Knowledge Graph와 Vector Search 기반 하이브리드 검색 시스템의 설계 및 성능 평가에 관한 연구

차보경, 김지원, 권병헌, 김은호*

인텔렉투스

bg.cha@int2.us, jiwon.kim@int2.us, bh.kwon@int2.us *eunho.kim@int2.us

A Study on the Design and Performance Evaluation of a Hybrid Search System Based on Knowledge Graph and Vector Search

Cha Bo Gyeong, Kim Ji Won, Kwon Byeong Heon, Kim Eun Ho*

Intellectus

요 약

본 연구는 공공데이터 및 빅데이터 플랫폼의 정보 탐색 효율을 향상시키기 위해, Knowledge Graph(KG) 기반 시맨틱 검색과 벡터 기반 유사도 탐색을 결합한 하이브리드 검색 시스템을 설계하고, 자율주행 데이터셋을 활용하여 성능을 실험적으로 검증하였다. 기존 키워드 중심의 검색 방식은 다양한 형식의 비정형 데이터를 다루는 데 한계가 있으며, 특히 도메인 지식이 요구되는 환경에서는 탐색 효율이 현저히 낮다. 이를 해결하기 위해, 본 시스템은 KG 기반의 구조화 질의 처리와 딥러닝 임베딩을 통한 벡터 유사도 검색을 통합하였으며, 의미 기반 탐색과 정확도 향상을 동시에 달성할 수 있도록 설계되었다. 본 연구를 통해 하이브리드 검색 구조가 대규모 비정형 데이터를 다루는 환경에서 높은 실용성과 기술적 유효성을 지님을 보여준다. 향후에는 멀티모달 질의 처리 및 분산형 실시간 검색 구조로 확장하여 시스템의 범용성과 확장성을 고도화할 계획이다.

Ⅰ. 서 론

4차 산업혁명 시대에 들어서며 빅데이터는 산업, 행정, 연구 등 다양한 분야에서 핵심 자산으로 간주되고 있다. 특히 정부는 행정 데이터를 기반으로 한 공공데이터 개방 정책을 통해 정보의 민주화와 민간 활용을 촉진하고 있으며, 그 대표 사례로 공공데이터포털(data.go.kr)이 존재한다. 해당 포털에는 교통, 보건, 환경, 자율주행 등 수많은 분야의 데이터셋이 축적되어 있으며, 이는 민간 기업의 서비스 개발 및 연구에 있어 중요한 기반으로 작용하고 있다 [1][2].

하지만 현재 공공데이터 플랫폼은 사용자 친화적인 탐색 기능이나 고도화된 의미 기반 검색 기능이 부족하다는 평가를 받고 있으며, 이는 비정형 데이터가 많고 도메인 지식이 필요한 경우 탐색에 큰 제약을 초래하여 데이터 재활용률을 떨어뜨리는 주요 원인이 된다. 전통적인 검색 시스템은 대부분 키워드 기반 검색(keyword-based search)에 의존하고 있으며, 이는 단순 문자열 매칭에 기반한 결과를 반환하는 구조다. 이러한 방식은 자율주행, 교통, 제조 등 복합적이고 비정형 데이터가 혼재된 도메인에서 유의미한 결과를 도출하기 어렵다는 한계가 있다 [3][4].

이러한 문제를 해결하기 위한 방안으로, 최근에는 Knowledge Graph(KG)를 기반으로 한 시맨틱(semantic) 검색 기술이 주목받고 있다. KG는 개체 간의 의미적 관계를 그래프 형태로 구조화하여, 단순 키워드 일치가 아닌 개념 중심의 탐색과 사용자 맞춤형 정보 추천을 가능하게 한다 [5][6].

특히 최근에는 KG 기반의 검색에 딥러닝 기반 임베딩 모델을 결합하여 구조화된 정보와 비정형 텍스트를 통합적으로 반영하는 하이브리드 검색(hybrid search) 기법이 등장하고 있다. 하이브리드 검색은 벡터 임베딩을 통한 의미 유사도 기반 검색과 KG 기반 정형 탐색을 융합하여, 검색 정확도와 탐색 다양성을 동시에 확보할 수 있다는 장점을 가진다.

본 연구에서는 이러한 기술적 접근을 바탕으로, 자율주행 빅데이터 기반 포털에 적용 가능한 KG 기반 시맨틱 검색 및 벡터 검색 기반 자연어 질의 처리 시스템을 구현하고, 실제 환경에서의 성능 평가를 통해 다양한 도메인으로의 실용적 확장 가능성을 검토하고자 한다.

