A Web Crawler and SPARQL-Based Search Engine to Expand and Navigate Brain Literature Records

Shiladitya Dutta, Carl Taswell M.D., Ph.D.

2019-September-14

Abstract: In this paper we describe CoVaSEA (Concept-Validated Search Engine Agent): an automated web crawler/query engine that is inter-operable with the Nexus-PORTAL-DOORS System. The Nexus-PORTAL-DOORS System (NPDS) is a data management system that manages repositories of the lexical metadata (in PORTAL servers) and semantic representations (in DOORS servers) of resources. Due to the purpose built hybridized nature of NPDS, it is well-placed to perform a variety of lexical-semantic data analysis tasks. However, many of these tasks require records of semantic descriptions which are labor-intensive to create and maintain due to the substantial and rapidly increasing quantities of brain-related literature available on the open web. To remedy this, we have created CoVaSEA with the intention of providing a method for users to navigate and expand the semantic records of brain literature in the NPDS directories. To this end, CoVaSEA integrates multiple features which benefit NPDS including: (A) An implementation of SPARQL query based search to allow retrieval and manipulation of RDF descriptions, (B) Targeted web-crawling for relevant articles from external biomedical literature databases to broaden NPDS records, and (C) Automated translation of free-form text into RDF triples to derive the semantic portrayals of lexical data. CoVaSEA consists of three principal components: the web-crawler, the lexical to semantic converter, and the SPARQL query engine. The web crawler retrieves articles along with their basic metadata (title, abstract, author(s), etc.) from several of biomedical literature databases via REST API. However, in order to create a full semantic description of the data in each article, key RDF triples which describe the abstracts are constructed. First, each of the unique nouns in the passage are registered via coreference resolution and pronomial anaphora. Then the sentences are parsed into constituency tree format so that the subject(s), verb(s), and object(s) can be extracted. Once the SVO triples are extracted, they are transformed into valid RDF by assigning unique resource identifiers (URI) to each part of the triples. This is accomplished by using various databases (i.e. MeSH) for terminology and select named entities, word sense disambiguation for standard words, and literals for any other sections. These triples are stored via the Scribe API in a DOORS directory where they can be retrieved via the SPARQL query engine. In order to create a more conducive user experience, the query engine supports the capability to construct SPARQL queries from expressions in conjunctive normal form, thus circumventing the need to know SPARQL syntax. With the distinct advantage that the system is automated, CoVaSEA presents the capability to search "externally" to furnish large numbers brain-related literature descriptions on a regular basis and search "internally" to provide a method of retrieving those descriptions, thus laying the groundwork for a variety of future NPDS applications for which semantic metadata stores of brain literature are a necessity.