SYNTHÈSE DE L'ATELIER 6 SUR LE WEB SÉMANTIQUE

M2 IA - Intelligence Artificielle & Cognition

Ismail CHAKRANE, Inès LIROULET, Mohamed Massamba SENE, Oumayma YAKOUBI

Le web traditionnel, conçu pour le partage d'informations entre humains, présente des limitations dans l'organisation et l'accessibilité des données. La difficulté de trouver et de connecter efficacement les informations pertinentes a donné naissance à l'idée d'un Web plus structuré. Cette problématique a notamment été au cœur des réflexions de Tim Berners-Lee, le créateur du World Wide Web, qui a identifié le besoin d'une évolution majeure.

Cette vision d'un nouveau Web vise à créer un réseau de données liées et structurées, rendant l'information compréhensible par les machines et facilitant l'automatisation des tâches complexes. Face aux défis qu'il a rencontrés, le Web Sémantique s'est plutôt imposé comme une extension du web traditionnel plutôt qu'un nouveau Web à part entière (du moins jusqu'à ce jour).

La réalisation de cette vision repose sur plusieurs technologies. Au niveau le plus fondamental, on retrouve le langage **XML (Extensible Markup Language)**, qui fournit une syntaxe pour structurer les documents sans imposer de sémantique. Cette base est donc complétée par **RDF (Resource Description Framework)** qui permet de décrire les informations via des triplets (sujet, prédicat (=verbe), objet), qui peuvent être interprétées sous forme de graphes de données. Les **URI (Universal Resource Identifier)** jouent également un rôle crucial en identifiant de manière unique les ressources sur le Web, notamment les éléments décrits par le RDF (sujet et objet).

Au niveau supérieur, on trouve les ontologies qui définissent formellement les concepts et relations présents dans un document. Pour exprimer ces relations complexes, le langage **OWL (Web Ontology Language)** permet d'établir des règles logiques entre les concepts. **SPARQL (SPARQL Protocol And RDF Query Language)** vient compléter cet ensemble en offrant un langage de requête pour interroger et manipuler les données RDF. D'autres technologies sont également intégrées, notamment les langages de règles comme **RIF (Rule Interchange Format)** et **SWRL (Semantic Web Rule Language)**, pour la formalisation d'inférences et de relations complexes entre données.

Tim Berners-Lee a proposé en 2006 les principes du **Linked Data**. Ces principes sont des bonnes pratiques à suivre plutôt qu'un formalisme du web sémantique, et abordent l'utilisation des URIs pour nommer les ressources, l'emploi d'URIs HTTP pour permettre leur recherche, la fourniture d'informations via les standards établis, et l'inclusion de liens vers d'autres URIs.

Au fil du temps, trois grandes visions pour le Web Sémantique se dégagent :

"Dompter le web" ou la bibliothèque d'Alexandrie : Cette vision veut structurer, organiser le web et donner du sens à son contenu, pour faciliter la recherche d'informations. Elle a perdu son importance avec l'apparition des moteurs de recherche actuels.

Base de connaissance et de données fédérée : Cette vision a été inspirée par la représentation formelle en IA, et cherche à faire un web où les bases de données sont interconnectées. Cela permettrait à des machines d'effectuer des opérations entre ces bases sans se soucier de l'authentification ou des protocoles, qui seraient gérés par le web sémantique. Cependant, cela nécessiterait de standardiser le format d'échange d'informations entre les bases.

Navigateur de connaissances (vision de Berners-Lee) : Cette vision veut déléguer les tâches complexes de recherche d'information à des agents intelligents. Ces agents pourraient comprendre et manipuler les données sur un web entièrement lisible par les machines, parallèle au web lisible par les humains. Cependant, cette vision est très ambitieuse et présente des obstacles importants, comme la représentation universelle des connaissances.

