Design and Implementation of Knowledge Graph-Based Intelligent Search

(Technical Paper)

The Ethical and Societal Implications of Data Practices: Privacy, Bias, and Governance in a Data-Driven World

(STS Paper)

A Thesis Prospectus Submitted to the

Faculty of the School of Engineering and Applied Science
University of Virginia • Charlottesville, Virginia

In Partial Fulfillment of the Requirements of the Degree
Bachelor of Science, School of Engineering

Trishal Muthan

Fall 2024

On my honor as a University Student, I have neither given nor received unauthorized aid on this assignment as defined by the Honor Guidelines for Thesis-Related Assignments

Signature ____ Date __________
Trishal Muthan

STS Advisor: Richard D. Jacques, Ph.D., Department of Engineering & Society

Introduction

There is perhaps no more valuable commodity in the modern digital era than data. During all of human history until 2003, 5 exabytes of data, or 5 billion gigabytes, had been produced. Eleven years later, that was the same amount of data produced every 10 seconds (Zwitter, 2014), with that amount having grown exponentially since then. With so much time being spent by people on devices and the fundamental integration of technology into humanity, it becomes obvious how the existing amount of data has reached such an exorbitant number. However, the amount alone is not what makes data so powerful. What makes data so powerful is that it yields the ability to make informed decisions. Taken at face value, this may not seem so extraordinary but essentially every company, government, institution, and even person relies on data for this very reason. Companies can use data to recommend better movies or personalize the advertisements users see for different products, governments can use data to keep track of national demographics or gather information on the needs of the electorate, and even regular people may use data to keep track of their sleep or get better at a video game.

Data is essentially everywhere and used for all sorts of things. Thus, it becomes essential to consider how said data is obtained, how it is stored, and how it is used. Various techniques can be used to store data including relational databases, hierarchical databases, and document databases. An interesting database structure that has risen in popularity recently is the graph database or knowledge graph, which stores information in a system that resembles human memory. My technical topic focuses on the design, implementation, and use cases of knowledge graphs specifically, while my STS topic analyzes the social issues of data privacy and ethical data use that often arise due to this abundance of data.

Technical Topic

In 2012, Google introduced the term knowledge graph, a feature designed to enhance the data and relevancy of that data returned by their renowned search engine (Singhal, 2012). The term has since grown in popularity to generally describe the storage of information in network-based structures rather than table-based structures.

Knowledge graphs are unique in that they store not only individual pieces of data but also the relationships between different pieces of data. Essentially, knowledge graphs consist of entities and relationships where each relationship represents the connection between any two particular entities (Hogan et al., 2021). For example, you might have in your knowledge graph the entities “Joe Biden” and “Democratic Party” along with a relationship “member of” that connects Joe Biden to the Democratic Party, representing the information that Joe Biden is a member of the Democratic Party. This kind of information about the relationships between different data is incredibly powerful; it can provide useful insights into the data you are working with and offers enhanced capabilities in technology ranging from recommendation systems to large language models.

The technical component of this project was completed during an internship from May to August of 2024 and was overseen by company employees on a day-to-day level. The goal of the project was to build a full-fledged application that at its core took advantage of a custom knowledge graph to deliver data to end users. Specifically, the application uses knowledge graphs to offer intelligent search, which allows users to search for particular people, places, and organizations and view their related entities along with recent news articles they have appeared in.

Implementing a knowledge graph is not a particularly simple task. The first step involves obtaining the data. It can be difficult to easily obtain or decipher information about the relationships between data and insert them into the knowledge graph. To accomplish this goal of building a custom knowledge graph, entity extraction can be used on text data to selectively retrieve important information. With nearly an endless amount of text data on the internet, entity extraction to pull out prominent entities and the contexts in which they appear can be employed to obtain fruitful relationships. By utilizing web scraping to get news articles off the internet and this entity extraction technique, a significant amount of data can be attained.

Additionally, the setup and infrastructure of the knowledge graph itself require the use of graph databases, typically via available open-source software. Cloud services such as Amazon Web Services offer services for the handling of all the data along with graph databases that can be used to create the knowledge graph. By implementing a pipeline by which text data can be extracted, transformed, and loaded into a particular database, a functioning knowledge graph can be produced.

On top of the knowledge graph itself, a web application is necessary that enables users to actually interact with the data. The web application can be separated into a back-end and front-end component. The back-end will handle the connection to the database, acting as an application programming interface (API) that allows requests and information to be queried from the knowledge graph. The front-end will handle the actual website display itself, incorporating user interface and experience (UI/UX) principles to display the information obtained from the back-end API in a way that is accessible and easy to understand.

This type of application can act as an incredibly powerful search tool and yield extremely insightful information about the data within the knowledge graph. For the aforementioned example of Joe Biden, such a tool may not be as helpful given his position in the public eye. However, take individuals or organizations who are not as well known or perhaps those in foreign countries. Such a tool would be capable of providing information about these entities and who/what they are related to in some way, information that can be useful to journalists, governments, and even regular people. By taking advantage of the interconnectedness of real-life data, something knowledge graphs are uniquely able to model, it is possible to procure a new and more effective viewpoint into the world.

STS Topic

In today's increasingly data-driven world, the ethical, social, and political implications of data collection, storage, and usage are of vital importance. Data shapes decision-making in essentially every aspect of human life, from healthcare to law enforcement. However, alongside the advancements being made in this area, serious concerns arise regarding privacy, inequality, agency, and ethical responsibility. Data is not merely a neutral byproduct of technological innovation; it has become a powerful resource that can shape social realities, reinforce systemic inequalities, and even influence individual freedoms. Personal data gathered via social media, our devices, government collection, or any of the innumerable other data sources can be used in ways to predict and shape behavior in ways that challenge pre-existing notions of privacy and autonomy.

