Name: Steven Piquito

Student ID: Piquito2

University E-Mail Address: piquito2@illinois.edu

Course: CS410 Text Information Systems

Semester: Fall 2020

Chosen Technology Review: Google Knowledge Review

Introduction

Data, in its many forms such as text, images, web-blogs, excel spreadsheets and others, is the life blood of the recent advances and improvements in the field of data science as information contained within is increasingly used to improve and implement modern algorithms in fields ranging from computer science through to medicine. In the last two decades, data access has begun to change from not being readily available or accessible to academia and the public at large, towards now being available in abundance and little to no cost. This has in many respects changed the informational challenges of extrapolating from a scarcity of data towards refinement and selection from a broad range of data, often in different forms and content.

During the early 20th century, advances in structured data models such as SQL databases were observed as the cost of technology reduced and larger database designs were conceivable. In recent years however, the amount of unstructured data readily available on the internet provides an untapped source of information and poses a challenge towards traditional data models.

Certain newer data models such as Wikipedia and Freebase have been designed to tap the distributed user model of inputting data to great effect, however even these models remain largely dependent on human interaction, vetting and other touch points that could constrain truly large scale models of range of facts [2].

What is Google Knowledge Review

Google Knowledge Review (GKR) is an extremely large-scale knowledge base constructed using fully automated web-based (think web crawling) methods coupled with machine learning techniques to create a probabilistic-based Knowledge Vault (KV) with minimal human (subject matter expert) interaction.

In order to overcome the constraint of using supervised or semi-supervised techniques often requiring human interaction (to establish the ground truth), GKR scraps from multiple existing readily available data sources and stores mined key information in the form of Resource Description Framework (RDF) triples (subject, predicate, object). Each RDF triple is ultimately assigned a probability of accuracy (since

there is no ground truth to compare too) in an approach similar to that defined in [2] and based on a number of statistical techniques employed specifically in the GKR model.

Information derived in GKR considers prior knowledge and probabilities of accuracy from underlying data sources (e.g. GKR uses data from another large data sources called Freebase) in order to further improve and refine accuracy of the KV. Information contained in the RDF triples are then assumed to be true above a certain confidence interval to arrive at a new knowledge base containing information not previously observed.

Knowledge contained in RDF triples with an accuracy estimate of greater than 0.9 (>90%) are assumed to be highly accurate and factually credible.

The three major components of Google Knowledge Review

KV contains three major components that make up the model, namely:

- Extractors: these systems extract triples from a huge number of Web sources. Each extractor assigns a confidence score to an extracted triple, representing uncertainty about the identity of the relation and its corresponding arguments.
- Graph-based priors: these systems learn the prior probability of each possible triple, based on triples stored in an existing KB.
- Knowledge fusion: this system computes the probability of a triple being true, based on agreement between different extractors and priors.

What are the key features of Google Knowledge Review?

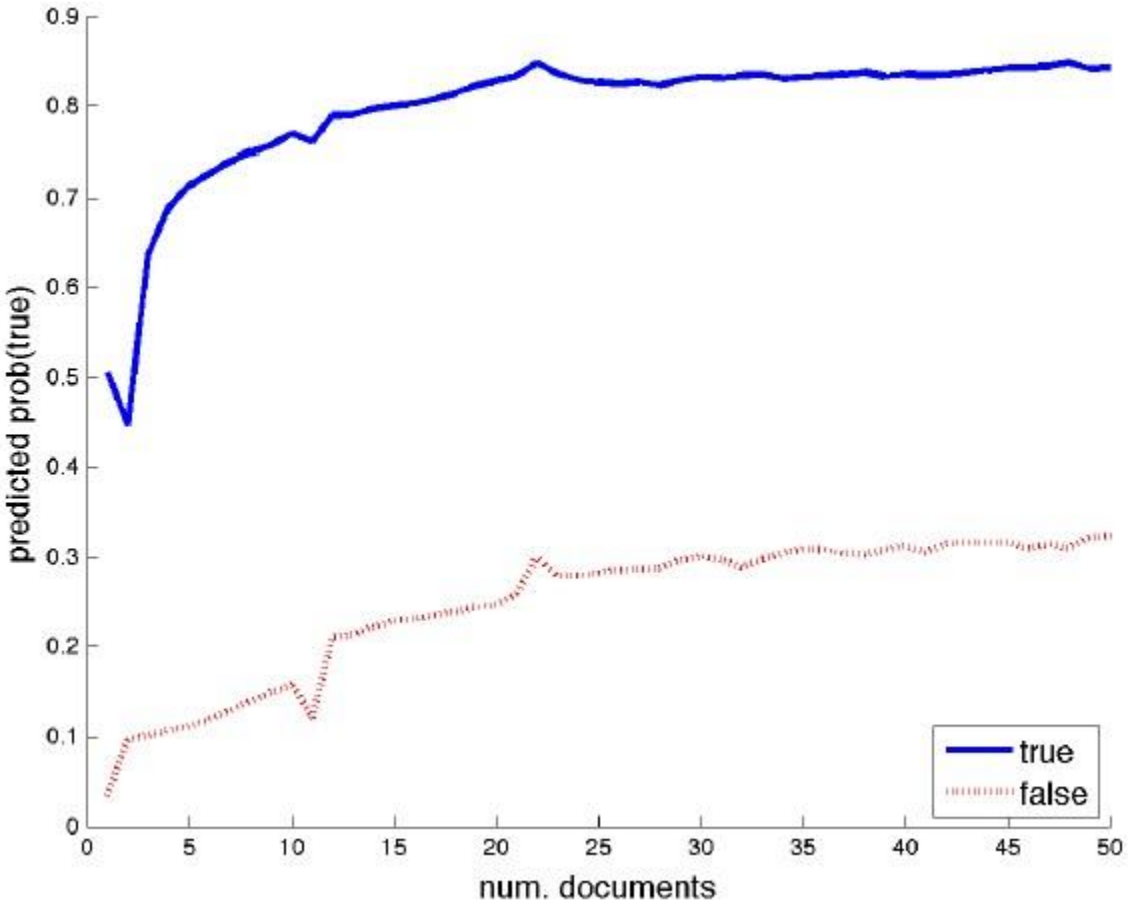
GKR has the following key features that make it quite unique amongst modern knowledge vaults and other repositories and include the following:

