**Programma di dettaglio della tesi**

**(da presentare al completamento del 20% delle attività di tesi)**

(*compilato dallo studente entro due mesi dalla comunicazione dell’assegnazione della tesi da parte della Commissione Tesi a seguito dell’incontro con il docente relatore e del relativo avvio delle attività*)

Da inviare utilizzando il seguente form:

<https://forms.office.com/Pages/ResponsePage.aspx?id=22cHw9o91E2KTQl9IsuZ03nJw8jHdCpNpV5x3k_J3dRUOFBDOFRXVDNOWEVZNUIxWVNLWFI0MjJQVS4u>

**Nome Cognome studente**

Luigi Russo

**Docente relatore**

Giuseppe D’aniello, Sabrina Senatore

**Insegnamento di riferimento**

Situation Awareness

**Titolo argomento di tesi**

Costruzione e specializzazione di un Knowledge Graph per la Cyber Situation Awareness

**Descrizione dell’attività**

1. ***Descrizione del problema che si vuole affrontare (circa 5 righe)***

Nella cybersecurity, la capacità di analizzare e rispondere in modo efficiente a vulnerabilità, debolezze, schemi di attacco e tattiche di minaccia è fondamentale per strategie di difesa efficaci. Con l’aumento della complessità e del volume dei dati sulla sicurezza informatica, i metodi tradizionali di interrogazione e recupero delle informazioni risultano spesso inadeguati.

L’obiettivo principale di questa attività di tesi è la modellazione e la progettazione di una pipeline di Retrieval-Augmented Generation (RAG) basata su Knowledge Graphs per la Cyber Situation Awareness.

Le informazioni di valore relative alla Cyber Situation Awareness spesso risiedono in report non strutturati, per cui è fondamentale l’estrazione di tali dati in tempo reale. La costruzione di un Knowledge Graph per strutturare e rappresentare le informazioni provenienti da report relativi a cyber threats e la capacità di poter recuperare queste informazioni con query in linguaggio naturale consente all’utente finale di raggiungere e mantenere un alto livello di Situation Awareness (SA).

La presente attività di tesi mira a supportare l’utente nell’interrogazione di una base di conoscenza relativa alla Cyber Situation Awareness elaborando la richiesta effettuata in linguaggio naturale. L’idea è quella di progettare uno strumento che permetta all’utente di formulare query in linguaggio naturale e di ricevere delle risposte contestualizzate e informate.

1. ***Analisi dello stato dell’arte, scientifico e/o tecnologico (circa mezza pagina)***

* *Se la tesi ha carattere prevalentemente progettuale – applicativo, identificare lo stato dell’arte tecnologico.*
* *Se la tesi ha invece carattere prevalentemente metodologico, fare riferimento a lavori scientifici o metodologie a cui la tesi è collegata.*
* *I riferimenti devono essere riportati nella forma [1] [2] ecc e devono essere inclusi nella sezione “Bibliografia”*

I grafi di conoscenza per la cybersecurity, che rappresentano il sapere informatico attraverso un modello di dati basato su grafi, offrono approcci olistici per elaborare enormi volumi di dati complessi sulla sicurezza informatica provenienti da fonti diverse. Possono aiutare gli analisti della sicurezza a ottenere informazioni sulle minacce informatiche, raggiungere un alto livello di consapevolezza della situazione, scoprire nuove conoscenze informatiche, visualizzare reti, flussi di dati e percorsi di attacco, e comprendere le correlazioni tra i dati aggregando e integrando le informazioni. L’articolo di Sikos [1] esamina i modelli di dati basati su grafi più rilevanti utilizzati in questo ambito. Ci sono molte caratteristiche relative alla sicurezza informatica e ai processi di rete che devono essere memorizzate quando si lavora con i grafi di conoscenza per la cybersecurity, e la semantica delle conoscenze catturate varia notevolmente a seconda del modello di dati del grafo utilizzato (RDF Graph, Labeled Property Graph, Hypergraph e Multigraph). Le diverse implementazioni basate su grafi presentano punti di forza e debolezze diffrenti. La provenienza dei dati, che può essere sfruttata per la cyber situation awareness, non è supportata da tutti i grafi di conoscenza. Il modello di dati RDF, ad esempio, non dispone di un meccanismo integrato per catturare la provenienza, ma la community di ricerca del Semantic Web ha introdotto formalismi avanzati che estendono il modello di dati RDF standard per questo scopo.

