Web信息处理习题解答

Chp.3 Text Processing

- 假设词典中包括词｛的确，王公，实在，在理，公子｝以及所有单字集合,请分别给出句子“王公子说的确实在理”的FMM和BMM分词结果

- FMM：从左至右尽可能查找最长的词，直到当前字符与已经处理的字符串不构成词，输出已经识别的词，并从识别出来的词后面接着查找

  王公 子 说 的确 实在 理

- BMM：从右至左尽可能查找最长的词

  王 公子 说 的确 实 在理

Chp.4 Indexing

- 考虑下面的文档：

  Doc 1　new home sales top forecasts
  Doc 2　home sales rise in july
  Doc 3　increase in home sales in july
  Doc 4　july new home sales rise

  （1）画出该文档集对应的term-document关联矩阵
  假定每个单词都作为一个索引词项
  （2）画出该文档集对应的倒排索引，假定每个单词都作为一个索引词项。要求每个词项包含
  document frenquency以及term frenquency

- 关联矩阵：关联矩阵的每一列都是0/1向量，每个0/1都对应一个词项

Chp.4 Indexing

	doc1	doc2	doc3	doc4
new	1	0	0	1
home	1	1	1	1
sales	1	1	1	1
top	1	0	0	0
forecast	1	0	0	0
rise	0	0	0	1
in	0	1	1	0
July	0	1	1	1
increase	0	0	1	0

Chp.4 Indexing

- 倒排索引:

new	2	1[freq.=1],4[freq.=1]
home	4	1[freq.=1],2[freq.=1],3[freq.=1],4[freq.=1]
sales	4	1[freq.=1],2[freq.=1],3[freq.=1],4[freq.=1]
top	1	1[freq.=1]
forecast	1	1[freq.=1]
rise	2	2[freq.=1],4[freq.=1]
in	2	2[freq.=1],3[freq.=2]
July	3	2[freq.=1],3[freq.=1],4[freq.=1]
increase	1	3[freq.=1]

Chp.5 Queries

- 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$\alpha = 1$，$\beta = 0.75$，$\gamma = 0.25$，请给出修改后的查询向量

- Rocchio 1971算法：

$$\text{query vector} = \alpha \cdot \text{original query vector} + \beta \cdot \text{positive feedback vector} - \gamma \cdot \text{negative feedback vector}$$

Typically, $\gamma < \beta$

Chp.5 Queries

- Term freq:

word	extremely	cheap	DVDS	CDS	software	thrills
freq	1	2	1	1	0	0

- Original query vector: <1,2,1,1,0,0>
- Positive feedback vector of d1: <0,3,1,1,1,0>
- Negative feedback vector of d2: <0,1,1,0,0,1>
- Query vector= Original query vector+0.75* Positive feedback vector-0.25* Negative feedback vector =<1,4,1.5,1.75,0,75,0>

✓ 若出现负的权重一律设为0

Chp.5 Queries

- 在实际的Web搜索引擎中我们很少使用相关性反馈技术，试分析一下其中的原因，给出至少3个原因

- Explicit Feedback：
  - 相关反馈开销很大
  - 相关反馈生成的新查询往往很长
  - 长查询的处理开销很大
  - 用户不愿意提供显式的相关反馈
  - 有时很难理解，为什么会返回(应用相关反馈之后)某
  - 篇特定文档

- Implicit Feedback
  - 对行为分析有较高要求
  - 准确度不一定能保证
  - 某些情况下需要增加额外设备

- Pseudo Feedback
  - 没有通过用户判断，所以准确率难以保证
  - 不是所有的查询都会提高效果

Chp.6 Ranking

- 假定已知文档d1和d2和查询q的词项以及词频如下：
  d1：(<2010,1>,<世博会,3>,<中国,1>,<举行,1>)
  d2：(<2005,1>,<世博会,2>,<1970,1>,<日本,1>,<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

$$w_{t,d} = \begin{cases} 1 + \log_{10} \mathrm{tf}_{t,d} & \text{if } \mathrm{tf}_{t,d} > 0 \\ 0 & \text{otherwise} \end{cases}$$

Chp.6 Ranking

- d1:

	2010	世博会	中国	举行	2005	1970	日本
$1+\log tf_{t,d}$	1	1+lg3	1	1	0	0	0
$\log N/df_t$	lg2	0	lg2	0	lg2	lg2	lg2
weight	lg2	0	lg2	0	0	0	0

