# RAG 시스템 성능 향상을 위한 웹문서 본문 정제 및 추출 시스템

김도현, 원일용, 유상현*, 김현정**

서울호서전문학교, *경민대학교, **건국대학교

kmds86312@kakao.com, clccclcc@shoseo.ac.kr, *simonyoo@kyungmin.ac.kr,
**nygirl@konkuk..ac.kr

# A Web Document Text Refinement and Extraction System for Enhancing RAG Performance

Kim Do Hyun, Won Il Yong, Yoo Sang Hyun*, Kim Hyun Jung**

Seoul Hoseo College, *Kyungmin Univ., **Konkuk Univ.

요 약

본 논문은 RAG 시스템의 성능을 향상시키기 위해 웹 문서에서 불필요한 데이터를 제거하고 핵심 본문을 정제하여 추출하는 시스템을 제안한다. 제안된 시스템은 사용자 쿼리와 웹페이지의 HTML 태그 구조를 분석하고, 코사인 유사도와 태그 문장화를 활용해 본문과 쿼리 간의 유사성을 측정하여 최적의 본문을 식별한다. 실험 결과, 다양한 웹사이트에서 본문을 높은 정확도로 추출하여 RAG의 최신 데이터 처리 성능을 강화할 수 있음을 확인하였다.

## Ⅰ. 서 론

대형 언어 모델(LLM)은 최신 데이터 처리에 한계가 있다. 이는 아직 학습되지 않은 정보에 대한 응답에 여러 가지 제한 사항이 있다는 것을 의미한다. 이 문제를 해결하기 위해 검색증강생성(RAG) 방법이 사용된다. RAG는 외부 지식을 참조하여 LLM의 응답 품질을 향상시킨다. 웹 문서를 외부 지식으로 활용함으로써 최신 데이터 처리에 대한 한계를 개선한다.[1][2][3]. 그러나 웹 문서에는 광고, 링크 등 RAG 성능에 부정적 영향을 미칠 수 있는 데이터가 다수 포함되어 있다. 본 연구는 웹 문서에서 불필요한 데이터를 제거하고 본문만을 효과적으로 추출하는 시스템을 제안한다. HTML 태그 구조를 분석하여 사용자 쿼리와 관련성이 높은 본문을 식별하고, 이를 기반으로 정제된 텍스트 데이터를 생성한다[4][5]. 코사인 유사도와 태그 문장화를 활용해 본문과 쿼리 간의 유사성을 측정하여 최적의 본문 추출을 목표로 한다. [6][7][8]. 본 연구는 LLM의 최신 데이터 반영 문제를 해결하는 데 기여하며, 검색증강생성(RAG) 방법의 실용성을 증명한다.

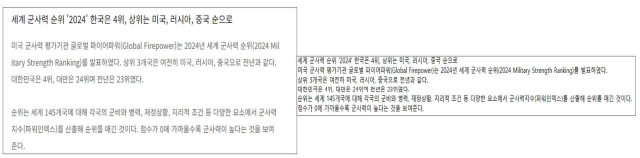
## Ⅱ. 시스템 구성

본 시스템은 사용자 쿼리를 기반으로 관련 웹 문서를 효과적으로 수집하고 웹 문서의 본문 내용을 추출하는 과정을 수행한다. 먼저, 사용자 쿼리를 활용하여 검색엔진을 통해 관련 웹사이트의 데이터를 수집하고, 이 데이터를 구조화하여 분석한다. 이후, 사용자 쿼리와 웹 문서 내 태그 간의 유사도를 측정하여 가장 관련성이 높은 태그를 식별한다. 불필요한 요소를 제거한 후, 핵심 정보를 추출한다. 최종적으로, 자동 추출된 결과와 사람이 수작업으로 추출한 결과를 비교하여 시스템의 성능을 검증한다.

## Ⅲ. 결론

### 1. 실험 데이터

본 연구에서는 시스템의 성능을 검증하기 위해 다양한 프롬프트를 생성하고, 이를 통해 수집된 웹사이트 데이터를 바탕으로 실험을 진행하였다. 각 프롬프트에 대해 관련 웹사이트의 데이터를 수집한 후, 제안된 시스템을 적용하여 본문을 추출하였다. 실험 데이터는 사용자 쿼리와 관련된 다양한 웹사이트를 포함하고 있으며, 이를 통해 시스템의 일반화 성능을 평가할 수 있었다.


세계 군사력 순위 '2024' 한국은 4위, 상위는 미국, 러시아, 중국 순으로

미국 군사력 평가기관 글로벌 파이어파워(Global Firepower)는 2024년 세계 군사력 순위(2024 Military Strength Ranking)를 발표하였다. 상위 3개국은 여전히 미국, 러시아, 중국으로 전년과 같다. 대한민국은 4위, 대만은 24위며 전년은 23위였다.

순위는 세계 145개국에 대해 각국의 군비와 병력, 재정상황, 지리적 조건 등 다양한 요소에서 군사력 지수(파워인덱스)를 산출해 순위를 매긴 것이다. 점수가 0에 가까울수록 군사력이 높다는 것을 보여준다.

