HW2

PB19030888 张舒恒

1.1

(1)记 $W_{t,d}$ 为词项 t 对于文档 d 的 tf–idf 值， $tf_{t,d}$ 为词项 t 对于文档 d 的 tf 值，N 为文档总数， df_t 为文档 t 的 df 值，则

$$W_{t,d} = (1 + \log tf_{t,d}) \cdot \log \frac{N}{df_t}$$

由此计算出各词项对于各文档的 tf–idf 值：

tf–idf	doc1	doc2	doc3
Car	4.135	3.109	4.092
Auto	3.476	5.601	0
Insurance	0	4.404	2.892
Best	2.947	0	2.763

(2)归一化运算为

$$\vec{W}_{t,d}^{'} = \frac{\vec{W}_{t,d}}{|W_{t,d}|}$$

由此计算出归一化后的文档向量：

doc1	(0.672, 0.565, 0, 0.478)
doc2	(0.400, 0.720, 0567, 0)
doc3	(0.715, 0, 0.505, 0.483)

(3)查询"Car Insurance"的向量 q=(1, 0, 1, 0)，查询与文档的相关程度以向量夹角余弦值计算为：

$$\cos(q,d)=\frac{\sum_{i=1}^{N} q_i d_i}{\sqrt{\sum_{i=1}^{N} q_i^2}\sqrt{\sum_{i=1}^{N} d_i^2}}$$

由此计算出文档相关程度：

cos(q, doc1)	0.475
cos(q, doc2)	0.683
cos(q, doc3)	0.863

则相关性排序为 doc3，doc2，doc1

1.2

邻接矩阵转置为：

$$M=\begin{bmatrix} 0 & 0 & 0 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{bmatrix}$$

归一化后：

$$R=\begin{bmatrix} 0 & 0 & 0 \\ 0.5 & 0 & 1 \\ 0.5 & 1 & 0 \end{bmatrix}$$

PageRank 转移概率矩阵为：

$$
\begin{aligned}
A &= dR + \left[\frac{(1-d)}{N}\right] ee^T \\
&= 0.85 \begin{bmatrix} 0 & 0 & 0 \\ 0.5 & 0 & 1 \\ 0.5 & 1 & 0 \end{bmatrix} + 0.05 \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix} \\
&= \begin{bmatrix} 0.05 & 0.05 & 0.05 \\ 0.475 & 0.05 & 0.9 \\ 0.475 & 0.90 & 0.05 \end{bmatrix}
\end{aligned}
$$

(2)

PageRank 值计算为：

$$
\mathrm{PR}\left(p_i\right) = \frac{1-d}{N} + d \sum_{p_j \in M(p_i)} \frac{PR\left(p_j\right)}{L\left(p_j\right)}
$$

取初始各节点的 PageRank 值为 $1/3$ ，迭代计算得

$$
PR = \begin{bmatrix} 0.05 \\ 0.475 \\ 0.475 \end{bmatrix}
$$

HITS 归一化迭代方法为：

$$
\begin{aligned}
a_{k+1} &= \frac{M^T h_k}{|M^T h_k|} \\
h_{k+1} &= \frac{M a_{k+1}}{|M a_{k+1}|}
\end{aligned}
$$

取初始 Hub, Authority 为全 1 向量，迭代计算得

$$Hub = \begin{bmatrix} \frac{\sqrt{6}}{3} \\ \frac{\sqrt{6}}{6} \\ \frac{\sqrt{6}}{6} \end{bmatrix}$$

$$Authority = \begin{bmatrix} 0 \\ \frac{\sqrt{2}}{2} \\ \frac{\sqrt{2}}{2} \end{bmatrix}$$

1.3

(1)

前 10 篇查询结果中有 4 篇相关，则 $P@10 = 0.4$

前 20 篇查询结果中有 7 篇相关，则 $P@20 = 0.35$

(2)

如果相关文档数小于 N，则查询结果中一定有非相关文档，此时 P@N 的理论上限必定小于 1

如果相关文档数大于 N，则一定有相关文档不在查询结果中，此时 R@N 的理论上限必定小于 1

(3)

