Query类目

Query类目指的是根据查询内容将查询词Query归类到某个特定的分类体系中。这个体系通常是多级的，能够将查询词从更广泛的类别逐渐细分到更具体的子类目，这个体系通常在电商搜索和推荐领域中有重要的作用。

Query和Doc一般共用一套类目体系，搜索系统可以根据类目信息对搜索结果进行相关性排序，基于Query和Doc的类目匹配程度提高召回Doc的精准度。通过类目体系，搜索引擎可以挖掘用户的搜索历史和偏好，为用户提供个性化的搜索结果。此外，类目体系在搜索系统中起到了组织和分类内容的作用，从而有助于进行数据分析，了解用户的行为和偏好，优化搜索和推荐策略。

类目体系

类目通常是**分层的**，具有**一级类目**和**二级类目**（甚至更多层级），每个层级从广义到狭义逐渐细化，具体如下：

- **一级类目**：代表更宽泛的领域或主题，通常为大类，比如“时尚”、“体育”、“影视娱乐”等
- **二级类目**：在一级类目的基础上，进一步细分到更具体的领域或话题，如“穿搭”、“户外运动”、“手游”等
- **三级类目**（可选）：如果需要更细致的分类，可以继续细分。例如，“帽子推荐”可能在“穿搭”下面有一个“帽子”类目

在设计类目层级时，通常考虑以下原则：

1. **层级清晰**：类目层级应清晰明了，便于理解和使用
2. **逻辑合理**：类目之间的逻辑关系应合理，避免出现交叉或重复的情况
3. **可扩展性**：类目体系应具备一定的可扩展性，以适应未来可能出现的新信息

Query类目预测

由于Query和Doc共用一套类目体系，彼此相关的Query和Doc之间的类目通常也是相关的，且Doc类目由于其内容量多、信息量大，因而预测难度较低。所以Query类目的预测除了从文本语义角度出发，一般还依赖于Query关联Doc的类目信息。

基于PMI的类目预测

由于Query与关联Doc之间类目强相关，则Query历史点击Doc的类目可以作为Query类目预测。PMI主要用来衡量两个事件（在这里是“查询词”和“类目”）之间的相关性，帮助我们通过观察查询词和类目（点击Doc的类目）之间的共现关系，预测查询词所属的类目。

PMI衡量的是两个词或事件同时发生的概率与它们独立发生的概率之间的比率，PMI的值越高，意味着查询词和类目之间的相关性越强，反之则较弱。基于PMI的Query类目预测流程如下：

1. 统计每个查询词与类目组合在语料库中出现的频率（即共现次数），记为$N(x,y)$
2. 统计每个查询词在语料库中出现的次数，记为$N(x)$
3. 统计每个类目在语料库中出现的次数，记为$N(y)$
4. 计算语料库中查询词和类目的总数N_{total}

5. 对于每一对查询词x和类目y，计算其PMI值:

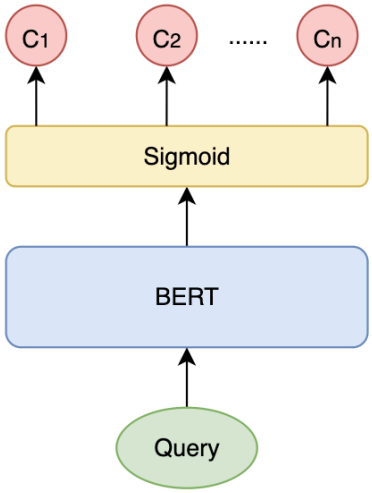
$$PMI(x,y) = \log \frac{P(x,y)}{P(x)P(y)} = \log \frac{\frac{N(x,y)}{N_{\text{total}}}}{\frac{N(x)}{N_{\text{total}}} \cdot \frac{N(y)}{N_{\text{total}}}} = \log \frac{N(x,y) \cdot N_{\text{total}}}{N(x) \cdot N(y)} \tag{1}$$

