|  |
| --- |
|  |
| HW4-Queries |
|  |
| PB15111604金泽文 |

[**Web Information Processing and Applications**](http://staff.ustc.edu.cn/~jpq/courses/webinfo.html)

HW4-Queries

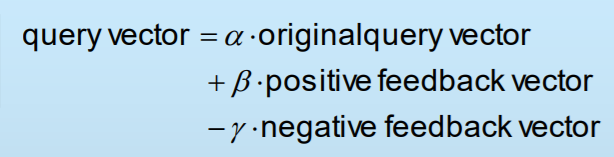
PB15111604金泽文

1. 假定初始查询Q为“extremly cheap DVDs cheap CDs”。文档d1包含词项“cheap CDs cheap software cheap DVDs”，文档d2包含“cheap thrills DVDs”。用户标记d1为相关文档，d2为不相关文档。假定我们直接使用词项频率作为文档向量中词项的权重，并采用Rocchio 1971算法进行相关性反馈，其中α＝1，β＝0.75，γ＝0.25，请给出修改后的查询向量。

根据题目中的词项，得到向量维度与词项的对应关系如下：

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| i | 1 | 2 | 3 | 4 | 5 | 6 |
| 词项 | extremly | cheap | DVDs | CDs | software | thrills |
| 初始q0 | 1 | 2 | 1 | 1 | 0 | 0 |
| 正反馈Dr | 0 | 3 | 1 | 1 | 1 | 0 |
| 负反馈Dnr | 0 | 1 | 1 | 0 | 0 | 1 |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| i | 1 | 2 | 3 | 4 | 5 | 6 |
| 词项 | extremly | cheap | DVDs | CDs | software | thrills |
| 初始q0 | 1 | 2 | 1 | 1 | 0 | 0 |
| 正反馈Dr | 0 | 3 | 1 | 1 | 1 | 0 |
| 负反馈Dnr | 0 | 1 | 1 | 0 | 0 | 1 |
| α\* 初始q0 | 1 | 2 | 1 | 1 | 0 | 0 |
| β\* 正反馈Dr | 0 | 2.25 | 0.75 | 0.75 | 0.75 | 0 |
| (-γ)\* 负反馈Dnr | 0 | -0.25 | -0.25 | 0 | 0 | -0.25 |
| 修改后的qm | 1 | 4 | 1.5 | 1.75 | 0.75 | 0 |

通过Rocchio 1971算法处理，

得到：

1. 查询扩展一般有几种实现方法？请比较一下它们之间的优点和缺点，并说明每一种方法分别适合于什么类型的信息检索应用。
   1. 人工构建同(近)义词词典
      1. 优点：经过人工处理，语义精准，命中率较高
      2. 缺点：代价大，成本高，速度慢，效率低
      3. 适用：专业名词（如医学名词等），网络词汇（如最新的流行词汇），初始词典的产生。
   2. 自动导出同(近)义词词典
      1. 优点：自动，快速，信息更具鲁棒性
      2. 缺点：涵盖的范围有限，不够精准
      3. 适用：根据现有词典、词条进行扩展，根据词条共现关系而得出对应关系。
   3. 基于查询日志挖掘出的查询等价类
      1. 优点：基于之前的查询结果，命中率高，针对性强，个性化效果好，
      2. 缺点：需要用户的参与，必须有前置的日志信息，刚开始效果不好。可能会对用户的隐私造成泄露。
      3. 适用：用户量大的查询系统，如搜索引擎。
2. 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>)

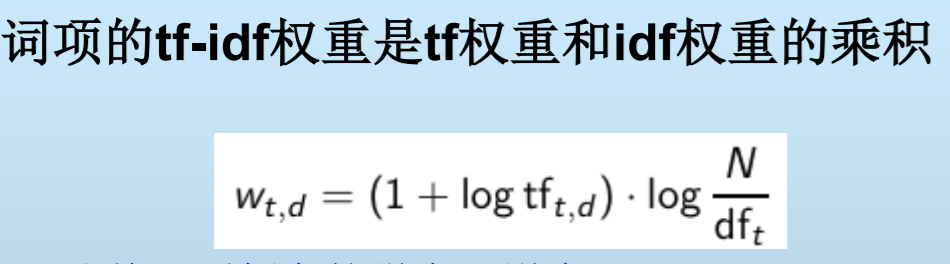
d2：(<2005,1>,<世博会,2>,<1970,1>,<日本,2>,<举行,1>)

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。

|  |  |  |  |
| --- | --- | --- | --- |
|  | d1 | d2 | q |
| 2010 | 2 | 0 | 1 |
| 世博会 | 3 | 2 | 2 |
| 中国 | 2 | 0 | 0 |
| 举行 | 1 | 1 | 0 |
| 2005 | 0 | 1 | 0 |
| 1970 | 0 | 1 | 0 |
| 日本 | 0 | 2 | 0 |

由题意，的表如下：

由

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

得tf-idf权值入下表：

所以基于tf-idf权值的向量表示为：

d1=(0.3916, 0, 0.3916, 0, 0, 0, 0)

d2=(0, 0, 0, 0, 0.3010, 0.3010, 0.3916)

q=(0.3010, 0, 0, 0, 0, 0, 0)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | d1 | d2 | q | qi \* d1i | qi \* d2i |
| 2010 | 0.3916 | 0.0000 | 0.3010 | 0.1179 | 0.0000 |
| 世博会 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| 中国 | 0.3916 | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| 举行 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| 2005 | 0.0000 | 0.3010 | 0.0000 | 0.0000 | 0.0000 |
| 1970 | 0.0000 | 0.3010 | 0.0000 | 0.0000 | 0.0000 |
| 日本 | 0.0000 | 0.3916 | 0.0000 | 0.0000 | 0.0000 |
| 平方和 | 0.3067 | 0.3346 | 0.0906 |  |  |
| 和 |  |  |  | 0.1179 | 0.0000 |
| 余弦相似度 |  |  |  | 0.7073 | 0.0000 |

所以

Cosine(d1, q)=0.7073

Cosine(d2, q)=0

q与文档d1更相关。