词权重（Term Weighting）

词权重（Term Weighting）指搜索引擎在处理用户查询时，用于衡量用户查询（Query）中每个词（Term）的重要程度。这种重要程度的评估对于搜索引擎准确理解用户意图、召回相关结果并排序至关重要。词权重信息主要应用在以下场景：

- 丢词重召：在召回结果较少或为空的情况下，通过识别并丢弃权重较低的词，对核心词重新进行召回，以达到扩召目的
- 相关性排序：词权重作为一个特征计算以查询与结果相关性程度，从而对召回Doc进行相关性分档排序。除了作为特征，也可以在相关性模型训练时用作数据增强（如丢弃核心词以构建负样本）

词权重依赖分词，即首先对Query分词获得Term序列，然后利用语料信息、点击日志等信息判断每个Term的重要程度。而如何确定词重要性决定了词权重的优化方向。本章节先介绍词权重的定义和标注规则，然后对词权重算法进行介绍。

词权重标注

词权重通常分为4档：

- 3档：核心词，为Query的核心意图词，当核心词丢弃时Query语义基本完全改变，无法检索符合原Query意图的Doc
- 2档：关键词，Query核心意图重要组成部分，丢弃时语义有较大变化，但仍然可以检索到符合原Query意图的Doc
- 1档：边缘词，通常在Query中作为修饰词，丢弃后Query主要意图基本不变，且可以检索到绝大部分符合Query意图的Doc
- 0档：冗余词，丢弃后Query意图不发生变化，且检索到的Doc基本都符合原Query意图

根据词权重的档位定义，我们可以得知，词权重信号判断依赖两个信号：先验的语义信息和后验的检索召回信息。在具体的词权重的标注过程中，可以结合具体搜索场景下的难/易Case，根据语义和检索信息完成词权重标注任务。

