阿里文娱在深度语义相关性计算中的探索

**导读：**如何帮助用户找到真正想要的内容，一直是阿里大文娱在深度语义相关性计算方面不断的探索。本次分享，将介绍阿里大文娱搜索在深度语义相关性计算中的探索，以及在业务场景中案例。本次分享会围绕下面四个方面的介绍：

* 文娱搜索业务简介
* 相关性和排序
* 多模态视频搜索
* 深度语义相关性的一些探索

**▌文娱搜索业务介绍**

文娱搜索业务介绍分别从三个方面介绍，优酷视频搜索的业务背景、搜索基本评估指标、以及搜索系统的算法框架。

1. **优酷视频搜索业务背景**

****

整个阿里文娱提供一站式的搜索推荐服务，服务范围不仅仅包括优酷的全部的搜索入口，包括APP端、OTT大屏端、PC端、Pad端，以及网页端等，同时还涉及到大麦、淘票票等领域，检索内容从传统的影剧、动漫等有版权的长视频影视库之外，同时也扩展到社会各个领域UPGC领域的视频内容，例如影人、小说、音乐、演出/场馆等的需要提供检索服务。

以优酷为例，有上亿级别的优酷视频资源，这不仅包括平台购买有版权的OGC视频，比如节目、电影、电视剧等资源，更多的是用户上传UGC的视频。

大家都知道，视频它作为4G以及5G时代最便捷的信息载体，它在给用户带来极大便利的同时，也会给我们带来了更大的挑战，因为视频不论从制作、存储、计算、以及分发，都比文字模态的信息要困难很多。

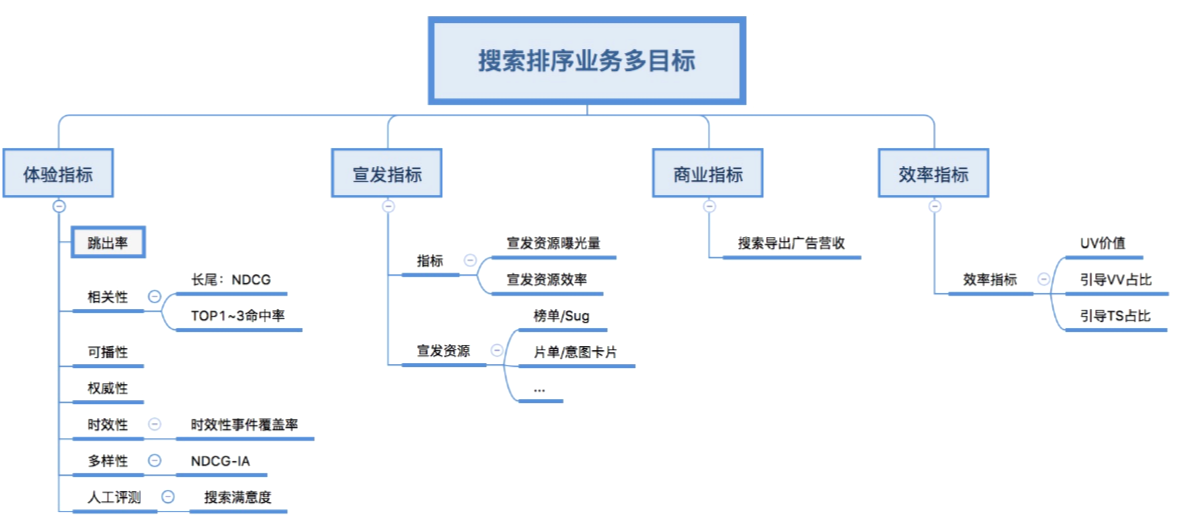
1. **搜索基本评估指标**

搜索技术是怎么做的呢？首先，跟大家介绍一下，视频搜索领域的用户价值是什么，以及它是如何影响搜索技术的评估指标的。

用户价值在这里主要体现在两个维度：

第一个维度是工具属性，那么所谓工具属性，也就是说用户将搜索服务作为寻找内容的工具，目标就是找准和找全，也就是搜索到和搜的准。这是用户价值的基础，同时也是搜索的基础属性。