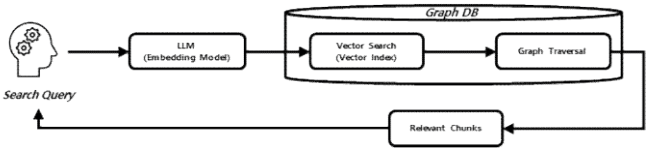

Fig.1.1 Hybrid Search 구조

Ⅱ. 본론

2.1 Knowledge Graph 성능 실험 설계

텍스트 질의어를 기반으로 Neo4j에서 관련 노드를 탐색하는 쿼리를 FastAPI 기반의 API를 통해 실행하였으며, 해당 Knowledge Graph는 클라우드 서버에서 수행하였다. 성능 실험 환경은 아래 Table.2.1.1과 같다.

Table.2.1.1 Test Spec.

No.	Item	Spec.
1	API	FastAPI
2	Query	Neo4j
3	OS	Ubuntu 24.04
4	vCPU	2
5	Clock Speed	2.3 GHz(Up to 3.3 GHz)
6	Memory	8GB
7	Storage	32GB

2.2 자연어 기반 검색 성능 실험 설계

자연어 기반 검색 성능 평가를 위해 자율주행 데이터셋 일부를 활용하여 정답/오답 이미지로 구성된 테스트셋을 구축하였다. 사용자의 자연어 질의를 컨텍스트 리스트로 변환한 후, 검색 결과에 대해 Precision, Recall, F1-score를 기준으로 성능을 측정하였으며, 테스트 환경은 KG 성능 실험 환경과 동일하다.

2.3 Knowledge Graph 성능 실험 방법

Knowledge Graph 성능 실험을 위한 방법은 아래와 같다.

1. 트랜젝션 정의

- 사용자가 텍스트 질의를 전송하고, Knowledge Graph에서 관련 노드를 탐색하여 결과를 반환받는 과정

2. TPS(Transactions per Second)

$$TPS = \frac{Nmber\ of\ Active\ Users}{Average\ Response\ Time(\sec)}$$

3. 시스템 설정 :

- 동시 사용자 수(Number of Active Users) : 350
- Ramp-up Period : 5s
- 총 10회 반복 수행, Time-stamp 기반으로 평균 TPS 산출

4. 검색 쿼리 구성

- 실험의 반복성과 일반화를 위해 검색어는 Pedestrian, Vehicle 두 가지로 설정

2.4 자연어 기반 검색 성능 실험 방법

자연어 기반 검색 성능 실험을 위한 방법은 아래와 같다.

1. 테스트셋 구성

- 총 100장의 자율주행 데이터셋 이미지 중, 보행자(pedestrian)가 포함된 정답 데이터 50장과 포함되지 않은 오답 데이터 50장 선별

2. 질의 처리 방식

- 사용자의 자연어 질의를 입력, 이를 사전에 정의된 컨텍스트 리스트 변환및 검색

3. 검색 결과 평가 방법

- 검색 결과를 바탕으로 Confusion Matrix를 구성하고, 아래 지표 계산

(1) Precision (정밀도) = TP / (TP + FP)

(2) Recall (재현율) = TP / (TP + FN)

(3) F1-score = 2 * (Precision × Recall) / (Precision + Recall)

2.5 Knowledge Graph 성능 실험 결과

KG 검색 성능 실험 결과 Pedestrian 검색 평균 TPS는 68.8, Vehicle은 69.0으로 나타났으며, 전체 평균 TPS는 68.9로 계산되었다. 이는 실험 환경 기준으로 실시간 검색이 충분히 가능한 수준의 처리 성능을 확보했음을 보여준다.

Table.2.5.1 Pedestrian 검색 결과

Label	Average	Std. Dev.	Through put	Received KB/sec	Sent KB/sec	Avg. Bytes
HTTP Request	191	36.72	**66.2502**	33.71	9.83	521
HTTP Request	177	11.81	**69.2658**	35.24	10.28	521
HTTP Request	217	44.52	**68.7758**	34.99	10.21	521
HTTP Request	191	23.21	**69.2247**	35.22	10.28	521
HTTP Request	181	13.42	**69.2795**	35.25	10.28	521
HTTP Request	189	23.72	**69.1153**	35.17	10.26	521
HTTP Request	190	19.67	**69.1017**	35.16	10.26	521
HTTP Request	186	23.96	**68.6006**	34.9	10.18	521
HTTP Request	179	12.1	**69.0335**	35.12	10.25	521
HTTP Request	189	20.54	**69.3619**	35.29	10.3	521
TOTAL	189	22.967	**68.8009**	35.005	10.213	521