In [2] viene introdotta la pipeline KG-RAG che costruisce un knowledge graph a partire da testo non strutturato e successivamente esegue il recupero di informazioni sul grafo appena creato per effettuare KGQA (Knowledge Graph Question Answering). La metodologia di recupero utilizza un nuovo algoritmo chiamato Chain of Explorations (CoE), che sfrutta il ragionamento dei Large Language Models per esplorare in modo sequenziale nodi e relazioni all’interno del grafo.

In [3] viene descritta l’implementazione di sistemi di Retrieval-Augmented Generation (CyRAG e GraphCyRAG) che integrano Large Language Models (LLM) con dati strutturati provenienti da database relazionali e grafi di conoscenza come Neo4j. CyRAG è progettato per gestire dati strutturati, concentrandosi sulle entità CVE (Common Vulnerabilities and Exposures) e CWE (Common Weakness Enumeration) per generare risposte accurate e ricche di contesto. Al contrario, GraphCyRAG sfrutta i grafi di conoscenza di Neo4j per recuperare informazioni interconnesse dai dataset CVE, CWE, CAPEC (Common Attack Pattern Enumeration and Classification) e ATT&CK (Adversarial Tactics, Techniques and Common Knowledge). Utilizzando il framework basato su grafi di Neo4j, GraphCyRAG consente un’analisi approfondita delle relazioni tra vulnerabilità e schemi di attacco, offrendo ai cybersecurity analysts una visione completa dei potenziali vettori di attacco e delle strategie di mitigazione.

I risultati preliminari dello studio fatto da Rahman et al. [3] dimostrano che l’integrazione di grafi di conoscenza con i sistemi RAG migliora significativamente sia l’accuratezza sia la profondità dell’analisi delle minacce, consentendo il recupero di dati dinamici in tempo reale e la generazione di risposte contestualmente consapevoli.

1. ***Finalità della tesi, contributo dello studente e descrizione dell’attività progettuale di tesi (circa una pagina)***

* *Descrivere dettagliatamente il progetto che si intende svolgere nell’ambito della tesi.* 
  + *Descrivere gli obiettivi, le metodologie/tecnologie adottate (che cosa lo studente farà durante la sua attività di tesi e come lo farà)*
  + *Dettagliare (se presente) il carattere di innovatività introdotto nel progetto rispetto allo stato dell’arte*
  + *Descrivere le fasi in cui si articolerà lo sviluppo del progetto di tesi*

L’obiettivo principale di questa attività di tesi è la modellazione e progettazione di una pipeline di Retrieval-Augmented Generation basata su Knowledge Graphs per la Cyber Situation Awareness.

In un contesto come quello della cybersecurity, in cui ci sono enormi quantità di informazioni complesse e veloci, un sistema del genere consenterebbe all’utente (es. analista della sicurezza) di raggiungere e mantenere un livello di Situation Awareness (SA) che gli permetta di identificare, comprendere e anticipare le minacce in evoluzione.

A tale scopo saranno utilizzati modelli di Named Entity Recognition (NER) per l’estrazione di entità e relazioni dai dataset di report APT, modelli di embedding per supportare il recupero di informazioni dalla base di conoscenza e un LLM esistente in letteratura per la comprensione della query in linguaggio naturale e generazione di risposte.

Un sistema di retrieval-augmented generation (RAG) costruisce un database indicizzando documenti (e parti di documenti) utilizzando encodings basati su Large Language Models. Successivamente, recupera i documenti più rilevanti (o parti di essi) per una determinata query. Questi documenti vengono poi utilizzati come contesto, insieme alla query fornita dall’utente (prompt), per un LLM. Le risposte generate da un LLM si basano quindi non solo sul prompt, ma anche sul contesto fornito dalla base di conoscenza. Poiché i documenti recuperati vengono forniti insieme alle risposte dell’LLM, l’utente può valutare le risposte nel contesto dei documenti recuperati. L’integrazione della retrieval-augmented generation con i grafi di conoscenza amplifica ulteriormente queste capacità. La RAG è un approccio avanzato che combina il recupero di informazioni con modelli generativi per fornire risposte accurate, contestualmente consapevoli e dettagliate a query complesse. I sistemi tradizionali di query-response si basano spesso su modelli pre-addestrati che potrebbero non avere accesso alle informazioni più aggiornate o complete. La RAG, invece, colma questa lacuna recuperando dati rilevanti da fonti esterne estese, come i grafi di conoscenza, e utilizzando queste informazioni recuperate come base per generare risposte informate. Questo rende il sistema altamente dinamico, capace di gestire informazioni in tempo reale e dataset complessi.

L’attività di tesi si articola nelle seguenti fasi:

* La prima fase riguarda un’analisi preliminare del dominio di interesse e l’analisi del relativo stato dell’arte. Verranno esplorate e comprese le metodologie e tecnologie esistenti nel campo della RAG, con particolare attenzione all’utilizzo di dataset relativi alla CyberSA ed a eventuali modelli di LLM già disponibili in letteratura. Verrà condotto uno studio dei vector database, dei linguaggi di interrogazione e delle ontologie, analizzando le loro applicazioni e limitazioni. Infine, verranno individuate le best practices e le soluzioni più accreditate nell’applicazioni di tali tecnologie.
* La seconda fase consiste nello studio dei dati e la raccolta di dataset rappresentativi del dominio di interesse. In particolare, dataset di report relativi ad attacchi avanzati e persistenti (APT) non strutturati, come quelli pubblicati da piattaforme di threat intelligence o società di sicurezza informatica (ad es. Mandiant, CrowdStrike). Questi report contengono descrizioni dettagliate delle campagne di attacco, Indicatori di Compromissione (IoC) e profili degli attaccanti.
* La terza fase consiste nell’implementazione di un modello di Open Information Extraction (OIE) e nell’utilizzo di un modello di Named Entity Recognition (NER) per estrarre entità e relazioni dai dataset di report APT. Successivamente, le entità estratte verranno integrate in un knowledge graph, creando una struttura coerente e navigabile. Saranno utilizzate ontologie standard del dominio, come OWL e specifiche come, ad esempio, STIX.
* La quarta fase del processo di sperimentazione consiste nella creazione di un vector database. Questo database conterrà le entità e le relazioni semantiche e sarà utilizzato per le interrogazioni da parte dell’LLM. Il vector database memorizza le rappresentazioni vettoriali (embeddings) dei dati semantici, per favorire un recupero delle informazioni in modo efficiente. Questa struttura permette di migliorare significativamente l’accuratezza e la rapidità nel fornire risposte pertinenti, sfruttando le capacità avanzate di comprensione del linguaggio naturale dell’LLM.
* Infine, verranno recuperati i documenti più rilevanti per una determinata query dal vector database utilizzando delle misure di similarità (fase di Retrieval). Questi documenti vengono poi dati al Large Language Model come contesto insieme alla query. A questo punto il Large Language Model potrà rispondere alla query in maniera contestualizzata e informata (fase di Generation).
* Testing: Svolgimento di test per valutare le prestazioni del sistema in condizioni realistiche. Il tesista itererà sul sistema, apportando miglioramenti e ottimizzazioni basate sui risultati dei test. In particolare, verranno valutate le performance sia in fase di Retrieval che in fase di Generation.

1. ***Descrizione del protocollo sperimentale o del setup sperimentale (circa mezza pagina)***

* *Sperimentazione che si intende effettuare (per tesi di tipo prevalentemente metodologico) o descrizione dettagliata del setup sperimentale (per tesi di tipo prevalentemente applicativo)*
* *Indici prestazionali che saranno utilizzati per misurare la bontà del lavoro svolto e riferimento ai valori di tali indici (se già disponibili) di sistemi esistenti o di metodi allo stato dell’arte*

In una prima fase verranno utilizzate tecnologie semantiche e moduli python tra cui un NER (Named Entity Recognition) per estrarre entità e relazioni dai dataset di report APT.

In una seconda fase verrà impiegata una rete di embedding per la creazione del vector database utilizzato per il recupero delle informazioni da parte del Large Language Model in fase successiva.

Si procederà poi con l’impiego di moduli Python per recuperare dal vector database i documenti più rilevanti in base ad una determinata query. Tali documenti verranno poi utilizzati dall’LLM per generare una risposta coerente con le richieste dell’utente.

