Bauhaus-Universität Weimar
Fakultät Medien
Studiengang Informatik
Schwerpunkt Security and Data Science

Einsatz von Knowledge-Graphen und LLMs zur Optimierung des Information Retrievals in wissenschaftlicher Literatur

Bachelorarbeit

Daniel Ocks

1. Gutachter: Prof. Dr. Benno Stein
2. Gutachter: Jun.-Prof. Dr. Maurice Jakesch

Datum der Abgabe: 10. April 2025

Erklärung

Hiermit versichere ich, dass ich diese Arbeit selbstständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel verwendet habe.

Sonnefeld, 10. April 2025

. . . . . .
Daniel Ocks

Zusammenfassung

Diese Bachelorarbeit untersucht den Einsatz von LLMs zur Umwandlung von Suchanfragen und wissenschaftlichen Arbeiten in Knowledge-Graphen, um zu prüfen, ob Knowledge-Graphen einen positiven Einfluss auf das Information Retrieval haben. Somit ist Ziel dieser Arbeit, herauszufinden, ob Knowledge-Graphen das Information Retrieval verbessern können. Grund für diese Untersuchung sind die Schwächen traditioneller Suchsysteme und die potenziellen Vorteile von Knowledge-Graphen. Während Standardsuchmaschinen bei komplexen Suchanfragen oft an ihre Grenzen stoßen, bieten Knowledge-Graphen die Möglichkeit, Informationen strukturiert darzustellen und eine semantische Ebene hinzuzufügen. Sie optimieren das Information Retrieval, indem sie kontextuelle und bedeutungsvolle Verbindungen herstellen, anstatt sich lediglich auf eine Schlüsselwortsuche zu stützen.

Die Experimente dieser Untersuchung bestehen aus zwei Hauptteilen. Im ersten Teil wird untersucht, wie gut wissenschaftliche Arbeiten in Form eines Rankings einer Suchanfrage zugeordnet werden. Der zweite Teil konzentriert sich darauf zu klassifizieren, ob eine wissenschaftliche Arbeit eine Suchanfrage beantworten kann oder nicht. Dabei wurden in beiden Teilen die gleichen Methoden eingesetzt, darunter der Vergleich von Knowledge-Graphen, der Vergleich eines Knowledge-Graphen mit einer wissenschaftlichen Arbeit, der Vergleich einer Suchanfrage mit einem Knowledge-Graphen, der direkte Vergleich einer Suchanfrage mit einer wissenschaftlichen Arbeit sowie einfaches Keyword-Matching. Mit Ausnahme des Keyword-Matchings wurden alle Methoden unter Verwendung eines LLMs zur Analyse der Eingaben genutzt. Die Knowledge-Graphen wurden ebenfalls mithilfe eines LLMs generiert.

Die Ergebnisse zeigen, dass der ausschließliche Einsatz von Knowledge-Graphen im Vergleich zu den anderen getesteten Methoden die schlechtesten Ergebnisse erzielt. Insbesondere weist diese Methode Schwächen in Bezug auf Konsistenz, Präzision und das Verständnis von Knowledge-Graph-Inhalten auf. Im Gegensatz dazu lieferten Methoden, die ohne Knowledge-Graphen arbeiten, die besten Resultate. Reine Knowledge-Graph Methoden im Information Retrieval bieten daher keine signifikanten Vorteile im Vergleich zu hybriden oder nicht auf Knowledge-Graphen basierenden Methoden. Obwohl sie in der Theorie großes Potenzial aufweisen, erfordert das praktische Verständnis von Knowledge-Graphen eine Weiterentwicklung der LLMs, um diese Strukturen korrekt zu erfassen und zu interpretieren.

Danksagung

An dieser Stelle möchte ich mich bei meinen Betreuern Dr. Tim Gollub und Maximilian Heinrich für ihre Unterstützung bedanken. Die wöchentlichen Meetings haben maßgeblich dazu beigetragen, meine Arbeit in die richtige Richtung zu lenken. Ihre wertvollen Anregungen und ihr Engagement haben mir sehr geholfen. Vielen Dank für Ihre Zeit und Ihre wertvolle Unterstützung!

Inhaltsverzeichnis

Kapitel 1

Einleitung

1.1 Problemstellung

In einer zunehmend digitalisierten und informationsgetriebenen Welt stellt die schnelle und präzise Identifikation relevanter Informationen eine zentrale Herausforderung dar, insbesondere in wissenschaftlichen Kontexten. Die IR Anthology[1], eine umfassende Sammlung wissenschaftlicher Publikationen im Bereich der Information Retrieval-Forschung, repräsentiert eine wertvolle Ressource, die jedoch aufgrund ihres umfangreichen und komplexen Inhalts schwer zu durchsuchen ist. Traditionelle Suchansätze, die auf keyword-basierten Algorithmen beruhen, stoßen bei der Bewältigung dieser Aufgabe an ihre Grenzen. *"Keyword suggestion, which consists in recommending keywords similar to the user's input, is critical in search scenarios where the completeness of the query clauses may dramatically affect the recall, such as academic or legal search. An incomplete query may end up in a null search session, that is, the system presents an empty result list to the user"* [Gabín et al., 2023]. Diese Einschränkungen führen dazu, dass relevante Informationen übersehen oder irrelevante Ergebnisse als relevant eingestuft werden. Insbesondere in Sammlungen mit hohem fachlichem und thematischem Detailgrad, wie der IR Anthology, kann dies die Effizienz und Qualität von Forschungsprozessen erheblich beeinträchtigen. Wissenschaftliche Arbeiten sind häufig nicht nur durch ihre Schlüsselbegriffe, sondern auch durch ihre inhaltlichen Verknüpfungen und den Kontext ihrer Forschungsergebnisse von Bedeutung. Eine rein textuelle Analyse kann solche semantischen Zusammenhänge jedoch nur begrenzt erfassen.

[1] `https://ir.webis.de/anthology/`

Angesichts dieser Problematik besteht ein dringender Bedarf an innovativen Methoden, die über traditionelle Suchalgorithmen hinausgehen. Insbesondere Verfahren, die semantische Beziehungen explizit modellieren und kontextbasierte Analysen ermöglichen, könnten dazu beitragen, die Sucheffektivität zu erhöhen und gleichzeitig tiefere Einblicke in die strukturellen Zusammenhänge wissenschaftlicher Werke zu ermöglichen. Dies schafft die Grundlage für qualitativ hochwertigeres und zielgerichteteres Information Retrieval, das den Ansprüchen moderner Forschung gerecht wird.

1.2 Zielsetzung und Motivation

Die Nutzung von Knowledge-Graphen für die Informationsgewinnung innerhalb wissenschaftlicher Literatur eröffnet neue Möglichkeiten, die über die traditionellen Ansätze hinausgehen. Ziel dieser Arbeit ist es, zu beantworten, ob LLMs mithilfe von Knowledge-Graphen das Information Retrieval in riesigen digitalen Sammlungen wie der IR Anthology signifikant verbessern können. Dabei werden thematische und semantische Zusammenhänge innerhalb wissenschaftlicher Arbeiten explizit gemacht, um eine tiefere, kontextbasierte Suche zu ermöglichen. Dies soll nicht nur die Effektivität der Suche steigern, sondern auch die Qualität der Ergebnisse erhöhen und die Identifikation wertvoller, aber schwer auffindbarer Informationen erleichtern. Ein zentrales Ziel dieser Arbeit ist es, zu testen, ob Knowledge-Graphen in Kombination mit LLMs zu einer besseren Genauigkeit der Suchergebnisse führen können im Vergleich zu Suchmethoden wie Keyword-Matching und einfachen LLM-Textvergleichen.

Die Motivation für diese Arbeit ergibt sich aus den klar erkennbaren Schwächen traditioneller Suchsysteme und den potenziellen Vorteilen, die Knowledge-Graphen bieten. Während Standardsuchmaschinen wie Google oder Bing bei allgemeinen Suchanfragen effektiv sind, stoßen sie bei der Verarbeitung komplexer, fachspezifischer Inhalte oft an ihre Grenzen. Die Vorteile der Informationsgewinnung mit Knowledge-Graphen gegenüber Standardsuchmaschinen verdeutlichen die Relevanz dieser Forschung:

Tiefere semantische Suche: Traditionelle Suchmaschinen basieren oft auf Schlüsselwortsuche und liefern Ergebnisse basierend auf einfacher Textübereinstimmung. Knowledge-Graphen nutzen semantische Beziehungen zwischen Entitäten und Konzepten, wodurch Suchergebnisse nicht nur relevante Begriffe, sondern auch kontextuell verbundene Informationen umfassen. Dies ermöglicht eine explorative Suche, bei der nicht nur direkte Treffer, sondern auch Zusammenhänge zwischen Themen gefunden werden.

Erklärung und Transparenz der Beziehungen: Knowledge-Graphen stellen die Verbindungen zwischen Konzepten explizit dar. Im Vergleich zu klassischen Suchmaschinen bieten sie eine nachvollziehbare Struktur der Beziehungen zwischen Begriffen und Entitäten. Aufgrund dieser Struktur wird ersichtlich, warum ein bestimmtes Dokument als relevant eingestuft wird.

Schlussfolgerungen und Wissensextraktion: Durch die Verknüpfung von Entitäten und Konzepten können Knowledge-Graphen versteckte Muster und Zusammenhänge aufdecken. Dadurch erhält man nicht nur direkte Antworten, sondern kann auch weiterführende Informationen erschließen und neue Erkenntnisse gewinnen. Dies ist besonders vorteilhaft für die wissenschaftliche Forschung, da es das Verständnis komplexer Themen erleichtert.

Spezialisierung auf Fachgebiete: Während allgemeine Suchmaschinen darauf optimiert sind, eine Vielzahl von Themen abzudecken, können Knowledge-Graphen spezifisch für bestimmte Fachgebiete angepasst werden. Sie können gezielt für wissenschaftliche Textsammlungen wie die IR Anthology optimiert werden, um relevantere und präzisere Ergebnisse für fachspezifische Suchanfragen zu liefern.

Resistenz gegen Manipulation: Klassische Suchmaschinen sind anfällig für SEO-Manipulation wie Keyword Stuffing, bei der Inhalte gezielt optimiert werden, um in den Rankings höher zu erscheinen. Knowledge-Graphen basieren auf der strukturierten Analyse von Beziehungen zwischen Entitäten und sind daher weniger manipulierbar. Dies führt zu einer höheren Verlässlichkeit der Ergebnisse, insbesondere in wissenschaftlichen Kontexten [Sarrafzadeh et al., 2020].

Diese Vorteile zeigen das Potenzial von Knowledge-Graphen, wissenschaftliche Recherchen präziser zu gestalten. Diese Arbeit trägt dazu bei, innovativ das Information Retrieval in wissenschaftlichen Kontexten zu verbessern.

1.3 Aufbau der Arbeit

Diese Arbeit ist in mehrere Kapitel unterteilt, die aufeinander aufbauen, um die Problemstellung, die Zielsetzung sowie die entwickelten Ansätze und Ergebnisse umfassend darzustellen. Der Aufbau gliedert sich wie folgt:

Kapitel 1: Einleitung
Dieses Kapitel führt in die Thematik ein und beschreibt die Problemstellung, die Motivation sowie die Zielsetzung der Arbeit. Es stellt dar, warum traditionelle Suchmethoden an ihre Grenzen stoßen und wie Knowledge-Graphen eine innovative Lösung für präzises wissenschaftliches Information Retrieval bieten können. Abschließend wird ein Überblick über den Aufbau der Arbeit gegeben.

Kapitel 2: Related Work
In diesem Kapitel werden der aktuelle Stand der Forschung und damit verbundene Arbeiten vorgestellt. Es wird erläutert, wie Information Retrieval und Knowledge-Graphen derzeit genutzt werden und welche Ansätze in der Literatur existieren, um semantische Beziehungen und Wissensstrukturen für das Information Retrieval zu verwenden.

Kapitel 3: Methodologie
In diesem Kapitel wird zum einen das Ziel der Experimente erläutert und zum anderen werden die für die Experimente relevanten Methoden beschrieben. Dazu gehören die verwendeten Datensätze, die Transformation von Suchanfragen und wissenschaftlichen Arbeiten in Knowledge-Graphen, die Kategorien und Metriken der Experimente sowie die Evaluierung der Ergebnisse.

Kapitel 4: Experimente
In diesem Kapitel werden die Ergebnisse der durchgeführten Experimente präsentiert, die darauf abzielten, die Präzision der entwickelten und angewandten Methoden zu bewerten. Es umfasst eine detaillierte Auswertung der Ergebnisse sowie die daraus gewonnenen Erkenntnisse.

Kapitel 5: Fazit und Ausblick
Das abschließende Kapitel fasst die Ergebnisse der Arbeit zusammen und reflektiert deren Bedeutung für die Wissenschaft und Praxis. Zudem werden potenzielle Weiterentwicklungen und offene Forschungsfragen aufgezeigt, die sich aus den Ergebnissen dieser Arbeit ergeben.

Kapitel 2

Verwandte Arbeiten

2.1 Information Retrieval und die IR Anthology

Information Retrieval (IR) ist ein zentraler Prozess in der Informationswissenschaft, der darauf abzielt, relevante Informationen aus umfangreichen Datenbeständen zu identifizieren und abzurufen. Der Hauptfokus von IR-Systemen liegt darin, die für eine Benutzeranfrage relevantesten Informationen bereitzustellen.

Dabei werden aus einer großen Sammlung von Dokumenten die ausgewählt, welche potenziell relevant für die Anfrage des Benutzers sind. Das geschieht häufig durch die Übereinstimmung von Begriffen zwischen der Anfrage und den Dokumenten. Anschließend werden die abgerufenen Dokumente nach ihrer Relevanz sortiert. Hierbei kommen verschiedene Modelle zum Einsatz, die darauf abzielen, die Relevanz der Dokumente zu bewerten und die relevantesten Ergebnisse an oberster Stelle anzuzeigen [Hambarde and Proença, 2023].