L'évolution du Web Sémantique a été marquée par des étapes clés :

- La publication de la vision initiale par Berners-Lee en 2001
- L'introduction des principes du Linked Data en 2006
- La standardisation de SPARQL en 2008, qui fournit un langage de requête unifié
- Le lancement de Schema.org en 2011 et le Google Knowledge Graph en 2012, qui marque les premières adoptions par l'industrie
- Depuis 2015, l'intégration croissante avec l'IA et les graphes de connaissances ouvre de nouvelles perspectives d'application

Malgré ces avancées, le Web Sémantique fait face à plusieurs défis majeurs. Sur le plan théorique, l'utilisation d'ontologies est remise en question du fait de son manque inhérent d'objectivité (classer des concepts en catégories définies et nommer ces catégories invite forcément la subjectivité humaine). La question de la gestion de l'incertitude des informations postées sur le web se pose aussi. Certaines solutions potentielles sont en cours d'exploration, comme l'utilisation de tagging plutôt que d'ontologies, ou la mise en place d'un système de gestion de l'incertitude.

On retrouve aussi des limites sur le plan pratique, avec les enjeux de la scalabilité et la performance du futur Web Sémantique, de la qualité et la fiabilité des données qui y seraient publiées, de la confidentialité et la sécurité par rapport aux données sensibles qui pourraient être utilisées, de la standardisation des pratiques et de l'adhérence des utilisateurs, et enfin du coût de mise en place et de la compétitivité du résultat sur la marché.

Pour ce qui est du lien avec le domaine de la cognition, le web sémantique s'aligne principalement avec une approche cognitiviste, reposant sur des représentations symboliques formelles. Bien qu'il présente des caractéristiques des systèmes complexes et dynamiques, il est assez éloigné des approches connexionnistes et énactivistes.

Le débat a d'abord mis en avant un optimisme autour de la vision originale de Tim Berners-Lee, visant à créer des agents autonomes capables d'accomplir des tâches complexes sur le web. Cet optimisme est principalement alimenté par l'émergence des grands modèles de langage. Toutefois, il reste beaucoup de doute et de scepticisme à cause des obstacles importants qui subsistent, par exemple en ce qui concerne le respect des standards RDF, indispensables à une structuration efficace du web sémantique.

Parmi les solutions explorées, l'idée d'utiliser les LLMs pour ajouter une couche RDF aux pages web automatiquement et rapidement a été proposée, bien que la faisabilité de cette approche reste à démontrer. Cela permettrait d'ouvrir la voie à une convergence possible entre les LLMs et le web sémantique, notamment grâce aux travaux sur la transformation du langage naturel en requêtes SPARQL.

Cependant, plusieurs défis persistent, notamment par rapport à la standardisation des structures et la sémantisation des données. La performance des systèmes utilisant l'inférence sur les vastes bases de données nécessaires au web sémantique dans la vision de Berners-Lee représente également un autre obstacle.

La discussion a ensuite porté sur la comparaison entre les LLMs et les approches basées sur le web sémantique dans le domaine de la compréhension du langage naturel. Contrairement aux LLMs, le web sémantique ne cherche pas à "comprendre" le langage, mais plutôt à organiser et structurer les données pour en faciliter l'exploitation. De ce fait, la question de la concurrence entre les deux approches ne se pose pas vraiment, car elles n'ont pas le même but. Il a été souligné que l'un des principaux avantages du web sémantique, à qualité de réponses similaires aux LLMs, réside dans la certitude de ses réponses et dans sa capacité à fournir des sources explicites et à expliciter le raisonnement derrière celles-ci.

Il est également important de noter que le web sémantique ne se limite pas à la simple structuration des données : il implique une organisation des connaissances, où le raisonnement humain joue un rôle crucial. Bien que des initiatives comme le "tagging" proposé par Clay Shirky aient cherché à permettre aux utilisateurs de contribuer à cette structuration et faciliter l'organisation, ces solutions présentent encore des limitations significatives, et le défi de la sémantisation n'est pas encore entièrement résolu.