지식그래프
시장현황과 사업모델

인포시즈
책임연구원 정이태

왜 지식그래프가 요즘 더욱 관심을 받게 되었을까요 ?

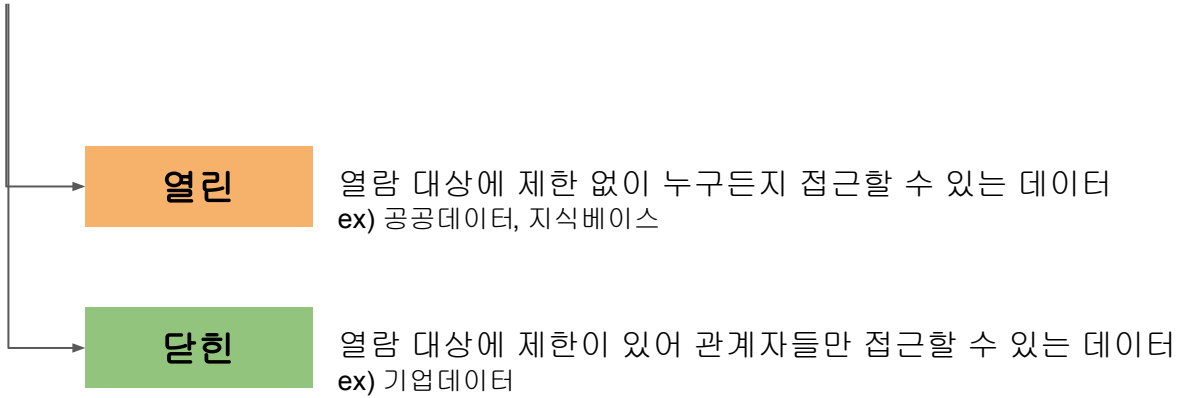
→ 이전대비 빨라진 변화(고객의 니즈)의 속도

데이터 수집・관리 기술이 발달하면서, 기업내부 데이터는 쌓여가고 있으나 어디에 어떻게 활용할지에 대해 고민이 많음.

트렌드를 파악하고 데이터를 접목하려는 순간 트렌드는 변해가고 있음.
흩어져 있는 데이터들을 통합하여 적재적소의 비즈니스 상황에 활용하는것.

What is a Knowledge Graph?

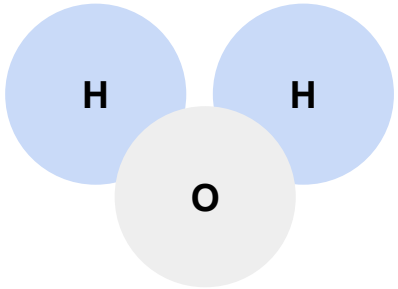
지식을 그래프 형태로 표현한 데이터.



What is a Knowledge Graph?

H_2O

H-O-H



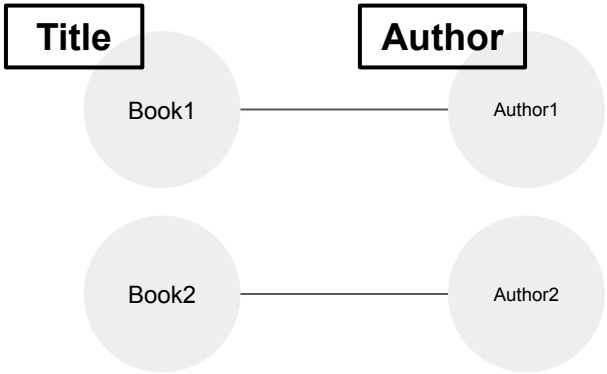
모두 물을 표현하는 방식임. 의미하는 바는 모두 동일하나 표현의 방식이 다름.

이처럼 현실세계 데이터들도 모두 의미하는 바는 동일하나 표현하는 방식이 다르기에,
이를 통합할 도구가 필요함.