Chp.6 Ranking

同理：

d2	2010	世博会	中国	举行	2005	1970	日本
weight	0	0	0	0	lg2	lg2	lg2

q	2010	世博会	中国	举行	2005	1970	日本
weight	lg2	0	0	0	0	0	0

Sim(q,d1)=0.707

Sim(q,d2)=0

故q与d1更相关

Chp.7 Evaluation

- 假设现在有2个检索系统S1和S2，它们在文档集上分别执行查询Q1和Q2，均返回了5个结果，如下表所示：

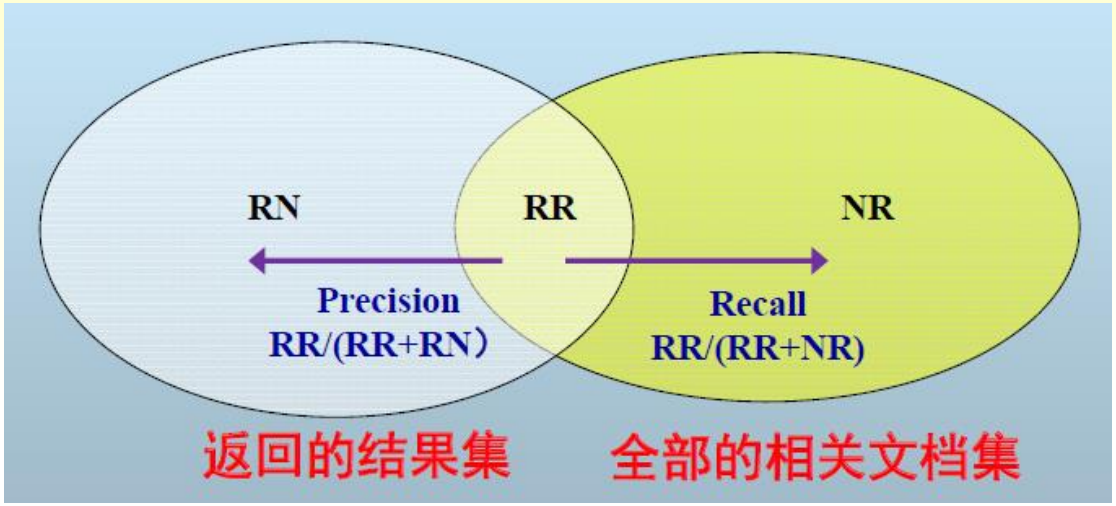
系统&查询	1	2	3	4	5
S1，Q1	d3	d5	d8	d10	d11
S1，Q2	d1	d2	d7	d11	d13
S2，Q1	d6	d7	d2	d9	d8
S2，Q2	d1	d2	d4	d11	d14

设Q1的相关文档为｛d3，d6，d7，d8｝，Q2的相关文档为｛d1，d4，d11｝

分别计算S1和S2对于查询Q1和Q2的正确率P、召回率R、F值、P@4和平均正确率AP，并计算S1和S2在所有查询上的MAP值

Chp.7 Evaluation

- 准略率与召回率：



S1：Q1　P=2/5(RN:d5,d10,d11)　R=2/4(NR:d6,d7)

Q2　P=2/5(RN:d2,d7,d13)　R=2/3(NR:d4)

F(Q1)=0.444　F(Q2)=0.5

同理

S2：Q1　P=3/5　R=3/4　F=0.667

Q2　P=3/5　R=1　F=0.75

Chp.7 Evaluation

- P@N:在第N个位置上的正确率:

S1: P@4(Q1)=2/4, P@4(Q2)=2/4
S2: P@4(Q1)=2/4, P@4(Q2)=3/4

- 平均正确率AP:这里使用简化的AP

✓ 若使用未插值的AP，则只需改变分母

S1：Q1 返回相关文档d3,d8 AP=(P@1+P@3)/2=5/6
Q2 返回相关文档d1,d11 AP=(P@1+P@4)/2=3/4

同理;

S2: AP(Q1)=13/15 AP(Q2)=29/36

- MAP:对所有查询的AP求算术平均反映在全部查询上的检索效果

S1: MAP=[AP(Q1)+AP(Q2)]/2=19/24
S2:MAP=[AP(Q1)+AP(Q2)]/2=301/360

作业提交

EX1：PB09000705 PB10000810 PB10011040 PB10203019 PB10203244
PL10215003

EX2：PB09000705 PB10000810 PB10011029 PB10011071 PB10011074
PB10011077 PB10011083 PB10203019 PB10203244 PB10207006
PL10215002 PL10215003

，

EX3：PB09000705 PB10000325 PB10000810 PB10009030 PB10009064
PB10203244 PB10207006 PB10210182 PB11210105 PL10215001
PL10215002 PL10215003

EX4：PB09000705 PB09210299 PB10000325 PB10000810 PB10009030
PB10009064 PB10011013 PB10011029 PB10203019 PB10203244
PB10207006 PB10210106 PL10215001 PL10215002 PL10215003

EX5：PB09000705 PB09210299 PB10000325 PB10000810 PB10009030
PB10009064 PB10011013 PB10011035 PB10203019 PB10203244
PB10207006 PB10207025 PB10210106 PL10215001 PL10215002
PL10215003

EX6：PB09000705 PB09210299 PB10000325 PB10000810 PB10009064
PB10011050 PB10011077 PB10030020 PB10203019 PB10203244
PB10207006 PB10210106 PL10215001 PL10215002 PL10215003

考试相关

- 考试方式：闭卷
- 5－6道大题
- 第一道为判断题（10小题）
- 其余都为问答或计算题
- 考试时间：1月9号，8:30AM-10:30AM
- Room 3C121 & 3C122

祝考试顺利