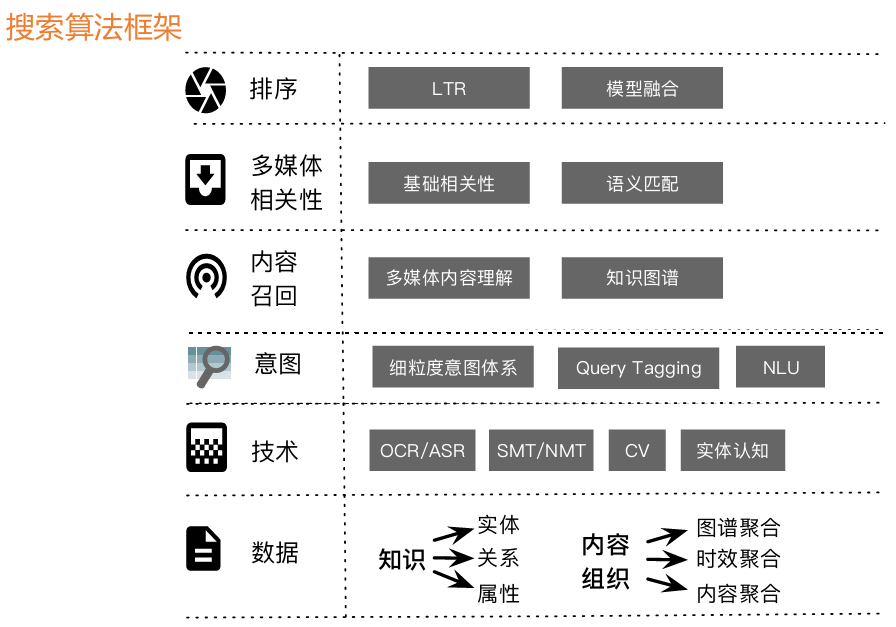
从这个维度去评估搜索效果的好坏，就需要一系列的体验类的指标，比如说：跳出率，还有一些相关性、时效性、多样性等等。



评估指标主要分为四大类：体验指标、宣发指标、商业指标和效率指标。

* 体验指标包括：跳出率、相关性、时效性、多样性等；
* 宣发指标包括：宣发资源曝光量、宣发资源效率等；
* 商业指标包括：搜索导出广告应收等；
* 效率指标包括：UV价值、引导VV占比、引导TS占比等。