(a) 웹 사이트 샘플

(b) 추출 본문 샘플

(그림 1) 본문 추출 표본 샘플

### 2. 실험 결과

추출된 본문 텍스트의 정확성을 평가하기 위해 텍스트 간 유사성을 측정하는 지표를 사용하였다.

<표 1> 자카드 유사도 결과

| | 자카드 유사도 | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| 프롬프트1 | 0.90 | 0.90 | 0.91 | 0.93 | 0.79 | 0.98 | 0.75 | 0.93 | 0.00 | 0.90 |
| 프롬프트2 | 0.94 | 0.00 | 0.73 | 0.89 | 0.91 | 0.65 | 0.04 | 0.92 | 0.99 | 0.89 |
| : | : | | | | | | | | | |

평가 결과, 제안된 시스템이 다양한 웹사이트에서 본문을 높은

정확도로 추출할 수 있음을 확인하였다. 자카드 유사도 결과에서 일부 낮은 값들(예: 0.04, 0.00)은 시스템이 추출한 본문과 Query 간의 내용적 차이가 크거나, 비교되는 텍스트의 길이 차이가 클 때 발생할 수 있다. 특히, 본문의 길이가 짧을 때 노이즈 및 불필요한 정보가 포함된 경우 유사도 값이 낮아지는 경향이 있다. 이러한 낮은 유사도 값은 시스템이 특정 상황에서 본문 추출의 정확도를 개선해야 할 필요가 있음을 시사한다.

결론적으로, 본 시스템은 대부분의 경우 높은 정확도를 유지하며, 최신 RAG기반 데이터 처리 성능을 유의미하게 향상시킬 수 있음을 보여준다. 그러나 특정 상황에서 발생하는 낮은 유사도 값을 개선하기 위한 추가적인 알고리즘 정교화 및 텍스트 정제 과정의 개선이 필요하다.

## 4. 결론 및 향후 과제

본 연구에서는 대형 언어 모델(LLM)의 성능을 극대화하기 위해 웹 문서에서 불필요한 데이터를 제거하고 핵심 본문을 효율적으로 추출하는 시스템을 제안하였다. 본 시스템의 목표는 웹 문서에서 의미 있는 정보를 정확하게 추출하는 것이었다. 실험 결과, 제안된 시스템이 다양한 웹사이트에서 높은 정확도로 본문을 추출할 수 있음을 확인했다. 제안된 시스템은 사용자 쿼리와 웹페이지의 HTML 태그 구조를 분석하여 본문을 식별하고, 코사인 유사도와 태그 문장화를 결합하여 최적의 본문을 추출하는 방식을 적용하였다. 이 접근 방식은 다양한 형식과 구조를 가진 웹사이트에서도 일관되게 높은 본문 추출 정확도를 달성하는 데 기여하였다.

향후 연구에서는 다음과 같은 과제를 해결할 필요가 있다. 첫째, 검색엔진 최적화 기법을 통해 사용자 쿼리에 가장 적합한 웹사이트를 더 신속하고 효율적으로 선택할 수 있는 방법을 모색해야 한다. 둘째, 더욱 정교한 본문 추출 알고리즘을 개발하여 다양한 웹 문서 형식에서도 일관된 정확도를 유지할 수 있도록 해야 한다. 셋째, 실시간 데이터 처리 성능을 개선하여 RAG의 응답 속도와 신뢰성을 높이는 것이 중요하다.

이러한 향후 과제들을 해결함으로써, 본 연구는 LLM의 최신 데이터 반영 능력을 더욱 강화하고, RAG 방법의 실용성을 증명하는 데 기여할 것으로 기대된다. 이를 통해 RAG의 응답 정확도와 신뢰성이 크게 향상될 수 있을 것이다.

## 참 고 문 헌


[1] 김휘군, 이지은, and 박상현, "인공지능 챗봇을 위한 검색 증강 생성 및 벡터 데이터베이스 최적화 연구 동향," 정보과학회지, Vol.42, No.3, pp.8-15, 2024.

[2] Nakhod, "Using retrieval-augmented generation to elevate low-code developer skills", Artificial Intelligence, 2023

[3] Melz, "Enhancing LLM Intelligence with ARM-RAG: Auxiliary Rationale Memory for Retrieval Augmented Generation", arXiv:2311.04177, 2023

[4]. H. J. Carey and M. Manic, "HTML web content extraction using paragraph tags," 2016 IEEE 25th International Symposium on Industrial Electronics (ISIE), Santa Clara, CA, USA, 2016, pp. 1099-1105, doi: 10.1109/ISIE.2016.7745047

[5] Manathunga & Illangasekara, "Retrieval Augmented Generation and Representative Vector Summarization for large unstructured textual data in Medical Education", arXiv:2308.00479, 2023

[6]. 모종훈, 유재명, "태그 서열 위치와 경사 부스팅을 활용한 한국어 웹 본문 추출" Journal of KIISE, Vol. 44, No. 6, pp. 581-586, 2017. 6

[7] Muneeswaran et al., "Minimizing Factual Inconsistency and Hallucination in Large Language Models", arXiv:2311.13878, 2023

[8] Asai et al., "Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection", arXiv:2310.11511, 2023