Semantic Web

서로 다른 사람들이 서로 다른 종류의 정보를 표현할 수 있도록 서로 다른 도구를 조합할 수 있음.
표현력의 깊이가 높고 낮음에 따라 사용하는 도구가 다르기에
목적에 맞는 도구를 선택하여 지식을 표현하는것이 핵심임.

What is a Knowledge Graph?



Turtle format 열린

```
@prefix ex: <http://example.org/> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix foaf: <http://xmlns.com/foaf/0.1/> .

ex:Book1 rdf:type ex:Book ;
    ex:title "Book 1" ;
    ex:author ex:Author1 .

ex:Book2 rdf:type ex:Book ;
    ex:title "Book 2" ;
    ex:author ex:Author2 .

ex:Author1 rdf:type foaf:Person ;
    foaf:name "Author 1" .

ex:Author2 rdf:type foaf:Person ;
    foaf:name "Author 2" .
```

RDF style

-Turtle은 구조화된 데이터를 가독성 좋게 표현하기 위한 형태

-@prefix는 긴 웹 주소의 짧은 이름(URI, 분별자)을 정의하는 데 사용함.

-**ex:, rdf:, foaf:** 같은 네임스페이스 접두사는 RDF 데이터의 카테고리를 지정하거나 엔티티 혹은 프로퍼티를 표현하는데 활용함.

csv format 닫힌

Title	Author	Rel
Book 1	Author 1	##
Book 2	Author 2	###

LPG style

- 그래프를 표현할 때 자주 활용하는 형태. 출발지, 도착지 2가지 컬럼으로 표현함.

- 관계에 대한 맵핑을 하고자할땐 3가지 컬럼으로 표현함.

What is a Knowledge Graph?

	장점	단점
열린	스키마가 구축되어 있어, 스키마에 적절하게 구축하면 되기에 지식 통합에 용이함.	스키마가 지정되어 있지않다면, 직접 지정**(온톨로지구축)**을 해주어야 함.
닫힌	기업 데이터를 저장 관리하기에 편리하기에 범용성이 높음.	스키마가 지정되어 있지않아, 데이터 통합시에 데이터 설계자 데이터 엔지니어 데이터 분석가 등 여러 이해관계자들을TF 로 배치하여 데이터 통합에 공수를 들여야함.

대다수 기업들은 ‘닫힌’데이터 형태로 데이터를 보유하고 있으며, 데이터간 통합을 통해 **비즈니스 가치를 도출**하고 싶어함.

Market Trends

은행별 생성형 AI 준비 현황 · 특징

KB국민은행	• 'KB-GPT' 데모 웹사이트 개설 • 검색, 채팅, 요약, 문서작성, 코딩 등 내부 직원 업무 효율성 제고에 적용 검토
신한은행	• 생성형 AI의 금융서비스 적용 전담 TF출범 • 글로벌 기업 외 KT, 네이버, LG 등 다양한 국내기업과 실증
하나은행	• 자체 금융 특화 버티컬 거대언어모델(LLM) 개발 • 하나금융융합기술원 언어모델 전문가의 참여 • 내년 고도화 예정인 하나은행 모바일 AI뱅커 등에 활용
우리은행	• 비정형데이터 자산화해 금융언어모델에 적용 • LG AI연구원과 금융언어모델 실증 위한 컨소시엄 운영함 • 연말에 챗봇 등부터 대고객 서비스 순차적 적용 시작
NH농협은행	• 구글 바드, 챗GPT 등 활용한 금융언어모델 실증 • 올바른 AI 활용 위한 AI거버넌스 수립 • 생성형 AI 기반 영업점 AI은행원 서비스 도입 고려