**3. 搜索系统的算法框架**

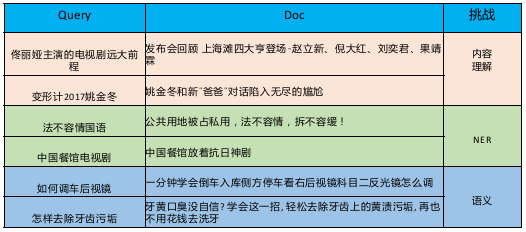


针对上述问题：

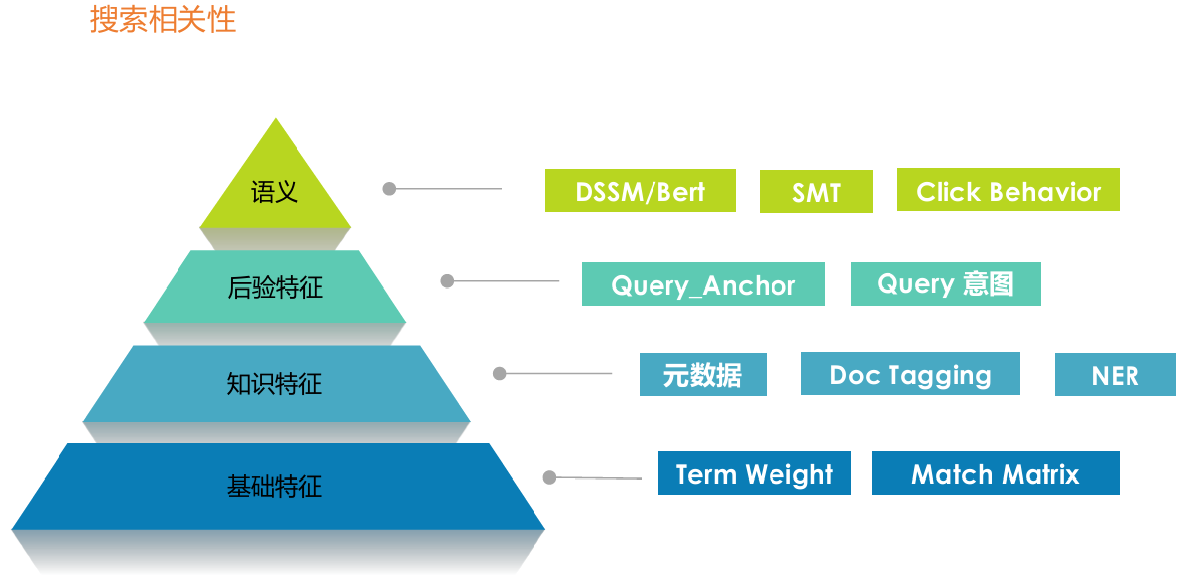
* 数据隔离：联邦学习的整套机制在合作过程中，数据不会传递到外部。
* 共同获益：无论数据源方，还是数据应用方，都能获取相应的价值。

**▌相关性和排序**

搜索相关性面临很多挑战，主要涉及：异构内容理解、实体知识匹配、深度语义计算等方面。

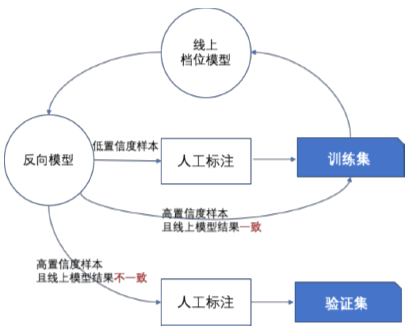


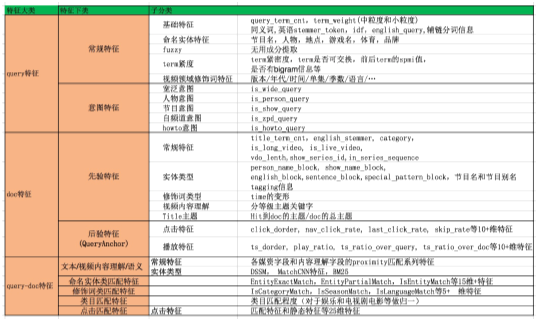
**1. 搜索相关性**



举个例子：

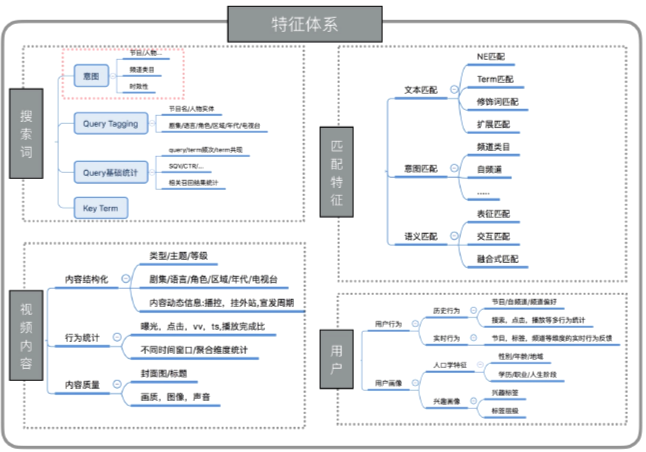
**2. 相关性数据集构建**





举个例子：

**3. 排序特征体系**



**▌多模态视频搜索**

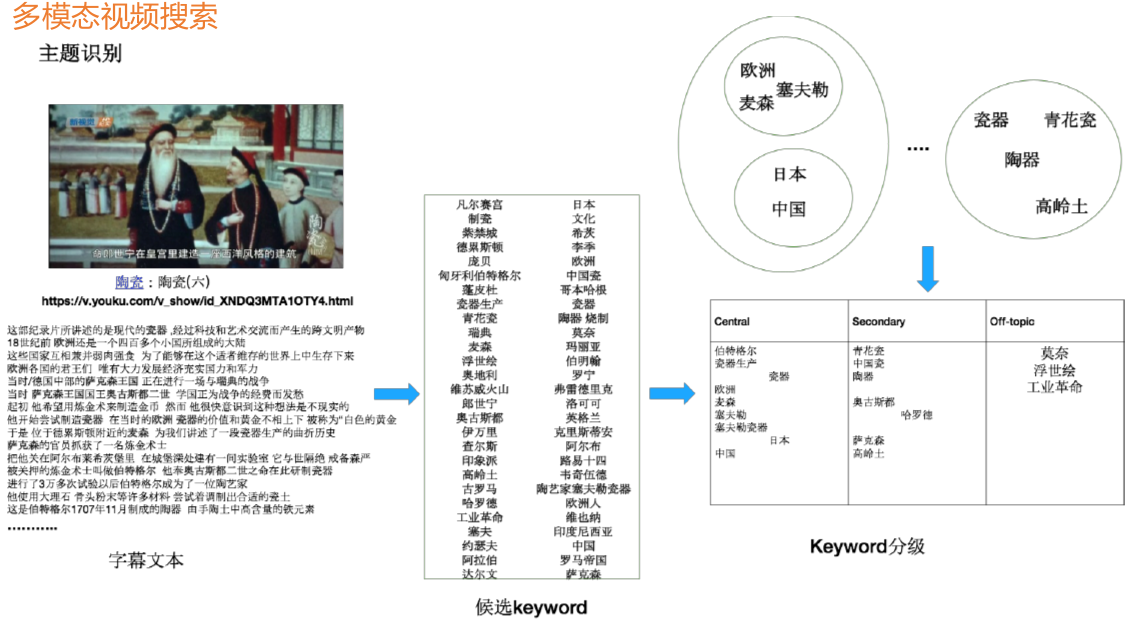
