Technology Review – Google Knowledge Vault

Alan Anderson alansa3@illinois.edu

Abstract

The purpose of this paper is to use the example of the research and publicity around Google's 'Knowledge Vault' to assess some of the limitations of automated data retrieval from the web as a means of identifying 'new' knowledge or assessing its credibility.

1. Introduction

Google's Knowledge Vault was the subject of a 2014 Google research paper, which promoted it as the next step beyond Google's Knowledge Graph, and declared that: "The goal of KV is to become a large-scale repository of all of human knowledge."[1] This lofty ambition was elevated still further in a contemporaneous New Scientist article, which declared: "It promises to let Google answer questions like an oracle rather than a search engine, and even to turn a new lens on human history."[2]

The motivation of the project was that voluntary human submissions to databases like Wikipedia have plateaued, and existing structured knowledge bases (e.g. the CIA World Factbook) have largely been exploited. The authors wanted to solve this problem by using the vast pool of 'knowledge' available on the public web, but confronted the issue that much of this data is noisy and unreliable. The premise of the research paper is that the automatic extraction of data from the web, combined with a 'prior' grounded in known facts from existing knowledge databases and trained with machine learning techniques, could enable an expansion of knowledge beyond what is available from existing knowledge databases (such as those used to compile Google's Knowledge Graph) and from voluntary human contributions (such as pages in Wikipedia). It would also enable a robust assessment of the probability that the 'facts' in the Knowledge Vault, including newly and automatically discovered ones, are true.

Eight years later, there has been little follow-up publicity on Knowledge Vault. Indeed, immediately after the New Scientist article, "Google indicated that the 'Knowledge Vault' was misrepresented or misinterpreted in the New Scientist article … Apparently this was a research paper (May 2014) and is not an active Google product in development."[3] Google Knowledge Graph, not Knowledge Vault remains its primary Knowledge Base, used to feed the "InfoBox" that comes up on the right of a Google search,[4] with a published API to access it.[5]

I will briefly outline the methodology of Google's Knowledge Vault, as outlined in the 2014 research paper. Then I will consider some of the issues which may have led to this initiative falling short of the hype, and the extent to which they may impede future attempts to create the Star Trek style 'oracle' foreshadowed by the New Scientist article.

[1] Xin Luna 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. Google, 2014.

[2] Hal Hodson. Google's fact-checking bots build vast knowledge bank. New Scientist, 20 August 2014.

[3] Greg Sterling. Google Knowledge Vault to Power Future of Search. Search Engine Land, 25 August 2014.

[4] Google Knowledge Graph, Wikipedia, https://en.wikipedia.org/wiki/Google_Knowledge_Graph retrieved 6 November 2022.

[5] Google Knowledge Graph Search API, https://developers.google.com/knowledge-graph retrieved 6 November 2022

2. Methodology of Google Knowledge Vault

Google Knowledge Vault is based on an entity-relationship model, and stores information as RDF triples (subject, predicate, object), each with a 'confidence' score of its correctness. It combines analysis of data from the Web with prior knowledge from existing structured knowledge repositories. It calculates a probabilistic inference of correctness.

Training and evaluating

To train the components of the algorithm requires a training data set, and to evaluate it requires a test data set. These were derived by making a "local closed world assumption". This was based on the Freebase database, derived from public databases such as Wikipedia as well as user contributions, which Google purchased and ultimately closed down (migrating the data to Wikidata and, indirectly, to Google Knowledge Graph). The assumption used to label the training and test data is that:

- If the 'subject, predicate, object' RDF triple (s, p, o) is in Freebase, it is correct
- If that RDF triple is not in Freebase, but some other RDF triple with the same subject and predicate (s, p, o') is present, then (s, p, o) is false
- If there is no RDF triple for that subject and predicate in Freebase, we exclude from the training/test data as we do not know whether it is false or merely missing from Freebase.

Extracting from the Web

Data was extracted from the Web using four means of extraction to obtain RDF triples:

- Natural Language Processing of text, including named entity recognition, part of speech tagging, dependency parsing, co-reference resolution (within each document), and entity linkage (to map to the Knowledge Base), looking for how a relationship is expressed between entities known to be linked by it, then using machine learning to search for similar expressions between other entities in order to build the classifier.
- Parsing of DOM trees, replacing the textual link between entities in the extractor above with the path between entities in the DOM tree to form the features against which to train the classifier.
- HTML tables (mapping the entities; assessing the predicate/relationship represented by the data column by comparing with data in Freebase)
- Human annotated tables (used for only 14 predicates, with a manual mapping based on schema.org)

The data from these four extractors is fused for each predicate, using a classifier that considers the square root of the number of sources and a mean score of confidence from the extractors.

Prior model

Two approaches were used to build the 'prior' for the RDF triples.

- A path ranking algorithm searched for paths that correlate with the predicate in Freebase. For instance, it identified that if a person had a sibling who went to a school, that makes it more likely that they too went to that school.

- A neural network model was used to build the prior, conceptualizing it as a graph of dimensions E x P x E (where E is the number of entities and P the number of predicates) with each entry set to a 1 or 0 depending on the truth of the relationship, then using the neural network model to associate a vector with each predicate to model the interactions (such that related predicates will be represented by similar vectors).

As with the extractors, the information from these two priors was fused using a binary classifier trained on a feature vector, this time representing the confidence values from each prior system and indicators showing whether each .prior system was able to predict or not.

Fuse priors with extractors

The data from the extractors was then fused with the data from the priors, using a similar approach. The paper notes that combining priors and extractors increases the number of high confidence facts (those with a probability greater than 0.9) from about 100M to about 271M.