내부 : 업무 효율성 제고를 위한 **챗봇**

외부 : 고객 만족도 제고를 위한 **챗봇**

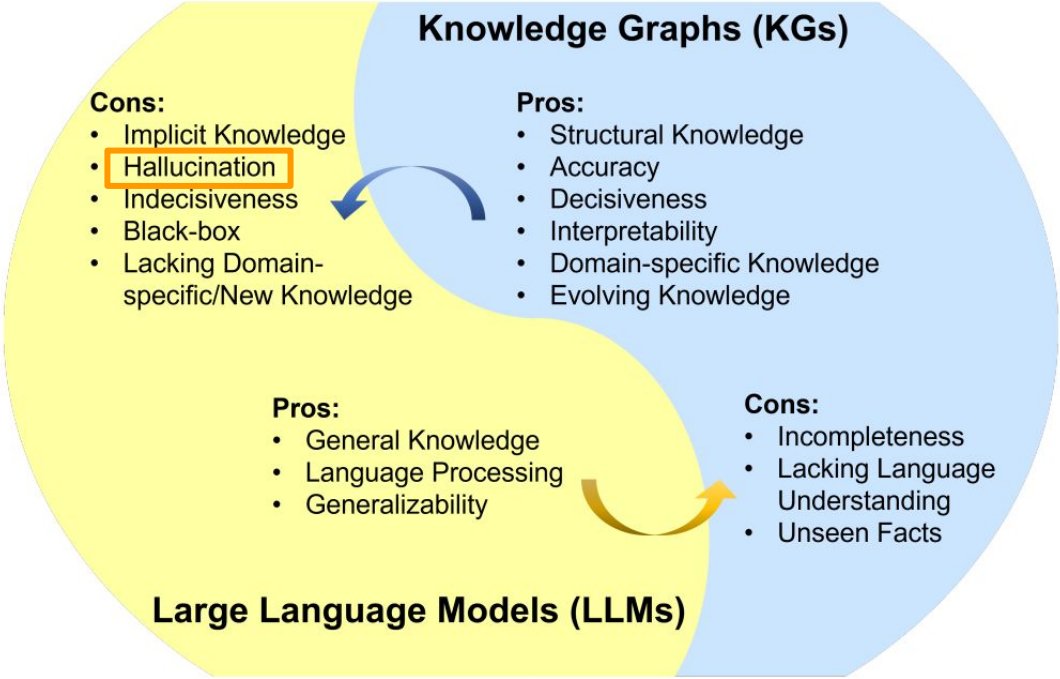
[스페셜리포트] 금융권에 부는 생성형 AI바람.. 5대 은행 '각양각색' , 2023-07-25, 정예린 기자, 전자신문

지식그래프와 생성형 AI (LLM)이 무슨 상관입니까?

LLM hallucination(환각현상)의 보완재

REASONING

Market Trends



환각현상은 모델이 실제로는 정확하지 않거나 **사실이 아닌 정보를 생성할 수 있는 경우**를 가리킵니다.

환각현상(Hallucination) 문제가 되는 이유 네가지.

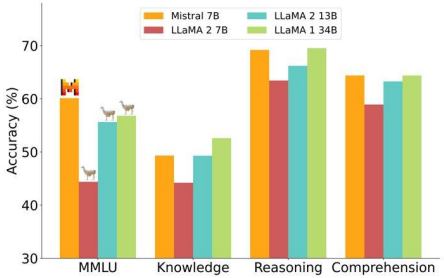
1. 정보의 신뢰성과 오류: 생성형 AI 모델은 훈련 데이터에 기반하여 텍스트를 생성하며, 훈련 데이터에 포함된 내용 중에서 나온 내용을 재생산할 수 있습니다. 이것은 **정보의 신뢰성에 대한 문제를 야기할 수 있으며, 잘못된 정보**를 퍼뜨릴 수 있습니다.

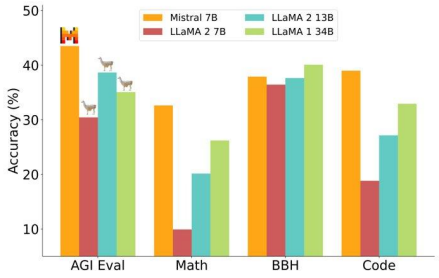
2. 보안 문제: 생성형 AI는 매우 큰 데이터베이스에서 학습하고 다양한 주제의 정보를 생성할 수 있습니다. 때때로 모델은 기반 데이터베이스의 정보에 제한 없이 접근하여 **민감한 정보나 심각한 주장을 생성**할 수 있습니다. 이러한 상황에서 환각현상은 모델이 허구 정보나 공격적인 주장을 만들 수 있는 잠재적 위험성을 내포합니다.

이러한 이유로 생성형 AI 모델을 개발하고 사용할 때 환각현상에 대한 주의가 필요하며, 모델이 생성한 내용을 신중하게 검토하고 필요한 경우 **인간의 감독을 통해 수정**해야 할 수 있습니다.

Pan, Shirui, et al. "Unifying Large Language Models and Knowledge Graphs: A Roadmap." arXiv preprint arXiv:2306.08302 (2023).

Market Trends





** https://mistral.ai/news/announcing-mistral-7b/

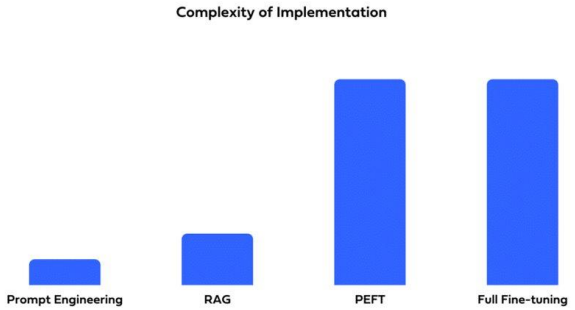
Model	Average
kyujinpy/KoR-Orca-Platypus-13B	50.13
42MARU/GenAI-llama2-ko-en-platypus	49.81
krevas/LDCC-Instruct-Llama-2-ko-13B-v4	49.58
kyujinpy/KoT-platypus2-13B	49.55
jyoung105/ko-platypus2-collective-13b	47.94
kyujinpy/KO-Platypus2-13B	47.9
kyujinpy/Korean-OpenOrca-13B	47.85
siryuon/KOEN-13B	47.84
kiyoonyoo/ko-platypus-13b-control	47.66
42MARU/GenAI-llama-2-ko-en-instruct-v1	47.28

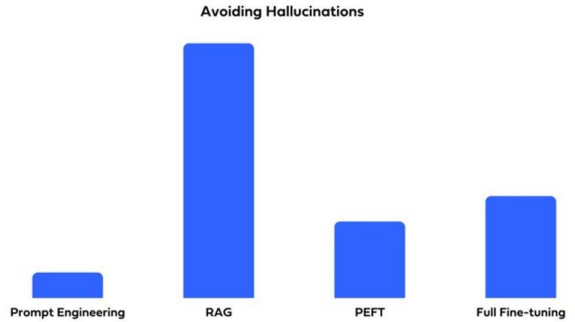
** KoLLM leaderboard

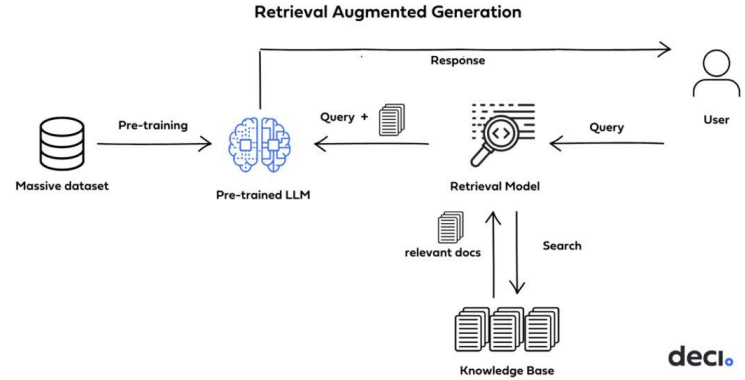
- 일반 ChatGPT API를 활용하면 고객, 기업 정보가 외부로 넘어가기 때문에 데이터 유출의 위험이 있음.이를 방지하기 위해서는 자체 모델을 구축해야함. 대다수 조직마다 자체 모델을 구축하거나 언어모델 기술이 있는 기업에게 의뢰를 함. 그러나, 이 두 방안 모두 모델 유지보수가 만만치 않음.

