EPUB 리더기의 GraphRAG 활용에 관한 연구

A Study on utilizing GraphRAG on EPUB Reader

박채원*, 옥지윤*, 한지운*, 성주연*, 황기태*

Chae-Won Park*, Ji-Yoon Ok*,Ji-Woon Han*, Ju-Yeon Soung*, Kitae Hwang*

chaewon24caley@gmail.com, 2271522@hansung.ac.kr, agn705@gmail.com,

tjdwndus1325@gmail.com, calafk@hansung.ac.kr

요 약

본 논문에서는 GraphRAG를 활용하여 설계 및 구현한 EPUB 리더기, Graph EPUB Reader에 대해 소개한다. 본 논문에서 구현한 Graph EPUB Reader는 기존 EPUB 리더기에 검색과 추론을 할 수 있는 질의응답 시스템을 결합한 것으로, GraphRAG를 사용하여 EPUB 파일을 지식 그래프로 생성하고, 생성된 지식 그래프에서 필요한 데이터를 추출한 후 자연어로 답변할 수 있게 한다. 본 논문에서 제안된 Graph EPUB Reader를 테스트한 결과, 사용자 질의에 대해 적절한 수준의 검색과 추론 결과를 보였다. 또한 Graph EPUB Reader의 성능을 평가하기 위해, EPUB 파일 크기에 따른 지식 그래프 생성 시간과 사용자 질의에 대한 검색 및 추론 시간을 실제 측정하여 평가하였다. 성능 평가 결과, 지식 그래프 생성 시간은 EPUB 파일 크기에 비례하였으며, 검색과 추론 시간은 질의의 유형에 관계없이 비슷한 것으로 평가되었다.

키워드 : GraphRAG, EPUB Reader, LLM, Knowledge Graph

I. 서 론

오늘날 사람들은 데이터가 폭발적으로 증가하고 있는 빅데이터 시대에 살고 있다. 이를 효과적으로 처리하기 위한 AI 기술도 함께 발전하고 있으며, 핵심 목표는 검색의 효율성과 추론의 정확성이다[1].

최근 ChatGPT와 같은 대규모 언어 모델(LLM)의 등장으로, 사용자가 자연어로 쉽게 검색할 수 있게 되면서 검색의 편의성이 높아져 많은 사람들이 사용하게 되었다[2]. 하지만 LLM은 인터넷의 많은 데이터를 학습하는 과정에서 검증되지 않은 정보까지 학습하여 잘못된 일반화가 발생하는 등, 부정확한 정보를 생성하는 할루시네이션 문제가 나타나고 있다[3].

이러한 문제를 해결하기 위해 RAG(Retrieval-Augmented Generation) 기술이 제안되었다[4]. RAG는 외부 지식 데이터를 벡터 데이터베이스에 저장한 후, 사용자 질문을 벡터화하여 벡터 데이터베이스에서 가장 유사한 데이터를 추출하고, 추출된 데이터를 기반으로 답변을 생성한다. 이 기술은 Microsoft의 대화형 인공지능인 Copilot에서도 활용되고 있다[5]. 그러나 RAG는 검색의 편의성과 효율성에 대한 문제를 해결했지만, 벡터 유사도 검색에만 의존하기 때문에 여전히 추론에는 부족함을 보이고 있다[6].

검색과 추론을 모두 해결하기 위해 최근 GraphRAG(Graph + Retrieval-Augmented Generation) 기술이 제안되었다[7]. GraphRAG는 지식 그래프와 RAG를 결합한 기술로, 데이터를 그래프 형태로 연결하여 검색과 추론을 용이하게 하며, 특히 복잡한 관계를 깊이 이해해야 하는 특정 분야의 지식을 포괄적으로 표현하는 데 적합하다.