3. Limitations of Google Knowledge Vault

The authors of the research paper candidly acknowledge and explain some of the shortcomings of their methodology, focusing on the limitations in its ability to draw more complex inferences. They identify examples of how they would like to improve its ability to assess truth, within the broad framework of their model (e.g. knowing that 'president' is a subset of 'politician', or that a child's birthdate is likely 15-50 years later than its parents'). I would assert that these are not the most fundamental limitations.

The gap between hype and reality for a system like Google Knowledge Vault is caused as much by the over-enthusiasm of the hype as by the limitations of the reality. Could a store of RDF triples, with a finite set of predicates upon which it was possible to find adequate training and test data, ever really "become a large-scale repository of all of human knowledge" [6] or "let Google answer questions like an oracle rather than a search engine, and even to turn a new lens on human history"? [7]

Versatility of the RDF triple

The RDF triple represents knowledge as a subject, predicate and object. With sufficient contortions, most human knowledge could probably be fitted into this format, but in many cases we would need some quite obscure or complex entities or predicates. The need to train the algorithm places a practical limit on how obscure and complex the entities and predicates can be. Google Knowledge Vault is noted in the research paper to be trained on 4469 predicates and 1100 types of entity. This is impressive, but the cap has most likely been set not by computational power, but by the availability of adequate training and testing data.

For simple examples, the RDF format it is perfectly adequate. "Barack Obama was born in Hawaii." "Bill Clinton is married to Hillary Clinton."

[6] Xin Luna 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. Google, 2014.

[7] Hal Hodson. Google's fact-checking bots build vast knowledge bank. New Scientist, 20 August 2014.

But human knowledge is much more complex than this. The authors, in acknowledging this issue, assert that RDF triples are probably less well-suited to representing the difference between running and jogging, or jazz and blues. Those examples seem surmountable to me. But what about representing that "High inflation in the year leading up to an election is correlated with large swings against the incumbent party", or "Approximately 25% of the price of a bleeding edge personal computer represents a premium for tech-enthusiasts that you can avoid if you are willing to wait a year" (incidentally, I have no idea if this second fact is true).

Real human knowledge is complex, and even if you could cram facts like those above into a contrived RDF triple (Subject: swing against incumbent party; Object: high inflation in the year leading up to an election; Predicate: is correlated with large swings against), clearly the entities and relationships will quickly multiply and fragment to the point that adequate training and testing data cannot be found.

Reliance on pre-existing and 'official' data sources

Google Knowledge Vault derives a prior from Freebase, and also uses Freebase to provide the training and test data for its algorithms. Freebase itself is derived from other 'reliable' structured data sources, and user contributions. So its accuracy and its assessment of confidence is dependent on the reliability of that pre-existing data.

While we cannot prove it, intuitively it seems reasonable to expect much of the 'bar trivia' knowledge is correct: birth dates, family relationships, locations of tourist attractions, etc. But if there are areas of knowledge where there is systemic bias or error in pre-existing sources, that will be transposed into Google Knowledge Vault.

By way of example, consider the question "Does wearing masks reduce your risk of Covid?", and imagine that there exists an RDF triple in our database to answer the question. In the early months of the pandemic, official advice was that it did not, from both the World Health Organization[8] and the United States Surgeon General.[9] Had an expanded Google Knowledge Vault been staying up to date on this area of knowledge, it would doubtless have attributed a reasonably high confidence to its assertion, based on the concurrence of official sources.

Yet many people did not accept this truth at the time, because they suspected that the authorities were lying, most likely to preserve limited stocks of masks for health care workers. This turned out to be true, but blind reliance on an 'official' data set would not have revealed it.

Subjectivity of knowledge

The masking example above takes us into politically controversial territory, at a point in history when the United States and other Western countries are experiencing a high polarization of opinion. While there will always be a grey area at the boundary of fact and opinion, recent years have seen contention about statements that would generally have been regarded as in the realm of fact.

[8] Jacqueline Howard. WHO stands by recommendation not to wear masks if you are not sick or not caring for someone who is sick. CNN, 31 March 2020.

[9] Deborah Netburn. A timeline of the CDC's advice on face masks. Los Angeles Times, 27 July 2021.

Scientific American published an article examining the extent to which bias infects the work of political 'fact-checkers', and proposing a regime of "adversarial fact-checking" (which the author somewhat optimistically believes will lead to "an infinite regress towards an uncertain truth").[10]

As we leave the 'bar trivia' categories of knowledge and move into more complex territory, the problem of subjectivity grows. In some cases, it may be that the machine-learning algorithms of a system like Google Knowledge Vault will genuinely uncover relationships and correlations which can better inform our assessment of probabilistic truth. But in other cases, it is likely that the algorithms will be dominated by the echo effect that drives noisy bubbles in human discourse on the Web.

4. Conclusion

The objectives of the authors of the Google Knowledge Vault research paper are admirable, and their results suggest that their methodology can meaningfully expand the body of 'knowledge' over which we can claim a high degree of confidence.

Yet we should be cautious of overselling the promise of this technology and its applications. Perhaps Google itself recognized this, leading to its response to the Scientific American paper in which it played own the significance of the Google Knowledge Vault work.

While automated extraction of data from the Web can help to supplement and check basic factual data, this is a long way from achieving the greater ambitions of recording all human knowledge, or answering questions like an oracle. The simplifications which are necessary to enable the machine-learning approach to be used – most notably the reduction of knowledge to a set of RDF triples – necessarily limit the application of the technology.

Perhaps the most important conclusion to draw is that, in assessing the promise of new technology, we should do our best to avoid hyperbole that is likely to be magnified in the media, and to emphasize the limitations and caveats along with the opportunities.

[10] Stephen J Ceci, Wendy M Williams. The Psychology of Fact-Checking. Scientific American, 25 October 2020.