Northwestern MSDS-459 Knowledge Engineering

Assignment 4: Application Planning and Development

Andrew Stevens

June 5, 2022

(1) Abstract

Tesla is often interpreted as more of a technology company than an automobile manufacturer. When considering Porter's 5 forces (1998), Tesla is a competitor and disruptor in the automotive industry, and for the purposes of this study this view will be the focus. Data has been collected through guided web scraping to catalog the competitors and suppliers and the entities within, then connecting the relationships between. A simple and fairly limited web app has been employed to search and return results relevant to keywords inputted by a user, hopefully providing knowledge that the user seeks.

(2) Introduction

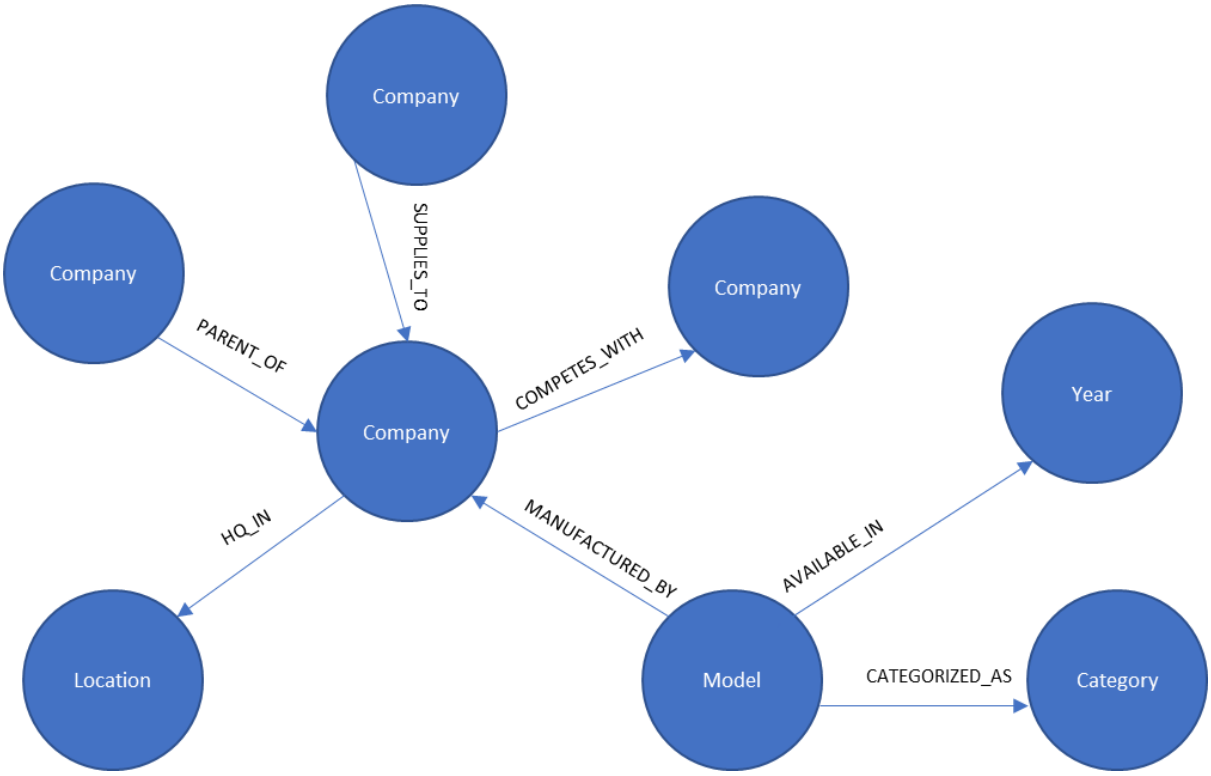
though there may be arguments surrounding its nature, implementation, motives and bias/ impacts (), the most complete, useful and used knowledge graph is Google's. It can answer most user queries with relevant references and correct context, providing it corrects to the correct entity. The foundation of the knowledge graph built for this study is the entities surrounding companies in the auto industry (see Methods section) and required the most thorough and intentional efforts

