HW4.0-WebInformation

PB16111485 张劲暾

1.假定已知文档d1和d2和查询q的词项以及词频如下：

- d1：(<2010,2>,<世博会,3>,<中国,2>,<举行,1>)
- d2：(<2005,1>,<世博会,2>,<1970,1>,<日本,2>,<举行,1>)
- q：(<2010,1>,<世博会,2>)

请给出文档d1、d2以及查询q的基于tf-idf权值的向量表示，然后分别计算q和d1、d2的余弦相似度，并说明q和哪个文档更相关

答：

doc	2010	世博会	中国	举行	2005	1970	日本
d1	0.3916	0	0.3916	0	0	0	0
d2	0	0	0	0.3010	0.3010	0.3010	0.3916
q	0.3010	0	0	0	0	0	0

cos(d1,q) = 0.80116

cos(d2,q) = 0

2.基于tf-idf的相关度计算方法有什么缺点？请给出两点以上，并加以解释。

答：

1. 词项之间的独立性假设不完全与实际相符，如：有的词项在特定的上下文中有不同的含义
2. 不能描绘词项之间的关系，如“北京地铁”会被按照“北京”“地铁”解读
3. 会有一些作弊手段，比如有意通过垃圾内容改变词项分布
4. 没有语义理解信息，对于文本主题描绘过于粗糙，比如一些高频相关词或术语可能会超过真正主题的重要性

3.在微博平台上每天都会出现一些热门微博和活跃用户，假设我们借鉴HITS算法的思想来实时检测热门微博和活跃用户，应该如何实现？请给出基本的算法思路，并给出算法伪码（须有适当注释）。

答：

可以用评论，点赞，转发等行为作为关系指向，还可以加权重，然后按照HITS算法迭代线上计算用户或微博的Authority和Hub值，返回两项指标均较高的用户或微博：

```
GodSpectofWeibo
    for each User and Weibo
        User.Aping = User.Azan = User.Azhuan = 1
        User.Hping = User.Hzan = User.Hzhuan = 1
        Weibo.Aping = Weibo.Azan = Weibo.Azhuan = 1
        Weibo.Hping = Weibo.Hzan = Weibo.Hzhuan = 1
    for each User and Weibo
        for UserR in User.Relation:
            User.Aping += UserR.Aping
            User.Azan += UserR.Azan
            User.Azhuan += UserR.Azhuan
            User.Hping += UserR.Aping
            User.Hzan += UserR.Azan
            User.Hzhuan += UserR.Azhuan
        for WeiboR in Weibo.Relation:
            Weibo.Aping += WeiboR.Aping
            Weibo.Azan += WeiboR.Azan
            Weibo.Azhuan += WeiboR.Azhuan
            Weibo.Hping += WeiboR.Aping
            Weibo.Hzan += WeiboR.Azan
            Weibo.Hzhuan += WeiboR.Azhuan
    for each User and Weibo
        User.key1 = User.Aping*Wping + User.Azan*Wzan + User.Azhuan*Wzhuan
        User.key2 = User.Hping*Wping + User.Hzan*Wzan + User.Hzhuan*Wzhuan
        Weibo.key1 = Weibo.Aping*Wping + Weibo.Azan*Wzan + Weibo.Azhuan*Wzhuan
        Weibo.key2 = Weibo.Hping*Wping + Weibo.Hzan*Wzan + Weibo.Hzhuan*Wzhuan
    Sort(User.key1)
    Sort(User.key2)
    Sort(Weibo.key1)
    Sort(Weibo.key2)
```