Table.2.5.2 Vehicle 검색 결과

Label	Average	Std. Dev.	Through put	Received KB/sec	Sent KB/sec	Avg. Bytes
HTTP Request	180	11	**69.1973**	41.29	10.07	611
HTTP Request	186	21.92	**68.857**	41.09	10.02	611
HTTP Request	179	15.19	**69.2658**	41.33	10.08	611
HTTP Request	177	11.26	**69.1836**	41.28	10.07	611
HTTP Request	189	26.25	**69.1836**	41.28	10.07	611
HTTP Request	178	11.66	**69.1017**	41.23	10.05	611
HTTP Request	192	19.84	**69.0744**	41.22	10.05	611
HTTP Request	181	12.54	**69.129**	41.25	10.06	611
HTTP Request	187	20.41	**67.9744**	40.56	9.89	611
HTTP Request	179	11.2	**69.2384**	41.31	10.07	611
TOTAL	182.8	16.127	**69.0205**	41.184	10.043	611

2.6 자연어 기반 검색 성능 실험 결과

자연어 기반 검색 성능 평가는 보행자(Pedestrian) 포함 여부를 기준으로 수행한 결과는 아래 Table. 2.6.1과 같으며, 이는 자연어 기반 질의 처리 방식이 실제 응용 환경에서 높은 정확도로 동작함을 의미한다.

Table. 2.6.1 자연어 기반 검색 성능 결과

Indicators	Value
Predicted	62
True Positive(TP)	45
False Positive(FP)	17
False Negative(FN)	5

Ⅲ. 결론

본 연구에서는 Knowledge Graph(KG) 기반의 시맨틱 검색 구조에 Vector Embedding 기반 유사도 탐색을 결합한 하이브리드 검색 시스템을 설계하고, 자율주행 빅데이터 환경에서의 성능을 실험적으로 검증하였다. Neo4j 기반 그래프 데이터베이스와 FastAPI 기반 검색 API를 연동한 시스템을 구축하고, "pedestrian" 및 "vehicle" 키워드를 활용한 구조화 질의 실험과 자연어 기반 질의 실험을 각각 수행하였다.

실험 결과, 평균 TPS는 약 68.9로 측정되어 동시 사용자 350명 환경에서도 안정적인 처리 성능을 보였으며, 자연어 질의 기반 검색에서는 Precision 0.7258, Recall 0.9, F1-score 0.8036의 성능을 기록하여 의미 기반 탐색의 실효성을 확인할 수 있었다. KG 기반 검색의 의미 확장성과, 임베딩 기반 벡터 유사도 검색의 정밀성이 결합되면서, 복합적인 검색 니즈를 충족시키는 통합 검색 구조로서의 가능성을 입증하였다.

이러한 결과는 하이브리드 검색 시스템이 대규모 비정형 데이터를 다루는 공공데이터 및 빅데이터 플랫폼에서 검색 정확도 향상, 사용자 경험 개선, 데이터 활용도 증대에 있어 유의미한 기술적 대안이 될 수 있음을 시사한다.

향후 연구에서는 자연어와 이미지 등 멀티모달 질의 처리, 벡터 DB 기반의 실시간 분산 검색 시스템 구현 등을 통해 시맨틱 검색 시스템의 범용성과 실효성을 한층 더 고도화할 예정이다.

ACKNOWLEDGMENT

이 논문은 정부(과학기술정부통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임 (RS-2023-00235293, 활용 목적에 따른 자율주행 데이터 제공을 위한 자율주행 빅데이터 가공/관리, 검색 및 공유 인터페이스 기술 개발)

참 고 문 헌

[1] National Information Society Agency (NIA), “Technical and Policy Trends Report for Promoting Public Data Utilization,” NIA Policy Research Report, 2020.

[2] OECD, "Open Government Data Report: Enhancing Policy Maturity for Sustainable Impact," OECD Publishing, 2019.

[3] OECD. (2021). The Path to Becoming a Data-Driven Public Sector.

[4] Wang, H., Zhang, F., Xie, X., and Guo, M., "A comprehensive survey of knowledge graph-based recommender systems," Information, vol. 12, no. 6, pp. 232, Jun. 2021.

[5] AIMultiple. (2024). Knowledge Graphs: Use Cases & Benefits.

[6] Stardog, "How to build a semantic search engine using a knowledge graph," Stardog Blog, 2024.