共有 10 篇相关文档

前 10 篇查询结果中有 4 篇相关，则 $R@10 = 0.4$

前 20 篇查询结果中有 7 篇相关，则 $R@20 = 0.7$

由于

$$F1 = \frac{2PR}{P+R}$$

计算得 $F1@10 = 0.4$ ，$F1@20 = 0.467$

(4)

简化 AP 值为

$$AP = \frac{\frac{1}{1} + \frac{2}{2} + \frac{3}{5} + \frac{4}{9} + \frac{5}{11} + \frac{6}{15} + \frac{7}{20}}{7} = 0.607$$

1.4

(1)用户打分的平均值为：

$$\begin{aligned}
\overline{r_1} &= \frac{1+2+1}{3} = 1.333 \\
\overline{r_2} &= \frac{4+2}{2} = 3 \\
\overline{r_3} &= \frac{3+5+4+4+3}{5} = 3.8 \\
\overline{r_4} &= \frac{4+1+3}{3} = 2.667 \\
\overline{r_5} &= \frac{2+5+4+3}{4} = 3.5 \\
\overline{r_6} &= \frac{5+2}{2} = 3.5 \\
\overline{r_7} &= \frac{4+3}{2} = 3.5 \\
\overline{r_8} &= \frac{4+2}{2} = 3 \\
\overline{r_9} &= \frac{5+4}{2} = 4.5 \\
\overline{r_{10}} &= \frac{2+3}{2} = 2.5 \\
\overline{r_{11}} &= \frac{4+1+5+2+2+4}{6} = 3 \\
\overline{r_{12}} &= \frac{3+5}{2} = 4
\end{aligned}$$

其他用户与用户 5 的相似度：

$$
\begin{aligned}
\mathrm{sim}(1,5) &= \frac{\left(2-4/3\right)(2-3.5)+\left(1-4/3\right)(3-3.5)}{\sqrt{(1/3)^2\times 2+(2/3)^2}\sqrt{1.5^2+1.5^2+0.5^2+0.5^2}} = -0.456 \\
\mathrm{sim}(2,5) &= \frac{\left(4-3\right)(2-3.5)+\left(2-3\right)(5-3.5)}{\sqrt{1^2\times 2}\sqrt{1.5^2+1.5^2+0.5^2+0.5^2}} = -0.949 \\
\mathrm{sim}(3,5) &= \frac{(4-3.8)(5-3.5)+(4-3.8)(4-3.5)+(3-3.8)(3-3.5)}{\sqrt{0.8^2+1.2^2+0.2^2+0.2^2+0.8^2}\sqrt{1.5^2+1.5^2+0.5^2+0.5^2}} = 0.214 \\
\mathrm{sim}(4,5) &= \frac{\left(1-8/3\right)(2-3.5)+\left(3-8/3\right)(4-3.5)}{\sqrt{(4/3)^2+(5/3)^2+(1/3)^2}\sqrt{1.5^2+1.5^2+0.5^2+0.5^2}} = 0.552 \\
\mathrm{sim}(6,5) &= \frac{(2-3.5)(4-3.5)}{\sqrt{1.5^2+1.5^2}\sqrt{1.5^2+1.5^2+0.5^2+0.5^2}} = -0.158 \\
\mathrm{sim}(7,5) &= \frac{(2-3.5)(3-3.5)}{\sqrt{0.5^2+0.5^2}\sqrt{1.5^2+1.5^2+0.5^2+0.5^2}} = 0.474 \\
\mathrm{sim}(8,5) &= \frac{\left(4-3\right)(5-3.5)+\left(2-3\right)(3-3.5)}{\sqrt{1^2+1^2}\sqrt{1.5^2+1.5^2+0.5^2+0.5^2}} = 0.632 \\
\mathrm{sim}(9,5) &= \frac{(4-4.5)(2-3.5)}{\sqrt{0.5^2+0.5^2}\sqrt{1.5^2+1.5^2+0.5^2+0.5^2}} = 0.474 \\
\mathrm{sim}(10,5) &= \frac{(3-2.5)(2-3.5)}{\sqrt{0.5^2+0.5^2}\sqrt{1.5^2+1.5^2+0.5^2+0.5^2}} = -0.474 \\
\mathrm{sim}(11,5) &= \frac{(5-3)(2-3.5)+(2-3)(5-3.5)+(2-3)(4-3.5)+(4-3)(3-3.5)}{\sqrt{1^2+2^2+2^2+1^2+1^2+1^2}\sqrt{1.5^2+1.5^2+0.5^2+0.5^2}} \\
&= -0.710 \\
\mathrm{sim}(12,5) &= \frac{(5-4)(4-3.5)}{\sqrt{1^2+1^2}\sqrt{1.5^2+1.5^2+0.5^2+0.5^2}} = 0.158
\end{aligned}
$$

