Task 2: Emerging Technologies Coursework

Charles Read | C1646151 CM3202

Introduction

'Getting the Most out of Wikidata: Semantic Technology Usage in Wikipedia's Knowledge Graph' is a research paper published by the Wikimedia Foundation. The paper discusses how semantic technologies have been used to develop, manage, access and expand the Wikidata knowledge graph. In this paper, tools and services such as Resource Description Framework (RDF), SPARQL are investigated and analysed in order to create conclusions based on their feasibility, efficiency and potential development into the future for both human and robotic requirements.

The authors have completed this research in order to find out if the technologies in place are enabling Wikidata to thrive and be the best knowledge graph it can be. This is an important topic as the semantic web is currently a very popular research field in computer science. With the cutting-edge research on AI and new technologies, having a web that contains meaning rather than just variables and strings will impact the way we interact with our life in the future as it will enable the technology around us to be much smarter. It will also defiantly influence the intelligence of our Internet of Things devices.

The environment that these semantic technologies operate in is not a casual one. This is partly due to the sheer amount of data available, "more than 400 million statements about more than 45 million entities". This proves that even with extreme conditions, the working methods and services could potentially be also used in more generally related computer science areas with large success and community popularity.

Semantic Technologies

"Unfortunately, the core of Wikidata is not well-adapted to the needs of data analysts and ontology engineers." [1] This is a problem for the researchers investigating on the knowledge graph, without the support of other technologies, it would be very hard to analyse. Wikidata is based on the methods of Wikipedia, which was designed for encyclopaedia text and individual strings, stored in a MySQL database. Luckily, Resource Description Framework (RDF) format is available which is much better at an analysis and data processors perspective.

RDF allows for the storage of data values as compound objects, for example, objects that are made up of two values, like a speed or a distance measurement that has a number as well as a unit (e.g. 100mph). Qualifiers, "auxiliary property-value pairs" are also used as *references*, which in turn are complex values characterised by many property-value pairs. For example, a qualifier could be a string paired with a town to briefly describe it and link it within the RDF format. The process of breaking these

down results in better efficiency for querying but substantially more data on the knowledge graph. Real-time linked data and RDF dumps are released by Wikimedia however the authors state RDF dumps are still considerably less popular than linked data requests.

The lack of a suitable querying language needed to be resolved by choosing a stable and efficient service that would work effectively even with millions of triples. SPARQL was chosen due to its "availability of well-supported free and open source tools for the main tasks" [1] and mature existence. This allows effective interaction with the Wikidata knowledge graph for either human or computer. The Wikidata Query Service (WDQS) (built upon BlaseGraph) offers, in the way of an extension, a easy to use web interface that made it very easy for developers to create their own queries with modern formatting options.

These tools offer Wikidata with documentation, formatting, ease of use and room for expansions that could be built in. Examples of these extensions include coordinates around a given point for 'find near me' type queries or "compute the distance between points on a globe" [1] queries. There are many more community developed expansions to the technologies used, a lot of which are open source.

Analysis and Findings

The authors of the research paper have found surprising statistics on the efficiency and the popularity of the Wikidata query service. "In 2018, there have been over 321 million requests within the twelve weeks from 1st January to 25th March" which is a massive amount of traffic being requested, performance was also analysed and the top 99% of requests stay below 40s on average, which shows the performance wasn't greatly affected during this traffic. This shows the Wikidata Query Service is very popular with the community however the authors stated the number of researchers on this specific topic are relatively small.

"The work has surpassed expectations, in terms of reliability and maintainability, as well as community adoption" [1]. The authors agree on the fact that the semantic technologies used have allowed the knowledge graph to grow and become even more stable and effective at storing knowledge within the web.

Conclusions

In conclusion, it would be considered a sensible to assume the semantic technologies being used to power Wikidata are appropriate and efficient as well as having the correct capabilities for future development. For the knowledge graph to grow into the future, more researchers would need to join the topic and work together to create new expansions, more efficient tools and to improve ease of use for developers.

References

1. Stanislav Malyshev, Markus Krötzsch, Larry González, Julius Gonsior, Adrian Bielefeldt. **Getting the Most out of Wikidata: Semantic Technology Usage in Wikipedia's Knowledge Graph.** In Proceedings of ISWC-18, 2018.