[搜索系统算法解析-OAA]

搜索召回

召回 是整个检索流程的核心组成部分之一，其主要任务是从大规模文档集合中初步筛选出一批可能与用户查询相关的文档。对于召回的整体定位和要求应当为：

- **快速缩小范围**：从海量文档中选出一个初始候选集，通常数量为几千到几万条
- **确保覆盖性**：尽量保证真正相关的文档被包含在候选集内，避免漏召
- **高效处理**：需要在低延迟下完成大规模检索，兼顾效果和计算开销

召回模块的**业务目标**主要围绕提高搜索系统的用户体验和业务价值，作为一个承上启下的环节，召回整体的目标要和搜索的最终目标对齐，即上下游协同发展。而召回作为一个功能模块，自然需要有其目标优化的侧重点：

- **覆盖率**
  - 定义：覆盖率是召回的核心目标，在候选集范围内需要尽可能全面地覆盖与用户查询相关的文档，避免漏掉高价值或高相关性的结果。具体的，可以通过多路召回的方式，从不同角度和方向去探索覆盖边界
  - 评估方式：
    - 离线评估：通过人工/LLM/规则等方式构建理想的标准测试数据集
    - 在线评估：通过线上监控出现零少召回的结果的情况，以衡量长尾查询的覆盖效果
- **效率**
  - 定义：召回面对的是百万/千万级的海量数据，需要尽可能小的开销实现快速筛选
    - 快速响应：以最低延迟完成初步筛选
    - 低计算成本：控制计算资源消耗，提高系统的性价比
  - 评估方式：在线耗时、资源消耗等监控
- **相关性**
  - 定义：虽然不是召回最关心的指标，但是仍然是一个约束项。召回的候选集需要与用户查询足够相关，为后续排序模块提供高质量输入
- **多样性**
  - 定义：确保候选集中的内容多样化，避免召回结果过于集中
  - 评估方式：冗余度
- **定制化**
  - 定义：
    - 个性化匹配：结合用户画像或上下文信息，召回与特定用户需求匹配的候选集
    - 场景适配：针对不同业务场景采用不同的召回策略
- **业务目标驱动**
  - 定义：
    - 提高转化率：通过召回高相关、高价值的文档或商品，推动用户进行点击、购买或其他目标行为

        - 提升用户粘性：通过更精准和有趣的召回结果，吸引用户持续使用平台
    - 评估方式：在线监控 CTR、CVR、搜索活跃用户数、用户留存率等
- **适应实时性需求**
    - 定义：实时调整召回策略或结果以适应变化，例如突发热点事件或快速变动的库存
- **数据可解释性**
    - 定义：确保召回模块能解释为何某些候选内容被选中，以便调优和迭代

召回算法的实现方式主要有：

- 文本召回：
    - 倒排索引：通过词项（Term）倒排索引，快速定位文档
    - KV召回：基于协同过滤等图学模型，通过用户行为或文档特征，推荐相似或相关的文档
- 向量召回：利用语义向量模型（如 Word2Vec、BERT、SimCSE）将查询和文档映射到向量空间，通过向量相似度（如余弦相似度、欧几里得距离）进行检索
- 混合召回：结合关键词召回、向量召回的优点，通过加权、交集或并集策略生成候选集

本章将具体介绍文本召回中的倒排召回、协同过滤召回和向量召回的实现方式，以及如何评估和优化这些召回策略。并在最后介绍如何在多召回通道下聚合多路召回结果。

倒排召回

协同过滤召回

向量召回

召回聚合