Ein gängiges Beispiel für diese Art der Suche ist das Conventional Term-based Retrieval, bei dem die Übereinstimmung von Suchbegriffen und Dokumenten im Vordergrund steht. Trotz ihrer weiten Verbreitung und Wirksamkeit stoßen solche Systeme jedoch oft an ihre Grenzen, wenn es um die präzise und kontextualisierte Suche in spezialisierten Fachgebieten geht.

Ein Beispiel für ein spezialisiertes IR-System ist die IR Anthology[1]. Diese ist eine spezialisierte Sammlung wissenschaftlicher Publikationen, die sich ausschließlich auf das Forschungsgebiet des Information Retrieval konzentriert; sie umfasst über 62.000 wissenschaftliche Arbeiten. Entwickelt wurde sie, um der wissenschaftlichen Gemeinschaft eine Plattform zur Verfügung zu stellen, die gezielt auf die Bedürfnisse und Herausforderungen dieses Fachbereichs zugeschnitten ist.

[1] `https://ir.webis.de/anthology/`

Traditionelle akademische Suchmaschinen, wie Google Scholar[2], decken eine Vielzahl von Disziplinen ab, bieten aber keine ausreichende Präzision bei der Suche nach spezifischen Themen innerhalb eines Fachgebiets. *"For instance, the query 'query processing' may yield publications from the perspectives of both databases and information retrieval"* [Potthast et al., 2021]. Dieser Mangel an thematischer Fokussierung führt dazu, dass für IR-relevante Arbeiten oft eine hohe Anzahl irrelevanter Ergebnisse angezeigt wird.

Die IR Anthology setzt diesem Problem eine spezialisierte Lösung entgegen. Sie kombiniert eine umfassende Sammlung von Publikationsmetadaten mit einem durchsuchbaren Volltextkorpus, der es Forschern ermöglicht, gezielte Anfragen zu stellen. Mit Hilfe der Suchmaschine `ChatNoir`[3] können Benutzer Volltextsuchen durchführen und die Ergebnisse mithilfe von Filtern nach spezifischen Metadaten, wie Autoren, Konferenzen oder Veröffentlichungsjahren, präzisieren. Dieser Ansatz führt zu einer höheren Präzision und Relevanz der Ergebnisse im Vergleich zu generischen Suchmaschinen.

Die Architektur der IR Anthology umfasst eine strukturierte Datenbank, die sowohl Metadaten als auch Volltextinhalte indiziert. Durch den Einsatz moderner Retrieval-Technologien, wie `BM25F`, werden Suchergebnisse basierend auf ihrer Relevanz für die Anfrage sortiert und angezeigt. Dies minimiert die Abhängigkeit von globalen Signalen wie Zitationshäufigkeit oder Publikationsdatum, die bei generischen Suchmaschinen häufig bevorzugt werden und dazu führen können, dass relevante, aber weniger häufig zitierte Arbeiten in den Ergebnissen untergehen [Potthast et al., 2021].

Neben der klassischen Suchfunktionalität bietet die IR Anthology auch innovative Ansätze zur explorativen Suche. Eine dynamische explorative Suchtechnologie wurde entwickelt, die es ermöglicht, nicht nur Dokumente zu finden, sondern auch analytische Aussagen über das zugrundeliegende Korpus zu treffen. Über die Schnittstelle *IR Anthology Analytics*[4] kann Facetten-basierte Filterungen vorgenommen werden. Durch eine interaktive Benutzeroberfläche können Forscher beispielsweise die wissenschaftliche Entwicklung eines Themas, Autorenverflechtungen oder Trends innerhalb der IR-Community nachvollziehen. Besonders bemerkenswert ist die Möglichkeit, Autoren auf Basis ihrer Veröffentlichungen bestimmten Konferenzen oder Zeiträumen zuzuordnen und daraus Erkenntnisse über Forschungsnetzwerke zu gewinnen [Gollub et al., 2023].

Die Entwicklung und Bereitstellung der IR Anthology sowie von IR Anthology Analytics stellen einen bedeutenden Fortschritt für die IR-Forschung dar. Sie bieten nicht nur eine zentrale Plattform für den Zugang zu relevanter

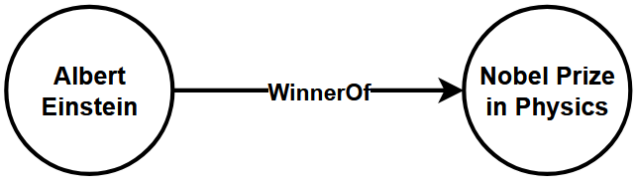
[2] `https://scholar.google.de/`
[3] `https://www.chatnoir.eu/`
[4] `https://ir-analytics.web.webis.de`

Fachliteratur, sondern auch ein leistungsstarkes Tool, um gezielt nach relevanten Dokumenten zu filtern und zu suchen. Doch was passiert, wenn eine präzise Antwort auf eine spezifische Frage erforderlich ist? In diesem Fall wäre die Entwicklung eines Systems notwendig, das Suchanfragen semantisch versteht und gezielt relevante Inhalte in der IR Anthology identifiziert. Während die IR Anthology und IR Anthology Analytics eine solide Grundlage für die Oberflächenrecherche bieten, sind sie derzeit nicht in der Lage, direkte Antworten auf konkrete Fragestellungen zu liefern.

2.2 Knowledge-Graphen: Definition und Konzept

Ein Knowledge-Graph ist eine strukturierte Repräsentation von Wissen, die Entitäten, Beziehungen zwischen diesen Entitäten sowie deren semantische Beschreibungen umfasst. Diese Entitäten können reale Objekte oder abstrakte Konzepte sein, während Beziehungen die Verbindungen zwischen den Entitäten darstellen. Wissen wird in Form von Tripeln modelliert, die aus einer Subjekt-Entität, einer Prädikat-Beziehung und einer Objekt-Entität bestehen, beispielsweise:

„Albert Einstein (Subjekt), WinnerOf (Prädikat), Nobel Prize in Physics (Objekt)“ [Shaoxiong et al., 2022].



Diese Struktur ermöglicht eine Visualisierung und Modellierung von zusammenhängenden Informationen. Darüber hinaus verfügen Knowledge-Graphen über eine klar definierte Semantik. Sowohl die Entitäten als auch ihre Beziehungen sind mit Typen und Eigenschaften versehen, die ihre Bedeutung präzise definieren und ein tiefes Verständnis der dargestellten Daten gewährleisten.

Die Anwendung und die Konzepte von Knowledge-Graphen umfassen mehrere wesentliche Aspekte. Ein zentraler Punkt ist die Darstellung und das Lernen von Informationen. Moderne Ansätze setzen hierbei auf Knowledge-Graph Embedding, um Entitäten und deren Beziehungen in niedrigdimensionalen Vektoren darzustellen. Diese Methode erleichtert maschinelle Lernprozesse und Analyseaufgaben, da sie semantische Interaktionen effizient modelliert. Ein weiterer wichtiger Bereich ist der Erwerb von Wissen, bei dem

Verfahren wie Relationsextraktion und Entity Linking eingesetzt werden. Diese Techniken ermöglichen es, Knowledge-Graphen durch das Hinzufügen neuer Informationen zu erweitern und vorhandene Datenlücken zu schließen. Schließlich finden Knowledge-Graphen in verschiedenen Anwendungsbereichen breite Verwendung. Sie kommen beispielsweise in Empfehlungsdiensten, Frage-Antwort-Systemen und bei Natural Language Processing zum Einsatz. Diese Vielseitigkeit unterstreicht ihre Bedeutung in der modernen Datenverarbeitung und -analyse [Shaoxiong et al., 2022].

Der Einsatz von Knowledge-Graphen ist besonders wertvoll bei datenintensiven und komplexen Problemstellungen, da sie eine präzise und strukturierte Wissensrepräsentation ermöglichen und somit die Grundlage für moderne KI-Systeme darstellen. Sie stellen ein essentielles Werkzeug dar, um maschinelles Lernen mit Hintergrundwissen zu bereichern und dadurch die Genauigkeit bei der Verarbeitung großer Datenmengen zu steigern. Knowledge-Graphen können Suchanfragen in eine Graphstruktur zerlegen und dabei den Kontext der Anfrage bewahren. Durch die Kontextbewahrung und die Unabhängigkeit von Schlüsselwörtern könnten sie die Basis für ein Suchsystem legen, das nicht nach Schlüsselwörtern, sondern nach Kontexten sucht. Voraussetzung hierfür wäre jedoch, dass auch die verfügbare Literatur in Graphen überführt wird, um den Kontext auch dort zu bewahren und gezielt relevante Antworten zu liefern.

2.3 Open Research Knowledge Graph

Der Open Research Knowledge Graph[5] (ORKG) ist eine digitale Infrastruktur, die darauf abzielt, wissenschaftliche Informationen gemäß den FAIR-Prinzipien (Findable, Accessible, Interoperable, Reusable) zu strukturieren, zu kuratieren und zugänglich zu machen. Er umfasst über 30.000 wissenschaftliche Arbeiten und hat das Ziel, eine neue Form der wissenschaftlichen Kommunikation zu fördern, bei der Inhalte von Publikationen maschinenlesbar und somit effizienter nutzbar werden [Stocker et al., 2023].

Traditionelle wissenschaftliche Kommunikation basiert hauptsächlich auf dokumentenbasierten Repositories, bei denen die Informationen oft in Texte eingebettet sind. Diese Struktur erschwert sowohl die maschinelle Verarbeitung als auch die systematische Suche nach spezifischen Inhalten und verlangsamt den Forschungsprozess, beispielsweise bei der Durchführung systematischer Reviews oder der Extraktion von Daten zur Beantwortung konkreter Fragestellungen. ORKG adressiert diese Problematik, indem es wissenschaftliche Inhalte graphbasierte repräsentiert und so Wiederverwendbarkeit und Vergleichbarkeit von Forschungsdaten verbessert [Stocker et al., 2023].

[5] `https://orkg.org/`

ORKG ermöglicht somit die semantische Erfassung und Darstellung von wissenschaftlichen Informationen, indem es eine strukturierte Beschreibung von Forschungsbeiträgen bereitstellt. Nutzer können mithilfe von Templates spezifische Eigenschaften und Werte für Forschungsprobleme, Methoden oder Ergebnisse definieren. Dabei unterstützt der ORKG sowohl automatisierte Verfahren, wie Natural Language Processing, als auch manuelle Ansätze, wie Crowd-/Expertensourcing, um Informationen aus Publikationen zu extrahieren und in den Knowledge Graph zu integrieren. Bei der Konstruktion der Graphen stehen die Kuratoren im Mittelpunkt, welche für die Verarbeitung der Daten verantwortlich sind. Sie sind für die Inhaltsanalyse zuständig. Hier stehen Strukturierte Daten in Form von Tabellen oder Diagrammen, Vergleichskriterien wie Benchmarks, Leistungszahlen und Datensätze, sowie Schlüsselbegriffe im Vordergrund. Mit Text Mining Tools wird nach Mustern oder feineren Informationen gesucht. Des Weiteren wird sowohl die Vorabproduktion von FAIR-Daten während der Datenanalyse als auch die Nachbearbeitung durch automatisierte und manuelle Informationsextraktion aus Artikeln ermöglicht [Stocker et al., 2023]. Die Autoren können in den Prozess der Gestaltung und Strukturierung des ORKG-Graphen ebenfalls eingreifen, durch die manuelle Eingabe und Organisation ihrer Beiträge. Sie haben dadurch einen gewissen Grad an Flexibilität und Mitbestimmung, müssen jedoch strukturelle und technische Rahmenbedingungen berücksichtigen.

Diese Mischung aus menschlicher und maschineller Intelligenz ist notwendig, um die erforderliche Qualität der Darstellung zu erreichen. Dadurch können neuartige Explorations- und Unterstützungsdienste für Forscher ermöglicht werden. Der ORKG kann verwendet werden, um einen komprimierten Überblick über den Stand der Technik in Bezug auf bestimmte Forschungsfragen zu geben, beispielsweise durch tabellarische Vergleiche der Beiträge nach verschiedenen Merkmalen der Ansätze [Auer et al., 2020].

Eine zentrale Funktion des ORKG ist der Comparison Service, der es Forschern ermöglicht, wissenschaftliche Beiträge zu bestimmten Fragestellungen automatisiert zu vergleichen. So können beispielsweise Unterschiede und Gemeinsamkeiten zwischen Methoden oder Ergebnissen verschiedener Studien auf einen Blick dargestellt werden. Diese Visualisierungen unterstützen Wissenschaftler bei systematischen Analysen und unterstützt die schnelle Identifikation relevanter Informationen [Stocker et al., 2023].

ORKG stellt ein innovatives Werkzeug dar, das die Art und Weise, wie wissenschaftliche Informationen erzeugt, kuratiert und genutzt werden, anders angeht. Indem es eine maschinenlesbare Struktur für wissenschaftliche Inhalte fördert, bietet das ORKG eine vielversprechende Lösung für die Herausforderungen moderner Forschungspraktiken. Knowledge-Graphen verbinden die einzelnen Informationen wissenschaftlicher Arbeiten zu umfassenden Netzwer-

ken, wodurch Zusammenhänge sichtbar werden. Durch die Konvertierung wissenschaftlicher Arbeiten in solche Graphen ermöglicht das ORKG einen halbautomatischen Ansatz, um ein Netzwerk von Informationen zu erstellen und miteinander zu verknüpfen. Dieser Prozess bleibt jedoch halbautomatisch, was ihn sehr zeitaufwendig und anspruchsvoll macht und tiefgehendes Fachwissen erfordert. Das langfristige Ziel besteht darin, ein ähnliches System vollständig zu automatisieren, indem LLMs nicht automatischen Prozesse übernehmen.

