Building Knowledge Graphs for Diagnostic Medicine — Cancer

Names and Uni: rh2962 Runyu Hao, lz2684 Lixin Zhang
Topic: Knowledge Graph, NLP, Cancer
Date: Feb 14, 2020

Abstract: To help doctors diagnose cancer-related diseases more efficiently, this project aims to build a knowledge graph for cancer diagnostics using text mining, which has many potential using areas. The raw data used to build the graph is got from PubMed, a medical literature abstract collection, and Wikipedia by web scraping. We use NLP (Natural Language Processing) to process the text data and add it to a graph database to build the knowledge graph.

1. Background :

Although Othology is a long-existing concept, the term, knowledge graph, was introduced by Google in 2012 to add information results for its search engine. Benefitted from the rapid development of NLP, a lot of research related to KG has been done since then in areas like knowledge extraction, error detection in knowledge graphs, completion of knowledge graphs and knowledge graph refinement.

Many KGs for specific areas have been built, in this project, we're going to build a KG for cancer diagnostics.

2. Introduction to the Project:

We are going to review related papers and others' work on building a knowledge graph. For our project, we are planning to use web scraping to retrieve the data from the web, use some NLP algorithms to do information extraction (IE) to extract entities, relations, and attributes, and then use a graph database to store the knowledge graph. The main expected challenge is IE, and we are going to learn and experiment with different methods to get the best results.

3. Introduction to the Dataset:

The data source of this project is PubMed and Wikipedia. PubMed is a medical literature abstract collection that contains numerous paper abstracts related to cancer. It is an unstructured data source that contains abundant semantic information related to cancer. Wikipedia is one of the biggest online encyclopedias and has many pages related to cancer. It's a semi-structured data source, which means it's easier to process compared to unstructured data. We obtain the data mentioned above using Web Scraping, and there are several popular packages in Python to do this job such as BeautifulSoup. The main difficulty is that the structure of web pages varies, and we need to figure out how to extract the needed information precisely.

4. Plan:

Milestone 1:

1. Learn the basic concepts about the knowledge graph and its applications.
2. Review others' work on building a knowledge graph.

3. Determine the steps to do this project: data retrieval, NLP processing, add to a graph database.
4. Learn the ontology relation in the cancer area.
5. Choose a proper cancer data source for the project.
6. Learn the basics of retrieving data from the web and start to do web scraping on PubMed and Wikipedia.

Milestone 2:

1. Learn and compare different information extraction(IE) methods like pattern-based methods, supervised learning methods, distant supervision, and Bootstrapping. For semi-structured data like Wikipedia, wrappers are also considered.
2. Start to do the named entity recognition(NER), Relation Extraction(RE) and Attribute Extraction(AE). This process is estimated to takes most of the time for the project.

Milestone 3:

1. Add the extracted entities and relations to a graph database to build the knowledge graph for cancer diagnostics.
2. We may consider developing some applications using the knowledge graph if time permits.

Reference:

1. Baker, S., Ali, I., Silins, I., Pyysalo, S., Guo, Y., Högberg, J., Stenius, U., & Korhonen, A. (2017). Cancer Hallmarks Analytics Tool (CHAT): a text mining approach to organize and evaluate scientific literature on cancer. *Bioinformatics (Oxford, England)*, *33*(24), 3973–3981. https://doi.org/10.1093/bioinformatics/btx454
2. Wang, R., Yan, Y., Wang, J., Jia, Y., Zhang, Y., Zhang, W., & Wang, X. (2018). AceKG: A Large-scale Knowledge Graph for Academic Data Mining. *CIKM '18*.
3. Ren, X., Wu, Z., He, W., Qu, M., Voss, C.R., Ji, H., Abdelzaher, T.F., & Han, J. (2016). CoType: Joint Extraction of Typed Entities and Relations with Knowledge Bases. *ArXiv, abs/1610.08763*.