

Sujet de thèse pour le projet CominLabs MiKroLog (2021-2024)

Titre : Un assistant en langue naturelle pour interroger le Web sémantique

Établissement : IRISA, équipe SemLIS <http://www-semliis.irisa.fr/>

Directeur : Sébastien Ferré

Encadrants : Sébastien Ferré et Peggy Cellier

Début de la thèse : 1er octobre 2021

Mots-clés : Web sémantique, IA, traitement de la langue naturelle, graphes de connaissances, *question answering*, recherche à facettes, construction interactive de requêtes, Sparklis

Contexte

Les technologies du Web sémantique [1] associent la précision et la puissance des bases de données à l'ouverture et l'interopérabilité du Web. Le W3C a standardisé plusieurs langages pour le Web sémantique, notamment :

- RDF pour la représentation des données ;
- RDFS pour la définition du schéma de ces données ;
- SPARQL pour l'interrogation de ces données.

L'écriture de requêtes SPARQL est au même niveau que l'écriture de requêtes SQL et n'est donc pas accessible au plus grand nombre. De nombreuses approches ont été proposées pour faciliter l'interrogation de données sémantiques [6]. La plupart sont des interfaces en langue naturelle [2,3] où l'utilisateur peut formuler sa question de façon spontanée. Cependant, en raison de la complexité de la langue naturelle, ces approches sont limitées aux questions les plus simples (ex., *"Qui est la femme de Barack Obama?"*) et manquent à la fois de robustesse (pas de réponse) et de fiabilité (mauvaises réponses). D'autres approches, telles que les constructeurs de requêtes et la recherche à facette [8], s'appuient sur l'interactivité pour guider pas à pas l'utilisateur dans la formulation de ses requêtes. Elles offrent une plus grande fiabilité en donnant davantage de contrôle à l'utilisateur mais sont d'utilisation moins spontanée.

L'équipe SemLIS a développé un outil de recherche sémantique, Sparklis [5], qui combine l'expressivité de SPARQL, le guidage de type recherche à facettes pour la construction de requêtes et la verbalisation des requêtes en langue naturelle. Néanmoins, la construction pas à pas des requêtes peut s'avérer fastidieuse, par exemple pour de grandes bases de données ou lorsque l'utilisateur commence à bien connaître le vocabulaire de la base.

Objectifs

L'objectif de cette thèse est d'explorer et d'évaluer l'ajout d'un dialogue en langue naturelle spontanée au-dessus d'un système interactif de construction guidée de requêtes, tel que Sparklis. L'avantage attendu de passer par ce système interactif pour construire une requête, plutôt que de produire directement une requête cible, est que celui-ci serve de guide dans la résolution des ambiguïtés inhérentes à la langue naturelle. Par exemple, si un utilisateur cherche des "scientifiques français en informatique" dans DBpedia, il pourrait d'abord demander les "scientifiques français". Le système demanderait alors s'il préfère ceux "nés en France" (propriété `dbo:birthDate`) ou ceux "de nationalité française" (propriété `dbo:nationality`). Après avoir consulté les résultats, l'utilisateur pourrait compléter sa recherche en demandant "ceux en informatique", etc.

Le principal verrou scientifique par rapport aux approches existantes en *question answering* est de parvenir à construire des requêtes, en une ou plusieurs étapes, sans que la robustesse et la fiabilité ne baissent sensiblement avec l'accroissement de leur complexité. Un autre verrou scientifique est de rendre l'interaction suffisamment fluide, c'est-à-dire avoir des délais de réponse compatibles avec

l'interaction et un nombre limité de questions du système à l'utilisateur.

L'étude bibliographique consistera à étudier les différentes approches de l'interrogation du Web sémantique [6,2] en approfondissant les techniques employées dans les approches d'interrogation en langue naturelle [3].

La stratégie envisagée consistera à concevoir, réaliser et évaluer différentes solutions, en partant de solutions simples et en évoluant vers des solutions plus complexes en fonction des résultats obtenus. Il s'agira de réutiliser des outils et ressources existantes dans le domaine du traitement de la langue [7] pour analyser les phrases de l'utilisateur, et de piloter le système de construction guidée en fonction des éléments reconnus. L'utilisateur doit pouvoir exprimer son besoin d'information en plusieurs étapes, alterner entre expression en langue naturelle et navigation dans l'interface. De plus, le système doit savoir poser des questions à l'utilisateur pour résoudre les ambiguïtés résiduelles.

Les solutions développées pourront être évaluées sur des jeux de questions-réponses tels que QALD [4] (questions simples) ou LC-QuAD 2.0 [9] (questions complexes). Ces questions portent sur les bases DBpedia et Wikidata, qui jouent un rôle central dans le Web sémantique.

Compétences attendues

- Web sémantique (RDF, RDFS, SPARQL 1.1)
- Traitement automatique de la langue (TAL)
- Algorithmes de recherche, de résolution de problème (IA)
- curiosité, autonomie, rigueur et ténacité
- goût pour l'expérimentation
- développement d'applications Web (la connaissance du langage OCaml serait un plus)

Références

1. Semantic Web, official W3C website: <https://www.w3.org/standards/semanticweb/>
2. V. Lopez, V. S. Uren, M. Sabou, and E. Motta. Is question answering fit for the semantic web?: A survey. *Semantic Web*, 2(2):125–155, 2011.
3. D. Diefenbach, V. Lopez, K. Singh, P. Maret. Core techniques of question answering over knowledge bases: a survey. *Knowledge and Information Systems*, 55:529-569, 2018.
4. QALD (Question Answering over Linked Data): <http://qald.aksw.org/>
5. S. Ferré. Sparklis: An Expressive Query Builder for SPARQL Endpoints with Guidance in Natural Language. *Semantic Web: Interoperability, Usability, Applicability*, 8(3):405-418, 2017.
6. J. Klímek, P. Škoda, M. Nečaský. Survey of tools for Linked Data consumption. *Semantic Web* 10.4 (2019): 665-720.
7. C. Barrière. Natural Language Understanding in a Semantic Web Context. Springer, 2016.
8. G.M. Sacco, Y. Tzitzikas. Dynamic taxonomies and faceted search. The information retrieval series. Springer, 2009.
9. Dubey M, Banerjee D, Abdelkawi A, et al. LC-QuAD 2.0: A large dataset for complex question answering over wikidata and DBpedia[C]//International Semantic Web Conference. Springer, Cham, 2019: 69-78.