- 다양한 채널을 통해 수집된 데이터를 일정 주기 혹은 실시간으로 반영하고자 재학습하자니 학습비용과 소요시간이 크기에 부담이 됨. 또한, 성능 개선을 위해 알고리즘을 적용 과정에서 소요되는 비용과 인력이 부담됨.

Market Trends



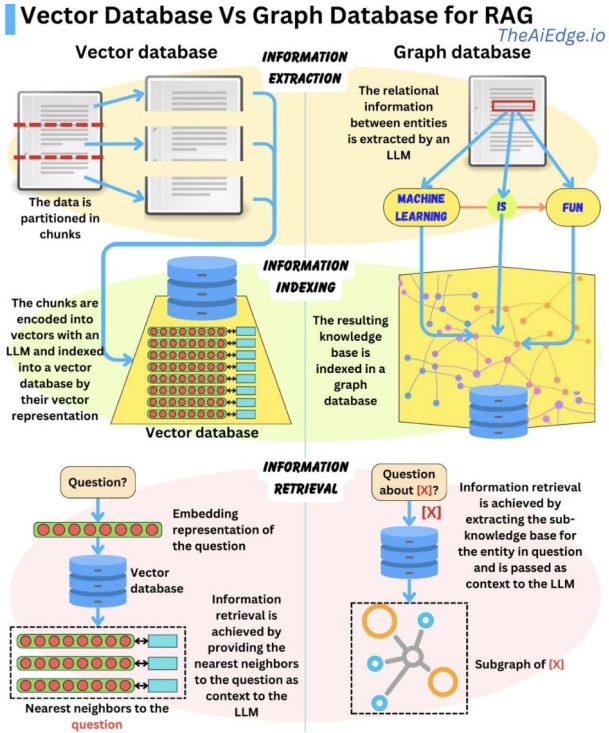




대형언어모델을 구축하는 대표 4가지 방법 중 **1.환각 현상에 강건하고, 2. 구현 복잡도가 낮은** RAG(Retrieval Augmented Generation)가 주목을 받는 중임.

* Full Fine-Tuning, PEFT, Prompt Engineering, and RAG: Which One Is Right for You?

