Tech Review - Knowledge Vault

Author: Zixuan Ge, zge7@

Intro

Google's Knowledge Vault, a probabilistic knowledge base, is a database of "facts" that have been scraped from the Web. Unlike the traditional ranking system we have seen from lectures which relies on incoming links and the number of links to determine the quality of the source, Knowledge Vault fuses together several extraction sources with prior knowledge obtained through existing KBs like Freebase, similar to that of querying our collective consciousness, to determine what is "true" or not and the confidence level of the derived facts.

Body

First and foremost, information in the Knowledge Vault (KV) is stored in the form of RDF triples (subject, predicate and object). Each triple is given a confidence score, which stands for the probability that the triple is correct given by the KV. KV is language independent, as it separates the facts from their lexical representation. [1] KV comprises of three components:

- Extractors: Information/Triples are extracted from the Web using various extraction methods. It can be extracted from various sources as well e.g. Text documents, DOM trees etc.
- Graph-based priors: KV learns from an existing knowledge base about the probability of the triple, based on the existing triples.

[1] https://research.google/pubs/pub45634/

- Knowledge fusion: The system combines extractors with priors to make predictions about the probability of a given triple, whether it's true or not with a certain confidence level.

Extractors

Several extraction methods are discussed in the paper when extracting from text documents (TXT), HTML trees (DOM), HTML tables (TBL), Human Annotated pages (ANO). Fact extraction is considered to be different in TBL, compared to that of TXT and DOM, as usually the column header contains the entity relations instead of in the text bodies. Comparing the result of the extraction, DOM extractions have the largest number of triples around 1.2 billion, of which 8% are considered with confidence higher than 90%. Whereas the TBL system extracts the least amount of triples of around 9.4 million, the reason being there are few columns in the table would map to a corresponding predicate in the Freebase. We can see from the ROC curve, which derives the AUC score, that the DOM system has the highest score among the four different methods. The overall fusion system score is also dominated by the DOM system, with 7% more high confidence triples compared to the individual extractor result.

Graph-Based Priors

Oftentimes, facts extracted from the Web are unreliable, thus we would need prior knowledge, derived from existing KBs, to correct the course and steer. Researchers classified this as a link prediction problem and two methods are discussed to address this problem: Path ranking algorithm (PRA) and Neural network model (MLP). PRA completes a random walk on the graph given pairs of entities that are connected by some predicates. On the other hand, MLP treats this problem as matrix completion and uses a standard multi layer perceptron to capture

interaction terms. [2] It turns out that both methods have about the same AUC score when compared using the ROC curves. By fusing the priors, we can achieve the highest AUC of 0.911.

Knowledge Fusion

When we combine the above mentioned techniques together with the fusing method, we can see that the number of triples that we are uncertain about has gone down, and false positive rates have been reduced. The paper gives an example showing the extracted triple's confidence rising from 0.14 to 0.61 with the fused system. The paper also discussed false negatives generated by using local closed world assumption (LCWA), though the difference is insignificant by comparing the result with human evaluators.

Differences with the Knowledge Graph

Knowledge Graph, considered as the predecessor of the new Knowledge Vault, is a knowledge base that Google uses to serve relevant information in a knowledge panel. It relies on crowdsourcing to collect and expand information. However, it involves immense human effort and there is a limit on how much humans can achieve. There comes the Knowledge Vault, where it uses algorithms to automatically collect information across the Web, using ML to connect data into usable information. The scale of KV is much larger than that of Knowledge graphs'. According to the paper, KV has 1.6 billion triples, and 271 million have a confidence interval of 0.9 and above. [3]

[2] https://research.google/pubs/pub45634/
[3] https://research.google/pubs/pub45634/

Conclusion

Knowledge Vault is the way Google implements the Web-scale knowledge base with a probabilistic model. By fusing together different extractions methods, along with sources of prior knowledge could the KV achieve high probability of predicting the “truth” and connect the missing links. Though there are still many areas the system could improve. In reality, many triples are correlated but we would treat them as independent in the Knowledge vault. KV could improve on modeling mutual exclusion/ soft correlation between facts. In a different scenario, some facts are only true for a period of time and how can we deal with that? Facts can be presented at different levels of abstraction. How would we classify those facts? Would they count as one single fact to save storage or multiple facts? No doubt that Knowledge Vault is the future for knowledge bases, but these are the questions we yet to resolve.