

<https://doi.org/10.7236/JIIBC.2024.24.6.29>

JIIBC 2024-6-5

# GraphRAG를 활용하여 뛰어난 검색과 추론 기능을 가진 EPUB 리더

## EPUB Reader with Advanced Search and Inference Functions using GraphRAG

박채원\*, 옥지윤\*, 한지운\*, 성주연\*, 황기태\*\*

Chaewon Park\*, Jiyeon Ok\*, Jiwoon Han\*, Juyeon Soung\*, Kitae Hwang\*\*

**요약** 기업과 조직 내 도메인 특화 데이터가 전자 문서로 저장되어 왔지만, 문서 간의 관계를 파악하고 정확한 검색과 추론을 수행하는 데 어려움이 있다. 이에 본 논문은 전자 문서를 대표하는 포맷 중 EPUB 포맷을 채택하여, EPUB 파일 속 내용을 검색하고 추론할 수 있는 Graph EPUB Reader 시스템을 제안하고 구현하였다. 이 시스템은 GraphRAG 기술을 활용하여 EPUB 파일 속 내용을 지식 그래프로 생성하고, 단순 사실 검색뿐 아니라 추론 과정을 거쳐야만 하는 검색에도 답변을 생성한다. 본 논문에서는 제안된 Graph EPUB Reader 시스템의 활용성을 검증하기 위해 지식 그래프 생성 성능과 검색과 추론 성능을 평가하였다. 성능 평가 결과, EPUB 파일의 크기에 따라 지식 그래프 생성 시간이 선형적으로 증가하는 것으로 측정되었으며, 검색과 추론은 약 90%의 높은 정확도를 보인 것으로 평가되었다. 본 논문에서 제안한 Graph EPUB Reader 시스템에 사용된 검색과 추론 기술은 EPUB 파일뿐 아니라 PDF 등 다양한 여러 전자 문서를 검색하고 추론하는데 활용될 수 있을 것으로 판단된다.

**Abstract** Domain-specific data within companies and organizations have been stored in electronic documents, but there are difficulties in understanding relationships between documents and performing accurate searches and inferences. Therefore, this paper proposed and implemented a Graph EPUB Reader system that can search and inference based on the contents of EPUB files, one of the representative formats for electronic documents. This system generates a knowledge graph from the contents of EPUB files and provides answers to both simple fact searches and inference-based searches. To validate the utilization of the system proposed in this paper, we analyzed the performances of knowledge graph generation, search, and inference. As a result of the performance evaluation, the knowledge graph creation time was measured to increase linearly depending on the size of the EPUB file, and search and inference were evaluated to have a high accuracy of about 90%. Finally, we conclude that the search and inference technology used in the Graph EPUB Reader system proposed in this paper can be used to search and infer not only EPUB files but also various electronic documents such as PDF.

**Key Words** : GraphRAG, Knowledge Graph, LLM, EPUB

\*준회원, 한성대학교 컴퓨터공학부

\*\*종신회원, 한성대학교 컴퓨터공학부, 교신 저자

접수일자 2024년 9월 24일, 수정완료 2024년 11월 14일

게재확정일자 2024년 12월 6일

Received: 24 September, 2024 / Revised: 14 November, 2024 /

Accepted: 6 December, 2024

\*\*\*Corresponding Author: calafk@hansung.ac.kr

Dept. of Computer Engineering, Hansung University

## I. 서 론

오늘날, 종이 문서들이 전자 문서로 전환되면서 문서 검색을 효율적으로 할 수 있게 되었다. 특히, 검색은 전자 문서를 관리하고 필요한 정보를 빠르게 찾는 핵심 기능으로 자리 잡았다[1]. 기업과 조직들은 PDF 등과 같은 다양한 형태의 전자 문서를 축적해왔지만, 여전히 문서 간의 관계를 밀접하게 파악하고 정확한 검색과 추론을 할 수 있는 시스템을 필요로 하고 있다[2].

