GOOGLE KNOWLEDGE VAULT

(sbs7@illinois.edu)

Introduction –

An interesting study to create knowledge base by not primarily focusing on text-based extraction as it could be very noisy but by combining extractions from Web content. These Web contents are obtained via analysis of text, tabular data, page structure and human annotations. This uses prior knowledge derived from existing knowledge repositories.

Supervised Machine Learning concepts are applied for fusing multiple distinct information sources. This is bigger than any previously published structured knowledge repository and features a probabilistic inference system that computes calibrated probabilities of fact correctness.

Overview –

This knowledge vault contains 3 major components namely –

a. **Extractors** – These extract triples from huge number of web sources. Each extractor assigns a confidence score to an extracted triple which also represents uncertainties about the identity of the relation and its corresponding arguments.

b. **Graph-based priors** – These systems learn the prior probability of each possible triple, based on triples stored in an existing knowledge base.

c. **Knowledge fusion** – This system computes the probability of a triple being true, based on agreement between different extractors and priors.

A weighted labeled graph will be constructed, as very sparse 3d matrix G, E X P X E (E – Entities and P – Predicates).
In knowledge fusion, there will be a condition between the text extractions and prior edges.

Fact Extraction from the Web –

a. **Text documents (TXT) –** A relatively standard methods for relation extraction from text is performed, this has been made at a very large scale than current systems. A suite of standard NLP tools are run over each document to perform entity recognition, part of speech tagging, dependency parsing, co-reference resolution (within each document) and entity linkage (maps proper nouns and their c-reference to the corresponding entities in Knowledge bases). Further processing is conducted.

b. **HTML trees (DOM) –** Web pages are parsed to their DOM tree structures. These are either obtained from text pages or from “deep web” sources, where data are stored in underlying

databases and queried by filling HTML forms. Lexicalized path is used between two entities as a feature vector. The score of the extracted triples is the output of the classifier.

c. **HTML tables (TBL) –** There are over 570 M tables on the Web that contain relational information. Fact extraction techniques developed for text and trees do not work very well for tables, because the relation between two entities is usually contained in the column header, rather than being close by in the text/ tree. Thus, heuristic technique is used. Firstly entity linkage is performed as in the text case, then the relation is identities that is expressed in each column of the table by looking at the entities in each column and reasoning about which predicate each column correspond to, by matching to Freebase as in standard schema matching methods. Ambiguous columns are discarded.

d. **Human Annotated pages (ANO) –** Many webpages where the webmaster has added manual annotations, here scehma.org annotation is used. Many of these annotations are related to events or products, such information is not currently included in the knowledge vault. The score of the extracted triple reflects the confidence returned by the named entity linkage system.

Graph-Based Priors –

Facts extracted from the web can be unreliable, thus prior knowledge based is used to combat this. Existing triples in Freebase to fit prior models are exploited, this assigns a probability to any possible triple, even if there is no corresponding evidence for a fact is found on the Web. This is thought to be like link predication in a graph.

Here the prediction is based on that, there are existing edges which are known, how can we predict if other edges are likely to exist. Two approaches are tried to solve this problem as discussed below –

a. **Path ranking algorithm (PRA) –** Similar to distant supervision, set of pairs of entities are connected by some predicate p. PRA then performs a random walk on the graph, starting a tall the subject nodes. Paths that reach the target nodes are considered successful. Since multiple riles or paths might apply for any given pair of entities, they are fitted by using a binary classifier (logistic regression). In PRA, the features are the probabilities of reaching O from S following different types of paths and the labels are devices using the local closed world assumption.

b. **Neural network model –** An alternate approach to building the prior model is to view the link prediction problem as matrix completion. A low rank decomposition of the tensor by association a latent low dimensional vector to each entity and predicate and then computing the elementwise inner product. This model is comparable to the state of the art. Neural model has about the same performance as PRA when evaluated using ROC curves.

Fusing the priors –

Different priors are together fused since there is no longer need of any extractions. A boosted classifier is trained using the signals and calibrated with Platt scaling. Fusing two prior methods helps performance since the have complementary strengths and weaknesses.

Conclusion/Discussion –

It has been concluded that this knowledge vault is found to be very useful but there are still many improvements. Some of the issues are discussed below –

a. Modeling mutual exclusion between facts
b. Modeling soft correlation between facts
c. Values can be represented at multiple levels of abstraction
d. Dealing with correlated sources
e. Some facts are only temporarily true
f. Adding new entities and relations
g. Knowledge representation issues
h. Inherent upper bounds on the potentials amount of knowledge that we can extract.

This knowledge vault is about 38 times bigger than existing automatically constructed Knowledge bases. The facts in the knowledge vault have associated probabilities, which are well calibrated, so that it can be distinguished what is known with high confidence from what we are uncertain about.