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

Introduction

Large-scale knowledge bases (KBs) store millions of facts about the world, such as information about people, places, and things. Despite their seemingly large size, these repositories are still far from complete. Standard methods for this task (cf. [44]) often produce very noisy, unreliable facts. The authors proposed a new way of automatically constructing awebb-scale probabilistic knowledge base.

Methods

The authors use relatively standard methods for relation extraction from text, but they do so on a much larger scale than previous systems.

They first run a suite of standard NLP tools over each document. These perform named entity recognition, part of speech tagging, dependency parsing, co-reference resolution, and entity linkage.

For each predicate of interest, they extract a seed set of entity pairs that have this predicate from an existing KB.

They then find examples of sentences in which this pair is mentioned and extract features or patterns from all of these sentences.

Results

Using the methods to be described, the authors have extracted about 1.6 billion candidate triples, covering 4469 different types of relations and 1100 different types of entities.

About 271 million of these facts have an estimated probability of being true above 90%; they are called "confident facts."

To evaluate the quality of the methods used, the authors randomly split this data into a training set (80% of the data) and a test set (20% of the data). The labels are inferred for these triples using the method described below.

To ensure that certain common predicates did not dominate the performance measures, at most 10,000 instances of each predicate were taken when creating the test set.

The samples were pooled from each predictor to get a more balanced test set.

Discussion

Knowledge Vault is a large repository of useful knowledge, but there are still many ways in which it can be improved.

For a functional relation such as born-in, there can only be one true value, so the (s, p, oi) triples representing different values oi for the same subject s and predicate p become correlated due to the mutual exclusion constraint.
A simple way to handle this is to collect all candidate values together and force the distribution over them to sum to 1. This is similar to the notion of an X-tuple in probabilistic databases.

We might have a fact that Obama was born in Honolulu and another one stating he was born in Hawaii. These are not mutually exclusive, so the naive approach does not work.

Conclusion

In this paper, a description of how to build a web-scale probabilistic knowledge base, called Knowledge Vault, has been proposed.

In contrast to previous work, the fusion of multiple extraction sources with prior knowledge derived from an existing KB has been implemented.

The resulting knowledge base is about 38 times bigger than existing, automatically constructed KBs. The facts in KV have associated probabilities, which we show are well calibrated so that we can distinguish what we know with high confidence from what we are uncertain about.

The authors hope to continue to scale KV, to store more knowledge about the world, and to use this resource to help various downstream applications, such as question answering, entity-based search, etc.

Google Knowledge Graph vs Google Knowledge Vault

The combining of several ways of extraction and validation to create a much more solid and trustworthy database entities is where the actual value of the methodology described in Knowledge Vault lies.

In other words, Google has figured out how to expand and strengthen the Knowledge Graph. In essence, Google is triangulating the apps that make sense out of all the entity applications.

The greater part of extractions originate from Google's crawler of websites' HTML trees, which is unstructured data that they transform into useful information. But the need for organized data remains. If you want to be a part of the search engine industry in the future, you must take this action.However, it might indicate that the markup isn't perfect and that some sites are using it incorrectly or in a self-serving manner. alternatively accurately enough to be utilized as an entity fact.

The fact that humans are involved in the Knowledge vault process is another noteworthy conclusion. This is mentioned in the report and adds some additional support to the notion that Google wants to have some form of author- or expertise-based ranking but wants to implement it via entities rather than explicit markup.

There is a component of link prediction and how Google is able to provide entities more context by knowing the kinds of information they should look for. They anticipate acquiring information about each individual such as "spouse," "date of birth," "place of birth," "gender," "parents," "children," etc. It is quite powerful to be able to produce these connection predictions and automatically search them out. The usage in these variety of scenarios is expected to increase as the Knowledge Vault develops, and Google's "trust" in it. This is moving Google closer to its ultimate goal of becoming a "ask me anything" Star-Trek computer. Although a substantial body of knowledge would be needed to accomplish this, if it were put together it would open up a world of new applications and computer possibilities.

The "Knowledge Vault" technology and corpus, may have consequences for third-party publications going forward as Google depends less and less on them. But to some extent, Google's business strategy modifies that.