한편, 데이터 처리 기술의 발전으로 검색과 추론 방식에 큰 변화가 일어나고 있다. 최근, 대규모 데이터 세트로 훈련된 인공지능 언어 모델인 LLM으로 OpenAI의 GPT-4, Meta의 LLaMA 등이 개발되어 간편하게 검색과 추론을 할 수 있게 되었다[3]. 하지만, 인터넷의 방대한 데이터를 학습한 LLM들은 할루시네이션 문제로 정확한 검색과 추론에 한계를 보인다[4]. 그러므로 도메인 특화 데이터들에 대해서만 검색하고 추론 할 수 있는 새로운 방법이 필요하였다. 이에 도메인 특화 데이터 세트를 벡터 데이터베이스에 저장하여 사용자 질의에 대해 의미적으로 유사한 정보를 찾아내는 RAG(Retrieval-Augmented Generation) 기술이 제안되었다[5]. 하지만 RAG는 연관 검색에서 뛰어난 성능을 보이지만, 복잡한 추론에는 한계가 있었고, 이를 보완하기 위해 최근 지식 그래프와 RAG를 결합하여 데이터 간의 관계를 명확히 표현하고 검색과 추론을 가능하게 하는 GraphRAG 기술이 제안되었다[6].

본 연구는 GraphRAG를 활용하여 전자 문서를 보다 정확하게 검색하고 추론할 수 있는 응용 시스템을 개발하여 GraphRAG의 활용성을 검증하고자 한다. 이에 대표적인 전자책 포맷인 EPUB을 기반으로 Graph EPUB Reader 시스템을 설계하고 구현하였다. EPUB 리더기는 전자책뿐 아니라 다양한 전자 문서 리더기로도 확장될 수 있기 때문에, 본 논문의 시스템 구축 경험은 다양한 전자 문서 관리 시스템에 적용될 것으로 판단된다.

본 연구팀이 설계 구축한 Graph EPUB Reader는 EPUB 파일을 지식 그래프로 생성하는 시스템과 EPUB 파일 속 내용에 대해 검색과 추론을 할 수 있는 시스템으로 구성된다. 또한, 사용자는 검색한 내용과 관련 있는 데이터의 상관관계를 지식 그래프 시각화를 통해 이해할 수 있다.

본 연구팀이 제안한 시스템의 활용성을 평가하기 위해, ‘피터팬’ EPUB 파일을 테스트 북으로 선정하여 EPUB 파일의 크기에 따른 지식 그래프 생성 시간, 검색

과 추론에 걸리는 시간과 정확성을 측정하여 평가하였다.

## II. 관련 연구

### 1. GraphRAG

벡터 데이터베이스는 고차원 벡터 공간에서 사용자가 입력한 쿼리에 대해 가장 유사한 데이터를 찾을 수 있는 시스템이며, 대표적으로 Pinecone과 Milvus 등이 있다 [7]. 검색의 정확도를 높이기 위해 벡터 데이터베이스에 LLM을 결합한 RAG가 제안되었다. LLM은 자연어로 질의를 입력받고 검색 결과를 자연어로 출력하는 용도로 국한된다. RAG는 유사성 검색을 통해 단순 사실 검색에는 뛰어난 성능을 보이지만, 데이터 간의 복잡한 관계를 심도 있게 파악하는 추론에는 부족함이 있다[8].

이러한 문제점을 극복하기 위해 지식 그래프와 RAG 기술을 결합한 GraphRAG가 제안되었다. 지식 그래프는 데이터를 그래프 형식으로 저장하는 데이터 모델로, 복잡한 정보를 직관적으로 표현할 수 있도록 한다[9,10]. 지식 그래프를 활용해 데이터를 관리하는 그래프 데이터베이스 시스템으로 Neo4j와 Amazon Neptune 등이 있다[11]. GraphRAG는 지식 그래프를 통해 정보 간의 관계를 명확히 파악하여 검색과 추론을 용이하게 하며, 검색에 대한 응답을 자연어로 생성하기 위해 LLM을 결합하여 사용한다.

최근, Microsoft에 의해 공개된 GraphRAG 라이브러리는 LLM의 구문 분석 기능을 활용하여 텍스트 문서에서 구문을 인식하고, 지식 그래프를 생성하여 Parquet 포맷으로 저장한다. Parquet 포맷은 빅데이터를 효율적으로 처리하기 위한 데이터 저장 포맷으로, Apache Spark, Apache Hive 등 빅데이터 처리 프레임워크에 사용된다[12]. 또한, 생성된 지식 그래프에서 Leiden 알고리즘을 통해 밀접하게 관련된 집합체인 커뮤니티를 감지하여 간결하게 요약한다[13]. 이를 통해 복잡한 데이터 간 관계를 파악하여 추론을 거쳐야만 하는 질의에 보다 정확한 답변을 생성한다. 본 연구에서는 Microsoft에서 제공하는 GraphRAG 라이브러리와 OpenAI LLM을 활용하여 Graph EPUB Reader 시스템을 개발하였다.