본 연구는 GraphRAG를 활용하여 특정 분야의 자료를 정확하게 검색하고 추론하는 응용 시스템 구축 방법을 체계화하기 위해 EPUB 리더기를 채택하여 Graph EPUB Reader를 설계하고 구현하였다. 본 연구팀이 개발한 Graph EPUB Reader는 EPUB 파일을 지식 그래프로 생성하는 시스템과 EPUB 파일을 가시화하고 사용자 인터페이스를 제공하는 기능을 갖추고 있다. 구체적으로, Graph EPUB Reader는 기존의 EPUB 리더기와 달리 검색과 추론이 가능한 질의응답 체계를 갖추었으며 EPUB 파일 속 데이터 간의 관계를 파악하여 사용자 질의에 효과적으로 대응한다. 또한 검색한 내용과 관련된 데이터의 상관관계를 지식 그래프 형태로 시각화하여 보여준다. 본 연구팀은 제안된 시스템의 활용성을 평가하기 위해 '이상한 나라의 앨리스' EPUB 파일을 테스트 북으로 선정하여, EPUB 파일을 지식 그래프로 생성하는 시간, 검색과 추론에 걸리는 시간을 측정하여 평가하였다.

II. 관련 연구

1. EPUB 리더기

EPUB(Electronic Publication)은 전자 문서로 표시하는 데 필요한 파일들을 하나로 압축한 ZIP 파일이다[8]. 기본적인 EPUB에는 문서의 메타데이터, 포함된 모든 파일의 목록을 담은 매니페스트, 그리고 읽는 순서를 정의하는 스파인이 포함된 OPF 파일, 책의 목차와 페이지 목록 등이 포함된 네비게이션 파일, 그리고 주요 콘텐츠가 담긴 XHTML, CSS, 이미지 파일 등을 포함한다. 또한, EPUB

*한성대학교 컴퓨터공학부

3.0에서는 텍스트 외에도 다양한 미디어를 표현하기 위해 Video와 Audio 등을 포함한다. 이러한 EPUB 파일을 읽기 위해 다양한 EPUB 리더기 소프트웨어와 앱들이 등장하였다.

초기 EPUB 리더기는 Adobe에서 출시한 전자책 리더 소프트웨어인 Adobe Digital Editions이다[9]. 이후, 스마트폰과 태블릿의 확산으로 Apple의 iBooks와 Google Play Books와 같은 앱들이 출시되어 사용자가 언제 어디서나 쉽게 EPUB 파일을 읽을 수 있게 되었으며 웹 기반 EPUB 리더기 또한 개발되어 웹브라우저를 통해 읽을 수 있게 되었다. 또한, 일반 개발자들도 PC에서 Calibre와 같은 다양한 오픈소스를 활용하여 EPUB 리더기를 어렵지 않게 만들 수 있다. 본 연구팀은 'React Reader' 오픈소스를 기반으로 Graph EPUB Reader를 구현하였다[10].

2. GraphRAG

GraphRAG는 지식 그래프와 RAG 기술을 결합하여 지식을 효과적으로 검색하는 방법론이다. GraphRAG에서 지식 그래프는 데이터를 그래프 형식으로 저장하는 저장소 역할을 하며, LLM은 사용자의 질의를 해석하여 지식 그래프에서 검색할 수 있는 쿼리문으로 변환하고 검색 결과를 기반으로 자연어 응답을 생성하는 역할을 한다.

현재 Microsoft는 GraphRAG를 응용 시스템 개발에 쉽게 활용할 수 있도록 라이브러리로 출시하였다. 최근 연구에 따르면 Microsoft에서 만든 GraphRAG는 벡터 데이터베이스 기반의 RAG보다 더 나은 답변을 제공할 뿐만 아니라 비용 측면에서도 더 효율적이라고 평가되었다[7]. Microsoft에서 공개한 GraphRAG 라이브러리는 LLM을 사용하여 텍스트 문서에서 자동으로 지식 그래프를 생성하고, 밀접한 관련이 있는 정보의 집합인 커뮤니티를 감지하고 요약한다. 이를 통해 전체 데이터를 종합적으로 이해하고 추론해야 답변이 가능한 질문에 대해 보다 정확한 답변을 생성할 수 있다. 예를 들어, "주인공은 착한 사람인가?" 와 같이 문서 전체의 내용을 파악해야 답변할 수 있는 질문에 대해, GraphRAG는 커뮤니티 요약을 사용하여 보다 신뢰성 있는 답변을 제공한다. 본 연구에서도 Microsoft의 GraphRAG 라이브러리를 사용하여 시스템을 개발하였다.