- Leverages existing a prior knowledge from underlying data sources where possible to improve the assigned probability of accuracy of RDF triples
- GKR is considerably larger than any other existing knowledge vault currently and has over 1.6B triples of knowledge (324M of which have a confidence of 0.7 or higher)
- GKR employs Graph-based techniques to improve and assign probabilities of accuracy on RDF triples based on extracted information from multiple sources coupled with prior knowledge
- GKR exploits inherent structure in modern web-based sources (such as Domain Object Models and HTML tables used in most webpages) to further improve the accuracy of determined knowledge
- GKR employs a heuristic known as the Local Closed World Assumption (LCWA) that inherently assumes a specific knowledge item's accuracy based on prior information on the existing data domain. This simplifying assumption is needed to establish a base line for new information contained in RDF triple form from which GKR's graphing model can improve over time (as new knowledge emerges and re-affirms the factual accuracy)

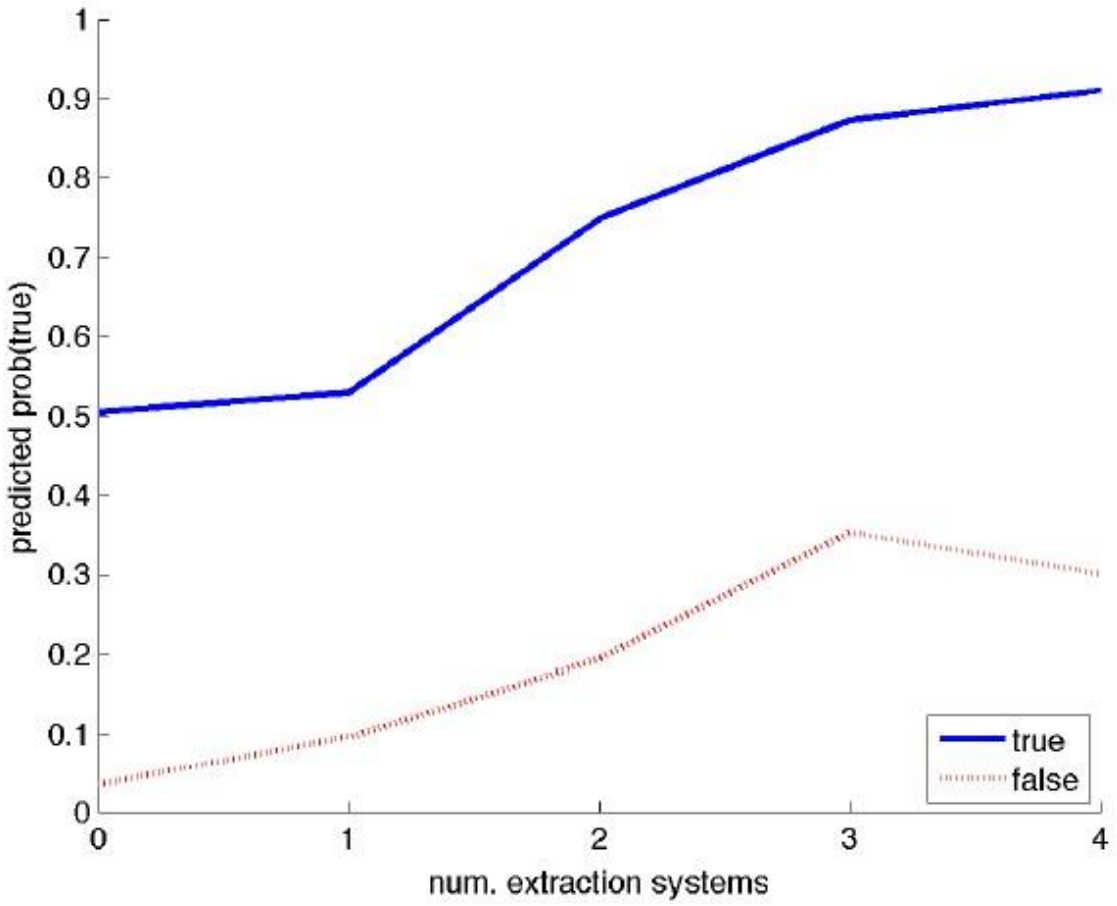
Probabilistic knowledge fusion

One of the key features of GKR is the use of probability of accuracy assigned to each knowledge element expressed as RDF triple in the model. The model utilises an approach called knowledge fusion whereby the underlying sources of data are combined with the key heuristic assumption that multiple instances of the same derived RDF triple increase the probability of the knowledge being accurate.

The follow tables show this model component as a function of multiple different underlying systems (data pools) as well as document instances.



Graph 1 – RDF triple predicted accuracy relative to the number of documents [1]



Graph 2 – RDF triple predicted accuracy relative to the number of underlying data systems [1]

The graphs above visually illustrate the key feature of the GKR model that is used to overcome manual ground truth assignment found in other models. Through the fusion of the above techniques with others (such as taking truth assignments from underlying data models), GKR is able to improve the overall accuracy assignment and achieve a new record of 271 million RDF triple knowledge elements with an accuracy greater than 90% confidence (out of a knowledge vault consisting of 1.6 billion RDF triples) [1].

How does it differ from traditional structured data models?

The below table performs a compare and contrast of the GKR model against traditional SQL database designs (used in the context of a Knowledge Vault) in order to better understand the nuances and differences of the two technologies and inform the correct use thereof:

Technology Aspect	Google Knowledge Review	Traditional SQL database KV's
Scalability	High – Cloud-based with existing capability to mine unstructured data from the web	Low – Usually local server-based with finite physical resources and a rigid data model
Data Model	Probabilistic Model	Deterministic (Pre-defined Relationships)
Knowledge Sizing	Up to 1.6B RDF triples of various accuracy with scope for more	Varies – typically determined as rows in a DB structure and can be in the millions
Data Sourcing	Fully automated – web-scraping of multiple underlying data sources across different data models	Manual – Requires manual input of underlying data