### 2. EPUB

EPUB(Electronic Publication)은 전자 출판을 위한 포맷으로, 전자 문서로 표현하는 데 필요한 파일들을 하

나로 압축한 ZIP 파일이다[14]. 기본적인 EPUB에는 문서의 메타데이터, 포함된 모든 파일의 목록을 담은 매니페스트, 그리고 읽는 순서를 정의하는 스파인이 포함된 OPF 파일, 책의 목차 등이 포함된 네비게이션 파일, 그리고 주요 콘텐츠가 담긴 XHTML, CSS, 이미지 파일 등을 포함한다. 또한, EPUB 3.0에서는 텍스트 외에도 다양한 미디어를 표현하기 위해 Video와 Audio 등을 포함한다.

인터넷에는 React Reader, epub.js, Radium등 다양한 종류의 EPUB 리더 라이브러리가 존재한다. 본 연구팀은 그중 React Reader 라이브러리를 사용하여 개발하였다[15].

### III. Graph EPUB Reader

#### 1. 시스템 구조

본 논문에서 제안한 Graph EPUB Reader 시스템은 그림 1과 같이 서버-클라이언트 모델로 구성된다.

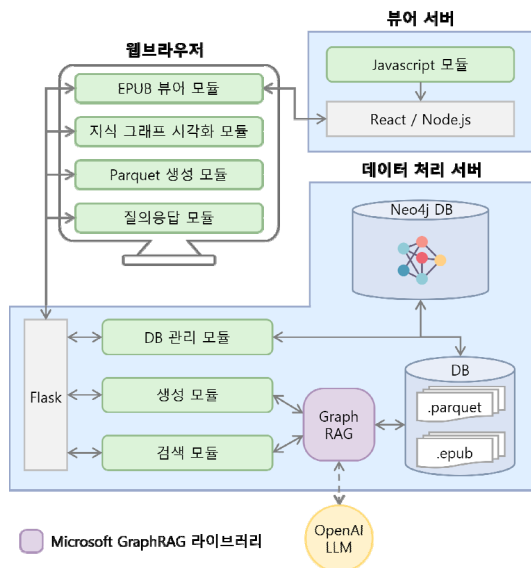


그림 1. Graph EPUB Reader 시스템 구조  
Fig. 1. System Architecture of Graph EPUB Reader

이 시스템은 크게 2가지 기능으로 나눌 수 있으며, 관리자가 EPUB 파일을 업로드할 때, 지식 그래프를 생성하는 기능과 EPUB 파일 속 내용에 대해 검색하고 추론하는 기능으로 구성된다.

사용자는 웹 브라우저를 통해 EPUB 파일을 다운로드하고 책을 읽거나 검색할 수 있다. 서버는 뷰어 서버와 데이터 처리 서버로 구성된다. 뷰어 서버는 React와 Node.js를 기반으로 구현되며, 사용자 웹 브라우저로 사용자 인터페이스를 위한 HTML 파일, 자바스크립트 파일과 EPUB 파일 등을 전송한다.

데이터 처리 서버는 Flask 웹 프레임워크를 사용하여 그래프 생성, 그래프 시각화, 검색 등 사용자 요청을 처리한다. Flask는 웹 브라우저로부터 EPUB 파일에 대해 그래프 생성 요청을 받으면, GraphRAG 라이브러리를 통해 EPUB 파일 속 내용을 지식 그래프로 생성하여 Parquet 파일로 저장한다. Flask가 웹 브라우저로부터 EPUB 파일과 지식 그래프 시각화 요청을 받으면, 데이터베이스로부터 지식 그래프 데이터를 전달한다. 사용자가 웹 브라우저를 통해 검색이나 추론을 포함하는 질의를 요청하면, Flask는 GraphRAG 라이브러리를 통해 질의에 대한 답변을 만들고, 생성된 답변을 전달한다.