Per avvalorare le performance del sistema potranno essere utilizzate alcune tra le metriche di valutazione selezionate in base allo stato dell’arte, quali Accuracy, Precision, Hits@N e MRR.

Il setup sperimentale comprenderà:

* Una fase di elaborazione del dataset iniziale per la creazione del knowledge graph e per la generazione del vector database.
* Una fase di ricerca per similarità delle informazioni più inerenti alla richiesta ricevuta.
* Generazione di una query SPARQL per interrogare la knowledge base e il modello LLM e la conseguente generazione di una risposta.
* La valutazione delle prestazioni confrontate con lo stato dell’arte e mediante metriche note citate precedentemente.

1. ***Dettaglio dei dataset che saranno utilizzati per la sperimentazione (con riferimento al tipo di dataset, se reale o sintentico, al numero campioni, ecc) (circa 5 righe)***

* *In tale sezione devono essere chiarite le motivazioni alla base della scelta di quel dataset evidenziando il fatto che il dataset sia rappresentativo di un’istanza reale del problema affrontato. Laddove non è possibile utilizzare un’istanza reale del problema, è necessario chiarire e motivare accuratamente le ragioni alla base di tale circostanza.*

1. ***Descrizione del dimostratore che sarà realizzato nell’ambito della tesi***
   * *Laddove non sia prevista la realizzazione di alcun dimostratore, specificare il motivo per cui non è possibile realizzare un dimostratore da mostrare live (anche nella forma di un video) durante la seduta di laurea*

**Bibliografia**

[1] Sikos, L.F. Cybersecurity knowledge graphs. *Knowl Inf Syst* **65**, 3511–3531 (2023).

[2] Sanmartin, Diego. "KG-RAG: Bridging the Gap Between Knowledge and Creativity." *arXiv preprint arXiv:2405.12035* (2024).

[3] Rahman, Moqsadur, et al. *Retrieval Augmented Generation for Robust Cyber Defense*. No. PNNL-36792. Pacific Northwest National Laboratory (PNNL), Richland, WA (United States), 2024.

**Proposta di sommario**

Nota: Deve contenere i titoli dei capitoli e delle sezioni. Questi capitoli sono obbligatori, ma è possibile integrare con capitoli aggiuntivi.

1. Introduzione
   * Definizione del problema
   * Rilevanza della problematica nel contesto dell’ingegneria informatica
2. Stato dell’arte
   * Analisi di dettaglio dello stato dell’arte
   * Individuazione di possibili avanzamenti rispetto allo stato dell’arte
3. Contributo originale alla soluzione del problema
   * Definizione della metodologia proposta
   * Innovazioni di carattere metodologico (eventuale)
   * Progettazione del sistema proposto
   * Innovazioni di carattere tecnologico e/o applicativo (eventuale)
   * Strumenti, tecnologie e modelli utilizzati per la realizzazione
4. Validazione sperimentale e aspetti applicativi
   * Descrizione delle metriche di valutazione dei risultati
   * Definizione del protocollo sperimentale o di verifica
   * Descrizione dei dati e/o del caso di studi utilizzato per la sperimentazione
   * Presentazione e analisi dei risultati
   * Valutazione della significatività dei risultati ottenuti e dei miglioramenti apportabili
5. Bibliografia

**Contributi di natura progettuale ed implementativa che il tesista dovrà fornire (circa 10 righe)**

XXX

**Tecnologie e materiali da impiegare durante l’attività di tesi**

* + *Dettagliare se il tesista avrà accesso a specifici strumenti (es: strumenti di misurazione, piattaforme robotiche, server messi a disposizione dal gruppo di ricerca, ad esempio equipaggiati con GPU)*

XXX

**Eventuali esami ancora da sostenere alla data di consegna del presente documento**

* + Indicare per ciascun esame la data (mese/anno) in cui si prevede di sostenerlo

XXX

**Data (mese/anno) in cui l’attività di tesi è iniziata con un impegno sostanzialmente a tempo pieno**

XXX

**Data (mese/anno) in cui presumibilmente sarà discussa la tesi di laurea**

XXX

***(Solo per le Tesi svolte in ERASMUS)***

**Sede**

XXX

**Periodo di svolgimento**

XXX

**Tutor**

XXX

Note