For instance, take the issue of algorithmic fairness. The use of historical data in algorithms and machine learning models can often reproduce existing societal biases resulting in discriminatory practices in fields like hiring, lending, and healthcare. An analysis was conducted in 2016 of the Correctional Offender Management Profiling for Alternative Sanctions

(COMPAS) tool, an algorithm used to predict the risk of arrested individuals to commit another crime in the future, a concept known as recidivism (Angwin et al., 2016). Such risk prediction tools have been widely used throughout the country to determine sentencing times and parole limits, among other things. The analysis revealed that black individuals were given higher scores at a disproportionate rate compared to white individuals for similar types of crimes. These algorithms were having direct impacts on the amount of time people were being sentenced to prison and, therefore, led to black individuals getting longer sentences than others. This phenomenon is not an anomaly. These algorithms, which have previously been touted for their supposed objectivity and fairness, are subject to the very same biases that affect real people, an example of an unintended consequence of technology. It is critical that algorithms such as these have sufficient oversight and that the data they are trained on do not serve to reinforce the very biases they were intended to stay away from.

Consent and autonomy in data practices are also often compromised when personal data is collected and used without informed or genuine consent. A prominent example is the Facebook-Cambridge Analytica scandal, where data from millions of Facebook users was harvested and shared without their explicit permission (Confessore, 2018). Cambridge Analytica used this data to build psychological profiles of users, ultimately leveraging it for targeted political advertising in elections, including the 2016 U.S. presidential campaign. This case highlights how vague or hidden terms of service can strip users of autonomy over their personal information, as most users were unaware of how their data was being used. The idea of critical data studies (CDS) critiques such practices by arguing that data ownership and consent should not be buried in legal fine print but should instead empower individuals to control how their data is utilized. Through this lens, CDS advocates transparent data policies that prioritize individual rights and informed consent, preventing exploitation and misuse of personal data for profit or political gain.

In an effort to explore and provide a more focused analysis of the aforementioned implications, my STS paper will attempt to answer a few key questions. Firstly, how do practices regarding the collection, storage, and usage of data challenge or reinforce societal power structures? Additionally, in what ways do algorithms and data-driven systems perpetuate societal biases, and how can these effects be monitored and mitigated? Lastly, what ethical frameworks and regulatory approaches can be used to ensure data is collected and used responsibly? To answer these questions, an analysis of existing case studies, literature, policy documents, and ethical guidelines related to data governance will be conducted.

Conclusion

The overarching goal of my thesis is to understand the power of data. For the technical component, I implement an application surrounding a knowledge graph, a novel method of structuring interconnected data. For the STS component, I aim to analyze some of the ethical considerations surrounding data and put forth some frameworks and approaches that can be used to establish the proper collection and use of the commodity going forward. Through this dual approach, I hope to bridge the technical and societal perspectives on data, highlighting both the immense potential of data-driven technologies and the responsibilities they entail. By examining how data practices influence societal power structures, individual privacy, and the fairness of algorithmic decisions, this work seeks to contribute to a more ethically aware approach to data innovation. Ultimately, it is important to not only advance technology but also prioritize human rights, transparency, and accountability in an increasingly data-dependent world.

References

Angwin, J., Larson, J., Kirchner, L., & Mattu, S. (2016, May 23). *Machine bias*. ProPublica. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Confessore, N. (2018, April 4). *Cambridge Analytica and Facebook: The scandal and the fallout so far*. The New York Times. https://www.nytimes.com/2018/04/04/us/politics/cambridge-analytica-scandal-fallout.html

Floridi, L., & Taddeo, M. (2016). What is data ethics? *Phil. Trans. R. Soc. A*. https://doi.org/10.1098/rsta.2016.0360

Hand, D. J. (2018). Aspects of Data Ethics in a Changing World: Where Are We Now? *Big Data*, *6*(3), 171–235. https://doi.org/10.1089/big.2018.0083

Hogan, A., Blomqvist, E., Cochez, M., D’amato, C., Melo, G. D., Gutierrez, C., Kirrane, S., Gayo, J. E. L., Navigli, R., Neumaier, S., Ngomo, A.-C. N., Polleres, A., Rashid, S. M., Rula, A., Schmelzeisen, L., Sequeda, J., Staab, S., & Zimmermann, A. (2021, July 2). *Knowledge graphs*. ACM Computing Surveys. https://dl.acm.org/doi/abs/10.1145/3447772

Ji, S., Pan, S., Cambria, E., Marttinen, P., & Yu, P. (2022, February). A survey on knowledge graphs: Representation, acquisition, and applications. https://ieeexplore.ieee.org/abstract/document/9416312/

Sagiroglu, S., & Sinanc, D. (2013). Big data: A review. *2013 International Conference on Collaboration Technologies and Systems (CTS)*, 42–47. https://doi.org/10.1109/CTS.2013.6567202

Singhal, A. (2012, May 16). *Introducing the knowledge graph: Things, not strings*. Google. https://blog.google/products/search/introducing-knowledge-graph-things-not/

Zwitter, A. (2014). Big Data ethics. *Big Data & Society*, *1*(2). https://doi.org/10.1177/2053951714559253