Query词权重方法（3） - 基于有监督学习

原创 XG数据 WePlayData 2019-04-03

本文继续介绍一种基于有监督学习的词权重计算方法。有监督学习相比于无监督学习的效果一般会更好，但也存在需要大量标注样本、模型难以更新，结果难以debug的等问题。有监督学习一般有样本构造、特征表示、模型训练、模型评估四部分。

1）样本构造

样本构造之前需要定义任务的形式。如果把词权重看成分类任务，则可以按二分类（term重要为1，不重要0）或者多分类（term的重要等级）进行标注；如果是排序任务，则可以按pair-wise去标注样本，即标注两个term间那个term更重要；如果是回归任务，则需要标注每个term的权重。但是人工标注需要大量的时间，并且可能存在模棱两可的情况，因此可以采用自动的方法去构建样本。这里介绍一种基于百度元搜的回归方法，去拟合query中的词在搜索结果里的term recall weight，

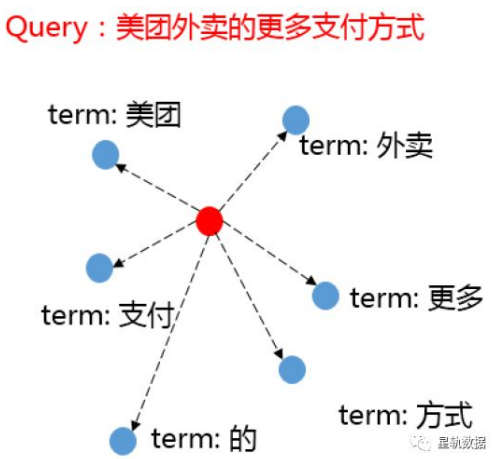
$$\text{term\_recall\_weight}(query, term) = \frac{\#doc_{query,term}}{\#doc_{query}}$$

#doc(query)指搜索结果中doc数目，#doc(query,term)是搜索结果中包含term的doc数目。

2）特征表示

特征表示现在比较流行的是embedding表示，在该任务中，可以分为把query和term表示低维embedding向量，然后用query embedding和term embedding的距离表示成向量，如下图，term和query距离比较近，则表示该term比较重要，相反比较远，则没有那么重要。

$$fea(term,query) = embedding(term) - embedding(query)$$



3）模型训练

因为是回归任务，这里可以采用linear regression或者GBRT，不过不同的模型对效果的影响并不大，因为任务的瓶颈还是在于样本构造和特征表示。

4）模型评估

模型评估一般分为直接评估和间接评估。直接评估采用模型本身的准确率或者人工评测，间接评估通常把词权重作为一种中间结果用于某个任务里面，通过任务的评价指标来反应词权重的质量。

相关阅读

1. Query理解 - 搜索引擎“更懂你”
2. 搜索引擎新的战场 - 百度、头条、微信
3. 当我们关注舆情系统时，我们忽略了什么？
4. 搜索引擎的两大问题（1） - 召回
5. 搜索引擎的两大问题（2） - 相关性
6. Query词权重方法（1） - 基于语料统计
7. Query词权重方法（2） - 基于点击日志