CS410 Technology Review: Google Knowledge Vault

As the breadth of human knowledge encompassed across the Web continues to grow at massive scale, it is becoming increasingly imperative that systems exist that can allow for simple gathering and production of this knowledge. Knowledge graphs have been built by various actors, including Google, Freebase, and Microsoft, in attempts to accumulate data from various sources and represent it in a format, typically RDF, that allows humans and machines to form programs and queries to extract pieces of information from them. Knowledge graphs typically represent data as a network of entities; nodes represent entities, such as objects, people, or places, while edges represent the relationships between them. Oftentimes a relationship is further enhanced with a label, denoting the type of relationship between the two entities. Together, the two entities and the label represent a Subject, Predicate, Object relationship. Knowledge graphs have utilized this data format with wide-ranging impact in many highly visible applications, including Google Search (for its search results pages) and Wikipedia.

Constructing a knowledge graph requires the formulation of extraction methods to glean knowledge from primarily textual data across the Web. Some methods have relied on human contributions to build a repository of mostly structured knowledge, such as is the case with Wikipedia. Alternatively, knowledge bases can be constructed over time by crawling the Web and parsing through text data to form insights about the data presented. The Google Knowledge Vault attempts to take the ideology behind knowledge graphs and enhance it to introduce the notion of prior knowledge. Specifically, the Knowledge Vault leverages previously stored knowledge in addition to existing knowledge extraction methods to better categorize and rank new pieces of knowledge. In addition to adding more confidence to the computed rankings of

pieces of knowledge, using prior knowledge helps make the Knowledge Vault more robust against the large amounts of misinformation that proliferate Web data.

The Knowledge Vault is classified as a probabilistic knowledge base in that it assigns a confidence score to each piece of knowledge, an RDF triple, based on the probability that the triple is believed to be correct. The Knowledge Vault works in three main phases, consisting of **extractors**, which create triples from Web data, **priors**, which utilize existing triples (primarily obtained from Freebase, an open-source knowledge base) to assign prior probabilities to each new possible triple, and **fusion**, which combines the extractors and priors to compute a probability score for each triple. The assigned probability score indicates how likely the knowledge graph believes the triple to be true. The Web data on which the Knowledge Vault depends on consists of a variety of formats, including text documents, HTML DOM trees, HTML tables, and human annotations on web pages. The Knowledge Vault operates at a much higher scale than many existing knowledge bases, maintaining confidence scores of over 1 billion triples.

The applications of the Knowledge Vault vary, ranging from search engines to virtual assistants. While the goal of this service is to provide a complete representation of human knowledge, I envision that the Knowledge Vault could be extended to specialize in a particular domain to provide highly relevant domain context to machine learning systems, such as NLP engines, performing complex tasks.

As it stands currently, the Knowledge Vault has a number of limitations, such as in how it deals with varying levels of abstraction (i.e. recognizing that a baby born in California was also born in the USA) and the means by which it can represent knowledge not found through the Web (i.e. common sense knowledge). However, given its ability to combine prior knowledge with

new information much like how the human mind processes knowledge, it is clear that the future of the Knowledge Vault is very promising and seeks to have a large impact on our collective knowledge.

References

By: IBM Cloud Education. (2021, April 12). *What is a knowledge graph?* IBM. Retrieved November 8, 2021, from https://www.ibm.com/cloud/learn/knowledge-graph#toc-how-a-know-0rYVxiNb.

Dong, X., Gabrilovich, E., Heitz, G., Horn, W., Lao, N., Murphy, K., Strohmann, T., Sun, S., & Zhang, W. (2014). Knowledge vault. *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. https://doi.org/10.1145/2623330.2623623

Fundamentals, O. (2021, July 13). *What is a knowledge graph?* Ontotext. Retrieved November 8, 2021, from https://www.ontotext.com/knowledgehub/fundamentals/what-is-a-knowledge-graph/.