#### 2. EPUB 파일로부터 지식 그래프 생성

지식 그래프 생성은 관리자가 EPUB 파일을 업로드할 때, EPUB 파일 속 내용을 기반으로 지식 그래프를 생성하는 과정으로, 알고리즘은 그림 2와 같다.

##### Algorithm 1. Knowledge Graph Generation

```
FUNCTION generate_process(epub_file):
    text_content = convert_epub_to_text(epub_file)
    knowledge_graph = GraphRAG.index(text_content)
    parquet_file = save_to_parquet(knowledge_graph)
    save_to_database(parquet_file)
```

그림 2. 지식 그래프 생성 알고리즘  
Fig. 2. Algorithm of Knowledge Graph Generation

관리자가 EPUB 파일을 업로드하면, 데이터 처리 서버에서 EPUB 파일 속 내용을 텍스트로 변환하고, GraphRAG 라이브러리를 통해 변환된 텍스트 내에서 의미있는 데이터를 추출하여 지식 그래프를 생성해 Parquet 파일로 저장한다. GraphRAG 라이브러리는 텍스트에서 의미있는 단어나 구문을 추출하기 위해 내부적으로 OpenAI LLM을 사용한다.

#### 3. 검색

본 시스템은 EPUB 파일 속 내용에서 단순 사실에 대한 검색을 할 수 있으며, 알고리즘은 그림 3과 같다.



CPU에 256GB 메모리를 가진 서버 컴퓨터를 사용하였다. LLM은 OpenAI에서 제공하는 GPT-3.5 Turbo를 사용하였다.

## 1. 지식 그래프 생성 성능

본 연구팀은 EPUB 파일로부터 의미 있는 데이터를 추출하여 지식 그래프를 생성하는 시간을 측정하였다. 측정 결과, EPUB 파일의 크기에 비례하여 지식 그래프 생성 시간도 그만큼 길어지는 것으로 나타났다. 이에 본 연구팀은 구체적인 성능 평가를 위해 책 한 페이지가 약 1,200자 분량이라는 점을 기준으로, 40페이지씩 분량을 증가시키며 지식 그래프 생성 시간을 측정하였다. 성능 측정 결과는 총 5번 실험의 평균값을 이용하였으며 그림 8과 같다.

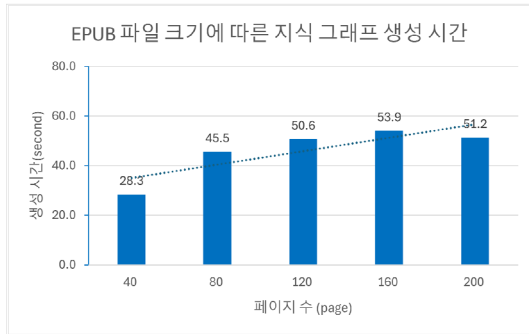


그림 8. EPUB 파일 크기에 따른 지식 그래프 생성 시간  
Fig. 8. Knowledge Graph Generation Time According to EPUB File Size

지식 그래프 생성 시간은 EPUB 파일의 크기에 선형적으로 비례한다. 책 200페이지 분량을 지식 그래프로 생성하는 데 1분 미만이며, 관리자가 EPUB 파일을 업로드할 때 한 번만 발생하는 작업이기 때문에 감내할만한 시간이라고 판단된다.

## 2. 검색 성능

본 연구팀은 EPUB 파일 속 내용과 관련된 단순한 사실에 대한 검색 시간을 측정하였다. 검색 시간은 데이터 처리 서버가 사용자 질의 메시지를 받았을 때부터 검색 결과를 생성할 때까지 걸린 시간으로 정의하고, 질의에 대한 예상 답변과 실제 생성된 답변을 비교하여 정확성을 판단하였다. 성능 평가에는 단순한 사실을 검색하는 총 20개의 샘플 질의를 만들어 실험하였으나, 지면 관계상 3가지만 표 1에 보인다.

표 1. 검색 샘플 질의와 검색 시간

Table 1. Sample Search List and Search Time

질의1 : 웬디는 피터팬의 그림자를 어떻게 붙여줬나요?

예상 답변	웬디는 바느질로 피터팬의 발에 그림자를 붙여주었습니다.
실제 답변	웬디는 피터팬의 그림자를 바느질로 발에 붙여주었어요. (성공)

질의2 : 피터팬이 날 수 있는 이유는 무엇인가요?

예상 답변	피터팬은 요정 가루를 사용하여 날 수 있습니다.
실제 답변	피터팬이 날 수 있는 이유는 요정 가루를 사용했기 때문입니다. (성공)

질의3 : 텡커벨은 어떤 언어로 대화하나요?

예상 답변	텡커벨은 종소리와 같은 요정 언어로 대화합니다.
실제 답변	텡커벨은 요정 언어로 대화합니다. (성공)

20개의 샘플 질의를 실험해본 결과, 약 90%의 확률로 예상 답변과 실제 답변이 일치하여 검색에 대한 정확성이 높은 것으로 판단되며, 검색 시간은 약 8초로 측정되었다.

## 3. 추론 성능

추론 시간은 데이터 처리 서버가 추론성 질의 메시지를 받았을 때부터 검색 결과를 생성할 때까지 걸린 시간으로 정의하고, 질의에 대한 예상 답변과 실제 생성된 답변을 비교하여 정확성을 판단하였다. 성능 평가에는 총 20개의 추론성 샘플 질의를 만들어 실험하였으나, 지면 관계상 표 2에 총 3가지만 보인다.

표 2. 추론 샘플 질의와 추론 시간

Table 2. Sample Inference List and Inference Time

질의1 : 피터팬은 어떤 성격을 가진 인물인가요?

예상 답변	피터팬은 도전적이고 적극적이면서 어린 아이같은 순수한 면을 가진 캐릭터입니다.
실제 답변	피터팬은 용감하고 모험적이며 도전적인 성격을 가지고 있다. ... (성공)

질의2 : 피터팬이 어른이 되고 싶어 하지 않는 이유는 무엇인가요?

예상 답변	피터팬은 어른이 되면 자유와 모험이 줄어들 것이라고 생각하기 때문에 어른이 되고 싶어 하지 않습니다.
실제 답변	피터팬은 어른이 되고 싶어 하지 않는 이유는 어린이의 순수하고 자유로운 마음을 지키고 싶어하기 때문입니다. 그는 어른이 되면 잃어버리게 될 상상력과 ... (성공)

질의3 : 피터팬이 웬디에게 키스의 개념을 몰랐던 이유는 무엇인가요?

예상 답변	피터팬은 어른들과 달리 사회적 개념이나 애정 표현에 대해 무지하기 때문에 키스의 개념을 몰랐습니다.
실제 답변	피터팬이 키스의 개념을 몰랐던 이유는 성장과 성인들의 세계에 대한 이해가 부족하여 성적인 개념을 이해하지 못했기 때문입니다. ... (성공)

추론성 질의에 대한 실험 결과, 20개의 샘플 질의에 대해 약 90%의 확률로 예상 답변과 실제 답변이 일치하는 것으로 나타났고, 추론 시간은 약 10초로 측정되었다. 이에 본 연구팀은 본 논문에서 구축한 시스템이 높은 추론 성능을 가진 것으로 판단한다.

## V. 결 론

본 논문은 EPUB 파일의 내용을 검색하고 추론할 수 있는 Graph EPUB Reader 시스템을 설계하고 구현한 내용을 소개하였다. Microsoft의 GraphRAG 라이브러리를 사용하여 기존의 EPUB 리더기에는 없던 검색과 추론을 가능하게 하였으며, EPUB 파일 속 데이터 간의 관련성도 지식 그래프를 통해 시각적으로 볼 수 있게 하였다.

성능 평가 결과, EPUB 파일의 크기에 비례하여 지식 그래프를 생성하는 시간이 선형적으로 증가하는 것으로 측정되었으며, 관리자에 의해 한 번만 발생하는 작업이므로 생성 시간은 크게 중요하지 않다고 판단된다. 또한, 검색과 추론 성능을 측정한 결과, 검색과 추론은 약 90%의 정확성을 보였으며, 검색 시간은 약 8초, 추론 시간은 약 10초 정도로 측정되었다.

