人才聊天搜索系統

- 基於 AI 的智能人才匹配解決方案
- 2025 年 10 月
- 人才分析系統開發團隊

專案背景

- 現狀挑戰:
- - 傳統人才搜索依賴複雜的篩選條件
- - HR 需熟悉系統術語
- - 搜索效率低
- - 缺乏智能匹配解釋

- 市場需求:
- - 精準匹配需求上升
- - 期望自然語言搜索
- - AI 技術成趨勢

解決方案概述

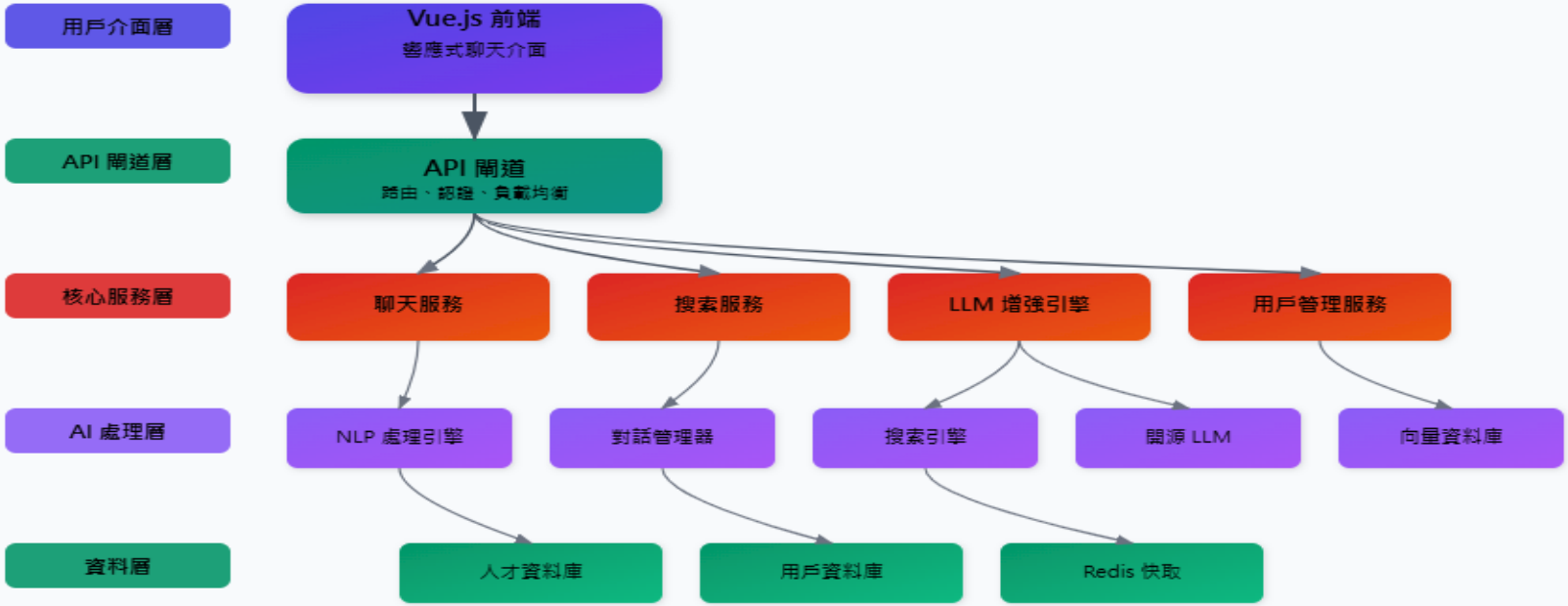
- 核心價值:讓人才搜索像聊天一樣簡單

- 主要功能:
- 自然語言搜索
- AI 智能匹配
- 個性化推薦
- 對話式體驗
- 學習優化

技術架構

- Vue 前端 + Node/Python 後端 + 向量資料庫 + 開源 LLM (Qwen-2.5)
- 採用 RAG 架構、語義搜索、Redis 快取

架構圖



系統架構說明

- Vue.js 前端提供響應式聊天介面，支援即時通訊和結果展示
- API 閘道負責路由、認證和負載均衡，統一管理 API 請求
- 微服務架構確保系統可擴展性和維護性，各服務獨立部署
- LLM 增強引擎整合開源大語言模型，提供智能搜索和結果解釋
- 向量資料庫支援語義搜索和 RAG 架構，提升搜索準確性
- Redis 快取提升系統響應速度和用戶體驗，減少資料庫壓力
- 所有組件通過清晰的數據流連接，確保系統高效穩定運行
- 系統採用分層設計，各層職責單一，便於維護和擴展

RAG 架構優勢

- 流程:
- 用戶查詢 → 向量化 → 語義搜索 → 結果生成 → 推薦

- 優勢:
- – 準確
- – 即時
- – 可解釋
- – 可擴展

RAG 系統架構流程


RAG 系統架構流程
Retrieval-Augmented Generation 檢索增強生成
1. 用戶查詢
"需要善於溝通的銷售人員"
自然語言輸入
2. 查詢向量化
Embedding 模型
文字轉換為向量表示
3. 向量搜索
相似度計算
檢索最相關候選人
4. 結果生成
LLM 生成解釋
提供推薦理由
資料儲存層
人才資料庫
結構化候選人資料
向量資料庫
Embedding 向量儲存
知識庫
職位描述、技能要求
RAG 架構優勢與實作細節
檢索增強 (Retrieval)：
• 使用向量搜索快速找到相關候選人資料
• 支援語義搜索，理解查詢意圖而非僅關鍵字匹配
• 多模態檢索支援文字和技能匹配
技術實作：
• 使用 BGE-M3 或 Text2Vec 進行中文向量化
• Pinecone 或 Qdrant 作為向量資料庫
• 實時更新候選人向量索引，確保資料新鮮度
生成增強 (Generation)：
• LLM 基於檢索結果生成個性化解釋
• 提供匹配理由和改進建議
• 支援多輪對話和上下文理解
效能優化：
• 向量快取減少重複計算
• 分層搜索策略提升響應速度
• 批次處理優化 LLM 推理效率

用戶體驗設計

- 聊天介面特色:
  1. 直觀操作
  2. 即時反饋
  3. 智能建議
  4. 結果展示

- 互動流程:
  1. 用戶輸入需求
  2. 系統澄清
  3. 展示匹配
  4. 細化條件
  5. 查看詳情

技術創新點

- AI 增強搜索: 語義理解 / 上下文感知 / 個性化學習
- 開源 LLM: 成本效益 / 隱私 / 微調
- RAG 架構: 混合搜索 / 動態權重 / 快取