Data Conflict Management	ML and statistical-based mechanisms to handle conflicting data	Rigid conflict management – SQL based KV's cannot have certain data fields being different due to strict referential integrity
Data Type	Unstructured	Structured

Current limitations of the GKR approach to a Knowledge Vault?

GKR attempts to solve one of the key constraints in modern information science being scalability and accuracy in the absence of the human intervention. The approach is a highly scalable multi-data source statistical technique that seeks to determine knowledge through continuous improvement over time as more data becomes readily available.

However employing such a technique is not without potential new caveats (read as future research areas) to consider as machine learning techniques such as NLP continue to struggle with deep contextual understanding of information contained in written (human legible) form.

Some of the interesting areas for improvement or consideration include:

- Modelling mutual exclusion between facts – currently GKR treats each knowledge item or fact as an independent binary random variable that is either true or false. Through the correct

modelling of mutual exclusion (e.g. if the knowledge item 'Former President Obama was born in Hawaii' is correct, then we can correctly assume that the item 'Former President Obama was born in Nairobi' as being false) the ability to improve GKR's factual accuracy is possible

- Soft correlations – GKR has the ability to potentially incorporate further statistical heuristics such as 'the average number of children a couple have is usually between 0 and 5' in order to improve the accuracy of knowledge given such a prior knowledge assumptions
- Multiple levels of value abstraction can and do routinely exist in data that, on an automated ML approach, complicate the determination of factual accuracy. As an example, 'Former President Obama was born in Hawaii' and 'Former President Obama was born in the US' are both factually accurate but at different levels of abstraction.
- Temporal factual accuracies are prevalent whereby a certain knowledge item is deemed to be accurate only for a certain time period. An example of temporal accuracy is 'Google's CEO is Larry Page' was only factually correct from 2001 to 2011. Such occurrences complicate the GKR probabilistic approach as the ability to determine temporal accuracy is extremely difficult programmatically.

Conclusion

The Google Knowledge Review represents an interesting approach to the continued development and improvement of Knowledge Vaults that seek to store knowledge in order to inform people, processes, and systems ultimately and improve decision making.

GKR employs the latest machine learning, NLP as well as other automated techniques to attempt to derive a largely fully automated KV with minimal manual intervention of factual accuracy. The results shown and summarised in this research document show the technique to be successful and improve over existing KV's and their respective approaches to achieve an extremely large web-scale level KV.

Further work and refinement are however noted and provides for further interesting areas of research to further improve on GKR's current factual accuracy and limitations.

References

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*, KDD 2014

2.

G. Angeli and C. Manning. *Philosophers are mortal: Inferring the truth of unseen facts*. In CoNLL, 2013.

3.

H. Ji and R. Grishman. *Knowledge base population: successful approaches and challenges*. In Proc. ACL, 2011.