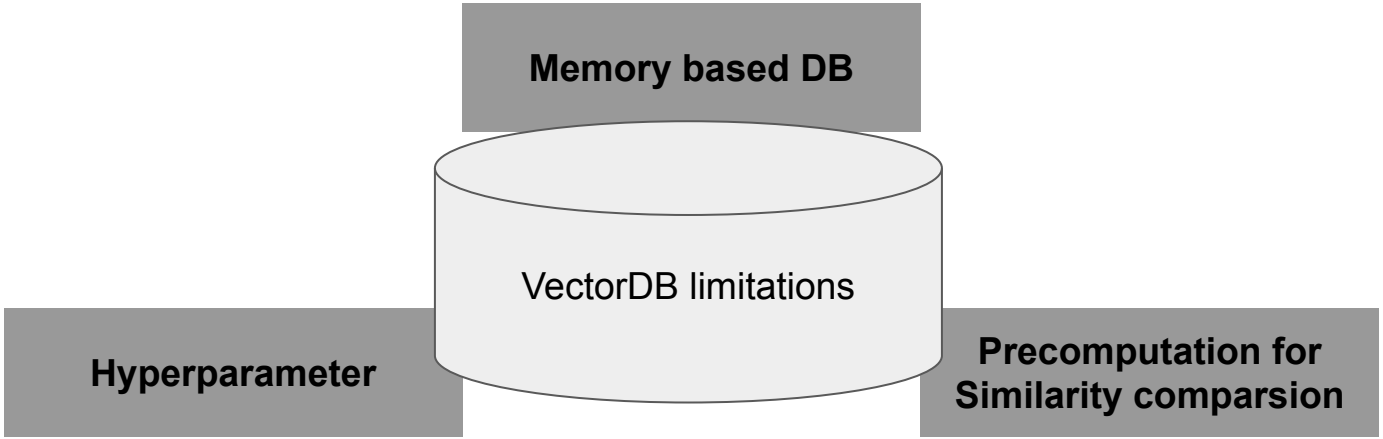
Reference architecture



LLM의 가중치 값을 보관하기 위해 **VectorDB**를 활용하는게 일반적임. 가중치를 효율적으로 관리 및 활용한다는 관점에서 유용함.

Linkedin, Damien Benveniste, PhD

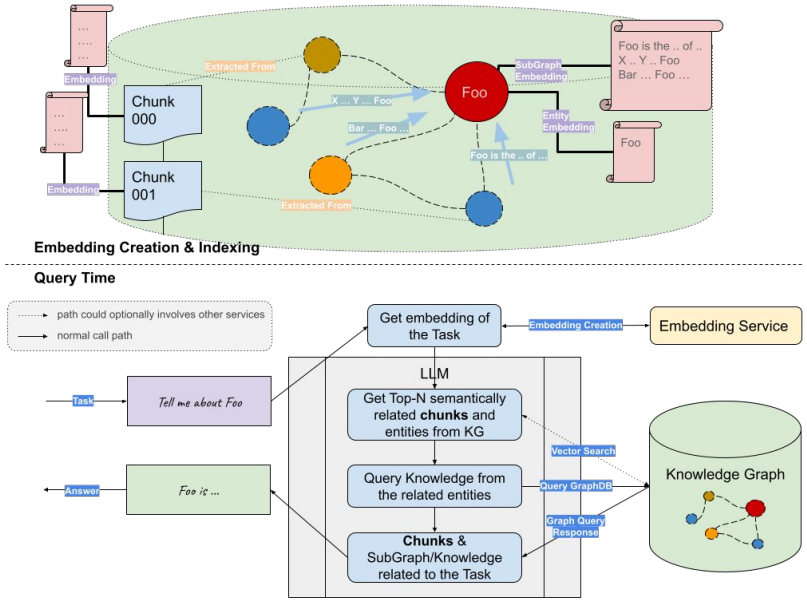
Reference architecture



Reference architecture

하지만, 정확한 답을 도출하는게 아닌 유사도에 기반하여 유사한 값을 도출하는것이기에 정확도에 대한 니즈가 있음.

→ 지식그래프



[NebulaGraph Launches Industry-First Graph RAG: Retrieval-Augmented Generation with LLM Based on Knowledge Graphs]

Challenge

1.한국어

한국어 고유 특성인 교착어 , 어순 , 띄어쓰기 의문문 평서문 등이 오히려 불규칙적인 패턴을 야기하여 데이터 학습하기에 까다로운 케이스 발생.

