Answer My Question

Sujitha Puthana

Dept. of Computer Science

University of Missouri Kansas City

Kansas City, USA

Spb4b@mail.umkc.edu

Jakkepalli Rama Charan Pavan

Dept. of Computer Science

University of Missouri Kansas City

Kansas City, USA

rjhg5@mail.umkc.edu

Yalamanchili Sowmya

Dept. of Computer Science

University of Missouri Kansas City

Kansas City, USA

syb7c@mail.umkc.edu

Nandanamudi Sreelakshmi

Dept. of Computer Science

University of Missouri Kansas City

Kansas City, USA

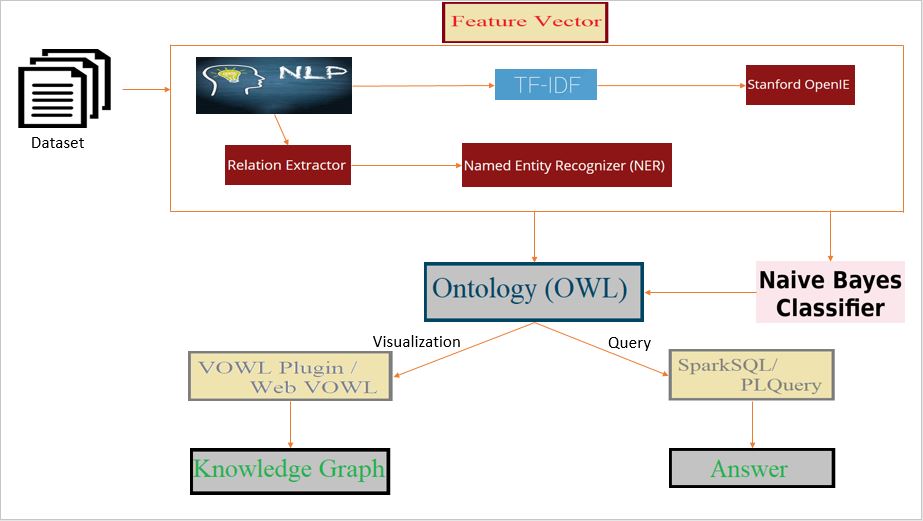
snhnc@mail.umkc.edu

***Abstract*— Data Science is a system to extract knowledge of data in various forms, either structured or unstructured from various domains, like Knowledge Discovery in Databases(KDD). Natural language processing is used for processing the text which is machine understandable and which will help for fast retrieval of data. Ontology plays an important role with respect to entity classification to answer questions. Visualization of the data by classification into classes, subclasses, data properties, object properties using the concept of ontology tool protégé. Protégé tool has its unique features for fetching the related information by using either spark SQL or DL query, which are the simplified query to fetch instances. This application helps in fetching the answer to questions by using NLP Process, word2vec, TF-IDF, N-gram. NLP, kmean, Classification of data, NLP algorithm is useful step for text processing and then we are extracting the relevant data. Visualization of the knowledge graph is also of great use. However, all the algorithm we are using in the project have its own significance. Comparing all these processes to find the best process with respect to time, accuracy, cost to select the best process. Query using the spark sql or DL sql can fetch the information from the entity classified. These queries are very fast to extract the information to answer the relevant questions.**

I.INTRODUCTION

In present days, the amount of data is increasing and this is leading to the difficulties in handling the data. So, we need the machine learning algorithms to handle these huge data. We are making use of artificial intelligence algorithm for machine learning to handle data and search the data.

As we are using these AI algorithms for handling data, this helps in getting through different algorithms available including TFIDF, NLP algorithms, word2vec algorithm, kmean algorithm, classification of data using all these processes and analyzing the accuracy. This tremendous process leads to precise answer of the question.



In the process of going through different AI algorithm to classify data and handle them. We could understand the importance of each algorithm with the specified uniqueness. By making using use of all these algorithms simplify the management of data.

Human are more prone to understand the visualized data than the text data. Visualization includes the presentation of the data in the form of knowledge graph. The text data is classified into classes, subclasses, properties are extracted. We generate an owl class and give it as input to protégé and visualize it using either plugin VOWL or webvowl.

WebVowl: <http://visualdataweb.de/webvowl/>

Spark sql is like normal sql commands that can be used to fetch information in the form of schema. In our application, we are using the spark sql commands to answer some questions. Protégé tool has its own query language DL query which is more simplified version of querying. DL query fetched the instances of the classes which can answer few questions.

II.RELATED WORK

In the present days, where the data is huge leading to data management issues. There are many algorithms already existing but the main problem in the existing algorithms are completeness and correctness. To solve this problem, we need to consider all these algorithm and judge wisely which all are the algorithms that we can use to easily maintain data and give us the high accuracy. But a single algorithms or approach cannot solve this problem. Hence, we should integrate multiple algorithm for high accuracy in designing the search engine.

Searching the huge amount of data is very difficult. Knowledge Graph represents the graphical representation of the entities and interrelated relationship. There is different knowledge graph available in the market but googles knowledge graph is the popular search engine algorithm. Best knowledge graph can be designed solving the completeness and correctness issue by integrating different approaches of knowledge graph available in the market.

Data sources that are available to us are limited. We can increase the accuracy to provide the best answer to any question is by considering all the data sources that are available on the web. The solution for this approach is the knowledge vault that was made available to us by google that takes the data in RDD triplets i.e., subject, object, predicate. After collecting the data and finding the entities our next problem would be organizing the data. We Deep Dive approach helps in resolving the problem of extraction of data and its integration to fetch accurate prediction making the training process easy.

After the data is represented in RDF triplets, the semantic relationship can be organized using the FehSen to merge the related information leading to more simplified data. It is known fact that structured data is easy to handle than unstructured data. Fonduer is the approach in focusing the construction of the structured data from the plain text. By using all these approach helps in improving the handling the data and solve the “completeness and correctness” problem.

Optimization of the questions is important to get high accuracy. Latent dirichet allocation is used to extract the topics. Applying the LDA algorithm on the question is used to cluster the question topic, measuring the similarity based on semantic between ssmultiple questions. OpenIE algorithm is also applied on the questions to generate the RDF triplets to understand the question.

Visualization of the data plays a keys role in understanding and process huge data. Visualization is done by extracting the key entities and relationship between them. Object properties defines the property relationship between two instances. Data properties defines the relationship between two entities. Modern algorithm “Concept Net” which is an improved version to visualize the data using the labels and edges.

In our world where there exists data in multiple languages. To achieve the high accuracy information, we need to consider data from all the available sources in all the languages. DBpedia algorithm is the best approach for this process. After completing the data extraction, data retrieval our main task is to improve the processing time and accuracy to fetch the most relevant answer to the question. One good approach is the query to fetch the relevant answer. Spark query and DL query are the highly used fast processes query languages. Thus, we are processing question and data through all the available algorithm to fetch the answer very fast.

III.PROPOSED WORK

IV.IMPLEMENTATION AND EVALUATION

V.DISCUSSION AND LIMITATION

VI.CONCLUSION

REFERENCE

1. <https://blog.algorithmia.com/introduction-natural-language-processing-nlp/><https://en.wikipedia.org/wiki/Question_answering>
2. <https://nlp.stanford.edu/>
3. <http://visualdataweb.de/webvowl/>
4. <https://rajpurkar.github.io/SQuAD-explorer/>
5. <https://cogcomp.cs.illinois.edu/page/resource_view/89>
6. <https://protegewiki.stanford.edu/wiki/Main_Page>