词权重算法

在基于词权重标注规则完成数十万量级的训练样本标注后，接下来的工作就是特征构造和模型打分了。

特征挖掘

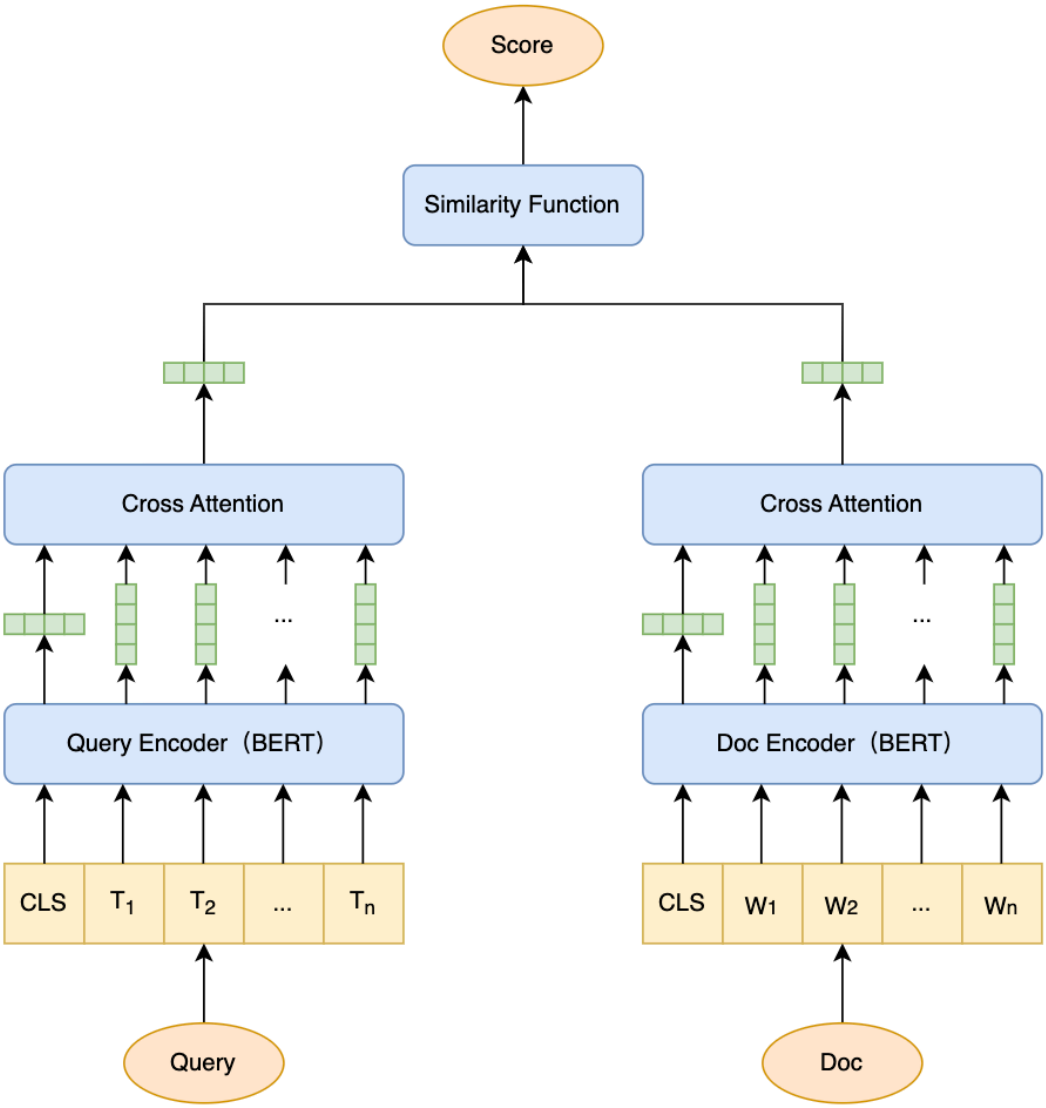
常见的词权重特征如下：

特征
Query长度
Query中Term数
Term在Query中的位置
Term是否为英文、数字
Term词性、NER、句法成分
TF（Term在Doc中出现次数/全部Doc中的Term数）
IDF（全部Doc数/包含Term的Doc数）
IQF（全部Query数/包含Term的Query数）
单Term作为Query的数量
Term在Query中的数量
独立索引占比（单Term作为Query的数量/Term在Query中的数量）
Term的左右临熵、紧密度
丢弃Term后Query语言模型概率/相似度变化（Query Embedding Cosine距离）
丢弃Term后Query语义损失度
丢弃Term前后检索Doc特征分布变化（Doc曝光、点击、类别分布的KL散度）
共现率（点击的Doc集合中term出现频次/点击Doc数）

模型打分

完成特征挖掘后，利用模型完成对词权重的打分。模型输入为Term粒度下的Term特征，输出为Term的标签，模型可以根据情况选择树模型或DNN模型。

除了上述常规的数据标注 + 模型预测的方法，也可采用点击数据训练一个双塔模型，利用Query侧塔的注意力层中间结果作为词向量的权重（也可作为监督模型的特征采用），具体的：



1. 对Query和Doc分词并分别输入到对应的Encoder模块（通常是BERT）获取词向量表征、CLS向量表征（作为全局文本语义信息表征）
2. 对于Query和Doc双塔，其对应CLS向量表征（Q）和词向量表征（K、V）经过交叉注意力层获得Query向量和Doc向量
3. 使用相似度度量来计算这两个向量的匹配度，常用方法有点积或余弦相似度
4. 最后经过激活函数（如 Sigmoid）将匹配度映射到[0,1]区间，目标是点击预测
5. 模型完成训练后，Query塔中注意力层的attention score（$\text{score} = \frac{\mathbf{q} \cdot \mathbf{k_i}}{\sqrt{d_k}}$）将作为最后的词权重

总结

综上，本章介绍了词权重在搜索系统里的应用，以及词权重分档定义和标注准则，另外在算法实现上介绍了常用的文本特征和统计特征，以及相关模型设计。

参考文献

1. Term-weighting approaches in automatic text retrieval
2. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding