

Contents

우리은행 비정형 데이터 자산화 시스템(2022 vs 2024) 기술 비교 종합 보고서	1
개요	1
1. 2022년 비정형 데이터 자산화 시스템: 기술 구성 및 특징	1
임베딩 모델	1
벡터 검색엔진 및 벡터 DB	1
시스템 구조 및 검색 파이프라인	2
성능	2
한계	2
2. 2024년 비정형 데이터 자산화 시스템: 기술 구성 및 특징	2
임베딩 모델	2
벡터 검색엔진 및 벡터 DB	2
Retrieval 파이프라인 및 고급 기능	3
시스템 구조 및 운영	3
특장점	3
3. 기술 변화의 이유 및 개선 효과	3
임베딩/검색 모델 변화 동기	3
시스템/비즈니스 효과	3
4. 구성 요소별 강점/특징 정리	3
임베딩 모델	3
벡터 검색엔진/벡터 DB	4
검색 파이프라인	4
5. 미기재/OA 구성요소 및 한계	4
결론	4
Sources	4

우리은행 비정형 데이터 자산화 시스템(2022 vs 2024) 기술 비교 종합 보고서

개요

이 보고서는 2022년과 2024년 우리은행의 비정형 데이터 자산화 시스템 제안서에 명시된 임베딩 모델, 벡터 검색엔진, 벡터 데이터베이스(Vector DB)의 기술 구성 요소를 정밀 비교한다. 각 연도별로 도입된 주요 기술의 선정 배경, 아키텍처적 특징, 성능 및 적용상의 특장점, 변경 원인과 개선효과, 공식 문서에 기재된 강점까지 총체적으로 분석하였다. 또한 상세한 기술 진화의 이유와 실제 업무·비즈니스 활용성 측면에서의 영향도 심층적으로 다루었다.

1. 2022년 비정형 데이터 자산화 시스템: 기술 구성 및 특징

임베딩 모델

- **Airbill BERT (한국어 특화)**
 - SK C&C의 한국어 전용 트랜스포머 기반 모델 도입.
 - 우리은행 업무 구문 및 금융 용어 이해에 최적화된 구조.
- **Accu.TA Transformer 기반 DPR/DHR**
 - Accu.TA 한국어 트랜스포머로 Dense Passage Retrieval(DPR) 모델 및 Dense Hierarchical Retrieval(DHR) 모델을 구성.
 - 도메인 파인튜닝을 통해 우리은행 실업무 질의·문서에 맞춤 학습.

벡터 검색엔진 및 벡터 DB

- **FAISS (Facebook AI Similarity Search)**
 - 오픈소스 고속 벡터 인덱싱 및 검색 라이브러리.

- 벡터+ID만 저장하는 단순 구조, 주기적 파일 관리로 실시간성/확장성 제한.
- 메타데이터 조건 검색, 복합 필터 불가.
- 별도의 전문 Vector DB는 미구축. FAISS가 벡터 인덱스와 검색을 담당.

시스템 구조 및 검색 파이프라인

- 사용자가 입력한 자연어 질문을 DPR 모델이 임베딩하여, Vector IndexDB(문단 단위)에 저장된 임베딩과 벡터 유사도 검색.
- 키워드 기반 텍스트 검색과 양상블(Ensemble) 및 Re-ranking 조합 처리.
- GPU 서버 기반으로 고속 연산을 지원함.
- 1시간 주기로 인덱스/벡터 동기화 운영.

성능

- BM25(전통 IR): Accuracy@Top20 = 54.08
- DPR(일반): 83.67
- DPR(우리은행 데이터 학습): 92.1로 38p 증가(BM25대비), 8p 증가(비학습 DPR 대비)
- 실제 업무 데이터에서 고정밀 자연어 검색 성능 확인

한계

- 전체 문장/문맥 이해 취약(키워드 기반).
- 메타데이터 조건 검색 미지원.
- Long-tail 질의, 복잡 쿼리 처리 한계.
- 인덱스/검색의 실시간성, 확장성에 제약.

