CS 410 Tech Review: Google Knowledge Vault

Mathew McDade

mmcdade2@illinois.edu

The term *knowledge base* refers generally to technology systems for the storage and retrieval of facts derived from large collections of complex and unstructured information. Knowledge bases are becoming increasingly relevant in the context of large web and text information systems where knowledge is more diffuse, complex, and abundant than data handled by traditional database systems. *Knowledge* in the context of knowledge base systems references “large networks of entities, their semantic types, properties, and relationships between entities” (Ehrlinger & Wolfram, 2016). The acquisition of knowledge from these sources, storage of relevant knowledge data, and the retrieval of stored knowledge each present unique challenges not present in typical database systems or even earlier attempts to achieve expert-knowledge systems.

The Google Knowledge Vault is an extension, or maybe an evolution, of Google’s proprietary knowledge base system, Google Knowledge Graph. Where the Knowledge Graph system relies on manually structured knowledge sources such as Wikipedia infoboxes and direct user feedback, the Knowledge Vault project seeks to use automated information retrieval and analysis techniques to provide a massive knowledge base system that relies not just on previously structured data but also unstructured, web-based and context data (New Scientist, 2014). The Knowledge Vault product allows Google to scale their knowledge base even beyond the already massive scale of the Knowledge Graph by utilizing automated text information retrieval and analysis techniques along with graph algorithms and supervised machine learning. These techniques are used to build a knowledge base that not only includes knowledge of entities and their relationships but also calculated probabilities for the correctness of the relationships (Dong et al., 2014).

One challenge encountered by other attempts to build automated, large-scale knowledge bases is that analysis of unstructured text information sources tends to produce noisy results from missing or contradictory data and difficulty asserting confidence even for true statements. The Knowledge Vault attempts to address these shortcomings by utilizing a prior model built from existing knowledge bases as a baseline for evaluating and estimating the probabilities for new knowledge

(Dong et al., 2014). The use of prior models in text analysis algorithms allows for the use of new data to reduce noise from errors in source data, account for unknown data, and improve model predictions. Because the prior model is based on existing, potentially incomplete knowledge bases, it's also important that the prior model contain something equivalent to smoothing in order to avoid calculating prior probabilities of zero for triples not discovered in the prior knowledge bases. While it's beyond the scope of discussion for this review, smoothing is addressed with a creative combination of graph algorithms and machine learning processing of the priors--a topic worthy of its own review.

Another important part of the Knowledge Vault implementation is the use of RDF triples to store relationships. Each RDF triple consists of a subject, predicate or relationship, and an object. Each of these RDF triples is associated with a probability representing the system's confidence in the trueness of the triple's assertion. A key feature of these RDF triples is that relationship predicates are separated from their lexical representations, meaning that variations of a relationship's lexical expression can be represented by a single RDF triple (Dong et al., 2014).

The architecture of the Knowledge Vault system can be divided into three major parts. The first part is the extraction of RDF triples and associated confidence probabilities from web-based source materials. This involves the processing of massive amounts of web data for entity extraction, identifying relationships, and determining naive probabilities. The second part is the building of RDF triples based on data found in an existing knowledge base or bases. This set of triples with probabilities serves as the prior model. The final part is referred to as "knowledge fusion." This part calculates the final probabilities of each RDF triple being true based on the sets of triples from the extraction model and the prior model (Dong et al., 2014).

The retrieval and analysis of web-based text documents during the extraction step of building the Knowledge Vault are of interest to us considering our course material. Relation extraction from text documents includes performing well-known text information analysis algorithms for named entity recognition, part of speech tagging, as well as other semantic and syntactic parsing such as dependency parsing and reference resolution, with entity extraction and part of speech tagging clearly

being essential to building the RDF triples which basically consist of (entity, predicate, entity) sets. The RDF triples and themselves are formed using the extracted entities and predicates as features in supervised machine learning.

In summary, the Google Knowledge Vault project represents a major leap in the extraction of knowledge from vast, unstructured web-based data. At its core, it uses many standard text information retrieval and analysis techniques, although modified for the structure and scale of the project. Some familiar techniques are web scraping, entity recognition, part of speech tagging, and use of prior models to improve predictions.

**References**

Wikimedia Foundation. (2021, October 23). *Google knowledge graph*. Wikipedia. Retrieved November 1, 2021, from https://en.wikipedia.org/wiki/Google_Knowledge_Graph.

Hodson, H. (2014, August 20). *Google's fact-checking bots build vast knowledge bank*. New Scientist. Retrieved November 1, 2021, from https://www.newscientist.com/article/mg22329832-700-googles-fact-checking-bots-build-vast-knowledge-bank/.

Ehrlinger, L., & Wöß, W. (2016). Towards a Definition of Knowledge Graphs. *SEMANTiCS (Posters, Demos, SuCCESS)*, *48*(1-4), 2.

Zhao, Z., Han, S. K., & So, I. M. (2018). Architecture of knowledge graph construction techniques. *International Journal of Pure and Applied Mathematics*, *118*(19), 1869-1883.

Dong, X., Gabrilovich, E., Heitz, G., Horn, W., Lao, N., Murphy, K., ... & Zhang, W. (2014, August). Knowledge vault: A web-scale approach to probabilistic knowledge fusion. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 601-610).