**Review - Knowledge Vault: A Web-Scale Approach to Probabilistic Knowledge Fusion**

Xin Luna Dong, Evgeniy Gabrilovich, Geremy Heitz, Wilko Horn, Ni Lao, Kevin Murphy, Thomas Strohmann, Shaohua Sun, Wei Zhang

**Brief summary:** This paper describes in detail Google’s knowledge database, *knowledge Vault*. Authors present a way of automatically constructing a Web-scale probabilistic knowledge base. The proposed method deals with the problem of limited knowledge of the traditional knowledge bases, wherein the information is mainly provided by the humans. Furthermore, KV creates low noise data by automatically leveraging already-cataloged knowledge.

**Motivation:** In the recent years, various large-scale knowledge bases like Wikipedia have emerged. These KBs have been very popular, however, the knowledge stored in these are mostly annotated by the humans which may yield limited amount of knowledge. Despite their seemingly large size, these knowledge repositories are far from complete. Thus, there is a need of new approach that could scale up the automatic construction of noise-free knowledge base.

**Literature review:** The previous approaches for creating a knowledge base were mostly text-based extractions or were mostly annotated directly by the humans. These approaches worked fine but could be very noisy and incorrect sometimes. The present work is based on the idea of fusing web contents with the prior knowledge and thereby generating less noisy knowledge base.

**Methods:** Knowledge Vault is a probability triple store database. Each entry in the database is of the form subject-predicate-object probability restricted to about 4500 predicates. The database was built by combining the knowledge base Freebase with the Wikipedia and approximately one billion web pages. The information from the Web was extracted using standard natural language processing (NLP) tools including named entity recognition, part of speech tagging, dependency parsing, co-reference resolution, and entity linkage. Probabilities for each triple store are calculated using logistic regression. The main components of KV include: a) *extractors* - to extract triples, b) *graph-based priors* – to learn prior probability of each triple, c) *Knowledge fusion* – to compute probability of a triple being true. The facts extracted using the various extraction techniques are fused with logistic regression and boosted decision stumps. Implications of the extracted knowledge are created using two techniques: the *path ranking algorithm* and a *modified tensor decomposition*. The path ranking algorithm can guess that if two people parent the same child, then it is likely that they are married. Tensor decomposition is just a generalization of singular value decomposition, a well-known machine learning technique.

**Results:** Through this paper the authors clearly presented the main contributions of the Knowledge Vault. Knowledge Vault combines noisy extractions from the Web together with prior knowledge overcoming the errors due to extraction process. The resulting knowledge base is much bigger with better confidence as compared to the other KBs. Moreover, KV uses multiple extraction sources, such as Free text, human annotations of Web pages and HTML DOM trees and Web tables, to perform a detailed comparison of the quality and coverage of different extraction methods.

**Discussion:** Although, Knowledge Vault gives good results but there is still scope of improvement. A) The current technique represents each fact as independent binary random variable but, in reality many triples can be correlated. B) The paper doesn’t show how to accommodate for the facts that are temporarily true, e.g. President of ABC country may not be same all the time. C) RDF representation is a good option for factual assertions but how can the difference between music types (e.g. Pop and Rock) can be represented? D) Though web has a lot of information available but still it’s not complete. Some crowdsourcing techniques can be incorporated to acquire more knowledge.

**Conclusion:** A web-scale probabilistic knowledge base called Knowledge Vault is introduced wherein the idea is to fuse together multiple extraction sources with prior knowledge derived from an existing KB. This resulted in a KB, 38 times bigger than existing KBs. The facts in KV have associated probabilities, that are well-calibrated, and can be distinguished by what is known with high confidence from what is uncertain.