결론적으로, GraphRAG를 EPUB 리더기에 적용한 결과, 검색과 추론 성능이 뛰어났으며, 본 시스템 구축 기술이 다양한 전자 문서를 검색하고 추론하는데 활용될 수 있을 것으로 판단된다.

## References

- [1] A. Gaudinat et al., "Health search engine with e-document analysis for reliable search results", International Journal of Medical Informatics, vol. 75, no. 1, pp. 73-85, Jan. 2006, doi: <https://doi.org/10.1016/j.ijmedinf.2005.11.002>.
- [2] S.F.M. Yatinet al., "Electronic Document Management System: Malaysian Experience", Australian Journal of Basic and Applied Sciences, vol. 9, pp. 82-89, 2015.
- [3] W. X. Zhao et al., "A Survey of Large Language Models", arXiv (Cornell University), Mar. 2023, doi: <https://doi.org/10.48550/arxiv.2303.18223>.
- [4] J.-Y. Yao, K.-P. Ning, Z.-H. Liu, M.-N. Ning, and L. Yuan, "LLM Lies: Hallucinations are not Bugs, but Features as Adversarial Examples", arXiv.org, Oct. 04, 2023. doi: <https://doi.org/10.48550/arXiv.2310.01469>
- [5] P. Lewis et al., "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks", arXiv.org, Apr. 12, 2021. doi: <https://doi.org/10.48550/arXiv.2005.11401>
- [6] D. Edge et al., "From Local to Global: A Graph RAG Approach to Query-Focused Summarization", arXiv.org, Apr. 24, 2024. doi: <https://doi.org/10.48550/arXiv.2404.16130>
- [7] Y. Han, C. Liu, and P. Wang, "A Comprehensive Survey on Vector Database: Storage and Retrieval Technique, Challenge", arXiv.org, Oct. 18, 2023. doi: <https://doi.org/10.48550/arXiv.2310.11703>
- [8] B. Peng et al., "Graph Retrieval-Augmented Generation: A Survey", arXiv (Cornell University), Aug. 2024. doi: <https://doi.org/10.48550/arxiv.2408.08921>.
- [9] A. Singhal, "Introducing the Knowledge Graph: things, not strings", Google, May 16, 2012. <https://blog.google/products/search/introducing-knowledge-graph-things-not/>
- [10] Ye-Won Seo, Yeon-Seo Bae, You-Jeong Jeong, Kitae Hwang, "Knowledge Graph of Smartphone Photos", The Journal of The Institute of Internet, Broadcasting and Communication, vol. 24, no. 5, pp.203-209, Oct. 31, 2024. doi: <https://doi.org/10.7236/JIIBC.2024.24.5.203>.
- [11] R. Angles and C. Gutierrez, "Survey of graph database models", ACM Computing Surveys, vol. 40, no. 1, pp. 1-39, Feb. 2008. doi: <https://doi.org/10.1145/1322432.1322433>.
- [12] <https://parquet.apache.org/>
- [13] V. A. Traag, L. Waltman, and N. J. van Eck, "From Louvain to Leiden: guaranteeing well-connected communities", Scientific Reports, vol. 9, no. 1, Mar. 2019. doi: <https://doi.org/10.1038/s41598-019-41695-z>.
- [14] <https://idpf.org/>
- [15] <https://www.npmjs.com/package/react-reader>

## 저 자 소 개

박 채 원(준회원)



- 한성대학교 컴퓨터공학부 재학
- 관심분야 : 인공지능, 빅데이터

옥 지 윤(준회원)



- 한성대학교 컴퓨터공학부 재학
- 관심분야 : 웹 공학, 인공지능

한 지 운(준회원)



- 한성대학교 컴퓨터공학부 재학
- 관심분야 : 웹 공학, 분산 시스템

성 주 연(준회원)



- 한성대학교 컴퓨터공학부 재학
- 관심분야 : 웹 공학, IoT

황 기 태(중신회원)



- 서울대학교 컴퓨터공학과 학사
- 서울대학교 컴퓨터공학과 석사
- 서울대학교 컴퓨터공학과 박사
- 현재 한성대학교 컴퓨터공학부 교수
- 경력  
University of Florida 방문 교수
- 관심분야 : 모바일 시스템, IoT, 인공지능

※ 본 연구는 한성대학교 교내 학술 연구비를 지원받았음.