2.4 LLMs im Information Retrieval

Während klassische IR-Modelle auf termbasierten Methoden wie der Booleschen Suche und Vektorraum-Modellen basierten, ermöglichten neuronale Modelle eine semantisch reichhaltigere Suche. LLMs, wie `GPT-4`, `Llama` oder `BERT` zeichnen sich durch ihre leistungsfähige Sprachverarbeitung und kontextbewusste Generierung aus, wodurch sie verschiedene Komponenten von IR-Systemen verbessern können. LLMs werden in IR-Systemen in verschiedenen Bereichen eingesetzt:

Query Rewriting: LLMs können Suchanfragen umformulieren oder erweitern, um eine höhere Relevanz und bessere Trefferquote zu erzielen. Ein zentraler Anwendungsfall für LLMs ist das Umformulieren von Anfragen, um Suchergebnisse zu verbessern. Traditionelle Methoden zur Query-Expansion basierten auf Wörterbüchern oder Kookkurrenz-Analysen, während LLMs direkt semantisch relevante Begriffe vorschlagen oder ganze Sätze umformulieren können. Beispielsweise kann eine einfache Anfrage wie „beste Kamera“ durch ein LLM zu „Welche Kamera hat die beste Bildqualität unter 1000 Euro?“ umgeschrieben werden, um genauere Suchergebnisse zu erhalten.

Retriever: Diese Modelle helfen dabei, relevante Dokumente effizienter aus großen Datenbanken zu extrahieren. Moderne Information Retrieval Systeme nutzen Vektor-Repräsentationen für Dokumente und Anfragen, um semantische Ähnlichkeiten effizient zu berechnen. LLMs können in diesem Bereich als Dense Retriever fungieren, indem sie Texte in hochdimensionale Vektoren umwandeln und damit die Genauigkeit der Suche erhöhen. Sie übertreffen oft klassische Modelle wie BM25, indem sie kontextuelle Abhängigkeiten berücksichtigen.

Reranker: Nachdem erste Suchergebnisse geliefert wurden, können LLMs diese in einer verbesserten Reihenfolge präsentieren. Reranking-Modelle verfeinern die Reihenfolge der Suchergebnisse, indem sie zusätzliche Signale zur Bewertung der Relevanz eines Dokuments verwenden. LLMs sind besonders nützlich, um feingranulare semantische Unterschiede zu erkennen, die reine Keyword-Matching-Modelle übersehen könnten. Beispielsweise könnte ein LLM erkennen, dass ein Dokument über „nachhaltige Energiequellen“ relevanter für eine Anfrage zu „erneuerbare Energien“ ist als eines, das nur das Wort „Energie“ enthält.

Reader: Sie ermöglichen eine tiefere Analyse der Ergebnisse, indem sie Zusammenfassungen oder direkte Antworten generieren. Reader-Modelle ermöglichen es, aus Suchergebnissen direkt zusammenfassende Antworten zu generieren, anstatt nur eine Liste von Dokumenten anzuzeigen. In Frage-Antwort-Systemen wie Google Search oder Bing Chat nutzen Suchmaschinen LLMs, um relevante Informationen direkt aus mehreren Quellen zusammenzufassen und dem Nutzer eine präzisere Antwort zu liefern [Zhu et al., 2023].

Es zeigt sich, dass LLMs das Information Retrieval verbessern. Ein vielversprechender Ansatz für ein Suchsystem mit Knowledge-Graphen könnte durch die Unterstützung von LLMs somit ermöglicht werden, wenn sie neben den genannten Eigenschaften noch dazu in der Lage sind, Suchanfragen und wissenschaftliche Arbeiten in Knowledge-Graphen zu transformieren. Durch diese Transformation wäre es möglich, gezielt relevante Informationen aus Texten verschiedenen Umfangs zu extrahieren und zu verknüpfen, um genauere und nachvollziehbare Suchergebnisse zu liefern.

2.5 LLMs zur Knowledge-Graph-Extraktion

Die automatisierte Erstellung von Knowledge Graphen aus Texten hat in den letzten Jahren erheblich an Bedeutung gewonnen. Durch den Einsatz von LLMs können semantische Beziehungen zwischen Entitäten effizient extrahiert und in strukturierter Form gespeichert werden.

Ein zentraler Aspekt der Extraktion von Knowledge Graphen aus Texten ist die Identifikation von Entitäten und deren Relationen. Traditionelle Ansätze setzen auf mehrstufige Pipelines mit Named Entity Recognition und Relation Extraction, die jedoch oft durch Fehlerpropagation beeinträchtigt werden. Neuere End-to-End-Ansätze nutzen generative Sprachmodelle zur direkten Extraktion von Entitäts-Relation-Triaden.

Bei der Analyse der Performance von `REBEL`, einem spezialisierten Mo-

dell zur gemeinsamen Entitäts- und Relationenextraktion, und dem Vergleich seiner Ergebnisse mit denen von `ChatGPT` zeigten sich folgende Erkenntnisse: Während `REBEL` für über 200 Relationstypen optimiert wurde, zeigt `ChatGPT` trotz seiner primären Funktion als Dialogmodell eine bemerkenswerte Fähigkeit, relevante Entitäten und Relationen zu extrahieren. Die Experimente zur Analyse wurden anhand von Nachrichtenartikeln zum Thema Nachhaltigkeit durchgeführt, wobei beide Modelle zur Erstellung eines Knowledge Graphs genutzt wurden. Ein wesentlicher Unterschied zwischen den beiden Modellen liegt in der Struktur der extrahierten Daten. Während `REBEL` präzisere und kontextuell kohärente Relationen extrahiert, neigt `ChatGPT` dazu, längere Phrasen als Entitäten zu identifizieren, was die Normalisierung der Knowledge Graphen erschwert. Durch gezieltes Prompt Engineering konnte allerdings die Qualität der von ChatGPT generierten Knowledge Graphen erheblich verbessert werden.

Darüber hinaus konnte die automatische Erstellung von Ontologien mit LLMs beobachtet werden. Hierbei wurde `ChatGPT` angewiesen, basierend auf den analysierten Texten eine Ontologie zu generieren. Die Ergebnisse zeigen, dass LLMs nicht nur Entitäten und Relationen extrahieren, sondern auch strukturierte Ontologien mit hierarchischen Konzepten und Instanzen erzeugen können [Trajanoska et al., 2023].

Zusammenfassend lässt sich sagen, dass der Einsatz von LLMs für die automatisierte Extraktion von Knowledge Graphen geeignet ist, jedoch weiterhin Herausforderungen bestehen, insbesondere in Bezug auf die Konsistenz und Normalisierung der extrahierten Daten, welche allerdings durch Prompt Engineering behoben werden können. *"By offering a mechanism to fine-tune model outputs through carefully crafted instructions, prompt engineering enables these models to excel across diverse tasks and domains"* [Sahoo et al., 2024]. Somit bieten LLMs eine Basis dafür, sowohl Suchanfragen als auch umfangreiche wissenschaftliche Arbeiten in Knowledge-Graphen zu transformieren.

2.6 Limitationen von Knowledge-Graphen im IR

Trotz der zahlreichen Vorteile, die Knowledge-Graphen im Bereich des IR bieten, gibt es einige signifikante Limitationen, die deren Anwendung und Effektivität einschränken können. Diese Herausforderungen lassen sich in verschiedene Kategorien unterteilen:

Herausforderungen bei der Wissensakquisition: Ein wesentliches Problem bei der Erstellung von Knowledge-Graphen ist die Akkuratheit und Vollständigkeit der extrahierten Informationen. Dessen Konstruktion erfordert eine umfangreiche Extraktion von Entitäten und deren Relationen aus Datenquellen. Häufig basieren diese Extraktionen auf maschinellem Lernen oder regelbasierten Methoden, die fehleranfällig sind und inkonsistente Daten liefern können.

Unvollständigkeit und Beschränkung durch fehlende Daten: Die meisten Knowledge-Graphen sind unvollständig, da sie nicht alle möglichen Relationen und Entitäten enthalten. Ein IR-System, das auf einem Knowledge-Graphen basiert, kann deshalb relevante Dokumente übersehen, wenn wichtige semantische Verknüpfungen fehlen. Methoden zur Knowledge-Graph Completion, wie Link-Prediction-Algorithmen, versuchen, diese Lücken zu schließen, haben jedoch oft Schwierigkeiten bei der Generalisierung auf unbekannte oder seltene Entitäten.

Einschränkungen durch Knowledge Reasoning: Knowledge-Graphen bieten oft semantische Inferenzfähigkeiten, die es ermöglichen, neue Relationen aus bestehenden Daten abzuleiten. Insbesondere bei großen Datenmengen kann dies die Effizienz von IR-Systemen verringern, da komplexe Abfragen zu langen Berechnungszeiten führen. Darüber hinaus können falsche Schlussfolgerungen dazu führen, dass Entitäten und Relationen falsch gewählt werden.

Obwohl Knowledge-Graphen erhebliche Vorteile für IR-Systeme bieten, gehen sie gleichzeitig mit einer Reihe technischer Herausforderungen einher. Insbesondere besteht die Schwierigkeit darin, sämtliche relevanten Informationen innerhalb des Graphen vollständig und präzise abzubilden. Diese Problematik lässt sich jedoch durch den Einsatz von LLMs, die in der Lage sind, große Datenmengen zu analysieren und zu interpretieren, in Kombination mit Prompt Engineering überwinden. *"Prompt engineering has emerged as an indispensable technique for extending the capabilities of large language models"* [Sahoo et al., 2024].

2.7 Fazit

Die analysierten Arbeiten zeigen, dass eine solide Grundlage für die Implementierung eines Systems besteht, das Suchanfragen und wissenschaftliche Arbeiten in Knowledge-Graphen überführen und anschließend vergleichen kann, um relevante Literatur gezielt zu identifizieren. Dies wäre insbesondere für große Sammlungen wie die IR-Anthology von Vorteil, da eine präzisere Suche ermöglicht würde, bei der gezielte Anfragen direkt zu den passenden wissenschaftlichen Arbeiten führen, die die gestellte Frage beantworten.

Es sind somit die nötigen Werkzeuge vorhanden, um zu beantworten, ob mithilfe von Knowledge-Graphen das Information Retrieval in riesigen digitalen Sammlungen wie der IR Anthology signifikant verbessert werden kann.

LLMs haben das Potenzial, sowohl Suchanfragen als auch wissenschaftliche Arbeiten in Knowledge-Graphen zu überführen (siehe Kapitel 2.4 und 2.5). Dabei treten Herausforderungen auf, die jedoch durch gezieltes Prompt Engineering minimiert oder überwunden werden können. Dies zeigt, dass LLMs in der Lage sind, komplexe Wissensstrukturen zu extrahieren und strukturiert darzustellen.

Besonders relevant ist die Möglichkeit, die derzeit von Kuratoren im ORKG übernommenen halbautomatischen Aufgaben (siehe Kapitel 2.3) durch LLMs zu ersetzen. Diese Modelle können bereits jetzt Graphen aus umfangreichen wissenschaftlichen Arbeiten generieren, die für die weiterführende Literatursuche von großem Nutzen sind.

Offen bleibt jedoch die Frage, ob LLMs nach dem Generieren von Knowledge-Graphen auch in der Lage sind, diese korrekt zu interpretieren und miteinander zu vergleichen. Es ist zu untersuchen, ob sie diesen Schritt präzise ausführen können oder ob alternative Methoden hier überlegen sind und Knowledge-Graphen keinen signifikanten Vorteil im Information Retrieval bieten.

Kapitel 3

Methodologie

3.1 Zielsetzung

Das Ziel der in folgendem Kapitel (siehe Kapitel 4) ausgeführten Experimente besteht darin, wissenschaftliche Arbeiten in Bezug auf eine gestellte Suchanfrage einzuranken und zu klassifizieren, um herauszufinden, ob Knowledge-Graphen im Bereich des Information Retrieval einen signifikanten Vorteil bieten. Dabei wird analysiert, wie gut LLMs mit Knowledge-Graphen für eine Verbesserung im Information Retrieval umgehen können. Im Rahmen dieser Untersuchung werden die Ergebnisse eines reinen Knowledge-Graph-Ansatzes mit den Ergebnissen hybrider und Knowledge-Graph-freier Ansätze verglichen. Ziel ist es, zu bewerten, ob die Knowledge-Graph-Methoden anderen Ansätzen:

- konkurrieren können.
- überlegen sind, indem sie präzisere Ergebnisse liefern.
- Schwächen aufweisen.

Es soll dadurch die Präzision von Knowledge-Graphen im Information Retrieval bewertet werden. Die Experimente verfolgen dabei zwei Ansätze:

Erstellung einer Rangfolge (Ranking): Der erste Ansatz zielt darauf ab, eine Rangfolge wissenschaftlicher Arbeiten zu erstellen, die ihre Relevanz in Bezug auf die Suchanfrage widerspiegelt. Dokumente, die entweder exakt in das Schema der Suchanfrage passen oder die Suchanfrage am besten beantworten, sollen in der Rangfolge höher eingestuft werden. Dieses Ranking ermöglicht eine priorisierte Darstellung der relevanten wissenschaftlichen Arbeiten.

Klassifizierung in relevant und nicht relevant (Klassifikation): Der zweite Ansatz hat das Ziel, wissenschaftliche Arbeiten in die Kategorien relevant und nicht relevant einzuordnen. Dabei wird geprüft, ob ein Dokument die gestellte Suchanfrage beantworten kann. Nur die wissenschaftlichen Arbeiten, die die Anfrage direkt oder indirekt beantworten, werden als relevant klassifiziert.

Beide Ansätze verfolgen das gemeinsame Ziel, die Suchanfrage mit passenden wissenschaftlichen Dokumenten zu beantworten und damit die Präzision des Information Retrievals zu verbessern.

