CS 846 Software Engineering for Big Data and AI

Knowledge Graphs for Social Good: An Entity-Centric Search Engine for the Human Trafficking Domain

Jun Lim

20870249

The paper addresses the issue of increasing trends in web advertising related to human trafficking (HT) activity. As these illicit activities on the web are on the rise, various agencies have been pressing and turning to technology to assist in combating this issue. The authors aim to address this issue by introducing a semantic search engine, built upon a query-centric knowledge graph, to assist and aid analysts and experts in the HT field.

This proposed search engine is able to answer entity-centric questions over web corpora associated with HT activities. This involves structuring a large corpus of web advertisements related to HT into an indexed knowledge graph, which subsequently allows investigators to ask specific questions relevant to human trafficking.

However, given web domains's heterogeneity, information is often obfuscated and diversified to evade detection; it is difficult to extract key data (such as phone numbers, age, email address, etc.) and infer from it due to the variety of styles of writing, symbols, and obfuscation techniques used. As such, traditional methods of keyword-based matching do not work effectively in such a use case.

The authors then proposed and built the search engine, which contains the following two components:

1. **Offline - construct a knowledge graph**: which takes a crawled web corpus as input and structures it into a semi-structured knowledge graph that is stored and indexed in a NoSQL database.
2. **Online - implement real-time entity-centric information retrieval**: which allows users to express their original needs in an intuitive query language, which is then processed using semantic execution plans to retrieve / extract relevant information.

The authors then evaluated the constructed knowledge graph on real-world data collected over 90,000 webpages related to HT activities, and the following results were obtained:

- **Query exeuction**: a set of queries were run and their queried results were compared with the prototype, and the mean average precision metric was found to be promising.
- **Scalability**: the scaling capabilities of the entity-centric search were evaluated, and it was found that the scalability performance was achieved due to the adoption of scaling techniques with Apache Spark and Elasticsearch.

The authors have also integrated their proposed engine into a GUI that is widely used by enforcement agencies in the US, while several extensions are currently in the process of exploration and development. The paper was then concluded with future goals to possibly include aspects like indexing and searching for non-HT domains, processing queries in natural language forms, and developing a more natural user interface.

Paper Commentary

It is great that the authors made a novel and innovative contribution on utilizing technology to combat illegal activities, bringing good progress to society as a whole. By leveraging a query-centric knowledge graph, they're not only providing a tool for investigators but also showcasing the potential of technology for social good. Overall, the paper was well-written in a concise manner with a comprehensive evaluation, and it has high relevance to the current state of the world (addressing the global challenge of human trafficking).

However, there are also some downsides to the paper; it contained too many irrelevant details and lengthy discussions. There could also be cases of false positives that have not been discussed thoroughly.