University of Illinois Urbana-Champaign

Computer Science - CS410 Text Information Systems

Technology Review

Tony Law

Google’s Knowledge Vault

Knowledge bases are a common structured system to store organized information about the world. The rapid increase in recent years in large scale knowledge bases such as Wikipedia, Freebase and Google’s Knowledge Graph was driven by the ongoing need of coming up with an efficient and structured way for automated reasoning and automating the construction of a web-scale knowledge base.

Google Knowledge Vault offers a novel way of automating the process KB construction that is backed by a probabilistic inference engine to give weights to the derived relationships of information entities. This approach offers more fact correctness and is helpful in minimizing errors that may come from spam sources on the web.

Knowledge Vault takes a different approach from other KBs by treating facts separately from the language that describes them. This method is very useful in eliminating some redundancy that results from the lexical representation of the same fact or entity. In addition to that, KV adds a weighted score to indicate the confidence of certain relationships between entities, which is very useful in calculating priors to new or unknown entities based on previously constructed known entities. This approach is vital to consuming large-scale information from the web as it automatically eliminates the noise and continuously evaluates new information based on existing knowledge.

Since Knowledge Vault is mainly designed for consuming web text data, building efficient text extractors that accommodate the majority of document formats is essential to build the KV by fusing four different extraction methods together, averaging over their sources and computing the different weights of each based on machine learning models using a training set of local world assumptions. The four extractors handle raw text data, HTML DOM trees, HTML tables and other sources for Human Annotated pages. This fusion of sources not only increases the confidence of evidence but also also helps the systems in classifying the information based on their features.

The core part of Knowledge Vault however is the way it constructs new knowledge entities based on priors. The information on the web, while it offers a great value of quantity, is also challenged by the fact that a great deal of it can be unreliable and inaccurate. To combat this problem KV also combines different weighting systems or priors. Using Path Ranking Algorithms and Neural Network Models offers a probabilistic way to learn the path or relationship between two graph entities. Neural Networks can derive the meaningful semantic representations of entities. The benefit of combining these two different ways increases the confidence values as well as increasing the overall performance and since each prior has its pros and cons, combining the two helps in minimizing their biases.

While Knowledge Vault follows the same direction of other systems in automating the extraction of web information, the main difference is how well it combines and fuses the facts using different extractors and then fusing and combining the priors with help of existing knowledge.

The result is that KV is almost 40 times bigger than other similar bases and offers better calibrated fact weights to support its construction of knowledge about the world.