Query词权重方法（4）- beyond 词粒度

原创 XG数据 WePlayData 2019-04-04

前文介绍的词权重方法都是预测单个term的权重，在实际使用中，可能也需要某个短语或n-gram的权重，比如query"我 的 前 半生 在线 观看"，可能也需要知道"我 的 前 半生"的权重，虽然从单个term角度来看，每个term的重要性都不大，但从整体来看，"我的前半生"的权重还是比较大的。

因此在idf的基础上，针对n-gram可以基于语料统计计算一个ngram-idf，相比于idf，ngram-idf可以在同一维度空间比较任意长度n-gram的重要性。这是因为idf计算中受限于ngram长度的影响，ngram越长，其出现次数越少，计算出的idf就越高。但是idf的高低和ngram长度并无直接关系，ngram-idf的计算中引入其他计算因子减轻了长度的影响。

$$IDF(\theta(g)) = \log\frac{|D|}{df(g)}$$

$$IDF_{N-gram}(g) = \log\frac{|D|df(g)}{df(\theta(g))df(\theta(g))}$$

上图给出了ngram-idf的计算方式，对于ngram g，df(g)表示g紧邻出现在语料中的次数，df(\thea(g))表示g非紧邻出现在语料中的次数，要求在一定窗口内。公式前半部分类似于idf的计算，表示ngram出现的次数越少，ngram的信息量就越大；后半部分表示ngram在文本紧邻和非紧邻出现的次数越接近，ngram的内凝度就越大。两者组成了ngram的重要性。ngram-idf在计算过程中，一个挑战是如何基于大规模语料统计ngram在预定义窗口内非紧邻出现的次数。

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
8. Query词权重方法（3） - 基于有监督学习