(3) Literature review

In determining which entities to focus on collecting data around, porter (1998) served as the impetus. Chakrabarti provides a thorough layout of how one may approach the retrieval of data required for a graph surrounding a company (1999)

(4) Methods

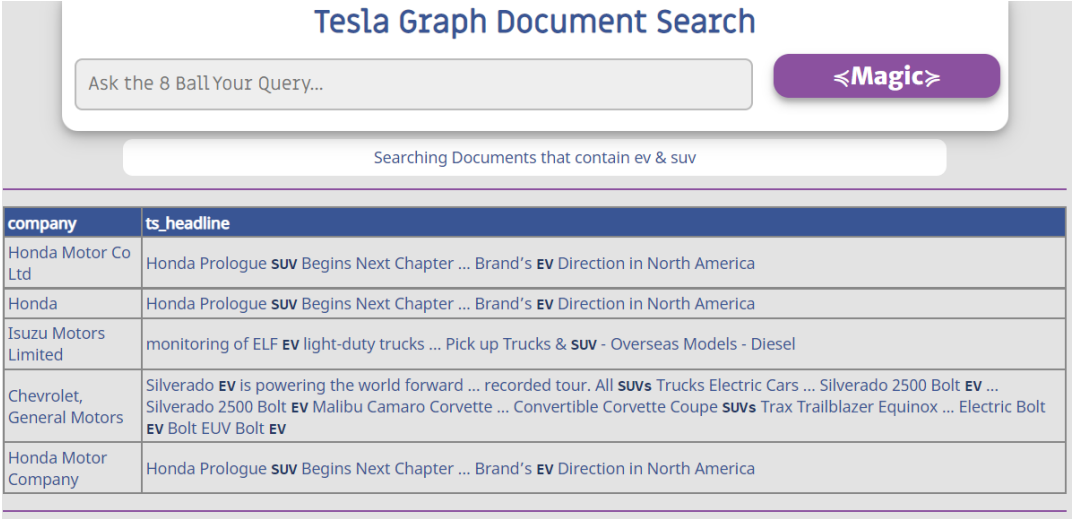
The final schema of the graph is shown in Figure 1.



The documents of websites were connected to companies through the following query

```
create or replace view public.sites2companies AS
select compet.*, company.*
from public.competitors compet
inner join "tesla5forces"."Company" company
on position(split_part(compet.parent,'.', 2) in
lower(replace(company.properties::varchar,' ','')))<>0
```

Allowing the “home” search page to fetch documents related to companies matching the input search terms. Huge thank you to Kelly Carlson for that leg up.



company	ts_headline
Honda Motor Co Ltd	Honda Prologue **SUV** Begins Next Chapter ... Brand's **EV** Direction in North America
Honda	Honda Prologue **SUV** Begins Next Chapter ... Brand's **EV** Direction in North America
Isuzu Motors Limited	monitoring of ELF **EV** light-duty trucks ... Pick up Trucks & **SUV** - Overseas Models - Diesel
Chevrolet, General Motors	Silverado **EV** is powering the world forward ... recorded tour. All **SUVs** Trucks Electric Cars ... Silverado 2500 Bolt **EV** ... Silverado 2500 Bolt **EV** Malibu Camaro Corvette ... Convertible Corvette Coupe **SUVs** Trax Trailblazer Equinox ... Electric Bolt **EV** Bolt EUV Bolt **EV**
Honda Motor Company	Honda Prologue **SUV** Begins Next Chapter ... Brand's **EV** Direction in North America

Relationships built through more structured generation of nodes, not extracted from unstructured text, were built into documents.

```
create or replace view public.v_textTriples AS
select n1.labl || ' ' || n1.val || ' ' || rel || ' ' || n2.labl || ' ' || n2.val
triples
from public.v_nodes n1
left join public.v_rels r
    on r.start_id::varchar = n1.id::varchar
left join public.v_nodes n2
    on r.end_id::varchar = n2.id::varchar
```

for recognition with PostgreSQL's built-in search function

```
    SELECT triples, ts_rank_cd(to_tsvector(triples), query) AS rank
    FROM public.v_textTriples, to_tsquery('{input}') query
    WHERE to_tsvector(triples) @@ query
    ORDER BY rank DESC
    LIMIT 5;
```

The triples were extended into quintuples in all directions (because directionality duplicates queries necessary to recognize when connecting) – unioning 4 queries together into a large tuples table

(see pg_steup.sql in git repo). The front end is hosted by flask and queries are sent through psycopg2 to the database to run search.