LLMs wie `GPT 3.5` sind in der Lage, in einem Zero-Shot-Setting mehrere Inputs zu vergleichen und sinnvoll zu ranken. Der Ranking-Prozess wird dabei über natürliche Sprache gesteuert und geschieht mithilfe von instruktionsbasierten Prompts. Diese Fähigkeit ermöglicht somit, die genannten Experimente in Skalenform durchzuführen [Hou et al., 2023].

3.2 Beschreibung der Datensätze

Für die folgenden Experimente wurden drei verschiedene Datensätze zusammengestellt, um unterschiedliche Aspekte der entwickelten Methoden zu evaluieren. Diese Datensätze umfassen wissenschaftliche Arbeiten aus verschiedenen Quellen und mit unterschiedlichen thematischen Schwerpunkten, um sowohl die Robustheit als auch die Spezifität der Experimente zu gewährleisten. Im Folgenden werden die Datensätze beschrieben.

Datensatz 1: Wissenschaftliche Arbeiten aus der IR-Anthology
Dieser Datensatz besteht aus 10 wissenschaftlichen Arbeiten, die zufällig aus der IR Anthology extrahiert wurden. Die Arbeiten behandeln verschiedene Themen im Bereich Information Retrieval. Dadurch bietet der Datensatz eine Grundlage für die Evaluierung verschiedener Ansätze.

Datensatz 2: Duplizierte wissenschaftliche Arbeit
Dieser Datensatz besteht aus zehn identischen Kopien der wissenschaftlichen Publikation *"Dynamic Exploratory Search for the Information Retrieval Anthology"* [Gollub et al., 2023]. Diese Arbeit, die bereits im ersten Datensatz enthalten ist, wurde bewusst vervielfältigt, um Konsistenz und Robustheit der angewandten Methoden zu bewerten.

Datensatz 3: Arbeiten zum Thema Knowledge-Graphen
Dieser Datensatz umfasst zehn wissenschaftliche Arbeiten aus verschiedenen Quellen, die sich alle mit Knowledge-Graphen befassen. Aufgrund ihrer thematischen Nähe eignet er sich besonders zur Untersuchung der Fähigkeit von LLMs, feine inhaltliche Punkte präzise zu identifizieren und zu analysieren.

3.3 Erstellung der Such- und Dokumentgraphen

Für die Darstellung und Analyse von Informationen werden Such- und Dokumentgraphen verwendet, die mithilfe von `Gemini 1.5-pro`[1] generiert werden (siehe Anhang C.2). Dieses Model wird genutzt, da es bereits als leistungsfähiges Modell für komplexe Zero-Shot-Aufgaben validiert wurde, bei denen es darum geht, Informationen aus umfangreichen Eingabekontexten gezielt zu extrahieren. Insbesondere die Benchmark „Counting-Stars“ demonstriert, dass `Gemini 1.5-pro` in der Lage ist, Informationen zuverlässig zu identifizieren und korrekt zusammenzuführen. Diese Fähigkeit ist essenziell, wenn es darum geht, Knowledge-Graphen zu erstellen [Song et al., 2023].

Die Knowledge-Graphen folgen einer standardisierten Syntax auf Basis von `Neo4j Cypher`[2], da diese Sprache eine effiziente Beschreibung, Abfrage und Modellierung komplexer Graphstrukturen ermöglicht. Die Integration in die Neo4j-Datenbank gewährleistet eine effiziente Speicherung und Verarbeitung großer Datenmengen und schafft eine strukturierte Grundlage für die Analyse und Experimente (siehe Anhang C.1).

Der Suchgraph wird aus einer gestellten Suchanfrage erstellt. Dazu analysiert `Gemini` die Anfrage, um relevante Entitäten und Beziehungen zu identifizieren. Der an `Gemini` übergebene Prompt beschreibt präzise, wie die Suchanfrage in einen Suchgraphen transformiert werden soll, und enthält ein konkretes Beispiel für die Umsetzung von Text in `Neo4j Cypher-Syntax`. Der resultierende Suchgraph bildet die Suchanfrage exakt ab, indem ausschließlich explizit genannte Begriffe und Verbindungen berücksichtigt werden. Diese strukturierte Darstellung erleichtert den Vergleich mit Dokumentgraphen und ermöglicht eine effiziente Analyse, ohne die Intention der Anfrage zu verfälschen. Erweiterungen durch Synonyme oder verwandte Begriffe werden bewusst vermieden, um die Präzision der Repräsentation zu gewährleisten.

Der Dokumentgraph wird ebenfalls mithilfe von `Gemini` generiert, wobei eine wissenschaftliche Arbeit und die zugehörige Suchanfrage als Eingabe dienen. Ziel ist es, die wichtigsten Entitäten und Beziehungen aus der Arbeit zu extrahieren und diese in einem Knowledge-Graphen darzustellen. Im Gegensatz

[1] `https://aistudio.google.com/prompts/new_chat?model=gemini-1.5-pro`
[2] `https://neo4j.com/docs/cypher-manual/current/introduction/`

zum Suchgraphen wird der Dokumentgraph durch Synonyme und verwandte Begriffe erweitert, um eine umfassendere Abdeckung des Wissensbereichs zu ermöglichen. Bei der Generierung verfolgt das LLM zwei Ansätze:

1. Antwort auf die Suchanfrage:
Wenn die wissenschaftliche Arbeit die gestellte Suchanfrage beantwortet, wird die Antwort explizit im Dokumentgraphen dargestellt.

2. Unabhängiger Dokumentgraph:
Falls die Antwort nicht in der wissenschaftlichen Arbeit enthalten ist, wird ein Dokumentgraph erstellt, der ausschließlich auf den Inhalten der Arbeit basiert, ohne die Suchanfrage weiter zu berücksichtigen.

Dieses Vorgehen soll gewährleisten, dass der Dokumentgraph nicht nur die Inhalte der Arbeit abbildet, sondern gezielt die für die Suchanfrage relevanten Punkte hervorhebt.

Zusätzlich zu der beschriebenen Methode wurden weitere Verfahren zur Graph-Generierung eingesetzt, jedoch aufgrund ihrer hohen Fehleranfälligkeit und Ungenauigkeit nicht weiter berücksichtigt (siehe Anhang B).

3.4 Metriken

SG und DG
In diesem Ansatz wird ein LLM verwendet, um einen Suchgraphen (SG) mit einem Dokumentgraphen (DG) zu vergleichen. Das LLM bewertet dabei, ob die Suchanfrage im Suchgraphen mithilfe des Dokumentgraphen beantwortet werden kann. Ziel dieses Ansatzes ist es, die Suche mit Knowledge-Graphen zu modellieren und aufzuzeigen, wie präzise eine Suche allein durch den Einsatz von Knowledge-Graphen erfolgen kann.

SG und WA / SA und DG
In diesem Ansatz wird ein Suchgraph (SG) direkt mit einer wissenschaftlichen Arbeit (WA) bzw. eine Suchanfrage (SA) mit einem Dokumentgraphen (DG) verglichen. Das LLM bewertet dabei das jeweilige Szenario. Dieser hybride Ansatz soll aufzeigen, ob Methoden, bei denen nicht die gesamte Suche in Knowledge-Graphen aufgelöst wird, sondern nur ein Teil, einen Vorteil gegenüber einer ausschließlich auf Knowledge-Graphen basierenden Methode oder einer Methode ohne Knowledge-Graphen bieten.

SA und WA
In dieser Methode wird die Suchanfrage (SA) direkt mit einer wissenschaftlichen Arbeit (WA) verglichen. Das LLM beurteilt, ob die Suchanfrage durch den Inhalt einer wissenschaftlichen Arbeit beantwortet werden kann. Ziel dieser Methode ist es, zu zeigen, ob LLMs ohne die Transformation von Texten in Knowledge-Graphen besser abschneiden oder nicht.

Die Ergebnisse dieser drei genannten Methoden sind numerische Werte, welche die Präzision der Suche beschreiben.

KWMS (KeyWord Matching Score) in %
Dieser Score gibt den Prozentsatz der Keywords an, die in der wissenschaftlichen Arbeit enthalten sind und dient der Bewertung der Präzision eines Dokuments auf Basis des implementierten Keyword-Matching-Verfahrens (siehe Anhang A.4). Dieser Ansatz fungiert als Vergleichspunkt und soll letztlich die Frage beantworten, ob Keyword-Matching, den bisher genannten LLM-gestützten Methoden mit und ohne Knowledge-Graphen, überlegen ist oder nicht.

Darüber hinaus wurden weitere ausgeführte, jedoch für die Kernbetrachtung dieser Arbeit nicht relevanten Metriken im Anhang A dokumentiert.

3.5 Evaluierung

Die Evaluierung der Ergebnisse basiert auf einer vorherigen manuellen Festlegung der erwarteten Ergebnismenge (**Erwartet**). Diese wurde für jede wissenschaftliche Arbeit individuell ermittelt, nachdem die Suchanfrage erstellt wurde. Ziel dieser Vorgehensweise ist es, die Genauigkeit der durch die entwickelten Methoden erzielten Ergebnisse mit den erwarteten Werten zu vergleichen. Dabei gilt:

1: Die Antwort auf die Suchanfrage ist in der Arbeit enthalten.

0: Die Antwort auf die Suchanfrage ist nicht in der Arbeit enthalten.

Diese binäre Klassifikation dient als Grundlage für die Evaluierung sowohl der Ranking- als auch der Klassifikations-Methoden.

Bei den **Ranking-Experimenten** wird für jede Suchanfrage oder jeden Suchgraph ein Wert zwischen 1 und 10 als Ergebnis ausgegeben. Ein höherer Wert wird für Arbeiten angestrebt, die die Suchanfrage präzise beantworten können (Erwartungswert: 1), während Arbeiten, die die Suchanfrage nicht oder nur unzureichend beantworten (Erwartungswert: 0), entsprechend niedrigere Werte erhalten sollen. Dabei wird jedes Ranking-Ergebnis vollständig unabhängig von anderen Suchanfragen, Suchgraphen oder bekannten Ranking-Informationen erstellt. Es gibt keine Abhängigkeiten oder Rückschlüsse auf vorherige Ergebnisse, um sicherzustellen, dass jede Bewertung ausschließlich auf der aktuellen Suchanfrage basiert. Die Bewertung der Ranking-Methoden erfolgt durch die Analyse der Übereinstimmung zwischen der erwarteten Rangordnung und den tatsächlich erzielten Ergebnissen.

Bei den **Klassifikations-Experimenten** wird jede Arbeit in die Kategorien relevant (1) oder nicht relevant (0) eingeordnet. Ziel ist es, dass das Ergebnis der Klassifikationsmethoden exakt der erwarteten Ergebnismenge entspricht. Eine korrekte Klassifikation bedeutet, dass die Methode den gleichen Wert (1 oder 0) liefert wie der manuell bestimmte Erwartungswert.

Die Ergebnisse des Rankings und der Klassifikation wurden mithilfe von `Gemini 1.5-pro` generiert. Die Inputs dafür folgen aus der jeweils angewandten Metrik (siehe Kapitel 3.4).

"The Counting-Stars mainly evaluates the long-context capability of LLMs from two perspectives, i.e., long-context multi-evidence searching and long-context multi-evidence reasoning."[Song et al., 2023] `Gemini` demonstriert seine Fähigkeit, mehrere Informationen im Kontext zu vergleichen, zu gewichten und gezielt auszuwählen – eine zentrale Voraussetzung für Ranking-Aufgaben. Das Modell erzielt überdurchschnittlich gute Leistungen bei Aufgaben, die Selektion, Gewichtung und Bewertung mehrerer Belege erfordern. Genau das ist für den Vergleich und das Ranking zweier Inputs entscheidend. Dies zeigt deutlich, dass das Modell nicht nur Informationen sammeln kann, sondern auch bewerten kann, welche davon wie relevant sind [Song et al., 2023].

Die Evaluierung dient dazu, die Präzision der entwickelten Methoden zu messen und ihre Übereinstimmung mit den manuell festgelegten Erwartungen zu überprüfen.

Kapitel 4

Experimente

Nach der Beschreibung der Methoden folgt nun die praktische Untersuchung. In diesem Kapitel werden die Ergebnisse der durchgeführten Experimente detailliert vorgestellt. Die Experimente umfassen sowohl das Ranking wissenschaftlicher Arbeiten als auch deren Klassifizierung in relevante und nicht relevante Ergebnisse.

Tabelle 4.1: Ergebnisse des 1. Experiments basierend auf Datensatz 1

WA	1	2	3	4	5	6	7	8	9	10
					Ranking					
SG und DG	10	1	10	1	1	1	1	1	1	1
SG und WA	10	1	1	2	1	1	1	1	1	1
SA und DG	2	1	1	1	1	1	1	1	1	1
SA und WA	7	1	1	1	1	1	1	1	1	1
					Klassifikation					
SG und DG	1	0	1	1	0	1	0	0	0	1
SG und WA	1	1	1	1	0	0	1	0	0	1
SA und DG	1	0	0	1	0	0	0	0	0	0
SA und WA	1	1	0	1	0	0	0	0	0	1
KWMS in %	4.54	0.41	0.0	0.0	0.04	0.0	0.02	0.04	0.03	0.10
Erwartet	1	0	0	0	0	0	0	0	0	0

Ergebnisse Tabelle 4.1:
Die Suchanfrage zu Experiment 1: *"Which authors have dealt with faceted search?"* hatte das Ziel, WA 1 zu identifizieren (siehe Tabelle 4.1). Im Folgenden werden die Ergebnisse der verschiedenen Ansätze dargestellt und bewertet:

1. Ranking SG und DG:
   Dieses Verfahren lieferte neben der gewünschten Zielarbeit einen zusätzlichen Treffer, der das Ergebnis verfälscht. Dies geschieht, obwohl der zugehörige Dokumentgraph keine Entitäten oder Relationen enthält, die auf dieses Ergebnis hindeuten könnten.
   Die Erwartungen an diese Methode wurden nicht erfüllt.