由于只有用户 1,3,6,9,11 对电影 1 有过评分，则此时用户 5 的 2–最近邻是用户 9 和用户 3，估计用户 5 对于电影 1 的评分为：

$$
Pred = 3.5 + \frac{0.214\times(3-3.8)+0.474\times(5-4.5)}{0.214+0.474} = 3.596
$$

基于物品的评分结果是 2.6，所以这里基于用户的预测评分高于基于物品的预测评分

(2)

认为用户 5 对电影 3 的评分未知，其他用户与用户 5 的相似度：

$$
\begin{aligned}
\operatorname{sim}(1,5) &= \frac{\left(1-4/3\right)(3-4)}{\sqrt{(1/3)^2\times 2+(2/3)^2}\sqrt{1^2+1^2}} = 0.289 \\
\operatorname{sim}(2,5) &= \frac{\left(2-3\right)(5-4)}{\sqrt{1^2\times 2}\sqrt{1^2+1^2}} = -0.500 \\
\operatorname{sim}(4,5) &= \frac{\left(3-8/3\right)(4-4)}{\sqrt{(4/3)^2+(5/3)^2+(1/3)^2}\sqrt{1^2+1^2}} = 0 \\
\operatorname{sim}(7,5) &= 0 \\
\operatorname{sim}(9,5) &= 0 \\
\operatorname{sim}(10,5) &= 0 \\
\operatorname{sim}(11,5) &= \frac{(2-3)(5-4)+(2-3)(4-4)+(4-3)(3-4)}{\sqrt{1^2+2^2+2^2+1^2+1^2+1^2}\sqrt{1^2+1^2}} = -0.408
\end{aligned}
$$

由于只有用户 1,2,4,7,9,10,11 对电影 3 有过评分，则此时用户 5 的 2,3,4,5–最近邻均是用户 1，估计用户 5 对于电影 3 的评分为：

$$
Pred = 4 + (2 - 4/3) = 4.667
$$

近邻数从 2 到 5 对于预测结果没有影响，这主要是因为数据量较小导致正相似度过少的结果。

在多数真实情况下，近邻数增大预测结果先变准确后变不准确，这是因为近邻数较小时受个别极端数据影响较大，近邻数较大时受多个相似度较低的数据影响。所以可以使近邻数从 2 逐渐增大，当预测结果出现明显拐点的时候即是最合适的近邻数。

1.5

归一化样本矩阵为：

$$
X = \begin{bmatrix} 1 & -1 \\ -1 & 1 \\ 2 & 2 \\ -2 & -2 \end{bmatrix}
$$

协方差矩阵为：

$$
C = \begin{bmatrix} \operatorname{cov}(x_1, x_1) & \operatorname{cov}(x_1, x_2) \\ \operatorname{cov}(x_2, x_1) & \operatorname{cov}(x_2, x_2) \end{bmatrix} = \begin{bmatrix} 3.333 & 2 \\ 2 & 3.333 \end{bmatrix}
$$

设 C 特征值为 λ，则

$$|\lambda I - C| = 0$$

解得特征值 $\lambda_1 = 5.333$， $\lambda_2 = 1.333$

对应的单位特征向量为 $\alpha_1 = [\frac{\sqrt{2}}{2}, \frac{\sqrt{2}}{2}]$，$\alpha_2 = [\frac{\sqrt{2}}{2}, -\frac{\sqrt{2}}{2}]$

则第一主成分为 $\alpha_1 = [\frac{\sqrt{2}}{2}, \frac{\sqrt{2}}{2}]$，第二主成分为 $\alpha_2 = [\frac{\sqrt{2}}{2}, -\frac{\sqrt{2}}{2}]$

降维后各样本点新特征表示为:

$$y_1 = \begin{bmatrix} 1 & -1 \\ -1 & 1 \\ 2 & 2 \\ -2 & -2 \end{bmatrix} \begin{bmatrix} \frac{\sqrt{2}}{2} \\ \frac{\sqrt{2}}{2} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 2\sqrt{2} \\ 2\sqrt{2} \end{bmatrix}$$

2.1

(a)

情境感知设备可以通过各种机制获取情境数据，例如通过特定的传感器、互联网、GPS，或者历史记录、过去的决策、位置或动作等。传感器类型和可用性已经增加并变得更加复杂，可以在平板电脑、可穿戴设备、智能手机甚至传统的笔记本电脑和台式机等设备上实现大量的情境感知计算。即使是基本的陀螺仪、加速度计和磁力计也可以获取方向和位置数据，从而预测用户当前所处的状态，比如检测到意外坠落时关闭设备，或者基于当前用户的位置来提示前方的餐厅或加油站等。

如果有多种传感器信息，则可以利用多传感器信息融合技术，比如现在较为成熟的方法:聚类分析法、证据理论法、DS 算法、最优理论法等。

(b)

在横向和纵向上均扩展传感器设备的信息采集源

横向上针对单个用户提出定制化服务，如允许用户自定义时间段开启景点位置感知，用户可定制冷门环境感知功能如行动不便的老人定制可全方位感知附近障碍的轮椅

纵向上针对常规热门情境感知服务尽可能搜集更多用户的反馈并改善之后的推荐系统，如新年期间年货是热门商品，则可以收集很多用户对于年货商店的评价，再对于每个用户开启情境感知，推荐附近综合评价最好的商店

(c)

可以使用多种传感器采集多维度的物理化学生物信息，在某一情境信息的临时缺失的情况下通过其他传感器计算当前所处状态，如 GPS 失效时我们可以测算当地温度，湿度，海拔，空气含氧量，地形，地貌等信息来判断用户所处位置。在这种情况下有可能可以填补缺失维度的情境信息。

2.2

(a)

- 用户表达采用的句式，语气，语调，如倒装句，反问句，诘问句往往是反话
- 相互矛盾或有悖常识的语义逻辑，如"你为什么不用膝盖思考这道题？"，常识是膝盖不能思考问题
- 用户的情绪，如果用户处于愤怒的状态，那么他所说的"正话"往往是想要表达相反的意思
- 用户的其他行为，比如一些购物平台上经常可以看见用户表面评价"正话"但实际打分却很低的现象

(b)

通过计算机视觉技术采集用户写下评价的面部特征并转换为情绪特征；检索用户的其他行为如是否在其他平台上给予了不相同的评价；计算用户评价中词项或短语同时出现的概率来初步判断是否存在不合常理的逻辑，比如"质量很好..."和"蒸坏这一台..."同时出现且相邻的概率很小。通过这三个方向综合判断用户是否在说反话。

2.3

不能使用马尔可夫链进行建模，因为马尔科夫链在任意时刻 t 的状态只依赖于前一时刻的状态，而用户的连续回退行为是和多个之前的状态相关的，所以不能使用马尔可夫链进行建模

2.4

(a)最优化目标修改为原目标函数和多样化性函数的加权值，多样化性函数可定义为搜索结果在主题, 受众, 热度等方向上的差异值

为保障算法的效率，可增大召回率，引入模型的正则化项避免过拟合，采用随机重启机制

(b)不需要保持多样化，因为用户已经表达了明确的搜索意图，此时如果再保持结果的多样化则会反馈很多违背用户明确意图的搜索结果，从而导致用户无法获取真正想要的信息