介绍

**1. 多模态视频搜索**

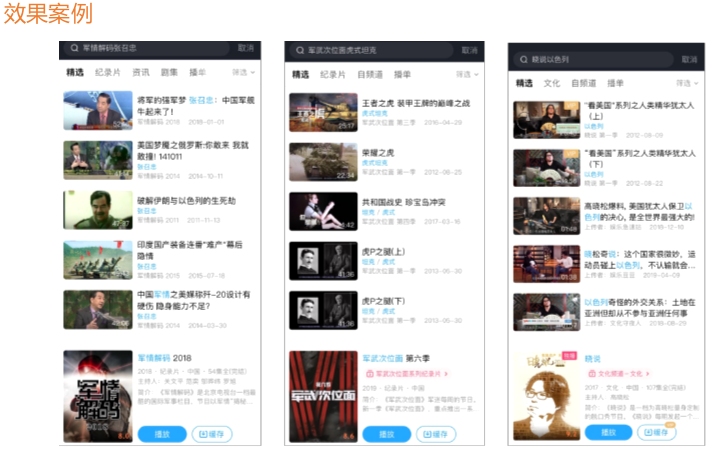


举个例子：

**2. 多模态视频搜索中主题识别**



效果案例：



**▌深度语义相关性的一些探索**

介绍

**1. 字面匹配vs语义匹配**

字面匹配和语义匹配的设计包括三个层面：系统设计层面、逻辑架构层面、模型算法层面。



**2. 相关性系统设计**

相关性老版系统设计：相关性提特征，LTR多目标优化；

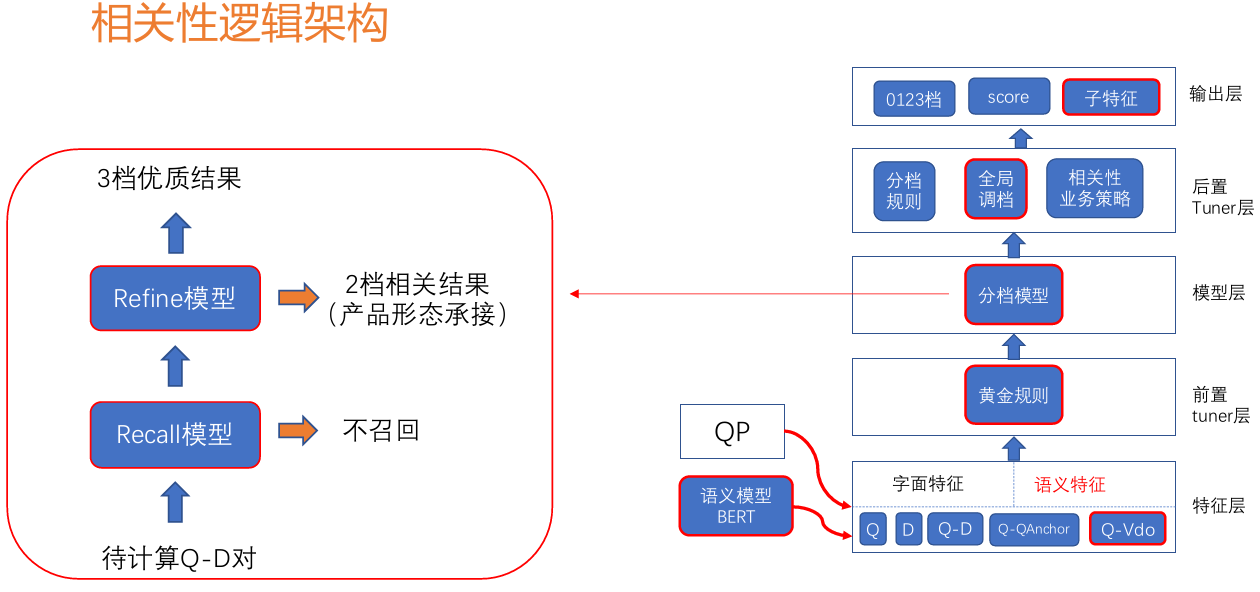
相关性新版系统设计：相关性解体验问题，LTR解分发问题。

相关性分档规范以及标注规范举例：

核心：识别用户query的主次意图。Query“步步惊心吴奇隆”，主意图是“步步惊心”，次意图是“吴奇隆”。

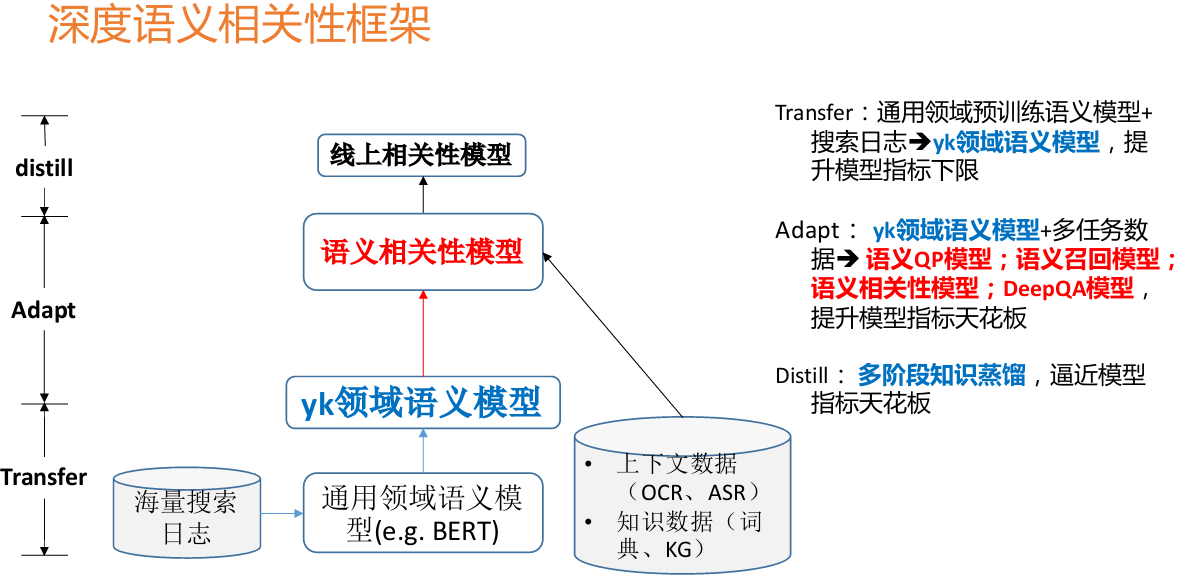
* 3档：用户全部意图满足，doc:“吴奇隆在步步惊心的现场花絮……”；
* 2档：用户部分意图满足，doc:“刘诗诗在步步惊心的现场花絮……”;
* 0/1档：用户意图不满足，term部分匹配，doc:“玻璃栈道上，每一步都走的步步惊心”.

