BERT 文本编码与相似度计算技术方案

方案概述

本方案旨在将历史 **FAQ**（常见问题解答）与用户实时查询纳入统一的向量检索流程，以通过语义匹配实现高准确率的自动问答。

1. 离线处理 (Offline)

- **数据清洗:** 对 **FAQ** 数据库进行清洗，包括去除冗余字符、统一格式、去重与文本归一化。
- **离线编码:** 使用预训练的 **Sentence-BERT (SBERT)** 或经过微调的 **BERT** 模型，将每条标准问题（**Question**）编码为固定维度的稠密向量（**Embedding**）。
- **索引构建:** 将生成的向量写入向量数据库（如 **FAISS, Milvus** 或 **Elasticsearch Dense Vector**），并构建高效的近似最近邻（**ANN**）索引。同时存储问题的元数据（类目、答案 **ID**、生效时间等）以便过滤。

2. 在线检索 (Online)

- **查询预处理:** 当用户发起查询时，系统首先进行分词、去除停用词和规范化处理。
- **实时编码:** 调用 **embedding** 服务（与离线使用同一模型）将用户查询转换为向量。
- **向量检索:** 在向量数据库中执行 **ANN** 检索（**Top-K**），快速召回最相似的若干个候选问题。
- **精排与决策:** 对召回的候选集进行精细排序（如计算余弦相似度或使用 **Cross-Encoder** 重排序），结合业务规则（相似度阈值、类目匹配、时效性）进行过滤。
- **结果返回:** 若最高分高于设定阈值，则返回对应答案；否则通过低置信度策略处理（如推荐相关问题、转人工客服或调用大模型生成）。

技术流程图