Représenter l'information et Web sémantique

Jérôme Euzenat, Marie-Christine Rousset

Représenter l'information et Web sémantique
Chapitres de l'ouvrage collectif du GDR IA "L'intelligence Artiificielle, de quoi s'agit-il vraiment ?"

Jérome Euzenat[1] and Marie-Christine Rousset[2]

[1]Université Grenoble Alpes, Inria, LIG, F-38000 Grenoble, France
[2]Institut Universitaire de France and Université Grenoble Alpes, CNRS, LIG, F-38000 Grenoble, France

1 Représenter l'information

Dans toutes ses tâches, l'intelligence artificielle (IA) manipule et exploite des informations. Ces informations doivent donc être représentées de façon à pouvoir être traitées par ordinateur. L'IA est en effet une des disciplines des sciences du traitement de l'information, à coté d'un certain nombre d'autres spécialités que sont notamment, les Bases de Données (qui permettent de stocker de l'information et de la retrouver), la Recherche d'Information (qui sélectionne des documents pertinents dans un corpus à partir de requêtes en termes de mots-clés), la Recherche Opérationnelle (qui cherche des bonnes solutions pour des problèmes posés en termes de contraintes à satisfaire et de critères à optimiser), les Interactions Humain-Machine (qui s'intéresse aux interfaces de communication), l'Automatique (qui porte sur la commande de systèmes dynamiques), et le Traitement du Signal et des Images (qui concerne leur analyse, interprétation, et transmission). La Reconnaissance des Formes, historiquement sœur de l'IA, et tendant maintenant à se fondre avec la partie de l'IA tournée vers l'apprentissage automatique, développe des méthodes afin de catégoriser des objets en identifiant des motifs caractéristisques dans des données les décrivant.

La représentation de connaissances est une branche de l'IA qui vise la formalisation de connaissances, produites et verbalisables par les humains, qui expriment des lois générales, universelles ou susceptibles d'exceptions (par exemple, "les hommes sont mortels", "les oiseaux volent"), de façon à pouvoir automatiser divers traitements par inférence logique. On parle de *connaissances* pour les distinguer des informations factuelles, qui se rapportent à des faits particuliers (par exemple, "Marie a 27 ans"), et qui correspondent à des données, souvent vus comme des triplets "attribut, objet, valeur" (par exemple, "l'âge

de Marie est 27 ans"). Les connaissances s'énoncent souvent avec des règles de la forme "si <condition(s)> alors <conclusion>", qui permettent d'exprimer des relations de conséquence qui peuvent se rapporter à une taxonomie ("si c'est un corbeau, alors c'est un oiseau"), à la causalité ("si le véhicule va trop vite, alors il dérapera"), ou à une recommandation ("si on est déshydraté, alors il faut boire"). Un type important de connaissances est constitué par les ontologies* de domaine, qui décrivent les relations taxonomiques entre les termes de vocabulaire d'un domaine spécialisé, par exemple médical. Un autre type d'information, qui ne concerne pas la description de l'état du monde, est constitué par les préférences d'un agent, ou de groupes d'agents.

La représentation des connaissances s'appuie souvent sur le formalisme de la logique classique, propositionnelle, ou du premier ordre (qui permet alors d'introduire des quantifications universelles (par exemple, $\forall x, \texttt{homme}(x) \Rightarrow \texttt{mortel}(x)$), ou existentielles des énoncés (par exemple, $\exists x, \texttt{homme}(x) \wedge \texttt{a\_marche\_sur\_la\_lune}(x)$. La logique classique s'avère cependant insuffisante pour raisonner en présence de règles avec exceptions, ou d'informations incohérentes (cf. le chapitre suivant "Raisonner, décider").

Une large part de l'information disponible sur laquelle on raisonne est incertaine : le cadre classique de traitement de l'incertitude est celui de la théorie des probabilités; mais si ce cadre est approprié quand on dispose de statistiques de bonne qualité sur la variabilité de quantités ou de traits caractéristiques (par exemple le nombre de jours où il pleut en avril à Toulouse), d'autres cadres de représentation plus récemment introduits, comme les probabilités imprécises, la théorie des possibilités, ou les fonctions de croyance peuvent s'avérer intéressants quand l'incertitude est épistémique, c'est-à-dire qu'elle est due à un manque relatif d'information plutôt qu'à la variabilité d'un phénomène (par exemple, ce que l'on sait de l'âge d'une personne déterminée sur laquelle on est peu renseigné). Les informations incertaines sont alors associées à des modalités (qui peuvent être une question de degré) dans l'ordre du probable, du crédible, du plausible, du possible ou du certain.

L'introduction de d'autres types de modalités sont utiles pour la représentation de relations temporelles ou spatiales, mais aussi des émotions. Par ailleurs, les énoncés avec des prédicats classiques ne peuvent être que vrais ou faux Mais si on utilise des propriétés de nature graduelle, comme "jeune" ou "grand" par exemple, alors l'énoncé "Jean est grand" peut éventuellement être considéré comme ayant un degré de vérité intermédiaire entre le vrai et le faux si Jean mesure 1,75m. c'est l'idée de départ de la logique dite *floue*.