[참고: 1, 2, 3, 4]

2. 2024년 비정형 데이터 자산화 시스템: 기술 구성 및 특징

임베딩 모델

- **M3**
 - Dense, Sparse, Multi-vector 혼합 임베딩 지원(하이브리드 검색 최적).
 - 100개 이상 다국어 지원, 도큐먼트마다 복수 벡터로 고차원 서치 가능.
- **E5 (Microsoft)**
 - 다국어 텍스트 쌍 사전학습, MTEB(Massive Text Embedding Benchmark) 리더보드 1위권.
 - 자연어 의미론적 정확성, Text Pair 검색 최적화.
- **성능 수치**
 - MIRACL 16개 언어 기준: nDCG@10(검색정확도) 66.5(MTEB 기준 최고 수준), R@100(재현율) 94.3

벡터 검색엔진 및 벡터 DB

- **Qdrant**
 - 오픈소스/엔터프라이즈 벡터 DB·검색엔진, 메타데이터 필터·복합조건 지원.
 - Mixed Retrieval(하이브리드: Dense+Sparse), 실시간성·확장성 우수.
 - Write Ahead Log, 클러스터 분산지원, 고가용성/운영 관리 기능 내장.
- **비교 벤치마크(자체 및 공식):**
 - 업로드/인덱스 처리, TPS, Latency 등 주요 항목에서 업계 최고 수준(RPS 1238, Latency 3.54ms, 5천만건 규모 실측).

Retrieval 파이프라인 및 고급 기능

• 고급 RAG(Advanced Retrieval-Augmented Generation)

- Mixed Retrieval(복합벡터), 최신성 필터(Recency Filtering), Re-rank(ko-reranker: 한글 특화 재정렬), Semantic Router(질의 유형별 서치경로 자동 선택), ARAGOG(결과 증강).
- 전체 프로세스: 사용자 질의 → 임베딩(M3/E5) → Qdrant 벡터검색 → 오거멘트/재정렬 → LLM 답변.

시스템 구조 및 운영

- 지식DB → 데이터 표준화/분할/메타데이터 적용 → 임베딩 → 벡터 DB/Qdrant → RAG 파이프라인 → LLM 서비스 연동.
- 멀티도메인, 멀티채널, 다양한 업무 서비스에 간편 확장, 실시간 신규 데이터 적용.

특장점

- 메타정보 저장/필터링(부서, 날짜, 키워드, 권한 등), 하이브리드 검색, 실시간/대용량 확장성, 엔터프라이즈 관리 편의.
- 최신 오픈소스 생태계 활용, 모델 및 시스템 유지보수와 서비스 확장 용이.

[참고: 5, 6, 7]

3. 기술 변화의 이유 및 개선 효과

임베딩/검색 모델 변화 동기

- 2022년의 국산(한국어 특화) DPR모델 및 FAISS 체계는 도메인 특화 성능은 우수했으나,
 - ① 멀티랭귀지 및 도메인 범용성 부족,
 - ② 문맥 이해 한계(키워드 치중, 의미론적 약점),
 - ③ 메타데이터/조건 기반 검색 미지원,
 - ④ 모델 확대·유지보수·확장성의 제약 등으로 협업 및 신규 서비스 확장에 한계가 있었다.
- 2024년 M3, E5 도입의 직접적 이유:
 - 한국어만이 아닌 다양한 언어, 다양한 업무 데이터, 복수 벡터 활용으로 복잡 질의·장문·특이도 높은 질의도 고정밀 검색 가능.
 - MTEB, MIRACL 등 공식 벤치마크에서 현저히 높은 성능, 글로벌 호환성, 최신 AI검색(RAG, LLM 연동)에 빠르게 대응 가능.
- Qdrant 도입 배경:
 - FAISS 대비 내장 운영관리, 메타데이터 저장/검색, 대량데이터 실시간 처리, 장애대응 편의 등 엔터프라이즈 업무의 요구를 충족.
 - 다중 조건 검색, 분산 확장, 서비스 SLA 강화, 고가용 등 실업무 인프라 요구를 충족.

시스템/비즈니스 효과

- 검색 정확도·재현율 대폭 향상(2022: DPR@Top20 92.1 → 2024: E5 nDCG@10 66.5/16개 언어, R@100 94.3).
- 신규 채널·서비스(예: 리서치, 보고서 생성, 사내외 포털 등) 손쉬운 연계.
- 사용자별, 업무별, 기간별·복합조건·권한 기반 검색, 분석보고 등 다양화 가능.
- 다운타임·장애율 저감, 모델 유지관리 용이성 및 체계적 서비스 품질관리.

