# Tech Review: Google Knowledge Vault

Early method of building a knowledge base relied on human volunteers to upload knowledge by themselves, such as Wikipedia. Now it is possible to build a knowledge base by automatically extracting information. This paper introduces a method for building a large-scale knowledge base called Knowledge Vault. This approach extracts information on the web and information already existing in the knowledge repository and builds a knowledge base by fusing different pieces of information through supervised learning. This tech review paper will review how the Knowledge Vault is built and what problems it still has.

In this paper, knowledge information is represented as a triple such as (subject, predicate, object). Knowledge Vault has three components: extractors, graph-based priors, and knowledge fusion.

Extractors is to extract triples from web dataset and assigned confidence scores to each triple. In this part, the authors initially proposed four extraction methods: text documents, HTML trees, HTML tables, and human annotated pages. For the above four approaches, the results need to be fused. The method is to build a feature vector for the result of each extractor and then make a classifier to determine which extractor's result is the right one. The feature vector is composed of the square root of the number of sources used to extract this triple and the average of the scores scored by all the sources extractors used to extract the triple. For each predicate, a separate classifier is fitted. Since it is not necessary to have the same standard for each extractor confidence score, the authors employ Platt Scaling. The text documents system extracted the largest number of triples and the highest confidence score, but the HTML tables system extracted the least. As more sources are available, the prior probability of the true triple is approximately close to 1, and the prior probability of the false triple is close to 0.5.

Graph-based priors learn the prior probability from the existing knowledge vault to verify if the information is reliable. In this paper, the authors use Freebase as the existing knowledge vault to assign a probability to each triple. This problem can be considered as link prediction in a graph, so the authors use two algorithms to solve it: path ranking algorithm and neural network model. For path ranking algorithm, it begins with subject node and finds paths to reach the object node based on a predicate. The algorithm fit a binary classifier to combine paths that have same start-end nodes. Neural network model considers the link prediction problem as matrix completion and calculate the prior of the relation on the graph. Furthermore, the different priors will be combined by the previous fusion method.

The final step is knowledge fusion, which combines extractor and prior results to calculate the probability of a triple being true. The method of fusion is similar to the previous extractors' fusion.

In addition, the authors present some issues of knowledge vault. The first problem is how to model mutual exclusivity between facts. In this paper, the authors consider each triple as an independent binary variable, but in real world many triples are mutually exclusive with each other. Secondly, there is soft correlations between facts, which means some facts are constrained by another facts. Thirdly, "values can be represented at multiple levels of abstraction." The same information can be expressed in several ways. Fourthly, with the increase of data, information will become large and overlapping. How to find a new approach to deal with related data is a problem. Fifthly, some facts are time limited. A fact may have been true in the past but is wrong now. Finally, finding new information and adding the new relation to the knowledge base is also a problem.

In conclusion, we review Google Knowledge Vault that is constructed by fusing facts extracted from the web with knowledge from an existing knowledge base and calculating confidence levels and review what other issues had not been addressed.

# Reference

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.