**3. 相关性逻辑架构**



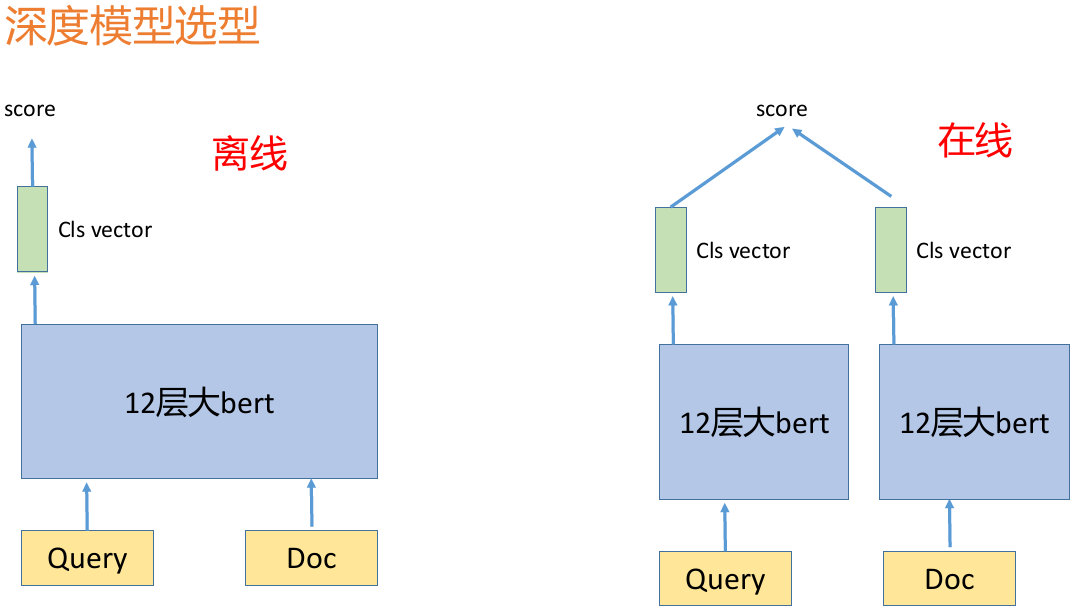
举个例子：

**4. 深度语义相关性框架**



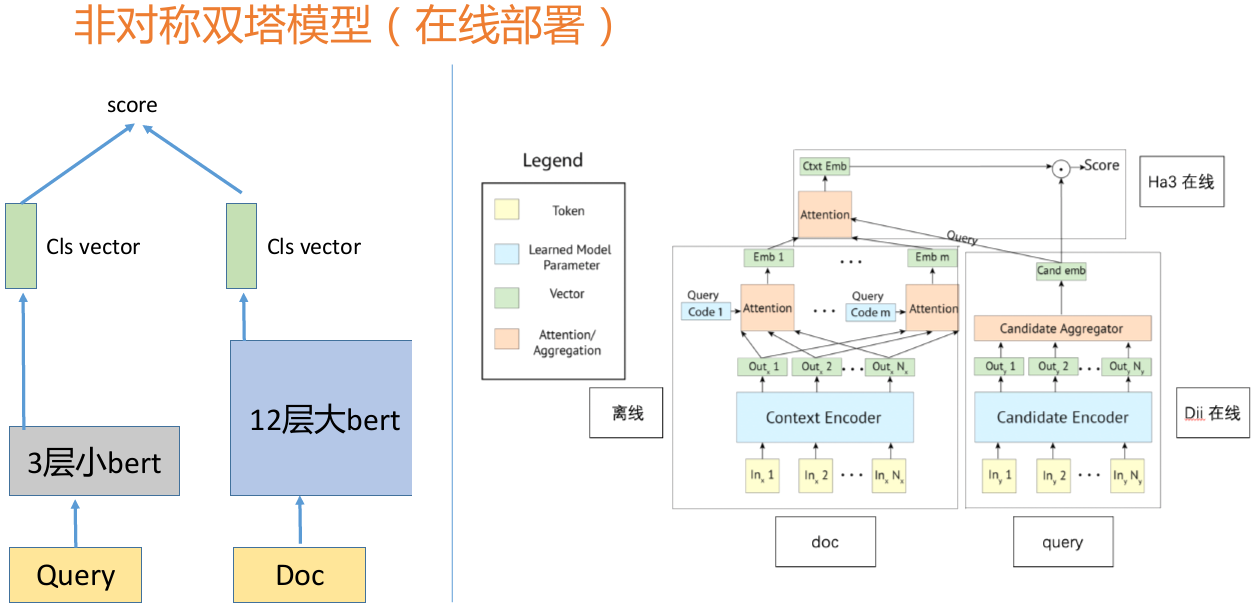
举个例子：

**5. 