Query词权重方法（1） - 基于语料统计

原创 XG数据 WePlayData 2019-04-01

query词权重（term weighting）是为了计算query分词后，每个term的重要程度。常用的指标是tf*idf（query中term的tf大部分为1），即一个term的出现次数越多，表明信息量越少，相反一个term的次数越少，表明信息量越多。但是term的重要程度并不是和term的出现次数呈严格单调关系，并且idf缺乏上下文语境的考虑（比如“windows”在“windows应用软件”中比较重要，而在“windows xp系统iphone xs导照片”的重要性就比较低）。词权重计算作为一种基础资源在文本相关性，丢词等任务中有着重要作用，其优化方法主要分为下面三类：

1）基于语料统计

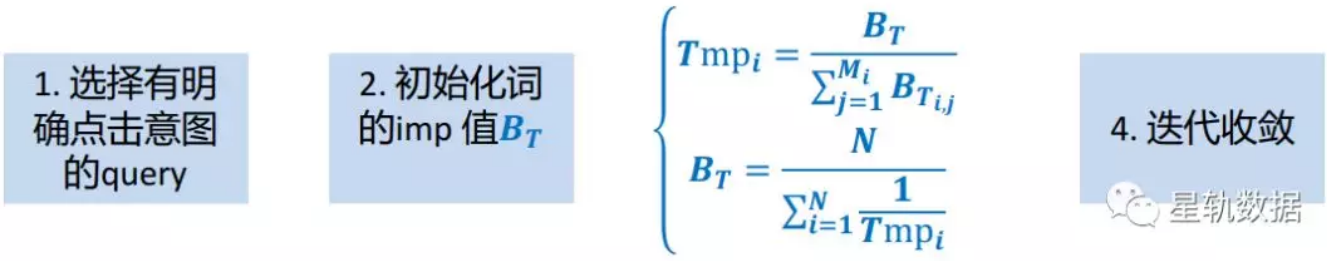
2）基于点击日志

3）基于有监督学习

本文首先介绍一些基于语料统计的计算方法。

一、imp（importance的缩写）

idf的一个缺点是仅仅依靠词频比较，imp从在query中的重要性占比基础上，采用迭代的计算方式优化词的静态赋权，其计算过程如下：



其中BT为term的imp值，初始值可设为1，Tmp_i是query中的第i个term的重要性占比，N指所有包含第i个term的query数目。

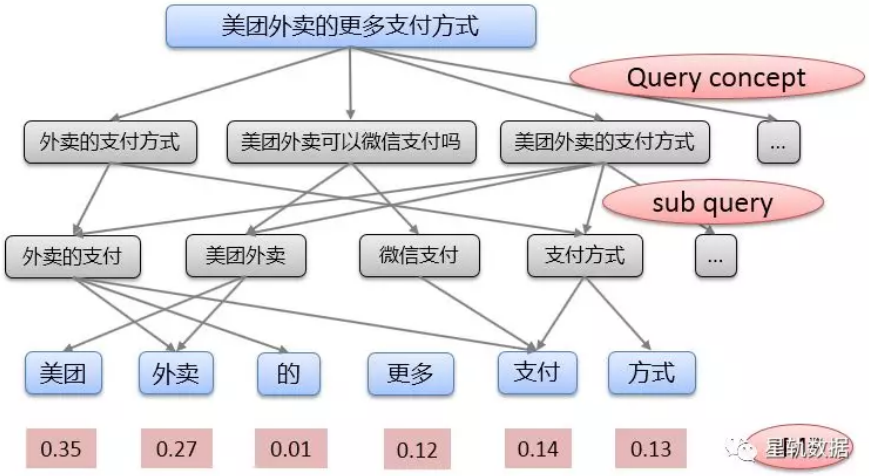
二、DIMP（Dynamic imp）

idf和imp的一个共同缺点是其都是静态的赋权。DIMP根据query的上下文计算每个term的动态赋权，其主要假设是任意query中的词权重可以由相关query 的词权重来计算，计算过程可分为两部分：

1） 自顶向下的query树构建

根据实际场景中采用不同的构建方法，这里介绍一种在搜索中的做法。如下图，给定query作为根节点，首先获取query的相关query作为第二层节点，在第二层的基础上，枚举相关query的子query作为第三层节点，最后一层为分词后的term节点。因此query树种的节点都

是不同粒度的文本串，边都是文本串间的相关关系。在拍卖词推荐任务中，用户query都是比较短的关键词，其可以通过拍卖词间的共同购买关系构建对应的query树。



2） 自底向上的term weight计算

在生成query树后，叶节点的term可以赋上静态imp值，然后自底向上的计算query中的某个term的权重，其递归计算公式如下：

$$w(t) = \sum_{i=1}^{n} w_{v_i}(t) f(v_i) g(v_i, v_r)$$

在imp向上传播中，不同的中间节点和边对词权重的贡献度是不一样的。因此需要一种策略去考虑节点和边的权重。同样不同的场景也有不同的计算方法，对应到上述公式中，分别对应叶子节点的初始权重，中间非叶子节点的权重，边的权重。

	节点	边
普通搜索	qv或satisfy=log（search）*（click/search）	两个query间的相似度或query间共点击次数；
拍卖词推荐	拍卖词的购买次数	两个拍卖词的共同购买次数

下表给出不同词权重方式的对比结果，从结果可以看出，IMP和DIMP相比于IDF的weight更合理一些。但是DIMP存在部分query的树因此数据比较稀疏长尾很难建立有意义的query树。实际业务中，可以先尝试imp值，在场景合适的情况下，可以尝试DIMP方法。

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