Technical Review on Google Knowledge Vault

Zhangzhou Yu

PageRank has been an integral part of web-scale text information retrieval tool. Even though the inflow and outflow relationships of web links well capture authoritative webpages and webpage hubs, it fails to identify whether all the information listed in webpages are facts or false information. Furthermore, a highly ranked webpage does not necessarily contain more correct information, whereas a lowly ranked webpage may include a lot of true and useful information. In order to tackle this problem, Google launched research on this topic and created a Web-scale probabilistic knowledge base that combines extractions from Web content (obtained via analysis of text, tabular data, page structure, and human annotations) with prior knowledge derived from existing knowledge repositories called Google Knowledge Vault. This paper gives a brief overview of the Google Knowledge Vault approach along with challenges it faces and remediations made to address those challenges.

There are three major components of Knowledge Vault: Extractors, Graph-based priors and knowledge fusion. Each extractor returns <subject, predicate, object> triples from a huge number of Web sources and assigns a confidence score to the extracted triple. The Graph-based priors learn the prior probability of each possible triple based on triples stored in existing Knowledge Bases such as Freebase. Finally, knowledge fusion computes the probability of a triple being true given an extractor and priors.

There are some findings from the Google Knowledge Vault research. First, the Google Knowledge Vault is much bigger than any other existing knowledge bases. A large variety of sources of Web data, including free text, HTML DOM trees, HTML Web tables, and human annotations of Web pages, are extracted; more usable triples are extracted from semi-structure data such as HTML trees and HTML tables than the free text data. In addition, both graphic based model (path ranking algorithm) and neural network model are tried out in the prior model and they both have similar performance. The prior model uses prior knowledge of the subject’s common grounds with another subject to make predictions. For example, a person is likely to attend the same school as their sibling.

However, when it comes to implementing the above components, there are a few underlying challenges. The biggest and foremost is extraction errors, for instance, mismatching entity, triple identification, linkage prediction and source data errors. An extractor could take a person’s place of birth as date of birth or misidentify the person as the person’s parent or child. In addition, an extractor could take another person’s nationality in the same webpage as that of the person in the triple causing an incorrect prediction linkage. If the source webpage has incorrect information to begin with, then the extracted triple is false information. Secondly, the probability of a triple being true may not be well-calibrated; predicted probability may not align with actual probability for the predicted population. Lastly, the data is sparse and of low quality. There are a lot of URLs including very few triples (not enough trustworthy information) while a few URLs contain a lot of triples which creates a bottleneck in parallelization.

In order to remediate the challenges listed above, in the bootstrapping phase a sufficient number of predicates are labeled and trained independently using logistic regression to capture patterns occurring between pairs of entities of the correct type. This additional step helps eliminate the extraction error of mismatching entity type.

Expectation-Maximation of Bayesian approach is used. In the Expectation step, probability of webpages providing triples given the extraction quality as well as probability of triples being true given source data is calculated. In the Maximization step, the source accuracy along with the extraction precision and recall is computed. The harmonic mean is used here instead of the maximum likelihood method. This is because of the computation complexity and costs. This process is iterated till the harmonic mean eventually converges.

Moreover, in the evaluation of triple accuracy, the final accuracy is weighted by extraction accuracy to get a true picture of the prediction. And to solve the cases for small and large data sources, webpages containing large data source are split while webpages containing small data source are combined to increase data extraction efficiency and model training effectiveness.

As it is very hard to judge if a triple is true or false, a heuristic approach is used for evaluation. The assumption of local closed world is applied; if there is no record of the <subject, predicate> pair in the existing knowledge base then the probability of the triple containing the pair is unknow, whereas if there is existing record of the <subject, predicate> pair but we are unable to find the connection of this <subject, predicate> pair with the object in the triple then this triple is false. This assumption exposes an obvious limitation: there can only be a truth for a <subject, predicate> pair. There are exceptional cases to counter this assumption, which we will discuss later.

Although the Google Knowledge Vault research is already very informative, there are some ongoing discussions around it. The assumption of triples containing the same <subject, predicate> pair but different objects cannot be true at the same time does not hold at all times. One exception is when the objects are in hierarchical order. For example, Obama was born in Honolulu and Obama was born in Hawaii are both true. Another exception relates to the evolving timeframe. The statement of the president of USA was Obama and the statement of the president of USA was Bush both hold true.

It becomes especially tricky to determine the correctness of political views. Different data sources would give different standards of true or false. For example, the CNN would give completely different points of view from Fox News' reporting article. The labeling of true or false in controversial areas such as politics and unproven sciences is extremely difficult. Additionally, there could be duplicate information extracted from different webpages since some data sources may be correlated; the current weighing algorithm gives higher weight to authoritative webpages, which can be biased since not all facts on those pages is correct; scalability is an area of further improvements for Google Knowledge Vault given the large amount of information on the Web.

In a nutshell, this paper gives a brief overview of the Google Knowledge Vault approach from its major components, modeling approach to evaluation methodology and assumptions. The challenges of setting up this system along with their remediations are highlighted and further discussion of areas for improvements is also included this paper.

Reference

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

Knowledge vault: A web-scale approach to probabilistic knowledge fusion

Proceedings of the ACM SIGKDD international conference on Knowledge discovery and data mining, ACM (2014), pp. 601-610