Knowledge Registry Address Expression Tool Suite (KRAETS)

Suresh K. Damodaran

MIT Lincoln Laboratory

General Purpose

Knowledge Registry Address Expression Tool Suite (KRAETS) contains a reference implementation of Knowledge Registry and address expressions or *A-Expressions* to demonstrate how Knowledge Query Language (KQL) works. This reference implementation includes (1) a compiler for A-Expressions, (2) implementation of KQL with embedded A-Expressions in SQL (Structured Query Language) that returns query results with provenance information, (3) Common Cyber Environment Representation (CCER) ontology schema, (5) an ontology for email corpus that conforms to CCER ontology, and its implementation in a Knowledge Registry, (6) Python based topic modeling algorithms with embedded KQL queries over Enron email data corpus stored in a MongoDB data store. The next section provides details of this reference implementation.

Technical Description

(U) A-Expression compiler: A-Expression compiler consists of a lexer and a parser, both of which are implemented using ANTLR, an open source software.

(U) Knowledge Query Language (KQL) implementation: KQL is an enhanced version of SQL that allows embedding of A-Expressions in SQL quer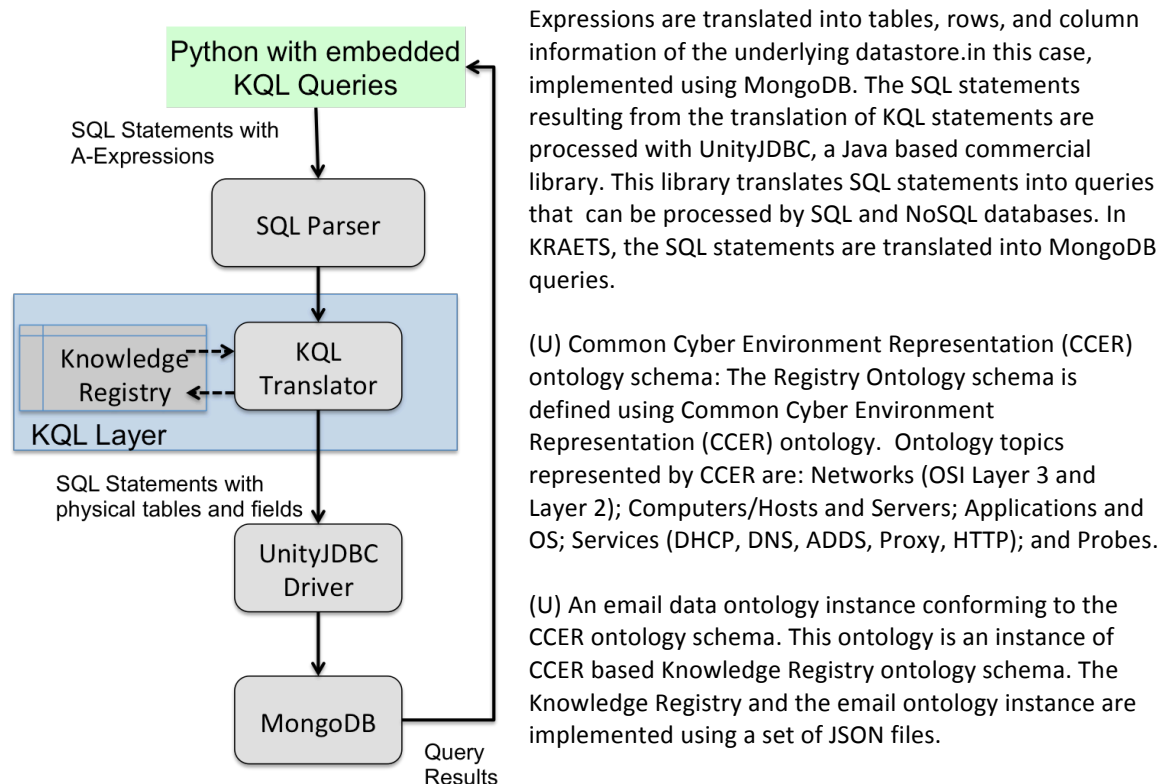ies. A KQL may have one or more A-Expressions. All of the A-Expressions are translated into tables, rows, and column information of the underlying datastore.in this case, implemented using MongoDB. The SQL statements resulting from the translation of KQL statements are processed with UnityJDBC, a Java based commercial library. This library translates SQL statements into queries that can be processed by SQL and NoSQL databases. In KRAETS, the SQL statements are translated into MongoDB queries.

(U) Common Cyber Environment Representation (CCER) ontology schema: The Registry Ontology schema is defined using Common Cyber Environment Representation (CCER) ontology. Ontology topics represented by CCER are: Networks (OSI Layer 3 and Layer 2); Computers/Hosts and Servers; Applications and OS; Services (DHCP, DNS, ADDS, Proxy, HTTP); and Probes.

(U) An email data ontology instance conforming to the CCER ontology schema. This ontology is an instance of CCER based Knowledge Registry ontology schema. The Knowledge Registry and the email ontology instance are implemented using a set of JSON files.

(U) KQL Examples: KRAETS includes several examples of KQL queries. These queries are embedded in Python scripts implementing topic modeling algorithms over Enron email corpus. For a general description of how the KQL queries are processed, see Figure 1. In Figure 1, the KQL Translator translates A-Expressions using the email ontology instance implemented in the Knowledge Registry. Based on the translation of A-Expressions into physical tables, fields (columns) of the data store, and the SQL query is rewritten using this information before it is passed on to the UnityJDBC driver that executes the SQL expression on the email data store implemented over MongoDB.

Figure 1: KQL Query Processing

Advantages of Using KRAETS Tool Suite

(U) KRAETS is a reference implementation of A-Expressions that are embedded in SQL statements. The A-Expressions are used to create declarative queries based on the ontology instance implemented in the Knowledge Registry. The KRAETS tool suite may be used to develop portable embedded queries or tools with portable SQL based algorithms or ad-hoc queries over a variety of data besides email.

(U) KRAETS tool suite also includes CCER ontology that can be used to create ontology instances for a variety of data sources.