut et doit jouer un rôle, en tant que formalisme permettant de stocker et retrouver l'expérience, mais encore en tant que formalisme permettant de confronter les expériences passées et présentes, en vue d'une adaptation potentielle pour le problème courant.

4. L'extraction de connaissances (ECBD) et la fouille de textes

4.1. Méthodes symboliques en extraction de connaissances

L'extraction de connaissances dans 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, système qui 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 de fouille de données pouvant s'appuyer sur des techniques symboliques ou numériques comme les classifications par treillis et par arbres de décision, l'induction, l'analyse des données ou les statistiques,
- une interface se chargeant des interactions et de la visualisation des résultats.

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 une base de connaissances ou encore une base de cas : les connaissances extraites sont stockées dans la base pour être réutilisées dans d'autres applications et mises à jour le cas échéant.

Parmi les méthode symboliques d'extraction de connaissances, nous nous intéressons plus particulièrement à la classification par treillis, à l'extraction e motifs fréquents et de règles d'association. La classification par treillis est une technique de fouille de données symbolique qui permet d'analyser une population, d'extraire des corrélations, des motifs et des règles, selon certains points de vue choisis. Elle relève de l'analyse de "tableaux booléens de données" ou de présence — absence de propriétés mono-valuées. Elle s'appuie sur la correspondance de Galois associée à une relation pour faire émerger un treillis de concepts formels (hiérarchie de concepts), où un concept est un couple (*intension,extension*) et des règles d'association exactes ou partielles [4, 14]. Ces hiérarchies de concepts particulières se construisent en fonction des connaissances disponibles sur le domaine des données et produisent des structures ordonnées interprétables et réutilisables.

Parallèlement à la classification par treillis, l'extraction de *motifs fréquents* correspond à l'extraction de motifs — un motif est un ensemble de propriétés apparaissant dans les données avec une certaine fréquence — dont le nombre d'occurrences dans les individus d'une population étudiée est supérieur à un seuil donné [5]. Sur la base des motifs extraits, il est possible d'extraire des règles d'association qui expriment des corrélations entre les propriétés qui composent les motifs fréquents.

L'extraction de motifs est en rapport avec les fermés d'une correspondance de Galois : certains motifs fréquents sont fermés et à partir de ces motifs fermés fréquents il est possible de retrouver tous les motifs fréquents. La recherche de motifs fréquents et l'extraction de règles d'association reposent sur la construction du treillis de Galois de la relation « l'objet `o` possède la propriété `p` », qui, à partir d'un tableau booléen, fait émerger un treillis de concepts formels, décrits par des ensembles de propriétés et des ensembles d'individus qui s'y rattachent.

La classification par treillis et la recherche de motifs fréquents pour la fouille de données se pratiquent classiquement sur des tableaux booléens de données. Certaines études montre que ce méthodes peuvent se pratiquer dans le cadre des SRBC pour pouvoir traiter des données plus complexes, composés d'ensembles de couples *(attributs,valeurs)*, où les attributs sont mono-valués, multivalués, ou relationnels [19]. Ce traitement nécessite de généraliser la notion classique de correspondance de Galois en prenant en compte la relation « l'objet `o` possède la propriété `p` dont les valeurs vérifie la contrainte `c` ». Les hiérarchies de concepts obtenus s'intègrent alors naturellement au modèle des systèmes de SRBC. La classification par treillis d'objets complexes peut se voir appliquée avec beaucoup plus de possibilités à divers domaines, dont la fouille de textes et la fouille de séquences génomiques.

Il est souvent fait un amalgame entre RÀPC et ECBD, où le RÀPC est cité comme un méthode de fouille de données : si l'on peut imaginer que que le RÀPC peut intervenir dans le processus d'ECBD, pour rejouer ou adapter des scénarios de fouille — comme il en est question pour la fouille du Web — ou retrouver des séquences de fouille similaires, il n'est guère possible d'envisager le processus de RÀPC lui-même comme une méthode d'ECBDà part entière. En effet, il n'est possible de produire des éléments de connaissances par l'intermédiaire du RÀPC que lorsque un cas adapté ayant permis de résoudre un problème a été jugé acceptable et digne d'intérêt pour être mémorisé et servir ultérieurement. Nous sommes là très loin des nécessités de l'extraction de connaissances dans les bases de données où les enjeux et les besoins sont tout à fait différents et d'une échelle totalement différente.

4.2. La fouille de textes

Un processus de fouille de textes doit, à partir d'un texte — ou d'un ensemble de textes — décomposé en groupes syntaxiques cohérents, fournir des éléments de synthèse permettant d'appréhender et de manipuler globalement le ou les textes étudiés. Ces éléments peuvent être un treillis de concepts, un ensemble de motifs fréquents ou de règles d'association explicatives. En outre, ces éléments synthétiques peuvent être (mis) en relation avec une ontologie du domaine ou un thésaurus.

Les perspectives d'utilisation de la fouille de données sur de grands ensembles de textes sont importantes, car les textes couvrent un spectre large d'information, mais la forme des textes et documents elle-même rend l'exploitation difficile et nécessite l'emploi de techniques liées au traitement du langage naturel. Bien qu'il commence à exister de plus en plus de travaux en fouille de textes, la confusion avec les problèmes d'accès à l'information reste fréquente. De nombreux travaux portent sur la classification de documents à partir de mots-clés, notamment sur la base de la classification par treillis, pour avoir des visions thématiques de collections de documents, mais les textes sont rarement utilisés pour découvrir de nouvelles connaissances à proprement parler. Il est possible de distinguer deux objectifs principaux dans la fouille de textes :

- chercher à automatiser au moins partiellement la construction de ressources linguistiques ou documentaires,
- chercher dans une grande collection de textes à faire émerger de nouvelles connaissances sous la forme de corrélations entre des faits ou des événements qui sont décrits dans les textes.

Ces objectifs nécessitent l'association de différentes compétences : gestion et représentation des connaissances, traitement du langage naturel et fouilles de données symboliques. De plus, la richesse et la complexité, des structures qui peuvent être extraites à partir d'un texte, comparées à des données booléennes classiques, posent de nouveaux défis pour la définition d'algorithmes de fouille de textes.

La fouille de textes passe par une annotation conceptuelle des textes, qui consiste à annoter — pour les manipuler ou les retrouver — des documents textuels par des structures conceptuelles reflétant le contenu des textes et extraites des textes eux-mêmes. Les annotations sont représentées par des concepts dans un SRBC avec lesquels il est possible de résoudre des problèmes. L'analyse du contenu des textes peut s'appuyer sur l'adéquation entre la syntaxe et la sémantique des textes, ce qui nécessite d'exploiter des connaissances du domaine. Ainsi, certains travaux prennent comme point de départ une analyse en constituants des phrases et identifient le rôle syntaxico-sémantique des différents constituants, avec l'hypothèse que, dans un domaine de spécialité pour lequel existe un modèle de connaissances, le fait de disposer de la représentation complète et correcte d'un ensemble de phrases de référence permet de mener à bien l'analyse de nouvelles phrases. C'est bien là le fait même du RÀPC : une première expérience a été menée en la matière, qui demande encore à être enrichie et étendue, mais qui est un point de départ indéniables à des travaux novateurs et d'importance, notamment en ce qui concerne la fouille du Web [7, 8].

5. Conclusion

Dans cet article, nous avons livré quelques réflexions sur la place du raisonnement à partir de cas dans trois domaines de recherche où les enjeux sont actuellement considérables : le Web sémantique, la gestion des connaissances et l'extraction de connaissances dans des bases de données. Cette étude est encore bien préliminaire et il reste encore beaucoup travail pour arriver à une réflexion stable qui puisse être exploitée opérationnellement : toutefois, l'avanir est radieux et nous ne désespérons pas.

6. Références

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

[2] R. Al-Hulou, A. Napoli, and E. Nauer. Une mesure de similarité sémantique pour raisonner sur des documents. In J. Euzenat and B. Carré, editors, *Langages et modèles à objets, Lille (LMO'04)*, pages 217–230. Hermès, L'objet 10(2–3), 2004.

