Tech Review: Google Knowledge Vault

Zachary Oldham (zoldham2)

Motivation

The goal of the Google Knowledge Vault is to facilitate the automatic construction of a "web-scale probabilistic knowledge base". The inspiration for this tool comes from the stagnation in the construction of existing, human-made knowledge vaults such as Wikipedia, in recent years. The thought is that we are reaching the limits of the knowledge that volunteers will freely contribute to such systems, and as such require a means of expanding our knowledge vaults automatically. The Google Knowledge Vault aims to accomplish this task.

Implementation

The Google Knowledge Vault stores data much differently than human-made knowledge vaults which typically consist of a series of articles about subjects that contain all relevant information about said subject (or links to articles that contain additional information). Rather, the Google Knowledge Vault makes use of so-called RDF triples that hold a subject, a predicate, and an object. The subject would be the entity that the tuple contains information about, the predicate would be the type of information, and the object would be the information itself. As an example, the object might be "Zachary Oldham", the predicate might be place of birth, and the object might be "Mason, Ohio, USA".

Each RDF triple also has an associated probability that represents the probability with which the system believes this RDF triple to hold accurate information. These probabilities are the reason the system is referred to as a "web-scale *probabilistic* knowledge base".

The process of actually obtaining these RDF triples is where the bulk of the work lies. The process involves several steps, the first of which is to extract candidate RDF triples from a

variety of text-based sources, including text documents (webpages), HTML trees, HTML tables, and webpages specifically annotated for tuple extraction. For each of these source types, preliminary candidates are generated using a variety of NLP techniques and are fed through binary classifiers, one per predicate, to create a set of candidates. The output candidates from each source type are combined using additional binary classifiers, again one for each predicate. The result of this process is a reduced set of candidate RDF tuples which are then assigned probabilities using Platt Scaling.

One of the major obstacles that this system faces is the fact that the internet is host to a wide array of misinformation, intentional or not. As such, these candidate tuples may be unreliable even if the system found plenty of "evidence" for them. To combat this, the Google Knowledge Vault incorporates prior knowledge about the world into the system before making final decisions about tuples. To accomplish this, it makes use of existing knowledge vaults and attempts to expand on them by identifying patterns in triple types, and using neural approaches. This process results in a set of "prior knowledge" candidate tuples which are merged with the extracted candidate tuples again using binary classifiers, one per predicate. The output of this fusion is the final set of RDF triples, which are again assigned approximate probabilities using Platt Scaling.

This system generated a vast number of RDF triples. In the end it generated over 1.6 billion triples. Of these, 160 million had a probability of over 0.7 and 100 million had a probability of over 0.9.

Evaluation

In order to evaluate their system, the researchers again exploited existing knowledge vaults, specifically FreeBase, along with manual human labeling, to generate a test set upon

which the performance of the system was evaluated. This test set consisted of 10 thousand ground truth triples. The resulting area under the ROC curve was 0.869, which is not perfect but still very impressive given the scale of the system.

Conclusion

The Google Knowledge Vault is not unique in its aspirations; many other systems have been created that attempt to automatically construct knowledge bases, some by looking at structured source such as Wikipedia, and others that also attempt to extract information from unstructured text. Unlike these systems, the Google Knowledge Vault is 38 times larger than the next largest automatically constructed knowledge vault, representing a huge leap forward in the field.

Despite its size, the Google Knowledge Vault is nowhere near exhaustive, as research by Razniewski Et al. suggests. A complete knowledge vault covering all predicates, subjects, and objects is impossible, and a knowledge vault covering even just a small subset of predicates is likely impossible, but as these systems mature and expand we may get closer to the ultimate goal of a complete representation of all human knowledge.

Systems like the Google Knowledge Vault could prove to be hugely valuable in the future. An obvious use case is like that of Wikipedia, aiding people in learning about the world, but any less obvious but hugely important use cases exist. As an example, a major limiting factor in many NLP tasks is that the systems do not have contextual knowledge about the world, something that is hugely important in understanding and generating natural language. A system like the Google Knowledge Vault could provide the context that these systems lack, massively improving them. Many other such use cases exist, making the future of the Google Knowledge Vault and other similar systems very exciting.

References

[1] Dong, Xin, et al. "Knowledge vault: A web-scale approach to probabilistic knowledge fusion." Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining. 2014.

[2] Razniewski, Simon, Fabian Suchanek, and Werner Nutt. "But what do we actually know?." Proceedings of the 5th Workshop on Automated Knowledge Base Construction. 2016.