深度模型选型**



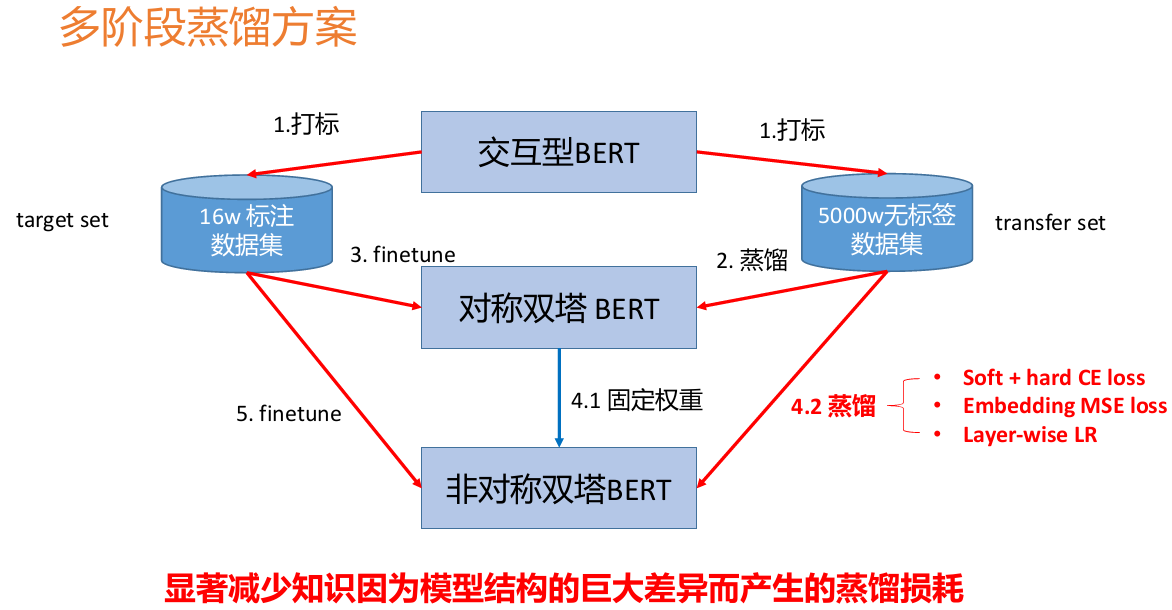
举个例子：

**6. 非对称双塔模型（在线部署）**



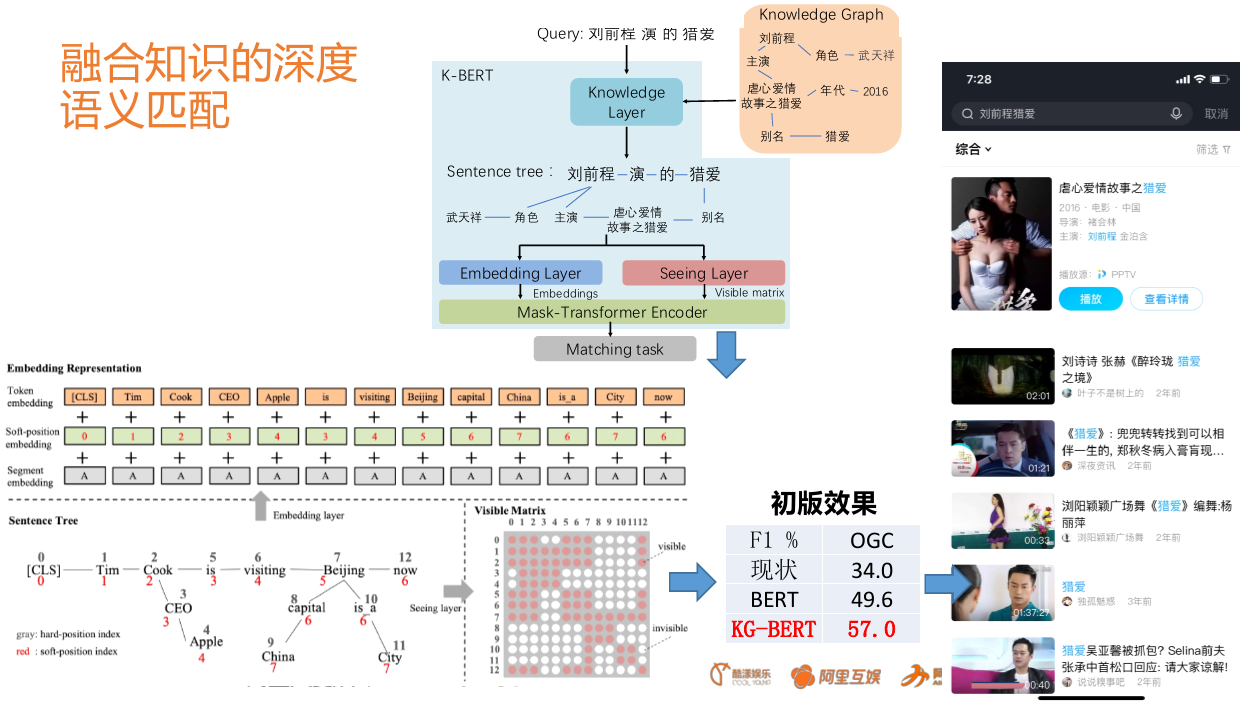
举个例子

**7. 多阶段蒸馏方案**



举个例子

**8. 融合知识的深度语义匹配**



效果案例：