2. Ranking SG und WA:
   Diese Methode erfüllte die Erwartungen vollständig und identifizierte die Zielarbeit korrekt.

3. Ranking SA und DG:
   Die Zielarbeit wurde als einzige mit einem höheren Ranking im Vergleich zu den anderen Arbeiten bewertet. Allerdings lag der Wert des Rankings insgesamt in einem zu niedrigen Bereich, um als zufriedenstellend zu gelten.
   Die Erwartungen an diese Methode wurden nicht erfüllt.

4. Ranking SA und WA:
   Diese Methode erfüllte die Erwartungen vollständig und identifizierte die Zielarbeit korrekt.

5. Klassifikations-Methoden:
   Alle getesteten Methoden waren in der Lage, die Zielarbeit zu identifizieren. Jedoch zeigten die Ergebnisse eine deutliche Streuung bei den übrigen Arbeiten. Es wurden Arbeiten als Treffer klassifiziert, die keine passenden Entitäten oder Relationen aufwiesen.
   Die Erwartungen an diese Methoden wurden nicht erfüllt.

6. KWMS:
   Die Arbeit mit der höchsten Keyword-Dichte entsprach der Zielarbeit. Dieses Ergebnis entsprach den Erwartungen und zeigt, dass die Keyword-Dichte ein hilfreicher Indikator für einen präzisen Treffer ist.

Die Methoden *Ranking SG und WA*, *Ranking SA und WA* sowie *KWMS* erzielten wie erwartet zufriedenstellende Ergebnisse. Im Gegensatz dazu zeigten die Methoden *Ranking SG und DG* und *Ranking SA und DG* sowie die Klassifikationsmethoden unerwartete Streuungen. Insbesondere wurden Arbeiten als relevant klassifiziert, die keine passenden Entitäten oder Relationen aufwiesen. Die Ursachen für diese Streuungen sind unklar.

Tabelle 4.2: Ergebnisse des 2. Experiments basierend auf Datensatz 1

WA	1	2	3	4	5	6	7	8	9	10
					Ranking					
SG und DG	7	1	2	7	1	1	2	4	4	7
SG und WA	10	1	1	2	1	8	1	2	2	1
SA und DG	7	1	1	1	1	1	1	1	1	1
SA und WA	10	1	1	1	1	1	1	1	1	1
					Klassifikation					
SG und DG	1	1	1	1	1	1	1	1	1	1
SG und WA	1	0	1	1	0	1	0	0	0	1
SA und DG	1	0	0	0	0	0	0	1	0	0
SA und WA	1	0	0	0	0	0	0	0	0	0
KWMS in %	0.59	0.42	0.03	0.04	0.0	0.5	0.04	0.03	0.2	0.10
Erwartet	1	0	0	0	0	0	0	0	0	0

Ergebnisse Tabelle 4.2:
Die Suchanfrage zu Experiment 2: *"Which researchers or authors have explored or published work on user interfaces with faceted filtering?"* hatte das Ziel, WA 1 zu identifizieren (siehe Tabelle 4.2). Im Vergleich zu Experiment 1 (siehe Tabelle 4.1) unterscheiden sich diese Experimente durch die Formulierung der Suchanfrage. Beide Suchanfragen verfolgen dasselbe Ziel, mit unterschiedlichen Formulierungen. Im Folgenden werden die Ergebnisse der verschiedenen Ansätze zusammengefasst und bewertet:

1. Ranking SG und DG:
   Diese Methode identifizierte die Zielarbeit, lieferte jedoch zusätzliche Treffer zu Arbeiten, die die Suchanfrage nicht beantworten können. Es kommt zu Treffern, obwohl die betroffenen Dokumentgraphen keine relevanten Entitäten oder Relationen aufweisen. Die Erwartungen an diese Methode wurden nicht erfüllt.

2. Ranking SG und WA:
   Neben der Zielarbeit lieferte diese Methode einen zusätzlichen Treffer zu einer Arbeit, die die Suchanfrage nicht beantworten konnte. Der Grund für diesen zusätzlichen Treffer bleibt unklar.
   Die Erwartungen an diese Methode wurden nicht erfüllt.

3. Ranking SA und DG:
   Diese Methode erfüllt die Erwartungen vollständig und identifiziert die Zielarbeit korrekt.

4. Ranking SA und WA:
   Diese Methode erfüllt die Erwartungen vollständig und identifiziert die Zielarbeit korrekt.

5. Klassifikation SG und DG:
   Diese Methode führte zu einer Klassifikation, bei der jede Arbeit als Treffer gewertet wurde. Dieses Verhalten deutet auf grundlegende Probleme beim Vergleich von Suchgraph und Dokumentgraph hin.
   Die Erwartungen an diese Methode wurden nicht erfüllt.

6. Klassifikation SG und WA:
   Diese Methode zeigte eine starke Streuung. Etwa die Hälfte aller Arbeiten wurde als Treffer klassifiziert, trotz nicht Relevanz.
   Die Erwartungen an diese Methode wurden nicht erfüllt.

7. Klassifikation SA und DG:
   Diese Methode identifizierte die Zielarbeit, zeigte jedoch eine geringe Streuung bei den übrigen Arbeiten, wodurch das Ergebnis verfälscht wurde. Der Grund für diese Streuung ist unklar.
   Die Erwartungen an diese Methode wurden nicht erfüllt.

8. Klassifikation SA und WA:
   Diese Methode erfüllte die Erwartungen vollständig und identifizierte die Zielarbeit korrekt.

9. KWMS:
   Die Arbeit mit der höchsten Keyword-Dichte entsprach der Zielarbeit. Dieses Ergebnis bestätigt, dass die Keyword-Dichte ein hilfreicher Indikator für die Relevanz einer Arbeit ist. Es ist jedoch zu beachten, dass der Unterschied zur zweithöchsten Keyword-Dichte (WA 2) gering war. Dies könnte bei größeren Datensätzen zu Problemen führen.
   Die Methode erfüllte die Erwartungen, jedoch mit der Einschränkung, dass eng beieinanderliegende Werte potenzielle Herausforderungen darstellen.

Die Methoden *Ranking SA und DG*, *Ranking SA und WA*, *Klassifikation SA und DG*, *Klassifikation SA und WA* sowie *KWMS* erzielten erwartungsgemäß zufriedenstellende Ergebnisse. Im Gegensatz dazu zeigten die Methoden *Ran-*

king SG und DG, *Ranking SG und WA* sowie die Klassifikationsmethoden auf Basis von SG unerwartete Streuungen. Insbesondere wurden Arbeiten als relevant klassifiziert, obwohl sie keine passenden Entitäten oder Relationen aufwiesen.

Tabelle 4.3: Ergebnisse des 3. Experiments basierend auf Datensatz 1

WA	1	2	3	4	5	6	7	8	9	10
					Ranking					
SG und DG	10	1	1	1	1	1	1	1	1	1
SG und WA	10	1	1	1	1	1	1	1	1	1
SA und DG	10	1	1	1	1	1	1	1	1	1
SA und WA	10	1	1	1	1	1	1	1	1	1
					Klassifikation					
SG und DG	1	1	1	1	1	1	1	1	1	1
SG und WA	1	1	1	1	0	0	0	0	0	0
SA und DG	1	0	0	0	0	1	0	0	0	0
SA und WA	1	1	0	1	1	0	0	0	1	0
KWMS in %	0.55	0.41	0.01	0.0	0.04	0.0	0.02	0.04	0.03	0.10
Erwartet	1	0	0	0	0	0	0	0	0	0

Ergebnisse Tabelle 4.3:
Die Suchanfrage zu Experiment 3: *"Which authors have published together with Tim Gollub?"* hatte das Ziel, WA 1 zu identifizieren (siehe Tabelle 4.3). Für einen Treffer in diesem Experiment musste die Entität *„Tim Gollub"* und eine Relation mit einer weiteren Entität, die darauf hinweist, dass die Arbeit von ihm veröffentlicht wurde, in der wissenschaftlichen Arbeit oder im Dokumentgraphen enthalten sein.

```
(publication)-[:published_by]->(tim_gollub);
```

Im Folgenden werden die Ergebnisse der verschiedenen Ansätze zusammengefasst und bewertet:

1. Ranking-Methoden:

   Diese Methoden erfüllen die Erwartungen vollständig und identifizieren die Zielarbeit korrekt.

2. Klassifikations-Methoden:

Die getesteten Klassifikationsmethoden konnten die Zielarbeit identifizieren, klassifizierten jedoch auch weitere Arbeiten des Datensatzes als relevant. Dabei enthielt lediglich die Zielarbeit WA 1 die Entität *„Tim Gollub“* und weitere nötige Relationen und Entitäten. Die übrigen Arbeiten enthielten dies nicht und hätten daher nicht als Treffer gewertet werden dürfen. Die fehlerhafte Streuung weist auf mögliche Probleme bei der Klassifikation von Dokumenten und Dokumentgraphen hin.
Die Erwartungen an diese Methoden wurden nicht erfüllt.

3. KWMS:

Diese Methode identifizierte die Zielarbeit korrekt und erfüllte damit die Erwartungen. Allerdings war der Unterschied im Keyword-Matching-Score zwischen WA 1 und WA 2 sehr gering. Dieser geringe Unterschied könnte zu einer Verfälschung der Ergebnisse führen, da der hohe Wert bei WA 2 nicht auf das Keyword *„Tim Gollub“* zurückzuführen war, sondern auf ein anderes, in der Suchanfrage nicht relevantes Keyword.
Die Erwartungen an diese Methode wurden erfüllt, zeigt jedoch eine potenzielle Schwäche bei eng beieinander liegenden Werten.

Die Ranking-Methoden erzielten erwartungsgemäß zufriedenstellende Ergebnisse, indem sie die Zielarbeit korrekt identifizierten. Im Gegensatz dazu zeigten die Klassifikations-Methoden unerwartete Streuungen. Insbesondere wurden Arbeiten als relevant klassifiziert, obwohl sie keine passenden Entitäten oder Relationen aufwiesen. Diese Streuungen beeinträchtigen die Präzision der Klassifikation und erfordern eine weitergehende Analyse, um die Ursachen der fehlerhaften Ergebnisse zu klären. Die Methode *KWMS* erfüllte die Erwartungen, wies jedoch ebenfalls potenzielle Schwächen aufgrund von geringen Unterschieden in den Ergebnissen auf.

Tabelle 4.4: Ergebnisse des 4. Experiments basierend auf Datensatz 1

WA	1	2	3	4	5	6	7	8	9	10
					Ranking					
SG und DG	6	1	2	1	1	1	2	2	1	1
SG und WA	8	1	1	1	1	2	1	1	1	1
SA und DG	7	1	1	1	1	1	1	1	1	1
SA und WA	7	1	1	1	1	1	1	1	1	1
					Klassifikation					
SG und DG	1	0	0	0	0	0	1	0	0	0
SG und WA	1	0	1	1	0	0	0	0	0	1
SA und DG	1	0	0	0	0	0	0	0	0	0
SA und WA	1	0	0	0	0	0	0	0	0	0
KWMS in %	6.45	0.14	0.46	0.39	0.32	0.77	0.19	0.11	0.28	0.28
Erwartet	1	0	0	0	0	0	0	0	0	0

Ergebnisse Tabelle 4.4:
Die Suchanfrage zu Experiment 4: *"Show me papers that present methods for computing relationships between facets or for providing a facet-based user interface, with a focus on interactions between exploratory and filtering search."* hatte das Ziel, WA 1 zu identifizieren (siehe Tabelle 4.3). Die Suchanfrage wurde speziell und umfassend auf WA 1 zugeschnitten, weshalb keine fehlerhaften Treffer zu erwarten waren. Im Folgenden werden die Ergebnisse der verschiedenen Ansätze zusammengefasst und bewertet:

1. Ranking-Methoden:
   Diese Methoden erfüllen die Erwartungen vollständig und identifizieren die Zielarbeit korrekt.

2. Klassifikation SG und DG und SA und WA:
   Diese Klassifikationsmethoden führten zu fehlerhaften Treffern, die das Ergebnis verfälschten. Die identifizierten Arbeiten enthielten keine relevanten Informationen, die die Suchanfrage hätten beantworten können.
   Die Erwartungen an diese Methoden wurden nicht erfüllt.

3. Klassifikation SA und DG und LLM SA und WA:
   Diese Methoden erfüllen die Erwartungen vollständig und identifizieren die Zielarbeit korrekt.

4. KWMS:

Diese Methode erfüllt die Erwartungen vollständig und identifizierte die Zielarbeit korrekt.

Die Ranking-Methoden, *Klassifikation SA und DG*, *Klassifikation SA und WA* sowie *KWMS* erfüllten die Erwartungen vollständig und identifizierten die Zielarbeit korrekt. Im Gegensatz dazu zeigten die Methoden *Klassifikation SG und DG* und *Klassifikation SG und WA* Schwächen, indem sie fehlerhafte Treffer lieferten, obwohl die identifizierten Arbeiten keine relevanten Informationen enthielten.