(5) Results

A major drawback of using Apache AGE as a knowledge graph is that under the hood it is still (just?) SQL. In attempting to engineer the ability to retrieve more complex information, I created a view for “quintuples”. If a Triple is defined as a node connected to another through a relationship:

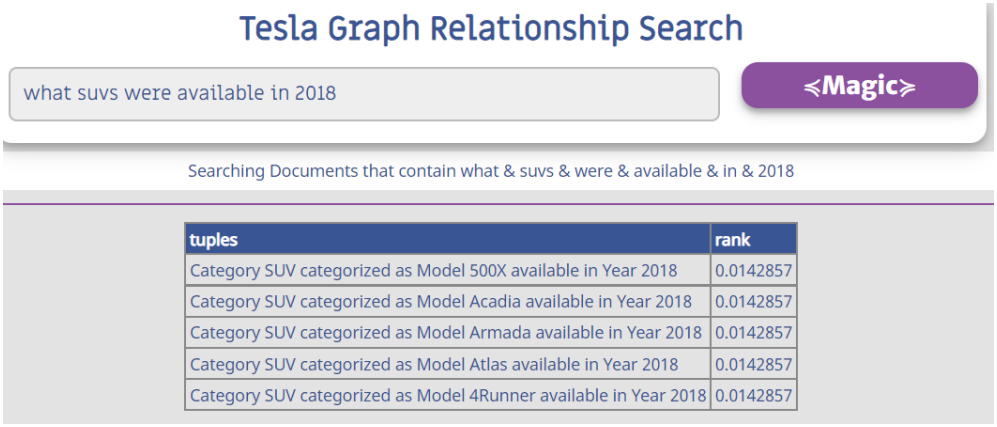
```
MATCH (n)-[r]->(n)
```

Then a quintuple would add another node with connecting relationship.

```
MATCH (n)-[r]->(n)-[r]->(n)
```

While executing a query to retrieve a triple takes approximately 30sec, retrieving a quintuple required 21x as long (10.5min). Graph database software written for the purpose of graphs and their complex queries (walks & hops) are much more performant. Running deep community or subgraph algorithms seem like execution would be impractical (though indexing keys and foreign keys would likely improve performance).

Aside from some performance issues, some basic queries are able to return interesting results. (though the language in the tuples table is not entirely clear when single direction relationships are flipped both ways [categorized in would need to be mirrored to `category includes`])



(6) Conclusions

I'd like to do a lot more work to make this graph more useful, but the amount of work required and the potential returns is why they are so valuable. More extensive NLP for knowledge extraction from the scraped documents would be helpful in building out knowledge in more areas of the domain.

(7) References

1. Porter, Michael E. Competitive Strategy: Techniques for Analyzing Industries and Competitors: With a New Introduction. 1st Free Press ed, Free Press, 1998.

2. "The Google Knowledge Graph: Information Gatekeeper or a Force to Be Reckoned With?" Strategic Direction, vol. 30, no. 4, Mar. 2014, pp. 15–17. DOI.org (Crossref), https://doi.org/10.1108/SD-04-2014-0049.

3. Chakrabarti, Soumen, Martin van den Berg, and Byron Dom. 1999, May 17. Focused crawling: a new approach to topic-specific Web resource discovery. Computer Networks: The International Journal of Computer and Telecommunications Networking, 31(11-16): 1623–1640.

4. CIGraphs. (n.d.). CIGRAPHS/Carlson-Adafruit. GitHub. Retrieved May 23, 2022, from https://github.com/CIGraphs/carlson-adafruit