# Technology Review on the Google Knowledge Vault

Luhao Wang

## Introduction

Knowledge bases or repositories are constructed to store facts of the world. For instance, persons, locations, and physical objects can all be referred to as entities in a knowledge base, where a substantial amount of information is involved. The Google Knowledge Vault is a web-scale probabilistic improvement of knowledge base that largely relies on supervised machine learning. The noisiness of text-based information extractions has been a major problem of previous approaches. The Google Knowledge Vault is therefore introduced to provide a fusion of extractions from web content. Besides text, the information extractions can be applied to the analysis of tabular data, page structure, and human annotations. Studies have shown that the scale of Google Knowledge Vault is significantly larger than any preceding structured or unstructured knowledge base.

## Motivation

Previously published knowledge repositories mainly require unsolicited contributions of human volunteers, with the support of integrations from other existing knowledge bases. This leads to a problem where portions of contents stored in repositories are inevitably biased, in another word, dominated by most popular entities and properties and largely affected by the publicity of events. Moreover, contributions from human volunteers are substantially unstable and unreliable. Researches indicate that the growth of knowledge bases, such as Wikipedia, has been stagnant, largely due to the inefficiency and randomness of relying on human contributions.

To resolve these issues in an effective way, Google Knowledge Vault aims to automate the process to eliminate the negative human factors of knowledge bases.

## Implementation

The GKV (Google Knowledge Vault) keeps information in an RDF triples (subject, predicate, object) manner, where each triple is associated with the confidence score indicating the probability of that information being correct. Compared to open information extraction approaches that normally maintain different entries with alternative word choices for an identical fact [2], such as <Bruce Lee, hometown, China> and <Bruce Lee, comes from, China>, GKV handles the facts and their lexical representation separately, thus being language independent. Therefore, the issue of having to keep redundant entries in the knowledge base as the given example can be resolved.

Generally, GKV can be divided into three major components: extractors, graph-based priors, and knowledge fusion.

The purpose of extractors is to extract information from web sources and form RDF triples along with a score of confidence level assigned to each of the triples. We consider a triple with a confidence of 0.9 or higher to be a confident fact. Before delving deeper, some statistics should provide a straightforward demonstration of the scale of GKV. The GKV contains 1.6 billion triples, where 324 million of them are associated with confidence 0.7 or higher and 271 million have the confidence of 0.9 or higher. To be specific, GKV has 38 times more a count of confident facts than the system previously known to be the largest [3]. To support such a large knowledge repository system, various cutting-edge extraction methods are applied to enable extraction from text documents, HTML trees (DOM), HTML tables (TBL), and human-annotated pages. For instance, to perform standard NLP analysis on documents, entity recognition, part of speech tagging, dependency parsing, co-reference resolution, and entity linkage have been used [1]. Another representative example is to use the Local closed world assumption (LCWA) to label the extractions, and then apply logistic regression to the labeled extraction set per predicate by using MapReduce. Another key part of extractions is the calibration of estimates. At the stage of the final extraction set, the confidence scores from different extractors can be on scales that mismatch. Thus, to be able to process the scores as valid probabilities, a method named Platt Scaling [4] is used. Platt Scaling ultimately completes the extraction process by fitting a logistic regression model to the confidence scores [1].

Graph-based priors provide safety, as facts of the real world extracted from web sources are sometimes unreliable. Specifically, knowledge existing in Freebase, one of many pre-existing large-scale knowledge bases, is adopted to assign a probability to any possible triple. Depending on the degree of precision of existing triples in Freebase, confidence scores can be adjusted accordingly to lower the possibility of incorrect facts having relatively high scores. However, from my perspective, information extracted from web sources is extremely unreliable and simply using prior knowledge is not sufficient to correct the errors or prevent incorrect entries at an acceptable level. I deem that a more effective way to handle inconsistency and unreliable extractions is to create distinct knowledge base models, each containing different extractors targeting diverse web sources through alternating paths, and then fuse the resulting knowledge triples to compute final confidence scores.

The fusion of knowledge is to combine implementations of extractor systems and prior systems by creating a feature vector for each extracted triple and then applying a binary classifier [1]. To train the systems of fusion, boosted decision stumps [5] provide sufficient performance for labeling the training set.

## Conclusion

This review mainly explains the motivation and implementation of GKV (Google Knowledge Vault), a web-scale probabilistic extension of pre-existing knowledge bases such as Wikipedia and Freebase. Compared to previously published knowledge bases, GKV has a substantially larger scale (specifically 38 times larger than the largest comparable system), automatic methods of construction, and enables extractions from various types of web content. The introduction of GKV is necessary and natural, as being limited to text extractions is nowhere near the destination of knowledge repositories. Though still not optimal nor exhaustive, the GKV has brought us a step closer to the goal of having one knowledge base to store all knowledge of this world. Through exploring its implementation and comprehending the concepts that lie behind it, I deem that GKV will have a promising future.

# References

[1] Dong, X., Gabrilovich, E., Heitz, G., Horn, W., Lao, N., Murphy, K., ... & Zhang, W. (2014, August). Knowledge vault: A web-scale approach to probabilistic knowledge fusion. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 601-610).

[2] A. Fader, S. Soderland, and O. Etzioni. Identifying relations for open information extraction. In EMNLP, 2011.

[3] F. Niu, C. Zhang, and C. Re. Elementary: Large-scale Knowledge-base Construction via Machine Learning and Statistical Inference. *Intl. J. On Semantic Web and Information Systems,* 2012.

[4] J. Platt. Probabilities for SV machines. In A. Smola, P. Bartlett, B. Schoelkopf, and D. Schuurmans, editors, *Advances in Large Margin Classifiers.* MIT Press, 2000.

[5] L. Reyzin and R. Schapire. How boosting the margin can also boost classi_er complexity. *In Intl. Conf. on Machine Learning*, 2006.