[참고: 6, 5, 2]

4. 구성 요소별 강점/특징 정리

임베딩 모델

- Airbill BERT/Accu.TA DPR (2022)

- 한국어 비정형 문서, 금융 도메인 특화, 실업무 질의에 최적.
- 전통 키워드 방식(예: BM25) 대비 자연어 질의 정확도 대폭 향상.
- **M3**
 - 하이브리드(복수 벡터) 임베딩, 100+ 다국어 지원, 도큐먼트별 다양한 검색 방식 지원.
 - RAG, ARAGOG, Semantic Router 등 AI검색 최신 파이프라인 적합.
- **E5**
 - Microsoft 오픈소스, Text Pair 기반 대규모 데이터 학습, 다국어·다도메인 확장성, MTEB 리더보드 상위.

벡터 검색엔진/벡터 DB

- **FAISS (2022)**
 - 백만~수억 벡터 고속 탐색, 대규모 데이터 실험 검증.
 - 메타데이터 조건·복합 필터 부재, 실시간성/확장성 한계.
- **Qdrant (2024)**
 - Dense+Sparse 하이브리드 검색, 메타/속성 동시 관리, 실시간성·대용량·분산·고가용 내장.
 - 엔터프라이즈급 관리 UI, RAG 파이프라인 연동, 신규 서비스 유연 확장.
 - 공식 벤치마크 상 업계 최고 수준 TPS, Latency.

검색 파이프라인

- 2022: Dense(임베딩) + 키워드(BM25) 양상블, GPU서버 기반 병렬 탐색, Re-ranking 모듈.
- 2024: Mixed/Hybrid Retrieval, Ko-reranker(한국어 특화 재정렬), 최신성 필터, ARAGOG 오그멘테이션, Semantic Router, LLM 연동.

5. 미기재/OA 구성요소 및 한계

- **2022년:** 표준 API 설계, 임베딩 벡터/메타 정보 실제 구조, 파인튜닝 세부 파라미터 등은 비공개 또는 미기재(오픈엔디드) 상태임.
- **2024년:** ARAGOG, Semantic Router, rerank 세부 구현·파라미터, 장기 Drift 대응, LLM 연동 커스텀 방식 등은 미상(오픈엔디드 Open Attribute).

결론

2022년의 우리은행 비정형 데이터 자산화 시스템은 한국어 도메인 특화임베딩과 FAISS 기반의 벡터 검색 조합으로 자연어 질의처리 성능을 크게 높였으나, 멀티링구얼·다도메인 확장, 조건/메타필터 검색, 하이브리드 검색, 실시간·확장성 등 엔터프라이즈 요구를 만족시키기에는 구조적 한계가 명확했다. 2024년에는 글로벌 표준 M3, E5 등 첨단 임베딩 모델과 Qdrant 벡터 DB를 도입하여, 고차원 하이브리드 검색, 복합 조건/메타데이터 검색, RAG 기반 LLM 서비스 연계, 대규모 데이터 실시간/분산 처리 등 거의 모든 주요 지표에서 혁신적 개선을 실현했다. 이는 검색정확도 향상, 업무 자동화 및 다양한 신규 서비스 연동 등 실업무·비즈니스 가치의 획기적 증대와 직결됨을 확인할 수 있다.

Sources

- [1] 2022-11-21_(주)우리은행_우리은행 비정형 데이터 자산화 시스템 구축_IV. 기술 부문.pdf
- [2] 2024-04-22_우리은행_비정형 데이터 자산화 시스템 2단계 구축_IV.기술부문.pdf
- [3] <https://qdrant.tech/benchmarks>
- [4] 2022-11-21_(주)우리은행_우리은행 비정형 데이터 자산화 시스템 구축_IV. 기술 부문.pdf (47~53페이지)
- [5] 2024-04-22_우리은행_비정형 데이터 자산화 시스템 2단계 구축_IV.기술부문.pdf (23, 24, 26, 31페이지)
- [6] <https://github.com/microsoft/unilm/tree/master/e5>
- [7] <https://faiss.ai/>