III. GraphRAG EPUB Reader

1. Graph EPUB Reader 구조 설계

본 논문에서 제안하는 시스템은 그림 1과 같이 서버-클라이언트 구조로 설계 구현하였다. 서버는 EPUB 파일과 Parquet 파일을 저장하는 데이터베이스, 지식 그래프를 저장하는 Neo4j, React를 사용하여 사용자 인터페이스를 제공하는 Main 모듈과 사용자 질의를 Microsoft의 GraphRAG 라이브러리를 사용하여 답변을 생성하는 검색 모듈로 구성된다. Parquet 파일은 데이터를 컬럼 방식으로 저장하여 대용량 데이터 처리에 적합한 저장 형식이다[11].

관리자는 DB 관리 모듈을 통해 서버 데이터베이스에 저장된 EPUB 파일에서 데이터를 추출하여 지식 그래프로 생성하고, 이를 Parquet 파일로 저장한다. 사용자는 EPUB 뷰어 모듈을 통해 웹브라우저에서 EPUB 파일을 확인할 수 있고 질의응답 모듈을 통해 검색을 한다면 서버의 검색 모듈로 사용자 질의를 전달하고 생성된 답변을 받을 수 있다. 또한, 지식 그래프 뷰어 모듈은 사용자가 검색한 내용과 관련된 데이터의 상관관계를 지식 그래프 형태로 시각화한다.

본 시스템은 크게 2가지 기능으로 구분되며 구체적으로 EPUB 파일로부터 지식 그래프를 생성하는 기능과 EPUB 파일 속 데이터에 대해 검색하는 기능이다.

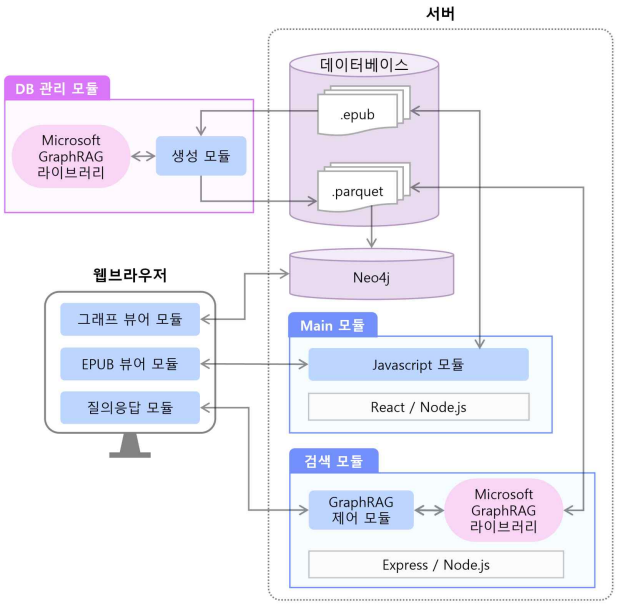

그림 1. Graph EPUB Reader 시스템 구조

2. EPUB 파일로부터 지식 그래프 생성

생성은 사용자가 EPUB 파일을 이용하기 전에, 관리자가 EPUB 파일을 지식 그래프로 변환하여 Parquet 파일로 저장하는 과정이다. 관리자는 DB 관리 모듈의 생성 모듈과 GraphRAG 라이브러리를 통해 자동으로 EPUB 파일로부터 데이터를 추출하고, 추출한 데이터를 기반으로 지식 그래프를 생성하여 Parquet 파일로 저장한다.

3. 검색

사용자는 웹브라우저에서 질의응답 모듈을 통해 EPUB 파일 내 사실 정보를 자연어로 검색하여 답변을 받을 수 있다. 예를 들어, 사용자가 "앨리스가 모험을 처음으로 시작한 곳은 어디인가?" 와 같이 검색하면, GraphRAG 라이브러리가 질의에 대해 분석하고 지식 그래프에서 의미적으로 연관된 데이터를 식별하여, 추출한 데이터를 기반으로 다음 그림 2와 같이 답변을 생성한다.

