Google Knowledge Vault:

Introduction:

What is a Knowledge Graph and Knowledge Vault?

Knowledge graphs and *knowledge vaults* are a new and developing technology currently being developed by the major search engines.

Google's Knowledge Vault is characterized as "the largest store of knowledge in human history", Knowledge Vault is being assembled from content across the internet without human editorial involvement. Knowledge Vault autonomously gathers and merges information from across the web into a single base of facts about the world, and the people and objects in it. Its predecessor is Google's Knowledge Graph. With this new advancement to the knowledge graph, Google is attempting to automatically acquire facts from the mass of unstructured information still on the web.

Body:

What was the previous approach for building knowledge base?

Previous approaches for building knowledge bases primarily relied on direct contributions from human volunteers as well as integration of existing repositories of structured knowledge (e.g., Wikipedia info boxes). It was observed that Wikipedia growth has essentially plateaued, hence unsolicited contributions from human volunteers may yield a limited amount of knowledge going forward.

Google's Knowledge graph relied on crowdsourcing. That is to say, human volunteers and researchers contributed facts to the project. While this did allow the company to amass a large database of reliable facts, Google quickly realized that the system would always be bottlenecked so long as it was powered by the human mind.

Source of data?

Knowledge Vault fuses all below signals together

1) Data from Web
   a. Unstructured text
   b. Semi-structured DOM trees
   c. Structured Web tables
2) "Prior" data from Freebase

Concept of Knowledge Vault?

There are 3 main components

1) Extractors – these systems extract triples from a huge number of Web sources
2) Graph Based priors - these systems learn the probability of each possible triple, based on triples scored in an existing KB (knowledge base)
3) Knowledge fusion - This system computes the probability of a triple being true, based on agreement between different extractors and priors.

The real value of the process outlined is in the combination of different methods of extraction and validation to produce a far more robust and reliable database of entities

Use of NLP and Machine Learning Algorithms?

With the help of advanced machine learning algorithm's, the presented system is able to calculate the probability of correctness of current information and is able to translate and merge new information at an enormous scale.

Another important part is the "people" aspect of Knowledge Vault - This is specified in the paper, and lends a bit more fuel to the idea that Google wants to have some sort of author or expertise-based ranking, but wants to do it through entities and not explicit markup.

Probabilistic Knowledge base?

The Knowledge Vault is an attempt to crawl the web for facts and properties, and by comparing to sources of reliable structured data already available, attempt to rank the facts that have automatically been acquired to give it a confidence level. This is known as a *probabilistic knowledge base*, and it is the only way that Google believe a web-scale knowledge base can be practically built.

Does it use Google's page rank algorithm?

Google has always relied on backlinks to rank sites. The number of backlinks a site has should be an accurate gauge of its overall quality.

Content?

Google has reportedly assembled 1.6 billion "facts" and scored them according to confidence in their accuracy. Roughly 16 percent of the information in the database qualifies as "confident facts."

Google also indicated that there are numerous models at the company and elsewhere about building a knowledge base with machines or extracting meaning from text.

How is the data stored?

Google are choosing to store all the facts and knowledge hoovered up by the knowledge crawler as RDF triples. Machines will be able to understand the data in this database by looking at the web ontology that Google is going to use to store the data.

Does it impact privacy?

As the primary function of Google Knowledge Vault is to gather facts, it remains to be seen how much personal data it will be able to collect from services which Google own's.

Alternatives?

Some of the prominent alternatives include:

1. Bing's Satori (From Microsoft)
2. Freebase
3. DBPedia
4. Wikidata

Conclusion:

How will it be useful?

This could have multitude of uses

1) The Knowledge Vault can be utilized by search engines to get relevant data. This should improve retrieval time.
2) The information gathered can be useful both in Content and Collaborative based filtering. This in turn will help recommender systems.
3) With personalized search becoming more useful and Artificial Intelligence systems being developed, predicting algorithms based on facts will work faster and better compared to invalidated information
4) Small and large organizations alike will be able to query and harness this knowledge base for themselves. Being able to rank the probability of a 'fact' being true and to gain new insights and facts using machines will provide as yet unknown, but potentially huge, benefits for businesses, governments, and third sector organizations alike.

Integration with other Google apps?

According to Tom Austin, an analyst at tech firm Gartner, Google may have plans to integrate this technology into their email service, Gmail. The upgraded email would be akin to a vastly more sophisticated Siri or Amazon Echo—able to interpret spoken commands, but also able to sort and prioritize email. It's reasonable to conclude the company's other services like Docs, Maps and Chrome will receive sizable intelligence upgrades as well.

References:

https://storage.googleapis.com/pub-tools-public-publication-data/pdf/45634.pdf

https://research.google/research-areas/data-mining-and-modeling/

https://gofishdigital.com/good-bye-knowledge-graph-hello-google-knowledge-vault/

https://www.searchenginenews.com/sample/update/entry/understanding-googles-knowledge-vault