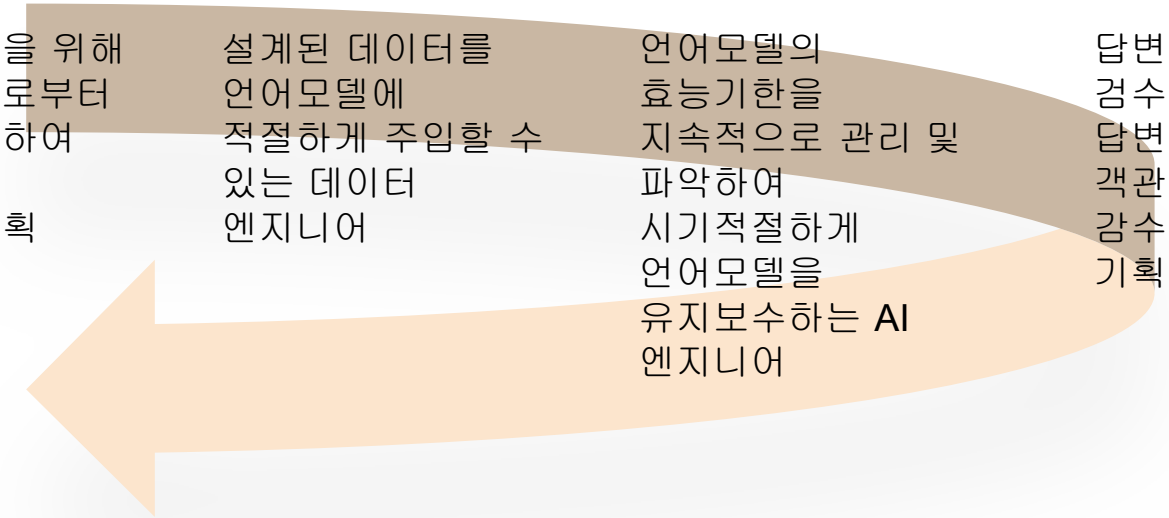
2.도메인 데이터 통합 체계 지속성 미비

조직마다 도메인에 대한 지식을 모두 사람이 기록해놓고 관리하는 형식으로 시스템이 구성되어 있음. 그렇기에, 데이터 통합시 데이터에 대한 인사이트나 노하우가 있는 직원이 참여하여 구축하는게 이상적이나, 현실은 1. 빠른 데이터 및 트렌드 변환 2. 담당직원 부서변경 등의 이유가 있어 지속적인 데이터 체계 구축에 어려움을 겪음.

3.지식 그래프 연결 끊김

실제 지식 노드간 연관이 있으나, 연관 되어 있는 지식 노드 사이에 관련없는 노드 혹은 관계가 있을시 이를 탐지하지 못하여 발견하지 못하는 경우 발생.

Business Model



온톨로지 설계자

데이터 통합을 위해 기존 데이터로부터 정보를 추출하여 온톨로지를 설계하는 기획 엔지니어

데이터 관리자

설계된 데이터를 언어모델에 적절하게 주입할 수 있는 데이터 엔지니어

언어모델 관리자

언어모델의 효능기한을 지속적으로 관리 및 파악하여 시기적절하게 언어모델을 유지보수하는 AI 엔지니어

답변 품질 검수 및 기획자

답변에 대한 품질을 검수하여 적절한 답변인지에 대해 객관적으로 감수하는 외부 기획자

Conclusion

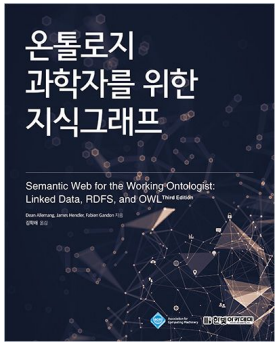
- 데이터 통합, 통합으로 부터 도출된 지식을 대형 언어모델에 추가 반복 주입하는 방식이 트렌드임.

- 유사도를 기반으로 언어모델의 답변을 추가반복 주입하기 때문에 답변의 부정확성 측면이 문제로 대두됨.

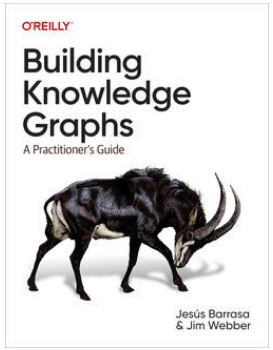
- 정확성을 향상하기 위해 지식을 체계화하여 관리하는 데이터 구조인 지식그래프 데이터가 각광받고 있음.

