Alexander Zurawski

azuraws2

CS410

11/4/2021

Tech Review: Knowledge Fusion for the Knowledge Vault

Relationships define how humans understand the world. Between two people can be the relationship of marriage. Between a city and the number 5000 can be the relationship of population. Knowledge bases hold millions of these relationships that have mostly been entered manually or taken from another existing knowledge base, whose new entries would have been entered manually. "Knowledge vault: A web-scale approach to probabilistic knowledge fusion[1]" explores a novel technique to increase the capabilities of such databases by combining diligent human effort with robust computer processes.

Existing knowledge databases are created using a mix of existing knowledge bases and direct human contribution; however, additions to public knowledge bases such as Wikipedia have severely dropped, which suggests new methods may be necessary to scale the construction of knowledge bases. One method is to automatically extract information across the ~1 trillion web pages on the internet; however, web exploration extracts noisy and sometimes unreliable facts. This will be compensated by prior knowledge extracted from existing knowledge bases constructed from direct human contribution. By mixing the two, a better balance of breadth and depth can be achieved.

The aspects that fuse together to define this specific approach are RDF triples, extractors, and priors. The RDF triples are how information in the knowledge vault is represented. Each triple has a subject, predicate, and object such as (subject: book, predicate: written by, object: author). There is also the confidence score that is a gauge of how sure the model is of the fact. A score close to 0.0 would indicate assured falseness, while scores close to one would indicate assured trueness. The amount of triple with a score in between 0.3 and 0.7 is to be minimized, as these would have high levels of unsureness.

Extractors are the gatherers of data and assigners of confidence scores. The four different extractors gather information from Document Object Models, Text Documents, HTML Tables, and Human Annotated Pages. The best result was a mixture of all extraction methods, as the confidence score of a triple increased both when extracted from more extractor methods and observed in more documents.

The prior knowledge methods in the implementation are path ranking algorithm and a neural network model. The fascinating path ranking algorithm finds relationships between two related entities other than their direct predicate. These new paths create rules that are applied on entities of the same types that the rules were constructed on to create new assumptions about previously unlinked entities. Instead of finding relationships between groups of entities, the neural network model finds clusters of relationships. Using a matrix representation, the neural network returns the nearest neighbors of relationships. As with the extractor, the best results occurs when mixing both strategies.

As seen by the below graphs, combining the extractor methods and prior methods improves results.

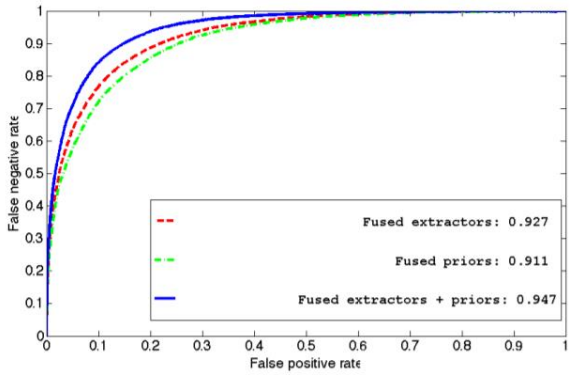

Figure 4: ROC curves for the fused extractor, fused prior, and fused prior + extractor. The numbers in the legend are the AUC scores.

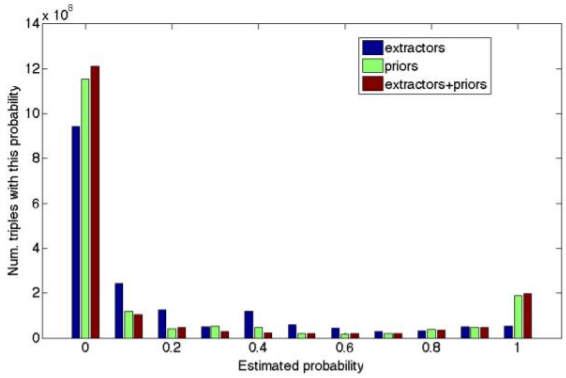

Figure 5: Number of triples in KV in each confidence bin.

Figure 4[1] shows the combined method produces a better ROC curve and Figure 5[1] shows the combined method reducing the number of triples in the unsure zone between 0.3 and 0.7.

The combination method has proven to be useful, but there are still areas that can be explored. Relationships are seen as a binary classifier in this model; however, they have the potential to tell a lot

more. Relationships can have correlations with others to assist in predictions. Relationships can be constrained, and they can change over time. Additionally, relationships cannot tell the whole story. They cannot articulate the differences between things or the question 'Why?'. Finally, not everything morsel of human knowledge is on the web, which limits how much the vault can hold. Regardless, mixing human intelligence with machine capabilities is an ingenious way to take on the knowledge vault task. The methodology explored imparts the importance of mixing techniques to solve complex problems.

[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.