Tabelle 4.5: Ergebnisse des 5. Experiments basierend auf Datensatz 2

WA	1	2	3	4	5	6	7	8	9	10
					Ranking					
SG und DG	7	7	2	7	7	1	7	7	7	7
SG und WA	10	10	10	10	10	10	10	10	10	10
SA und DG	10	10	7	7	10	4	7	7	7	7
SA und WA	7	7	7	7	7	7	7	7	7	7
					Klassifikation					
SG und DG	1	1	1	1	1	1	1	1	1	1
SG und WA	1	1	1	1	1	1	1	1	1	1
SA und DG	1	1	1	1	1	1	1	1	1	1
SA und WA	1	1	1	1	1	1	1	1	1	1
KWMS in %	4.55	4.55	4.55	4.55	4.55	4.55	4.55	4.55	4.55	4.55
Erwartet	1	1	1	1	1	1	1	1	1	1

Ergebnisse Tabelle 4.5:
Die Suchanfrage in Experiment 5 entspricht der Suchanfrage aus Experiment 1. Ziel dieses Experiments war es, die Konsistenz der Graphgenerierung durch das LLM sowie die Ausführung der angewandten Methoden zu überprüfen (siehe Tabelle 4.5). Konkret sollte das Ranking Ergebnisse mit Werten erzeugen, die eng beieinander liegen, während die Klassifikation ausschließlich identische Werte liefern sollte. Im Folgenden werden die Ergebnisse der verschiedenen Ansätze zusammengefasst und bewertet:

1. Ranking SG und DG und Ranking SA und DG:
   Diese Methoden zeigten inkonsistente Ergebnisse. Es wurden teilweise starke Abweichungen beobachtet, obwohl die erforderlichen Entitäten in den Dokumentgraphen enthalten waren.
   Die Erwartungen an diese Methoden wurden nicht erfüllt.

2. Ranking SG und WA und Ranking SQ und WA:
   Die Erwartungen an diese Methoden wurden erfüllt.

3. Klassifikation aller Methoden:
   Die Erwartungen an diese Methoden wurden erfüllt.

Während die Ranking-Methoden Schwierigkeiten bei der konsistenten Bewertung von Dokumentgraphen aufwiesen, lieferten die Klassifikationsmethoden sowohl bei Dokumentgraphen als auch bei wissenschaftlichen Arbeiten präzise und konsistente Ergebnisse. Die Inkonsistenzen im Ranking deuten darauf hin, dass das LLM Schwierigkeiten hat, Knowledge-Graphen zu verstehen. Im Gegensatz dazu erzielten die Klassifikationsmethoden erwartungsgemäße Resultate und bestätigten ihre Robustheit im Rahmen dieses Experiments. Besonders die Ergebnisse der Methoden ohne Knowledge-Graphen verdeutlichen, dass das LLM beim Ranking oder der Klassifikation auf festen Prinzipien beruht, die im Ranking berücksichtigt werden. Dies fiel während der Durchführung der Experimente besonders auf: Auf den gleichen Input folgt stets die gleiche Bewertung.

Tabelle 4.6: Ergebnisse des 6. Experiments basierend auf Datensatz 3

WA	1	2	3	4	5	6	7	8	9	10
					Ranking					
SG und DG	2	10	2	1	8	7	10	10	1	2
SG und WA	10	10	8	8	10	1	10	10	1	10
SA und DG	1	10	1	1	1	1	10	1	1	1
SA und WA	2	8	10	2	1	1	10	2	1	1
					Klassifikation					
SG und DG	0	0	0	0	0	0	0	0	1	0
SG und WA	1	1	1	0	0	0	1	1	0	1
SA und DG	0	1	0	0	0	0	1	0	0	0
SA und WA	0	1	0	0	0	0	1	1	0	0
KWMS in %	0.15	1.60	0.09	0.06	0.95	0.25	0.66	0.25	0.09	0.20
Erwartet	0	1	0	0	0	0	1	1	0	0

Ergebnisse: Tabelle 4.6:
Die Suchanfrage zu Experiment 6: *"Scientific papers on knowledge graphs published in 2021 or later."* hatte das Ziel, WA 2, 7 und 8 zu identifizieren (siehe Tabelle 4.6). Als Kriterium für einen Treffer genügte es, dass das Veröffentlichungsjahr und eine Relation mit einer weiteren Entität, die darauf hinweisen, dass die Arbeit 2021 oder später veröffentlicht wurde, in der wissenschaftlichen Arbeit oder im Dokumentgraphen enthalten sind.

```
(publication)-[:published]->(year);
```

Im Folgenden werden die Ergebnisse der verschiedenen Ansätze zusammengefasst und bewertet:

1. Klassifikation SA und WA:
   Die Erwartungen an diese Methode wurden erfüllt.

2. Restliche Ranking- und Klassifikations-Methoden:
   Keine der übrigen Methoden konnte ein Ergebnis liefern, das den Erwartungen vollständig entsprach. Die Abweichungen haben mehrere Gründe:
   1. Fehlende Treffer trotz vorhandener Entität.
   2. Falsche Treffer trotz fehlender Entität.
   3. Fehlende Treffer trotz vorhandener relevanter Informationen.
   4. Falsche Treffer trotz vorhandener irrelevanter Informationen.

   Die Erwartungen an diese Methoden wurden nicht erfüllt.

3. KWMS:
   Die Methode basierend auf der Keyword-Dichte lieferte kein aussagekräftiges Ergebnis. Die höchsten Werte der Keyword-Dichte stimmten nicht mit den erwarteten Treffern WA 2, 7 und 8 überein. Die Ergebnisse wurden durch die Häufigkeit irrelevanter Keywords in der Suchanfrage verfälscht.
   Die Erwartungen an diese Methode wurden nicht erfüllt.

Von den getesteten Ansätzen erfüllte einzig die Methode *Klassifikation SA und WA* die Erwartungen vollständig. Die restlichen Methoden, sowohl Ranking- als auch Klassifikationsansätze, zeigten deutliche Schwächen und lieferten keine präzisen Ergebnisse. *KWMS* zeigte sich aufgrund der Häufigkeit irrelevanter Keywords in der Suchanfrage als nicht präzise und führte zu einer Verfälschung der Ergebnisse.

Tabelle 4.7: Ergebnisse des 7. Experiments basierend auf Datensatz 3

WA	1	2	3	4	5	6	7	8	9	10
					Ranking					
SG und DG	6	1	2	7	1	1	7	1	1	7
SG und WA	10	2	7	8	2	1	8	1	1	1
SA und DG	2	1	1	7	1	1	1	1	1	1
SA und WA	2	1	2	2	1	1	2	1	1	2
					Klassifikation					
SG und DG	1	1	1	1	1	1	1	1	1	1
SG und WA	1	1	1	1	0	0	0	1	0	1
SA und DG	1	0	0	1	0	0	0	0	0	0
SA und WA	0	1	0	0	0	0	1	0	0	0
KWMS in %	0.15	1.15	0.09	0.6	0.95	0.25	0.43	0.08	0.07	0.21
Erwartet	1	1	1	1	0	1	1	0	1	1

Ergebnisse Tabelle 4.7:
Die Suchanfrage zu Experiment 7: *"Ranking Methods for evaluating knowledge graphs."* hatte das Ziel, die Arbeiten WA 5 und 8 nicht zu identifizieren (siehe Tabelle 4.7). Diese Suchanfrage stellt eine komplexere Herausforderung dar, da sie den Kontext der Arbeiten verstehen und interpretieren muss, um die Relevanz korrekt zu bewerten.
Keine der getesteten Methoden konnte ein Ergebnis liefern, das vollständig den Erwartungen entsprach. Obwohl einige Methoden in der Lage waren, die irrelevanten Arbeiten korrekt zu klassifizieren, wiesen sie dennoch entscheidende Schwächen auf:

1. Falsche Negative:
   Relevante Arbeiten wurden fälschlicherweise als nicht relevant eingestuft.

2. Verfälschungen:
   Die Ergebnisse wurden durch die fehlerhafte Klassifikation relevanter Arbeiten beeinträchtigt, was die Präzision der Methoden in diesem Szenario erheblich einschränkte.

Die mangelhafte Leistung der Ranking- und Klassifikationsmethoden bei der korrekten Beantwortung der gestellten Suchanfrage weist auf Schwierigkeiten bei der Interpretation und dem Verständnis komplexer, kontextbezogener Anforderungen hin. Auch die Ergebnisse von *KWMS* entsprachen nicht den Erwartungen. Die Zählungen dieser Methode waren wenig aussagekräftig.

Kapitel 5

Fazit

Diese Arbeit hat gezeigt, dass der Einsatz von Knowledge-Graphen in Kombination mit LLMs für die Verbesserung von Information Retrieval sowohl Chancen als auch Herausforderungen mit sich bringt. Die Ergebnisse der Experimente zeigen, dass reine Knowledge-Graph-Ansätze in Bezug auf Präzision Knowledge-Graph-freien oder hybriden Ansätzen[1] unterlegen sind.

Semantische Modelle können theoretisch komplexe Zusammenhänge erfassen, haben aber praktische Einschränkungen beim Umgang mit Graphstrukturen. Anhand folgenden Beispiels soll dies nochmal verdeutlicht werden:

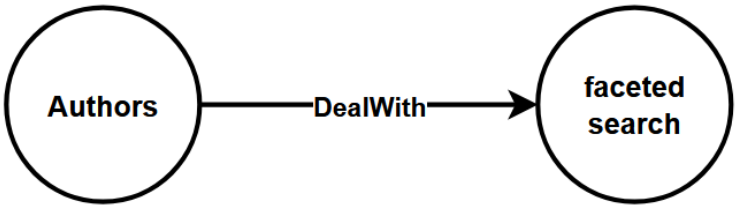


Abbildung 5.1: Suchgraph zu Experiment 5: „Which authors have dealt with faceted search?“

Betrachtet man die Dokumentgraphen des 5. Experiments (siehe Abbildung 5.2 und Abbildung 5.3), die aus der gleichen wissenschaftlichen Arbeit erstellt wurden – da der Datensatz aus identischen Arbeiten bestand – sowie den zugehörigen Suchgraphen (siehe Abbildung 5.1) mit dem Wissen, dass diese Graphen die wissenschaftlichen Arbeiten und die Suchanfrage präzise repräsentieren, so zeigt sich, anhand der grün markierten Bereiche, dass die

[1] Methoden, bei denen das LLM sowohl einen Knowledge-Graph als auch einen Text als Eingabe erhält (*SA und DG* und *SG und WA*).

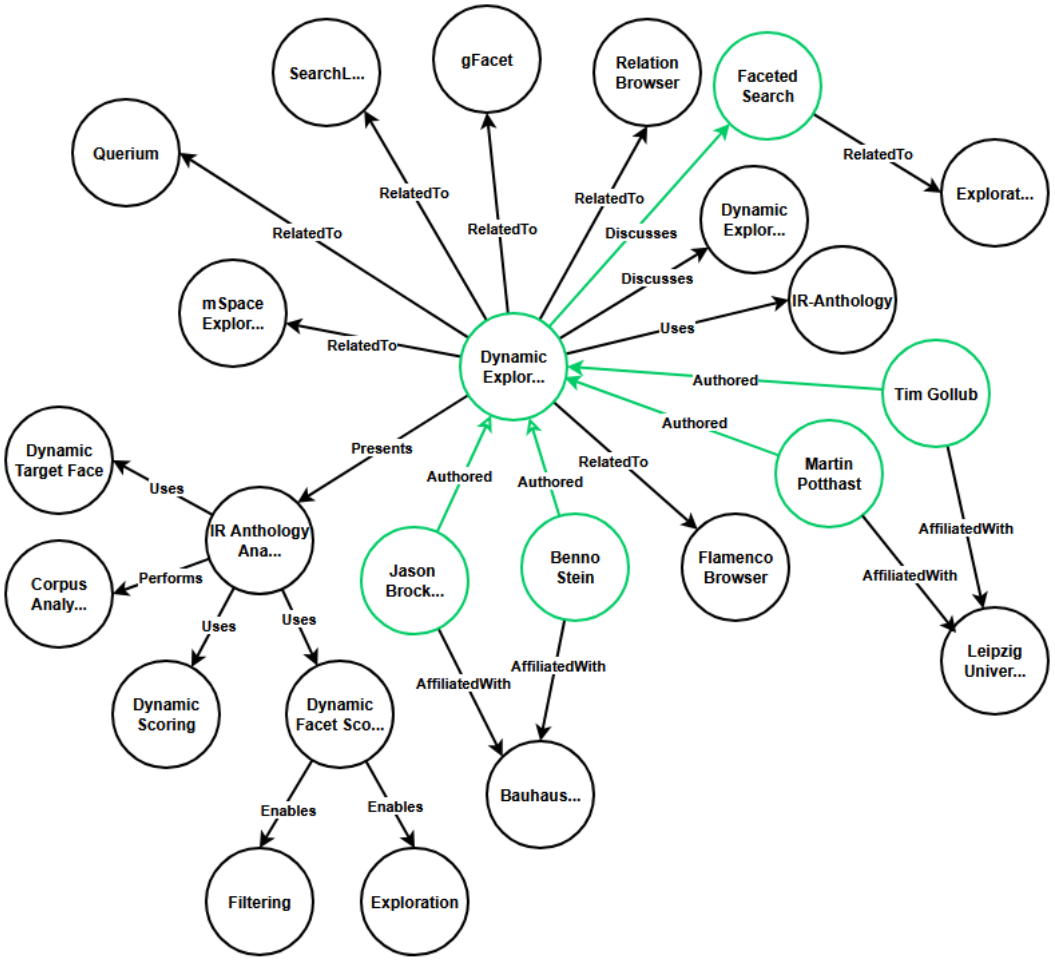


Abbildung 5.2: Dokumentgraph, zu WA 1 Datensatz 2.

Dokumentgraphen die Antwort auf den Suchgraphen enthalten. Die relevanten Entitäten und Relationen sind vorhanden, um die Suchanfrage korrekt zu beantworten und entsprechend hoch gerankt zu werden. Tatsächlich jedoch wird PDF 1 hoch eingerankt, während PDF 6 ein sehr niedriges Ranking erhält (siehe Tabelle 4.5). Trotz der hohen Qualität der Knowledge-Graphen wiesen alle im Rahmen dieser Arbeit verwendeten LLMs solche Inkonsistenzen auf. Obwohl beide Graphen die notwendigen Entitäten und Relationen enthalten und diese sehr ähnlich zueinander sind, führten die LLMs zu unterschiedlichen Ergebnissen, obwohl die Antwort auf die gestellte Frage offensichtlich ist. Dieses und ähnliche Probleme traten wiederholt in der Experiment-Auswertung auf.

Auf Grundlage der in Kapitel 4 durchgeführten Experimente lassen sich die Ergebnisse tabellarisch wie folgt zusammenfassen (siehe Tabelle 5.1).

Abbildung 5.3: Dokumentgraph, zu WA 6 Datensatz 2.

Tabelle 5.1: Zusammenfassung der Ergebnisse

Tabelle	4.1)	4.2)	4.3)	4.4)	4.6)	4.7)
			Ranking			
SG und DG	0	0	1	1	0	0
SG und WA	1	0	1	1	0	0
SA und DG	0	1	1	1	0	0
SA und WA	1	1	1	1	0	0
			Klassifikation			
SG und DG	0	0	0	0	0	0
SG und WA	0	0	0	0	0	0
SA und DG	0	0	0	1	0	0
SA und WA	0	1	0	1	1	0
KWMS in %	1	1	1	1	0	0

Die Ansätze bestehend aus *Suchanfrage und wissenschaftlicher Arbeit* erzielten in beiden Experimentkategorien die besten Ergebnisse, während der *Keyword-Matching Score* meist zufriedenstellende Resultate zeigte. Im Gegensatz dazu lieferten graphbasierte Methoden meist schwache Resultate. Dies deutet darauf hin, dass LLMs Schwierigkeiten haben, Antworten aus einem Graphen in einem anderen Graphen zu extrahieren. Die Ursachen dafür lassen sich in folgende Fehlerkategorien unterteilen:

1. Falsche Treffer:

- Arbeiten wurden als Treffer gewertet, obwohl die erforderlichen Entitäten oder Relationen im Dokumentgraph fehlten.

2. Fehlende Treffer:

- Arbeiten mit den passenden Entitäten oder Relationen wurden nicht als Treffer erkannt.

3. Unstimmige Rankings:

- Zu niedrige Rankings trotz passender Entitäten.
- Zu hohe Rankings, obwohl die entsprechenden Entitäten fehlten.

Diese Inkonsistenzen weisen auf grundlegende Schwächen der Knowledge-Graph Methoden hin. Ein möglicher Grund dafür ist, dass LLMs primär für sequenzielle Textverarbeitung optimiert sind, während der direkte Vergleich komplexer Graphstrukturen eine explizite semantische Interpretation erfordert.

Zusammenfassend zeigt diese Arbeit, dass LLMs die theoretischen Vorteile von Knowledge-Graphen bislang nicht ausschöpfen können und daher keine Verbesserung des Information Retrievals durch deren Einsatz erzielt wird. Hybride Ansätze, bei denen eine Komponente aus Klartext[2] besteht, liefern bessere Ergebnisse als reine Knowledge-Graph-Ansätze. Die beste Performance erreicht das LLM jedoch, wenn sowohl die Suchanfrage als auch die wissenschaftliche Arbeit in Klartext vorliegen. Dennoch schneidet auch diese Methode nur so gut ab wie die in dieser Arbeit implementierte *KWMS*-Methode, die hier in einer einfachen Form umgesetzt wurde und nicht das volle Potenzial von Keyword-Matching ausschöpft. Trotz der teilweise hohen Trefferquote treten erhebliche Probleme bei Suchanfragen auf, die sich beispielsweise auf Autoren oder Veröffentlichungsjahre beziehen. Dies unterstreicht die Notwendigkeit weiterer Forschung zur Entwicklung leistungsfähigerer Systeme.

[2]Text der Suchanfrage oder Text der wissenschaftlichen Arbeit.

Um die Präzision von IR-Systemen auf Basis von Knowledge-Graphen zu verbessern, sollte sich zukünftige Forschung auf die Entwicklung neuer Algorithmen konzentrieren, die Such- als auch Dokumentgraphen besser erfassen.

Es sollten Modelle geschaffen werden, die Knowledge-Graphen nicht nur strukturell analysieren, sondern auch deren semantische Beziehungen besser interpretieren. Dies könnte erreicht werden durch die Entwicklung neuer Graph-Matching-Algorithmen, die speziell darauf ausgelegt sind, inhaltliche Übereinstimmungen zwischen Such- und Dokumentgraphen präziser zu erkennen. Algorithmen, die nicht nur exakte Übereinstimmungen von Knoten und Kanten prüfen, sondern auch semantische Ähnlichkeiten berücksichtigen. Beispielsweise könnten Verfahren entwickelt werden, die alternative Bezeichnungen oder Synonyme für Knoten erkennen und gewichten. Des Weiteren könnten Graph-Matching Techniken genutzt werden, die nicht nur direkte Verbindungen zwischen Knoten betrachten, sondern auch weiter entfernte Beziehungen einbeziehen, um tiefere inhaltliche Zusammenhänge zu verstehen.

Ein weiterer vielversprechender Ansatz ist die gezielte Optimierung von LLMs auf das bessere Verstehen und Interpretieren von Knowledge-Graphen. LLMs könnten dafür mit Knowledge-Graphen trainiert werden. Dadurch könnten sie dazu befähigt werden, die komplexen Zusammenhänge in Knowledge-Graphen effektiver zu erfassen und darauf basierende Antworten präziser zu generieren. Dafür könnten Trainingsmethoden entwickelt werden, bei denen LLMs nicht nur auf großen Textkorpora, sondern auch gezielt auf strukturierten Graph-Daten trainiert werden. Zusätzlich könnten LLMs während des Trainings explizit darauf optimiert werden, logische Schlussfolgerungen aus Graphstrukturen zu ziehen, anstatt sich nur auf die probabilistische Vorhersage von Wörtern zu verlassen.

Durch diese Optimierungen könnten LLMs nicht nur präzisere Antworten auf Basis von Knowledge-Graphen liefern, sondern auch gezielt zur Verbesserung des Information Retrievals eingesetzt werden, indem sie relevante Informationen aus Graphstrukturen effizient extrahieren und sinnvoll verknüpfen. Diese Erkenntnisse bilden eine solide Grundlage für die nächsten Schritte in diese Richtung.

Literaturverzeichnis

Sören Auer, Allard Oelen, Muhammad Haris, Markus Stocker, Jennifer D'Souza, Kheir Eddine Farfar, Lars Vogt, Manuel Prinz, Vitalis Wiens, and Mohamad Yaser Jaradeh. Improving access to scientific literature with knowledge graphs. *IEEE transactions on neural networks and learning systems*, 44:516–529, 2020. doi: 10.1515/bfp-2020-2042. URL https://www.degruyter.com/document/doi/10.1515/bfp-2020-2042/html#:~:text=Improving%20Access%20to%20Scientific%20Literature%20with%20Knowledge%20Graphs,...%203%203%20Knowledge%20Graph%20Use%20Cases%20.

Jorge Gabín, Eduardo M. Ares, and Javier Parapar. Keyword embeddings for query suggestion. 2023. doi: 10.48550/arXiv.2301.08006. URL https://arxiv.org/abs/2301.08006.

Tim Gollub, Jason Brockmeyer, Benno Stein, and Martin Potthast. Dynamic exploratory search for the information retrieval anthology. 13982:242–247, 2023. doi: 10.1007/978-3-031-28241-6_21. URL https://link.springer.com/chapter/10.1007/978-3-031-28241-6_21.

Kailash Hambarde and Hugo Proença. Information retrieval: Recent advances and beyond. *IEEE Access*, 2023. URL https://arxiv.org/pdf/2301.08801.

Yupeng Hou, Junjie Zhang, Zihan Lin, Hongyu Lu, Ruobing Xie, Julian McAuley, and Xin Wayne. Large language models are zero-shot rankers for recommender systems. 2023. doi: 10.48550/arXiv.2305.08845. URL https://arxiv.org/abs/2305.08845.

Martin Potthast, Sebastian Günther, Janek Bevendorff, Jan Philipp Bittner, Alexander Bondarenko, Maik Fröbe, Christian Kahmann, Andreas Niekler, Michael Völske, Benno Stein, and Matthias Hagen. The information retrieval anthology. 2021. doi: 10.1145/3404835.3462798. URL https://dl.acm.org/doi/10.1145/3404835.3462798.

Pranab Sahoo, Ayush Kumar Singh, Sriparna Saha, Vinija Jain, Samrat Mondal, and Aman Chadha. A systematic survey of prompt engineering in large language models: Techniques and applications. 2024. doi: 10.48550/arXiv.2402.07927. URL https://arxiv.org/abs/2402.07927.

Bahareh Sarrafzadeh, Adam Roegiest, and Edward Lank. Hierarchical knowledge graphs: A novel information representation for exploratory search tasks. 2020. doi: 10.48550/arXiv.2005.01716. URL https://arxiv.org/abs/2005.01716.

Ji Shaoxiong, Pan Shirui, Erik Cambria, Pekka Marttinen, and Yu S. Philip. A survey on knowledge graphs: Representation, acquisition, and applications. *IEEE transactions on neural networks and learning systems*, 33:494–514, 2022. doi: 10.1109/TNNLS.2021.3070843. URL https://arxiv.org/abs/2002.00388.

Mingyang Song, Mao Zheng, and Xuan Luo. Counting-stars: A multi-evidence, position-aware, and scalable benchmark for evaluating long-context large language models. 2023. doi: 10.48550/arXiv.2403.11802. URL https://arxiv.org/abs/2403.11802.

Markus Stocker, Allard Oelen, Mohamad Yaser Jaradeh, Muhammad Haris, Omar Arab Oghli, Golsa Heidari, Hassan Hussein, Anna-Lena Lorenz, Salomon Kabenamualu, Kheir Eddine Farfar, Manuel Prinz, Oliver Karras, Jenniger D'Souza, Lars Vogt, and Sören Auer. Fair scientific information with the open research knowledge graph. *IEEE transactions on neural networks and learning systems*, 1:19–21, 2023. doi: 10.3233/FC-221513. URL https://content.iospress.com/articles/fair-connect/fc221513.

Milena Trajanoska, Riste Stojanov, and Dimitar Trajanov. Enhancing knowledge graph construction using large language models. 2023. doi: 10.48550/arXiv.2305.04676. URL https://arxiv.org/abs/2305.04676.

Yutao Zhu, Huaying Yuan, Shuting Wang, Jiongnan Liu, Wenhan Liu, Chenlong Deng, Haonan Chen, Zheng Liu, Zhicheng Dou, and Ji-Rong Wen. Large language models for information retrieval: A survey. 2023. doi: 10.48550/arXiv.2308.07107. URL https://arxiv.org/abs/2308.07107.

Anhang A

Weitere Metriken und Details

1. SS Entities (Similarity Score der Entitäten)

Dieser Score misst die Übereinstimmung der Entitäten zwischen einem Suchgraphen und einem Dokumentgraphen. Der Wert liegt im Bereich von 0.0 bis 1.0, wobei 1.0 für eine vollständige Übereinstimmung der Entitäten steht.

Diese Metrik könnte als effizienter und ressourcenschonender Indikator genutzt werden, in Kontexten, in denen der Dokumentgraph die gleichen Entitäten wie der Suchgraph enthält. Durch den Abgleich von Entitäten bietet dies eine Möglichkeit des Pre-Filterings, das die Notwendigkeit einer aufwendigen Verarbeitung durch ein LLM reduzieren könnte. Dadurch ließe sich die Relevanz von Dokumentgraphen in frühen Phasen der Analyse beurteilen.

$$\textbf{SS Entities} = \frac{|N_1 \cap N_2|}{|N_1 \cup N_2|}$$

N_1 = Menge der Knoten im Suchgraph
N_2 = Menge der Knoten im Dokumentgraph
Falls $|N_1 \cup N_2| = 0$, wird die Ähnlichkeit auf 0 gesetzt.

2. SS Relations (Similarity Score der Relationen)

Der Score bewertet die Übereinstimmung der Relationen zwischen einem Suchgraphen und einem Dokumentgraphen. Auch hier liegt der Wert zwischen 0.0 und 1.0, wobei 1.0 für eine vollständige Übereinstimmung der Relationen steht.

Diese Metrik kann nützlich sein, um zu bestimmen, ob ein Dokumentgraph die gleichen Relationen wie der Suchgraph enthält. Relationen bieten Hinweise darauf, ob ein Dokumentgraph die gestellte Suchanfrage thematisiert und möglicherweise beantworten kann. Darüber hinaus könnte diese Metrik in verschie-

denen Anwendungsbereichen wertvolle Unterstützung leisten, wie beispielsweise der Analyse der Struktur und Logik von Relationen in Knowledge-Graph, etwa bei der Validierung von Datenmodellen oder der Identifikation von Dokumenten mit spezifischen semantischen Beziehungen, wie beispielsweise „Autor schreibt Buch“.

$$\textbf{SS Relations} = \frac{|E_1 \cap E_2|}{|E_1 \cup E_2|}$$

E_1 = Menge der Relationen im Suchgraph
E_2 = Menge der Relationen im Dokumentgraph
Falls $|E_1 \cup E_2| = 0$, wird die Ähnlichkeit auf 0 gesetzt.

3. Graph Similarity

Die Graph Similarity kombiniert die Werte von SS Entities und SS Relations und liefert einen absoluten Score zwischen 0.0 und 1.0, der die Gesamtsimilarität zwischen Suchgraph und Dokumentgraph angibt. Ein Wert von 1.0 bedeutet eine vollständige Übereinstimmung.

Diese Metrik könnte genutzt werden, um die Ähnlichkeit zwischen zwei Graphen umfassend zu bewerten. Dadurch ermöglicht sie den Vergleich der Gesamtaussage von Texten oder Dokumenten anhand ihrer zugrunde liegenden Knowledge-Graphen und unterstützt die Ermittlung relevanter Kontexte.

Diese Metriken wurden im Rahmen der Experimente ebenfalls berücksichtigt. Allerdings lieferten sie keine aussagekräftigen Ergebnisse, weshalb sie in der Arbeit nicht weiter thematisiert wurden. Gelegentlich traten Übereinstimmungen auf, die jedoch meist in einem sehr niedrigen Bereich von 0.01 bis 0.04 lagen. Diese Werte entsprachen nicht den Erwartungen und zeigten kein konsistentes Muster, das auf einen Treffer oder Nicht-Treffer hätte hindeuten können.

$$\textbf{Graph Similarity} = \frac{\textbf{SS Entities} + \textbf{SS Relations}}{2}$$

4. KWMS in % Berechnung

$$\textbf{KWMS} = \left(\frac{\sum_{i=1}^{n} K_i}{W_{\text{PDF}}} \right) \times 100$$

K_i = Anzahl der Vorkommen des i-ten Keywords im PDF
W_{PDF} = Gesamtzahl der Wörter im PDF

Anhang B

Alternative Ansätze zur Graphgenerierung

Neben den in den Experimenten eingesetzten Methoden zur Graphgenerierung wurden weitere Ansätze untersucht, um die Präzision der Graphgenerierung zu verbessern. Die Ergebnisse dieser alternativen Ansätze werden im Folgenden dargestellt.

1. Freie Generierung der Graphen

Bei diesem Ansatz wurden sowohl Such- als auch Dokumentgraphen ohne Einschränkungen erstellt. Es gab keine festen Vorgaben, in welche Richtung sich die Graphen entwickeln sollten. Die Generierung basierte ausschließlich auf der Suchanfrage oder der wissenschaftlichen Arbeit.

Obwohl der Suchgraph die Suchanfrage präzise widerspiegelte, wiesen die generierten Dokumentgraphen häufig Ergebnisse auf, die nicht mit den Anforderungen des Suchgraphs übereinstimmten. Statt die im Suchgraphen thematisierten Aspekte aufzugreifen, fokussierten sich die Dokumentgraphen auf andere Inhalte der wissenschaftlichen Arbeit. Dies führte dazu, dass die spezifischen Fragen des Suchgraphs unbeantwortet blieben, da die Dokumentgraphen ihre Schwerpunkte auf irrelevante oder nicht gefragte Bereiche legten. Infolgedessen konnten einige Dokumente, die eigentlich als Treffer hätten erkannt oder hoch eingestuft werden müssen, nicht korrekt identifiziert werden, da die dafür erforderlichen Informationen in den Dokumentgraphen fehlten.

2. Binden der Dokumentgraphen an die Relationen der Suchanfrage

In diesem Ansatz wurden die Relationen der Suchanfrage als feste Grundlage für die Generierung der Dokumentgraphen verwendet. Dadurch konnte der Similarity Score (siehe Anhang A) zwischen Such- und Dokumentgraphen erhöht werden.

Allerdings zeigten die Ergebnisse keine klaren Muster, die für eine zuverlässige Bewertung hätten herangezogen werden können. Zudem wurde den Dokumentgraphen durch diese Einschränkung ihre Flexibilität genommen. Häufig versuchten die Dokumentgraphen, Informationen so umzuformulieren, dass sie in die vorgegebenen Relationen passten. Dies führte zu inkonsistenten und verfälschten Ergebnissen.

3. Binden der Dokumentgraphen mit zusätzlichen Relationen

Als Erweiterung des vorherigen Ansatzes (siehe B.2) wurden neben den festen Relationen der Suchanfrage zusätzliche Relationen erlaubt. Dennoch zeigte sich, dass dieser Ansatz ähnliche Einschränkungen aufwies wie die reine Bindung an feste Relationen. Es wurden keine zusätzlichen Vorteile erzielt, und die Ergebnisse blieben unverändert.

4. Gesamter Suchgraph zur DG-Generierung

Bei diesem Ansatz wurde der gesamte Suchgraph an das LLM übergeben, um dessen Struktur direkt zu übernehmen. Dies führte jedoch dazu, dass sowohl die Entitäten als auch die Relationen des Suchgraphs in nahezu jeden generierten Dokumentgraphen integriert wurden. Die resultierenden Dokumentgraphen ähnelten dem Suchgraph stark, da dessen Entitäten und Relationen fast vollständig übernommen wurden. Dies verfälschte die Ergebnisse erheblich, da die Dokumentgraphen kaum eigenständige Informationen aus der wissenschaftlichen Arbeit repräsentierten.Die besten Ergebnisse wurden mit der Methode erzielt, bei der die Dokumentgraphen basierend auf der Suchanfrage generiert wurden, wie in den Experimenten (siehe Kapitel 4) beschrieben. Dabei wurden keine zusätzlichen Einschränkungen wie feste Relationen oder die direkte Übernahme des Suchgraphs angewendet. Dieser Ansatz bot eine präzisere und unverfälschte Darstellung der relevanten Inhalte, ohne die Flexibilität oder Aussagekraft der Dokumentgraphen zu beeinträchtigen.

Anhang C

Implementierung und Probleme

1. Bibliotheken

Für die Durchführung der Experimente wurde die Programmiersprache Python verwendet, da sie eine Vielzahl nützlicher Bibliotheken bietet, die für die Anforderungen dieser Arbeit optimal geeignet sind. Die im Folgenden beschriebenen Bibliotheken kamen zum Einsatz.

Neo4j[1]
Zur Verwaltung der Knowledge-Graphen wurde die Neo4j-Datenbank genutzt. Diese Plattform ermöglicht das Speichern, Laden und Vergleichen von Graphen. Durch die Unterstützung von Cypher, einer Graph-Abfragesprache, bietet Neo4j robuste Werkzeuge zur Darstellung und Analyse von Graphstrukturen. Diese Funktionalitäten spielen eine wichtige Rolle bei der Evaluation der entwickelten Methoden in dieser Arbeit.

Google Generative AI[2] **und Google API Core**[3]
Die Bibliotheken `google.generativeai` und `google.api_core.exceptions` ermöglichen den Zugriff auf die generativen KI-Modelle `Gemini 1.5-pro` und `Gemini 1.5-flash` über die Google API. Diese Modelle wurden zur Verarbeitung von Texteingaben und zur Generierung von Knowledge-Graphen oder Antworten auf Suchanfragen eingesetzt.

Die Wahl von `Gemini 1.5-pro` erfolgte aufgrund seiner hohen Leistungsfähigkeit, insbesondere durch die Unterstützung von bis zu 1 Million Tokens, was die Verarbeitung langer Texte ermöglicht. Die Modelle wurden mit einer

[1] `https://neo4j.com/`
[2] `https://cloud.google.com/ai/generative-ai?hl=de`
[3] https://pypi.org/project/google-api-core/

Temperatur von 0.1 konfiguriert, um präzise und fokussierte Ergebnisse zu erzielen, wobei die kreative Variabilität reduziert wurde. Diese Konfiguration war entscheidend für die Anforderungen an Genauigkeit und Relevanz der Experimente. Andere Modelle wie `Chat-GPT-4o` und `Llama 3.1 8B` wurden auch berücksichtigt, konnten jedoch nicht die gleiche Leistungsfähigkeit und Genauigkeit bieten. Aus diesem Grund wurden in den finalen Analysen ausschließlich die Ergebnisse der Experimente mit `Gemini 1.5-pro` ausgewertet.

PyPDF2[4]
Zum Extrahieren von Text aus wissenschaftlichen Arbeiten, die typischerweise im PDF-Format vorliegen, wurde die Bibliothek `PyPDF2` verwendet. Sie bietet eine einfache Möglichkeit, Text aus den Dokumenten zu extrahieren, um diesen für die Erstellung von Knowledge-Graphen oder das Keyword-Matching weiterzuverarbeiten.

FlashText[5]
Die Bibliothek `flashtext.KeywordProcessor` ermöglichte eine schnelle und ressourcenschonende Extraktion von Keywords aus Texten. Im Vergleich zu regulären Ausdrücken zeichnet sich FlashText durch eine deutlich höhere Effizienz aus, insbesondere bei der Analyse großer Textmengen.

Rake (`rake_nltk`)[6]
Rapid Automatic Keyword Extraction wurde zur Extraktion von Schlüsselwörtern aus gestellten Suchanfragen verwendet. Diese Methode benötigt keine Trainingsdaten und ist ein etabliertes Werkzeug zur Analyse unstrukturierter Texte.

nltk[7]
Die Bibliothek nltk (Natural Language Toolkit) unterstützte die Textverarbeitung durch Funktionen wie Tokenisierung, Stopword-Entfernung und weitere NLP-Operationen. Als Standardbibliothek im Bereich Natural Language Processing sorgte sie für eine qualitativ hochwertige und effiziente Textanalyse.

[4] `https://pypi.org/project/PyPDF2/`
[5] `https://github.com/vi3k6i5/flashtext`
[6] `https://pypi.org/project/rake-nltk/`
[7] `https://www.nltk.org/`

2. Ablauf der Experimente

Jedes Experiment begann mit der Erstellung einer Suchanfrage, die zunächst manuell verifiziert werden musste. Dabei wurden die Erwartungen an das Experiment durch die Bearbeitung der wissenschaftlichen Arbeiten präzise definiert. Nach Abschluss der Suchanfrage und der Festlegung der Erwartungen wurde aus der Suchanfrage ein Suchgraph generiert. Daraufhin wurde für jede wissenschaftliche Arbeit im Datensatz ein entsprechender Dokumentgraph erstellt. Zur Generierung der Graphen erhielt das LLM spezifische Prompts. Für den Suchgraph einen Prompt mit einer Beispielgenerierung der Suchanfrage und für den Dokumentgraph einen Prompt mit einem Beispiel, der Suchanfrage (abhängig von der angewandten Methode) sowie der jeweiligen wissenschaftlichen Arbeit. Die erstellten Such- und Dokumentgraphen wurden in Textdateien gespeichert und anschließend in den Prompts der jeweiligen Teilexperimente weiterverwendet. Die Ergebnisse der Experimente wurden abschließend in Tabellenform ausgegeben.

3. Herausforderungen und Lösungen

Verbindungsprobleme zur Neo4j-Datenbank: Während der Implementierung traten Herausforderungen bei der Verbindung zur Neo4j-Datenbank auf. Diese resultierten häufig aus unsachgemäß konfigurierten Verbindungsdetails, wie einem fehlerhaften URI, falschen Benutzernamen oder Passwörtern. Darüber hinaus konnten Netzwerkbeschränkungen die Kommunikation zur Datenbank behindern.

Um diese Probleme zu lösen, wurde sichergestellt, dass die Verbindungsdetails korrekt hinterlegt und die Netzwerkverbindungen ordnungsgemäß eingerichtet waren. Zudem kam stets die aktuellste Version der Neo4j-Datenbank und der zugehörigen Python-Bibliothek zum Einsatz, um Kompatibilitätsprobleme zu vermeiden und von den neuesten Funktionen und Optimierungen zu profitieren.

Fehlerhafte Cypher-Abfragen: Ein weiteres Problem ergab sich aus Syntaxfehlern in den Cypher-Befehlen, die zu Fehlermeldungen oder fehlerhaften Ergebnissen führten. Zusätzlich beeinträchtigten fehlende Indizes oder Constraints in der Datenbank die Abfrageleistung. Ein spezielles Problem war, dass die vom LLM generierten Graphen häufig nicht in reiner Cypher-Syntax erstellt wurden. Stattdessen enthielten sie ergänzende Texte, optische Verzierungen oder kleine Syntaxfehler, die eine direkte Nutzung in der Neo4j-Datenbank verhinderten.

Dieses Problem wurde durch ein Post-Processing der generierten Graphen gelöst, bei dem fehlerhafte oder überflüssige Elemente entfernt wurden. Außerdem wurden die generierten Abfragen manuell auf Fehler überprüft und optimiert, um ihre Korrektheit und Effizienz sicherzustellen.

Umgang mit API-Beschränkungen bei der Nutzung des LLM: Die Nutzung der generativen KI brachte Einschränkungen wie begrenzte API-Aufrufe pro Zeitintervall und eine maximale Tokenanzahl pro Anfrage mit sich, was die Verarbeitung langer wissenschaftlicher Texte erschwerte. Zusätzlich führten hohe Serverauslastungen während Spitzenzeiten zu Latenzen, Verbindungsproblemen oder Zeitüberschreitungen.

Zur Bewältigung dieser Herausforderungen wurde eine Fehlerbehandlungslogik implementiert, die bei API-Fehlern wie Zeitüberschreitungen automatisch erneut Anfragen sendet. Durch die Einführung einer Warteschleifen-Logik konnten API-Nutzungslimits eingehalten werden. Zudem ermöglichte die Verwendung mehrerer API-Keys eine gleichmäßige Verteilung der Anfragen.

Fehlerbehebung bei der Textverarbeitung: Die Verarbeitung wissenschaftlicher Arbeiten im PDF-Format stellte eine weitere Herausforderung dar. PDFs mit mehreren Spalten, Tabellen oder eingebetteten Bildern führten häufig dazu, dass der Text nicht in der korrekten Reihenfolge extrahiert wurde. Darüber hinaus erschwerten eingebettete oder verschlüsselte Schriftarten die Verarbeitung. Weitere Probleme traten bei beschädigten oder fehlerhaften PDFs sowie bei der Erkennung von Formeln, Sonderzeichen oder Quellcode auf, die von Standardbibliotheken wie PyPDF2 nicht korrekt verarbeitet werden konnten.

Um diese Herausforderungen zu bewältigen, wurden spezialisierte Bibliotheken wie `pdfplumber`[8] eingesetzt, die auch mit komplex formatierten oder bildbasierten PDFs umgehen können. Dazu wurden Regeln zur Bereinigung und Umformatierung von Sonderzeichen und Fußnoten hinzugefügt, um die Lesbarkeit und die Analysefähigkeit der Daten zu verbessern.

Diese Maßnahmen sorgten dafür, dass die Textverarbeitung robust und präzise durchgeführt werden konnte, selbst bei schwierigeren Dokumentenformaten.

[8] `https://pypi.org/project/pdfplumber/`