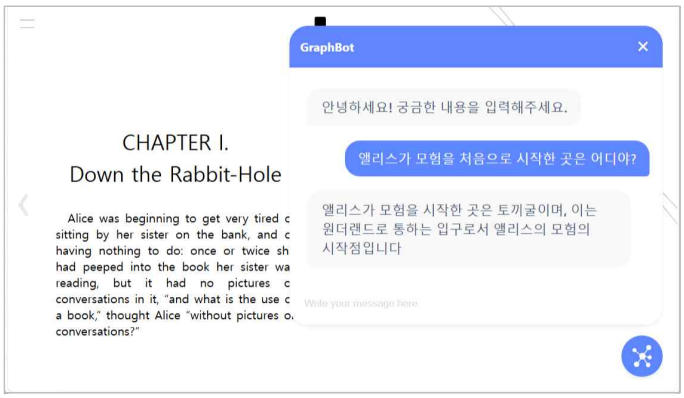

그림 2. 검색에 대한 응답

4. 추론

Graph EPUB Reader는 단순한 사실 검색뿐만 아니라 EPUB 파일의 전체 내용을 이해하고 추론 과정을 거쳐야만 하는 검색에 대한 응답도 제공한다. Microsoft의 GraphRAG 라이브러리는 지식 그래프 내에 의미적으로 유사한 정보 집합을 커뮤니티로 묶은 후, 요약하여 저장하기 때문에 EPUB 파일의 내용을 종합적으로 추론할 수 있게 한다. 예를 들어, "앨리스는 어떤 성격을 가진 캐릭터인가?"와 같이 추론 과정을 거쳐야 하는 질의에 대해서는, 지식 그래프 내에서 커뮤니티 요약을 기반으로 데이터를 추출하여 그림 3과 같이 답변을 생성한다.

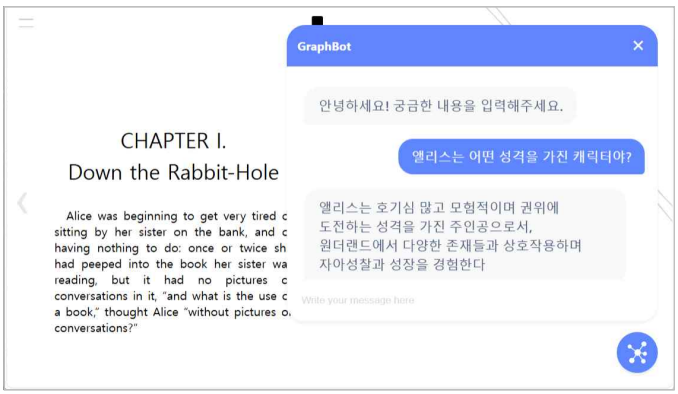

그림 3. 추론성 질의에 대한 응답

5. 지식 그래프 시각화

본 시스템은 지식 그래프 시각화를 제공하여 사용자가 데이터 간의 관계를 파악할 수 있게 하며, 이를 통해 검색과 추론의 결과를 시각적으로 확인할 수 있도록 한다. 서버에 저장된 Parquet 파일을 분석하여 Neo4j에 저장한 후, 해당 EPUB 파일의 지식 그래프를 그림 4와 같이 시각화할 수 있다.

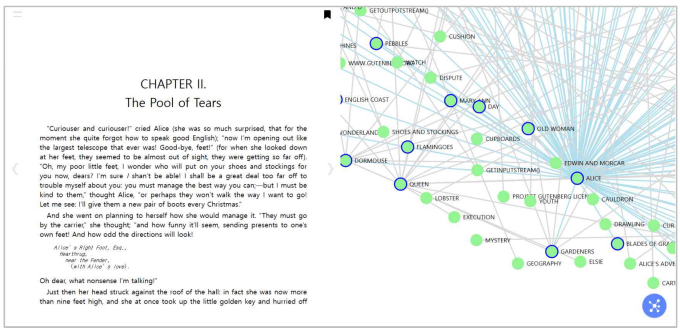

그림 4. 지식 그래프 시각화

IV. 성능 평가

이 절에서는 GraphRAG를 활용한 Graph EPUB Reader의 성능을 평가하였다. 이 성능 평가의 목적은 지식 그래프 생성 시간과 EPUB 파일 크기 간의 상관관계를 확인하고, 질의가 단순 검색인지 추론인지에 따라 검색 시간의 차이를 확인하는 것이다. 성능평가에 사용된 시스템은 Core-i7 CPU에 16GB 메모리를 가진 범용 노트북을 사용하였고 LLM은 gpt-3.5-turbo를 사용하였다.

1. 지식 그래프 생성 시간

이 실험에서는 한 EPUB 파일로부터 지식 그래프를 생성하는 데 걸리는 시간을 측정한다. 실험 결과, 지식 그래프 생성 시간은 단어 개수에 비례하는 것으로 나타났다. 이에 본 연구팀은 EPUB 파일의 단어 개수를 6,000개씩 증가시키며 지식 그래프 생성 시간을 측정하였다. 성능 평가 결과는 총 5번 실험의 평균값을 이용하였으며 그림 5와 같다.

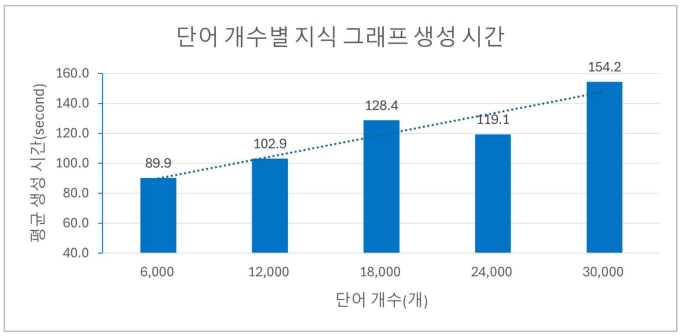

그림 5. 단어 개수 별 지식 그래프 형성 시간

실험 결과 그림 5와 같이 지식 그래프 생성 시간은 데이터 양에 선형적으로 비례한 것으로 판단된다. 다만, 단어 개수 18,000개는 약 13페이지 분량으로, 생성에 약 2분이 소요되어 다소 길게 느껴질 수 있지만, 생성 작업은 관리자에 의해 한 번만 이루어지기 때문에 시간의 크기는 문제되지 않는다고 판단된다.

2. 검색 시간

본 연구팀은 검색 시간을 측정하기 위해 여러 샘플 질의를 만들어 실험하였고, 그 중 하나는 표 1과 같다. 표 1과 같이 실험에 사용된 샘플 질의는 추론 없이 단순한 사실을 검색하는 목적을 가진 질의이다. 검색 시간은 서버가 질의 메시지를 받았을 때부터 검색 결과를 생성할 때까지 걸린 시간으로 정의한다.

성능 평가 결과, 여러 샘플 질의 중 표 1과 같은 질의의 검색 시간은 15.9초로 측정되었다. 또한, '앨리스가 처음 모험한 곳'을 묻는 질의에 대해 '토끼굴'이라는 적합한 답변을 생성하였다.

표 1. 검색 샘플 질의와 검색 시간

	샘플 질의	검색 시간
질의	앨리스가 처음 모험한 곳은 어디인가요?	15.9초
답변	앨리스가 처음 모험한 곳은 토끼굴이다. 이곳은 원더랜드로 향하는 입구로서 앨리스의 모험의 시작점이 되었다.	

3. 추론 시간

본 연구팀은 추론 시간을 측정하기 위해 여러 샘플 질의를 만들어 실험하였고, 그 중 하나는 표 2와 같다. 표 2와 같이 실험에 사용된 질의는 내용을 종합적으로 이해하고 추론을 해야 하는 질의이다. 추론 시간은 서버가 질의 메시지를 받았을 때부터 검색 결과를 생성할 때까지 걸린 시간으로 정의한다.

성능 평가 결과, 여러 샘플 질의 중 표 2와 같은 질의의 추론 시

간은 17.2초로 측정되었다. 또한, '앨리스의 성격' 을 묻는 질의에 대해 '호기심이 많고 모험을 즐기는 성격' 이라는 적절한 답변을 생성하였다.

