Query词权重方法（2） - 基于点击日志

原创 XG数据 WePlayData 2019-04-02

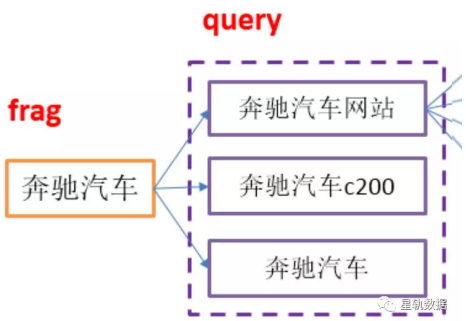
本文继续介绍一种基于点击日志的词权重计算方法。点击日志在搜索、nlp任务中起着不可或缺的作用。点击日志是系统独有数据，区别于公开数据，是系统的有效反馈数据。几乎所有的搜索、nlp任务都可以见到点击日志的影子。

基于点击日志的词权重计算方法，其主要假设是长query的中term的权重可以由短query中的term的权重近似计算得到。假设要求Query：奔驰 汽车 发电 机 故障 怎么 办？中每个term的权重。如果能够分别知道子片段中哪个term比较重要，发电 机 故障、汽车 发电 机 故障、奔驰 汽车 故障、奔驰 汽车，那么query中的term权重可以由这些子片段中term的权重推导得到。问题转化成求frag中的词权重p(term|frag)和词丢弃概率pvte(term|frag)。

$$p(\text{term}|\text{query}) = \left(\frac{1}{n}\sum_{\text{frag}} p(\text{term}|\text{frag})\right) * \left(1 + \frac{1}{n}\sum_{\text{frag}} pvte(term|frag)\right)$$

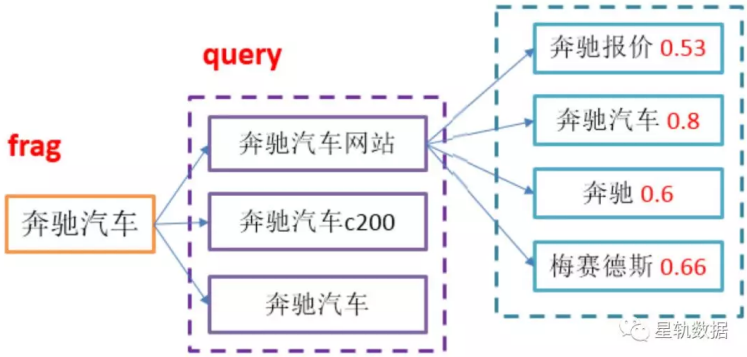
p(term|frag)是子片段frag中term的权重，表示term在子片段的权重越高，那么term在query的权重就越高；
pvte(term|frag)是子片段frag中term的丢弃概率，表示term在子片段中越不重要，那么term在query中就越不重要；

p(term|frag)和pvte(term|frag)的计算方式比较类似，下面以“奔驰 汽车”为例介绍下p(term|frag)的计算方法：
1) p(term|frag)：如下图，对于“奔驰 汽车”，首先找到包含该frag的query集合，如果term在这些query中的权重比较重要，那么term在该frag中的权重也会比较大，此时问题转化成计算query中的term的权重p(term|frag)；



$$p(\text{term}|\text{frag}) = 1/n \sum_{\text{query}} p(term|query) * \text{if}(\text{frag} \in \text{query})$$

2）p(term|query)：如下图，给定query，首先获取query的相关query集合(称为qanchor)，那么在qanchor集合中出现次数较多的term一般会是query中比较重要的term。问题转化成寻找query的qanchor，也就是计算query的qanchor的相似度p(qanchor|query)。



$$p(term|query) = 1/n \sum_{query} p(qanchor|query) * if(term \in qanchor)$$

3）p(qanchor|query)：这里的计算方式就比较灵活，可以抽象成两个短文本的语义相似性计算。一种方式根据query-doc的点击信息计算query和qanchor的相似度，即query和qanchor的点击doc分布越相似，query的qanchor的相关性就越大。

$$p(qanchor|query) = \sum_{doc} p(qanchor|doc) * p(doc|query)$$

最后根据依次计算得到的条件概率值反推去求query中的term的权重。这种方法的优势是完全依赖点击日志，因为点击日志是实时更新，因此很适合离线滚动更新p(term|frag)，然后近实时的影响在线query中term权重计算。

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