Text Mining and Analysis in Knowledge Base Construction

Introduction

The goal of this paper is to review existing text mining approaches to building knowledge bases on the web and to explore future methods of improvement in this field. In recent years, there has been an emerging number of large-scale knowledge base applications such as Wikipedia, Freebase and Google's Knowledge Graph. The construction of knowledge repositories is becoming more in demand as people are figuring out the best ways of accessing and utilizing information from the web.

Text Mining Approaches

There are several different approaches to constructing automatic knowledge bases. The main approach is to build knowledge bases based on Wikipedia infoboxes and other structured data sources. Freebase is one of the representations of this approach. Freebase is essentially a database system that stores knowledge and supports highly diverse data with high scalability. It has four major components to it: a scalable tuple store, an http/json-based API, a lightweight and collaborative typing system and a large diverse dataset. Freebase currently contains more than 125 million tuples, more than 4000 types and more than 7000 properties. Another approach is to create a graph-based representation of knowledge. Google's Knowledge Graph is one of the representatives of this approach. Back in 2019, the concept of linked data came out and the idea of linking datasets together from the web to create one consolidated global knowledge graph became more and more appealing and plausible. In 2012, Google proposed Knowledge Graph to use semantic knowledge in web search. It basically identifies and disambiguates entities in text to enrich search results with semantically structured summaries and to provide links to related entities in exploratory search. One of the applications of Knowledge Graph is to improve search results from web-based search engines because Knowledge Graph contains human knowledge about real-world entities.

Web-Scale Approach to Probabilistic Knowledge Fusion

There are some shortcomings in the traditional approaches of knowledge base construction. Mainly a lot of traditional approaches rely heavily on human input of knowledge in structured database systems. Therefore a new way of automatic knowledge base construction has been proposed and researched on, which is a web-scale probabilistic knowledge base called Knowledge Vault. In this approach, text is still

extracted from the whole web but on a greater scale. It also combines extracted information with prior knowledge derived from existing knowledge repositories and uses a probabilistic inference system to assist with fact checking. There are three components to Knowledge Vault: extractors, graph-based priors and knowledge fusion. In a test conducted by a group of researchers from Google, Knowledge Vault was able to extract 1.6B candidate triples, covering 4469 different types of relations and 1100 different types of entities. What's more, about 271 million of those facts have an estimated probability of being true above 90%. By comparison, Knowledge Vault outperformed all the other traditional approaches in this test. In order to verify the facts extracted, Knowledge Vault utilizes existing knowledge bases such as Freebase to link existing data with the new data to create graph-based priors. Lastly, Knowledge Vault combines the extractors and priors together using a fusion method.

Conclusion

With the ever-increasing data on the web, the need for better automatic knowledge base construction will become even higher. In the meantime, there has been a lot of studies in this field so that we can improve our information extraction from the web and use that information to assist with our decision making and knowledge gaining. To look into the future, there are many areas of improving the process of constructing knowledge bases such as modeling mutual exclusion between facts, modeling soft correlation between facts and representing values at multiple levels of abstraction.

Citations

X. L. Dong, E. Gabrilovich, G. Hertz, W. Horn, N. Lao, K. Murphy, T. Strohmann, S. Sun, W. Zhang. Knowledge Vault: A Web-Scale Approach to Probabilistic Knowledge Fusion. The 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '14, New York, NY, USA - August 24 - 27, 2014, pp. 601-610

K. Boolacker, C. Evans, P. Paritosh, T. Sturge, J. Taylor. Freebase: A Collaboratively Created Graph Database For Structuring Human Knowledge.

R. West, E. Gabrilovich, K. Murphy, S. Sun, R. Gupta, D. Lin. Knowledge Base Completion via Search-Based Question Answering. WWW, 2014.

X. Zou. A Survey on Application of Knowledge Graph. Journal of Physics: Conference Series, 2020.