표 2. 추론 샘플 질의와 추론 시간

	샘플 질의	추론 시간
질의	앨리스는 어떤 성격을 가진 캐릭터인가요?	17.2초
답변	앨리스는 호기심 많고 모험을 즐기며 권위에 도전하는 성격을 가지고 있다. … (생략)	

V. 토 의

본 성능 평가 결과, EPUB 파일에 대한 지식 그래프 생성 시간은 EPUB 파일 크기에 선형적으로 비례한 것으로 평가되었으며, 검색과 추론 시간은 질의 유형에 따라 큰 차이를 보이지 않는 것으로 평가된다. 명확하지 않지만, 질의 유형과 관계없이 질의에 응답하기 위해 지식 그래프 속 데이터를 검색해야 하기 때문인 것으로 사려된다. 추후 더 면밀한 분석이 필요할 것으로 판단된다. 이번 성능 평가를 통해 본 시스템은 성능뿐만 아니라 검색에 대한 적절한 답을 생성하는 데 잘 설계되었다고 판단된다.

VI. 결 론

본 논문은 EPUB 파일의 내용을 검색하고 추론할 수 있는 Graph EPUB Reader를 설계하고 구현한 내용을 소개하였다. GraphRAG 라이브러리를 사용하여 기존의 EPUB 리더기에서 할 수 없었던 검색과 추론을 가능하게 하였으며, EPUB 파일 속 데이터 간의 관련성도 가시적으로 볼 수 있게 하였다. 성능 평가에서 지식 그래프 생성 시간을 평가한 결과, EPUB 파일의 크기와 관련이 있다는 것으로 판단된다. 단순 검색과 추론이 필요한 질의를 한 결과, 검색과 추론에 걸리는 시간에 큰 차이가 없었다. 또한, 질의에 대한 적절한 답변을 얻는 데 성공하여 EPUB 파일을 읽는 다양한 독자들의 요구를 충족한다고 판단한다.

※ 본 연구는 한성대학교 교내 학술 연구비를 지원받았음

참 고 문 헌

[1] 송사광, 이승우, 정한민. (2013). 베이지안 추론망을 이용한 검색엔진 세부 모듈의 상세 분석 방법. 정보과학회논문지 : 소프트웨어 및 응용, 40(5), 277-282.

[2] Li, Y., Wen, H., Wang, W., Li, X., Yuan, Y., Liu, G., Liu, J., Xu, W., Wang, X., Sun, Y. and Kong, R., 2024. Personal llm agents: Insights and survey about the capability, efficiency and security. arXiv preprint arXiv:2401.05459.

[3] G. Perković, A. Drobnjak and I. Botički, "Hallucinations in LLMs: Understanding and Addressing challenges" in 2024 47th MIPRO ICT Electron. Conv., pp. 2084-2088, 2024.

[4] Lewis, Patrick, et al. "Retrieval-augmented generation for knowledge-intensive nlp tasks." Advances in Neural Information Processing Systems 33 (2020): 9459-9474.

[5] https://analyticsindiamag.com/ai-trends-future/rag-with-microsoft-copilot/

[6] Peng, B., Zhu, Y., Liu, Y., Bo, X., Shi, H., Hong, C., Zhang, Y. and Tang, S., 2024. Graph retrieval-augmented generation: A survey. arXiv preprint arXiv:2408.08921.

[7] D. Edge, H. Trinh, N. Cheng, J. Bradley, A. Chao, A. Mody, S. Truitt and J. Larson, "From local to global: A graph rag approach to query-focused summarization," arXiv preprint arXiv:2404.16130, 2024.

[8] Marinai, Simone, Emanuele Marino, and Giovanni Soda. "Conversion of PDF books in ePub format." In 2011 International Conference on Document Analysis and Recognition, pp. 478-482. IEEE, 2011.

[9] Garrish, M. and Gylling, M., 2013. EPUB 3 best practices. " O'Reilly Media, Inc.".

[10] https://github.com/gerhardsletten/react-reader

[11] https://parquet.apache.org/