[3] F. Baader, D. Calvanese, D. McGuinness, D. Nardi, and P. Patel-Schneider, editors. *The Description Logic Handbook*. Cambridge University Press, Cambridge, UK, 2003.

[4] M. Barbut and B. Monjardet. *Ordre et classification – Algèbre et combinatoire (2 tomes)*. Hachette, Paris, 1970.

[5] Y. Bastide, R. Taouil, N. Pasquier, G. Stumme, and L. Lakhal. Pascal : un algorithme d'extraction des motifs fréquents. *Technique et science informatiques*, 21(1):65–95, 2002.

[6] F. Le Ber, A. Napoli, J.-L. Metzger, and S. Lardon. Modeling and comparing farm maps using graphs and case-based reasoning. *Journal of Universal Computer Science*, 9(9):1073–1095, 2003.

[7] F. Chakkour, A. Napoli, and 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*, pages 1–6. Rapport IRIT/00-11-R, IRIT, Toulouse, 2000.

[8] F. Chakkour, A. Napoli, and Y. Toussaint. Extraire des structures prédicatives à partir des textes, vers une indexation conceptuelle des textes. In B. Fuchs and A. Mille, editors, *Actes du* IX^{ème} *séminaire français de raisonnement à partir de cas*, 2001.

[9] R. Dieng, O. Corby, A. Giboin, J. Golebiowska, N. Matta, and M. Ribiere. *Méthodes et outils pour la gestion des connaissances.* Dunod, Paris, 2001.

[10] R. Ducournau, J. Euzenat, G. Masini, and A. Napoli, editors. *Langages et modèles à objets — État des recherches et perspectives.* Collection Didactique D-019. INRIA, Le Chesnay, 1998.

[11] J. Euzenat. Représentation de connaissances par objets. In R. Ducournau, J. Euzenat, G. Masini, and A. Napoli, editors, *Langages et modèles à objets — État des recherches et perspectives*, Collection Didactique D-019, pages 293–319. INRIA, Le Chesnay, 1998.

[12] D. Fensel, J. Hendler, H. Lieberman, and W. Wahlster, editors. *Spinning the Semantic Web.* The MIT Press, Cambridge, Massachusetts, 2003.

[13] B. Fuchs, J. Lieber, A. Mille, and A. Napoli. Vers une théorie unifiée de l'adaptation en raisonnement à partir de cas. In R. Teulier, editor, *Actes de IC'99 – Ingénierie des Connaissances, Plate-Forme AFIA, École Polytechnique, Palaiseau*, pages 199–207. AFIA, Chambéry, 1999.

[14] B. Ganter and R. Wille. *Formal Concept Analysis.* Springer, Berlin, 1999.

[15] A. Hatchuel and B. Weil. *L'expert et le système.* Economica, Paris, 1992.

[16] J. Lieber and A. Napoli. Raisonnement à partir de cas et résolution de problèmes dans une représentation par objets. *Revue d'intelligence artificielle*, 13:9–35, 1999.

[17] A. Napoli, B. Carré, R. Ducournau, J. Euzenat, and F. Rechenmann. Objets et représentation, un couple en devenir. *L'objet*, 10:61–81, 2004.

[18] G. Schreiber, H. Akkermans, A. Anjewierden, R. de Hoog, N. Shadbolt, W. van de Velde, and B. Wielinga. *Knowledge Engineering and Management: the CommonKADS Methodoloy.* The MIT Press, Cambridge, MA, 1999.

[19] A. Simon and A. Napoli. Building Viewpoints in an Object-based Representation System for Knowledge Discovery in Databases. In S. Rubin, editor, *Proceedings of the First International Conference on Information Reuse and Integration (IRI'99), Atlanta, Georgia*, pages 104–108. The International Society for Computers and Their Applications, ISCA, 1999.

[20] S. Staab and R. Studer, editors. *Handbook on Ontologies.* Springer, Berlin, 2004.

[21] M. Zacklad and M. Grundstein, editors. *Management des connaissances (modèles d'entreprise et applications).* Hermès, Paris, 2001.