Des langages de représentations basés sur des fragments de la logique classique, ou des extensions limitées de cette logique qui maintiennent la complexité des algorithmes de raisonnement à un niveau acceptable, font l'objet d'études spécifiques pour différentes tâches de raisonnement. Les cadres de représentation offerts par la logique classique ou par les différents modèles de l'incertain ont des équivalents graphiques, qui présentent l'intérêt de visualiser des relations taxonomiques pour les premiers, ou des relations d'indépendance conditionnelle pour les seconds. Un autre cadre graphique de représentation des connaissances très différent est celui des réseaux de neurones artificiels, où l'information réside

dans les poids associés aux noeuds du réseau, mais qui sont moins directement interprétables en terme intelligible par l'humain.

2 Web sémantique et fouille de données

Le web sémantique est né au cours des années 90 dans le but de faire un 'web pour les machines', c'est-à-dire un web dans lequel l'information n'est pas simplement destinée à être lue par un humain mais puisse être exploitée directement par les ordinateurs. L'ambition initiale est que les textes et éléments multimédia présents dans les pages web soient complétés par des éléments de connaissance sur lesquels l'ordinateur va pouvoir raisonner pour fournir des réponses pertinentes à des questions complexes.

Le web sémantique est une application à grande échelle des travaux de recherche en représentation de connaissances (cf. chapitre 'Représenter l'information'). Il est nourri par un web de données liées (ou 'Linked Data') offrant de manière distribuée de grandes quantités d'information. Elles sont décrites dans le langage RDF qui permet d'exprimer des (méta-)données sous forme de graphes dont les noeuds sont identifiés, comme les pages du web, par des IRI, et les classes et relations peuvent être définies par des termes d'une ontologie* utilisant le langage OWL.

Bien sûr, cela ne signifie pas que l'ordinateur comprend le sens des symboles utilisés: ce n'est pas parce que les termes 'escalade' et 'mont Granier' sont attachés à une image que l'ordinateur sait ce qu'est l'escalade, ni ce qu'est un mont. Mais il est possible d'utiliser des techniques d'intelligence artificielle de manière à ce que l'ordinateur se comporte de manière pas si éloignée que s'il comprenait.

Des techniques fondées sur la fréquence de cooccurence de motifs (et en particulier de mots) permettent de déterminer une proximité assez fiable entre ces motifs. Elles associeront par exemple les mots 'mont', 'montagne', 'massif', 'escalade', 'avalanche', mais aussi 'neige' et 'données'. Cependant, dans un contexte ou sont présents 'massif' et 'escalade', 'neige' sera plus proche de 'montagne' que 'donnée'. Pour définir cette similarité, on considère ces termes comme des dimensions d'un espace vectoriel, dont les points caractérisent des documents. Ces dimensions sont réduites en rapprochant celles associées le plus souvent: c'est la base de 'latent semantic analysis' et 'word embeddings'. Cela fonctionne entre documents (textuels ou multi-média) de langues différentes si on dispose d'un corpus annoté multilingue.

La fouille de données permet d'extraire des motifs fondés sur des propriétés structurelles. Par exemple, on va trouver une classe d'objets, les montagnes, décrits par leurs noms, leurs altitudes, leurs massifs et parfois leurs coordonnées géographiques. Organiser le résultat de la fouille en une représentation explicite de la connaissance est l'objet de la découverte de connaissance qui peut organiser les motifs extraits en une véritable ontologie* des objets géographiques comme les montagnes, les rivières, les massifs et les relations entre eux (qu'une montagne fait éventuellement parti d'un massif montagneux et qu'elle peut être localisée

dans un ou plusieurs pays).

Enfin, à partir d'une telle représentation de la connaissance, il est possible de déduire de nouvelles informations, d'identifier les descriptions d'un même objet ou de répondre à des requêtes. Ainsi, quelqu'un cherchant des images d''alpinisme dans le massif de la chartreuse', pourra se voir retourner notre image étiquetée 'escalade' et 'Mont Granier'. Pour cela, il enchainera des pas de raisonnement sur différents éléments de connaissances, comme que l''escalade' est une catégorie d''alpinisme' et que le 'Mont Granier' est situé dans le 'massif de la chartreuse'.

3 Petit Glossaire de l'IA

Agent : Toute entité susceptible de réagir à l'arrivée d'information et présentant une certaine forme d'autonomie. Il s'agit donc d'un terme générique pouvant recouvrir des réalités très différentes allant d'un programme informatique, à un robot, ou un être vivant, en particulier un humain.

Ontologie : Vocabulaire partagé de classes et de propriétés définies à l'aide d'un formalisme logique compréhensible par des humains et traitable par des machines, qui décrit comment les concepts d'un domaine d'application sont structurés et reliés entre eux.