지식그래프 공부하기 좋은 레퍼런스

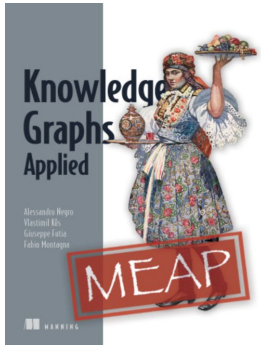
Book



온톨로지 기본서,
온톨로지에 대해
기본부터 심화까지
공부하고 싶다면



자연어처리와
그래프
데이터베이스
기술을 곁들여서
지식그래프를
형성하는 방법을
공부하고 싶다면



지식그래프를
속성으로 공부하고
실무에 어떻게
적용할지 코드와
설명을 토대로
공부하고 싶다면

Medium

Tomaz Bratanic

6.3K Followers

Data explorer. Turn everything into a graph. Author of Graph algorithms for Data Science at Manning publication. http://mng.bz/GGVN

실무와 트렌드를
겸비한 지식그래프
테크닉을
보고싶다면

Ontotext

1.1K Followers

Ontotext is a global leader in enterprise knowledge graph technology and semantic database engines.

산업계에서 어떻게
지식그래프가
활용되는지 실제
기업적용사례를
보고싶다면

지식그래프 플랫폼

Name	Company Description	Funding	Difference from Other Companies
Stardog Union	Cloud and AI based data fabric and knowledge management platform for decision making	$32.5M	Data graph platform, data lake acceleration, operational risk management, semantic data search. Serves financial services, life sciences, and manufacturing.
Maana	Provider of analytics and knowledge graph platform for multi-industry applications	$90.2M	Crawls, indexes, mines, enriches, joins, classifies, analyzes, clusters, connects, and correlates data. Serves logistics, finance, oil & gas, risk & compliance, and more.
Cambridge Semantics	Data Analytics and Data management platform	$8.8M	Configures a data lake for scalable graph storage, model-driven analytics, and more. Used by pharma, financial services, government, healthcare, retail, and media.
metaphacts	Online software for knowledge graph management	-	Enables knowledge workers to create and gain insight into data. Clients include Siemens, Sanofi, The British Museum, and others.
Shenqing Information Tech	AI company offering financial content management and technology solutions	$15.5M	Collaborates with data giants to provide differentiated products and services to enterprise customers based on industry big data.
Capsenta	Enterprise Data Platform for Virtual Data Integration	-	Uses advanced graph representation and semantic technologies to virtually integrate diverse data sources. Recognized by Gartner.
Franz	Enterprise graph, knowledge and database management software		Uses artificial intelligence for semantic search, builds a graph database, and provides actionable insights. Serves healthcare, defense, IT, finance, manufacturing, and pharmaceutical industries.
PoolParty	AI-based platform for knowledge graph management	-	Provides software to manage taxonomy, ontology, metadata, and structured data. Offers tools for information extraction, classification, content search, and semantic analysis.
Cubeit	Content organization and management tool	$3M	Allows users to collect, organize, share, and save data. Converts data into cards, uses proprietary algorithm. Acquired by Myntra.
DeepReason.ai	Enterprise AI-based knowledge graph management platform provider	-	Applies reasoning to enterprise data to discover insights and opportunities. Can connect to relational and graph databases in multiple data formats.

데모 케이스가 잘 작성되어 있는 지식그래프 플랫폼.

** https://tracxn.com/d/trending-themes/startups-in-knowledge-graph-platforms/__IOc2CE9ngaQqBbDOinaHyr3Q6Ta8lPnpb0l87z3TnKI

ETC

- 공급사슬망 빠른 의사결정을 위한 지식그래프.
- 이상거래탐지 및 분석을 위한 지식그래프.

Q&A