6. 对于给定的查询，计算查询中每个词与每个类目的PMI值，选择PMI值TopN的类目作为预测结果
7. 基于上述流程，通过搜索日志的语料库挖掘生成Query-类目映射表

基于PMI统计得到的类目表是Query和类目的映射关系，当线上Query没有命中类目表时，需要对Query进行分词，将分词后的Term分别召回对应类目，并可结合Term权重分取加权平均值，筛选加权平均值最高的类目作为预测结果。

基于BERT的类目预测

对于PMI类目预测中缺失统计信息造成置信度不高Query，可以通过基于BERT的类目预测调整结果。对于类目预测，这是一个多标签分类问题，目标是对Query预测多个类目的概率。



在多标签分类中，采用的激活函数是**Sigmoid**，而非Softmax。因为每个标签都是独立的二分类问题，**Sigmoid** 为每个标签分别计算独立的概率。常用的损失函数采用**二元交叉熵损失（BCE Loss）**，对于一个标签 y_i，其真实标签是$y_i \in \{0,1\}$，预测的概率是$\hat{y}_i$，损失函数的形式为：

$$\text{BCE}(y_i, \hat{y}_i) = -\left[y_i \cdot \log(\hat{y}_i) + (1-y_i) \cdot \log(1-\hat{y}_i)\right] \tag{2}$$

对于所有标签的总损失，取每个标签的损失的平均值，得到最终的损失：

$$\text{Total Loss} = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i \cdot \log(\hat{y}_i) + (1-y_i) \cdot \log(1-\hat{y}_i)\right] \tag{3}$$

其中：

- N 是标签的数量。
- y_i 是标签 i 的真实标签（0 或 1）。
- $\hat{y}_i$ 是标签 i 的预测概率（通过 Sigmoid 激活函数得到的值）

考虑到一个合格的类目体系下，各类目标签的分布应当符合正态分布。在构建训练集时也会不可避免的出现标签类别不平衡的问题，此时除了平衡训练集的类别分布，损失函数上也可以选择**Focal Loss**。

Focal Loss引入了一个焦点因子 $(1-\hat{y})^{\gamma}$，并且可能会为正负样本引入不同的权重因子 α 来缓解类别不平衡。
Focal Loss 的公式如下：

$$FL(y,\hat{y}) = -\alpha \cdot (1-\hat{y})^{\gamma} \cdot \left[y\log(\hat{y}) + (1-y)\log(1-\hat{y}) \right] \tag{4}$$

其中：

- y 是真实标签（0 或 1）
- $\hat{y}$ 是模型的预测概率（Sigmoid 输出）
- α 是平衡因子，用于调整类别不平衡问题，通常 α 可以为正负样本分配不同的权重
- γ 是焦点因子的指数，控制模型对难分类样本的关注程度

焦点因子$(1-\hat{y})^{\gamma}$在 **Focal Loss** 中的作用是 **降低易分类样本** 的损失影响，使得模型更多关注 **难分类样本**：

- 当 $\hat{y}$ 接近于 1 时（即正类容易分类），焦点因子的值接近于 0，降低了易分类样本的损失权重
- 当 $\hat{y}$ 接近于 0 时（即负类容易分类），焦点因子同样会使损失值变小，避免了对负类样本的过多关注
- γ 参数控制焦点因子的强度，通常 γ 取值为 2，用以显著降低对易分类样本的损失权重

BERT训练样本构造可以利用基于PMI的类目映射表，为了减少长尾类目由于点击行为较少带来训练不足的问题，可以挖掘长尾类目的Doc，利用Query-Title-类目构建训练集，从而提高模型在长尾类目上的预测效果，增强模型泛化能力。

总结

类目体系在搜索系统中起到了组织和分类内容、提高搜索效率和准确性、优化用户体验、支持数据分析和管理的重要作用。通过合理设计和维护类目体系，可以显著提升搜索系统的性能和用户满意度。

参考文献

1. Focal Loss for Dense Object Detection
2. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding