Intelligent Infrastructure Facilitating Sequence Recommendation for Cybersecurity Education Systems

Eric L. Brown
Department of Computer Science
Tennessee Tech University

Douglas A. Talbert
Department of Computer Science
Tennessee Tech University

Abstract

The ability to incorporate original and adapted data into query-based storage structures to provide dynamic and timely service to sequence recommendation systems is a continuous goal of learning management systems. This can be a challenging goal when data integrity and student privacy are paramount. We are developing a hybrid machine learning-assisted system (CyberTaliesin) for cybersecurity educational support. In this poster, we present the early building blocks of the system involving the use of federated knowledge graphs as a trusted knowledge source capable of learning from "less restricted" models such as large language models. How can integrating these tools yield a flexible system that improves sequence recommendations, facilitates concepts such as adaptive and personalized learning, and achieves improved competency-based educational outcomes?

Introduction

Like many domain areas, the cybersecurity knowledge domain is a collection of siloed specialized information that official organizations with recognized credentials in the space have highly curated. In other words, the datasets from these groups are considered trustworthy. One example of such a data set would be the Workforce Framework for Cybersecurity (aka NICE Framework) (Petersen et al. 2020). The National Initiative for Cybersecurity Careers and Studies dataset classifies the body of knowledge into categories/specialty areas, work roles, tasks, knowledge, skills, and abilities. The goal of the NICE Framework is to provide a common vocabulary to describe cybersecurity work roles in a way that can be commonly understood across other sectors where cyber is required (and that would be all of them). Other groups have published related datasets to describe cybersecurity educational outcomes, including the National Centers of Academic Excellence in Cybersecurity via their Knowledge Units documents in cyber defense education (CD) and cyber operations education (CO) (NCAE-C 2023). Additionally, ISC2 publishes a proprietary Common Body of Knowledge (ISC2 2023) in support of its CISSP certification.

These trusted datasets are the foundation for training programs. While widely known, the materials are siloed with no convenient mechanisms to correlate them individually to create personalized educational outcomes. Consider a high school student with a limited STEM background who is finding an interest in cybersecurity. What tools are available to guide that student through a successful path to a certification or post-secondary education? How could we automatically and effectively combine the trusted sources to inform a college freshman what classes to take in what order, based upon their individual experience?

Given the recent renaissance of chatbots and large language models (LLMs), machine learning and AI have become interesting for large dataset processing and the ability to rapidly incorporate unstructured data into semantic-preserving storage and retrieval systems. However, LLMs are naturally distrusted outside the confines of extreme training. How could we use known, stable machine learning structures to maintain data integrity and improve outcomes while leveraging LLMs' natural "dreaming" abilities? Can a feedback loop system be created to enforce data integrity standards?

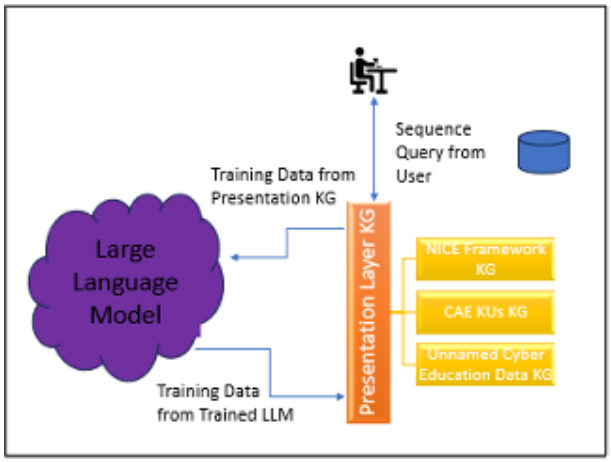

Figure 1: High-level conceptual view of the proposed system.

Course Sequencing Recommendations in Cybersecurity Education

As established, the data to support course and topic sequencing is available. The challenge is providing solutions based on that data and the individual's current knowledge status. How can current machine learning techniques be combined to support this process?

This poster outlines an experiment creating a collection of highly curated knowledge graphs (KGs) extracted from known, trusted datasets. A federated learning layer will use these knowledge graphs to provide a personalized course sequencing plan based on the student's prior educational background. The federated learning layer will also be informed by training data from queries to LLM instances.

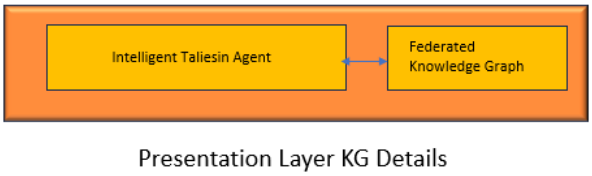

Figure 2: Breakout of the Presentation Layer KG.

Internal System Overview and Use Case

As pictured in Figure 1, the system consists of immutable KGs created from trusted sources and evaluated by subject matter experts. Once the curation process is complete, a date stamp and hash are made for each KG. The presentation layer, via the intelligent agent, as outlined in Figure 2, is responsible for checking the date stamps of the immutable KGs against the baseline date stamp of the federated knowledge graph (FKG). If these dates are out of sync, the agent will initiate a rebuild of the FKG based on the current state of the immutable KGs. The rebuilt FKG will be the basis for future queries.

The agent will use data from the FKG to send training data to an LLM to improve response reliability for future queries. Once a confidence level has been achieved, the agent will interrogate the LLM to create training data to be sent back to the FKG, further improving its response abilities. This continuous feedback loop repeats until confidence levels are achieved.

A user will provide a sequence query to the presentation layer of the system. The intelligent Taliesin agent will accept a sequence query from a client. On receipt and validation of the query, the agent will verify the current status of the federated knowledge graph (FKG), forcing a rebuild of the FKG if its date stamps are out of sync with the KG date stamps. Once cleared, the agent will query the FKG and respond to the user. The user can request a more speculative response, which would force the agent to conduct an additional query of the LLM to enhance the response to the user.

Federated Knowledge Graphs as a Basis for Federated Learning Structures

(Muniasamy and Alasiry 2020) explore the impact of deep learning methods for eLearning applications. These authors also identify Recurrent Neural Networks (RNN) as a deep learning model suited to sequence prediction, highlighting the data relationship connectivity and the ability to model changes in data over time. (Chen et al. 2022) proposes a structured federated learning (SFL) approach using a Graph Convolutional Network (GCN) to create a relation graph among the knowledge graphs. (Munir and Ferretti 2023) proposes a custom federated querying scheme by crafting SPARQL queries to Neo4j instances. The resulting knowledge graph "shards" could be consolidated into a semantic-preserving response.

The experiment's first phase will evaluate the construction of a federated learning layer leveraging current knowledge presented by the NIST Framework, CAE KUs, and related works. The presentation layer FKG is created from the results of the structured queries to the trusted KGs. We will also evaluate the ability of curated KGs at the presentation layer to contribute an archive graph, a collection of learned elements from prior transactions.

LLMs as a Training Tool in Federated Learning Structures

The experiment's second phase will evaluate the interconnection between the Phase 1 presentation layer FKG and new knowledge gained from queries against an LLM. This approach will allow us to use LLMs' "dreaming" abilities to evaluate missing relationships/links. We will also assess the effectiveness of LLM queries in quickly introducing new material.

Creating a monitored feedback loop may contribute to the improved training of the LLM and the KG at the presentation layer. This concept has been determined merited by (Lee et al. 2023). While that work focused on improving LLMs to improve QA systems, we will propose using structured data from the trusted KGs to assist in further training the LLM, thus reducing the effects of hallucination.

Once an acceptable confidence level is established, the LLM can contribute to the presentation layer FKG, providing additional training data. The goal is to achieve a completely trusted feedback loop between both layers (presentation and LLM) that can provide improved responses while quickly adapting to new information in cybersecurity.

Conclusion

The proposed work seeks to understand better and utilize the inherent relational abilities of knowledge graphs combined with the "dreaming" abilities of LLMs to create a better sequence recommendation model for cybersecurity education students. While this model could work for any domain space, we are specifically looking at cybersecurity education as an example of a dynamic changing field based in solid fundamentals with continually changing applications. The proposed hybrid machine learning approach will address this challenge.

References

Chen, F.; Long, G.; Wu, Z.; Zhou, T.; and Jiang, J. 2022. Personalized federated learning with graph. *arXiv preprint arXiv:2203.00829*.

ISC2. 2023. The isc2 cbk — common body of knowledge.

Lee, D.; Whang, T.; Lee, C.; and Lim, H. 2023. Towards reliable and fluent large language models: Incorporating feedback learning loops in qa systems. *arXiv preprint arXiv:2309.06384*.

Muniasamy, A., and Alasiry, A. 2020. Deep learning: The impact on future elearning. *International Journal of Emerging Technologies in Learning (Online)* 15(1):188.

Munir, S., and Ferretti, S. 2023. Towards federated decentralized querying on knowledge graphs. In *Computational Science and Computational Intelligence*.

NCAE-C. 2023. Cae documents library – dod cyber exchange.

Petersen, R.; Santos, D.; Smith, M. C.; Wetzel, K. A.; and Witte, G. 2020. Workforce framework for cybersecurity (nice framework). *National Institute of Standards and Technology Special Publications Series 800*.