Azim Keshwani

Introduction

Over the last couple decades, the rise of the internet has been accompanied by the ability to quickly and efficiently find information. Large-scale knowledge bases, such as Wikipedia, Freebase, YAGO, and others have become more prominent. These knowledge bases allow users to gain access to an incredible amount of information about various topics in a matter of seconds. However, as effective as these knowledge bases are, they do have some drawbacks. The information in these knowledge bases can sometimes be incomplete; there can be people with missing birth dates, nationalities, family info, etc. Part of the reason for this is that these knowledge bases heavily rely on humans to supply the information. Humans manually maintain these knowledge bases, and it is very difficult for every piece of information to be captured given human limitations and the amount of information. To solve this issue, Google has developed an automatically constructed Web-scale probabilistic knowledge base called the Knowledge Vault (KV).

What is the Knowledge Vault?

Essentially, the KV is a knowledge base that, like many other knowledge bases, stores information in the form of RDF triples. Each triple contains a relation and also a confidence score, which is the probability that the KV believes the triple is correct. The KV has three main features that differentiate it from other knowledge bases. First, KV combines noisy extractions from the Web together with prior knowledge, which is derived from existing knowledge bases. This approach can help KV overcome inaccuracies in web extracted information or information from the source. For example, if we have in the source the fact that a person is the president of the United States, then we can infer that this person must have been born in the US. So, if the web extractor returns results that suggest otherwise, then we can mark this as inaccurate.

The second feature of KV is that it is much larger than other knowledge bases. KV has 1.6 billion triples, 324 million of which have a confidence of 0.7 or higher. This is about 38 times more than the largest previous comparable system, which has 7 million confident facts. KV is able to achieve this size by extracting facts from a large variety of web data. This means that KV is able to potentially alleviate the issue that plagues current knowledge bases, which is the incompleteness of information.

The third key feature of KV is the fusion of extractors and priors. As aforementioned, KV compares the triples received by the extractors to the prior, or current, information. It also uses multiple extraction sources and methods. The benefit of doing this is that the confidence of certain facts can be increased or decreased accordingly. For example, compared with considering only extractions, combining priors and extractors increases the number of high confidence facts (those with a probability greater than 0.9) from about 100 million to about 271 million. In addition, combining extractions with priors decreases the amount of low confidence triples. KV is able to make a more definitive decision about whether or not a fact is true or false.

Conclusion

Overall, KV is an innovative upgrade to the current knowledge bases that currently exist. Current knowledge bases often contain inaccurate or missing information. They also heavily rely on humans to maintain the information. KV can alleviate both of these issues. It can perform quick comparisons between new and existing information to ensure that the facts are as accurate as possible, while at the same time assigning a confidence score to each fact. In addition, it does not rely on humans so it can hold much more information. Of course, this does not mean that KV is perfect, but it is a good first step to improving the current form of knowledge bases.

References

Dong, Xin Luna. *Knowledge Vault: A Web-Scale Approach to Probabilistic Knowledge Fusion*. https://storage.googleapis.com/pub-tools-public-publication-data/pdf/45634.pdf.