观点聚类任务小规模分析

刘咏彬

SNA&DS Research

April 9, 2019

目录

数据模型概述

- 对于一个问题，下面的点赞者构成 A 类点（称为 $voter$），回答构成 B 类点（称为 $answer$）。若 $voter_a$ 赞同了 $answer_b$，则在其之间连一条边，从 $voter$ 指向 $answer$。由此可见，这是一个二分图。
- 沿用图论定义，若某个 $voter$ 给 n 个回答点赞，那么称该 $voter$ 出度为 n。
- “观点一致假设”：一个 $voter$ 点赞的多个回答的观点总是类似的。
- “黏连现象”：若回答 a 和回答 b 共享很多点赞者，则称 a 与 b 是黏连的。
- “反义黏连现象”：若回答 a 和回答 b 观点并不相同，但是黏连，则称为反义黏连。

采样方法

- 首先过滤出点赞数量大于等于 3 的用户。
- 选择一个支持计划生育的回答，从过滤出来的用户中找到点赞该回答的，随机找 5 个。
- 选择一个反对计划生育的回答，从过滤出来的用户中找到点赞该回答的，随机找 5 个。
- 将这 10 个用户点赞的所有回答进行观点总结、分析。

用户 1 分析

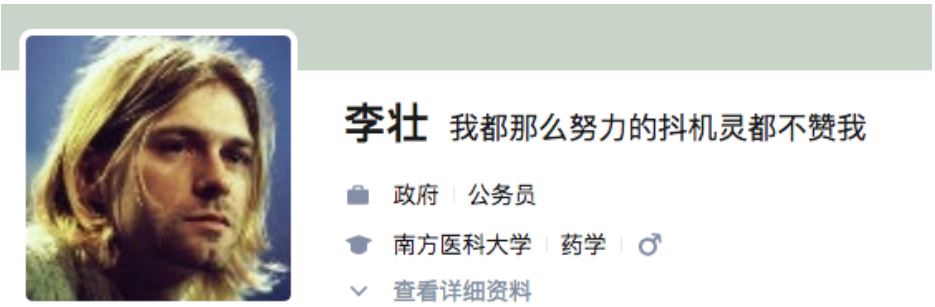


Figure 1: 用户 1 的资料

回答 id	观点总结
87068374	强烈支持计划生育
55907286	支持计划生育
14715172	认为计划生育是恶政
20439234	认为计划生育是好政策，但是执行有问题
87068374	强烈支持计划生育

Table 1: 观点总结

用户 2 分析

Figure 2: 用户 2 的资料

回答 id	观点总结
20752882	认为计划生育是理智的，支持计划生育
87068374	强烈支持计划生育
20439234	认为计划生育是好政策但是执行有问题

Table 2: 观点总结

用户 3 分析

Figure 3: 用户 3 的资料

回答 id	观点总结
15520133	认为计划生育功大于过
55907286	支持计划生育
87068374	强烈支持计划生育

Table 3: 观点总结

用户 4 分析

Figure 4: 用户 4 的资料

回答 id	观点总结
15519190	认为计划生育有害人权，但是有利于经济和社会发展
14836554	反对计划生育
23724601	认为计划生育是猛药，但是副作用大
87068374	强烈支持计划生育
23685562	支持计划生育

Table 4: 观点总结

用户 5 分析

Figure 5: 用户 5 的资料

回答 id	观点总结
23724601	认为计划生育是猛药但是副作用大
87068374	强烈支持计划生育
51767261	支持计划生育，认为底层人口倾向于多生育而中产则不想多生育

Table 5: 观点总结

用户 6 分析

Figure 6: 用户 6 的资料

回答 id	观点总结
13862730	反对计划生育，认为计划生育导致人口衰减
14684487	反对计划生育
14715172	强烈反对计划生育
56347405	强烈反对计划生育
14663415	强烈反对计划生育
62041147	反对计划生育，认为没有必要

Table 6: 观点总结

用户 7 分析

Figure 7: 用户 7 的资料

回答 id	观点总结
24913564	反对计划生育，认为计划生育对现在经济只有负面作用
14684487	反对计划生育
14715172	强烈反对计划生育，认为计划生育是恶政
27539755	强烈反对计划生育

Table 7: 观点总结

用户 8 分析

Figure 8: 用户 8 的资料

回答 id	观点总结
24913564	反对计划生育，认为计划生育对经济没有正面作用
14715172	反对计划生育，认为计划生育是恶政
14684487	反对计划生育

Table 8: 观点总结

用户 9 分析

Figure 9: 用户 9 的资料

回答 id	观点总结
24913564	反对计划生育
14684487	反对计划生育
14715172	强烈反对计划生育，认为是恶政

Table 9: 观点总结

用户 10 分析

Figure 10: 用户 10 的资料

回答 id	观点总结
24913564	反对计划生育
14663415	强烈反对计划生育
14715172	强烈反对计划生育，认为是恶政
27505244	反对计划生育，认为政策造成了很多家庭悲剧
14836554	反对计划生育，认为计划生育的预期效果缺乏科学依据

Table 10: 观点总结

总结

- 在这一问题下，voter 的观点大体是一致的，可以认为在这一问题下的“观点一致性”假设成立。

采样方法

- 同上问题，先分别找两个相反观点，然后找点赞该回答、且点赞回答大于等于 3 的用户各 5 个，分析其点赞情况。
- 但是，有的用户点赞答案太多（7-8 个），因此他们往往给双方观点都点赞了。这种用户的态度缺乏严谨性，应该过滤掉。

用户 1 分析

Figure 11: 用户 1 的资料

回答 id	观点总结
556066383	支持废除死刑
551431050	支持废除死刑，但认为目前中国尚未达到成熟时机
550336163	支持废除死刑，但认为目前中国尚未达到成熟时机

Table 11: 观点总结

用户 2 分析

Figure 12: 用户 2 的资料

回答 id	观点总结
552238861	支持废除死刑
551431050	支持废除死刑，但认为目前中国尚未达到成熟时机
555563538	支持废除死刑，认为法律的威慑力在于确定性而非重刑

Table 12: 观点总结

用户 3 分析

Figure 13: 用户 3 的资料

回答 id	观点总结
550507095	支持死刑，但是认为应该理性讨论
551431050	支持废除死刑，但认为目前中国尚未达到成熟时机
568607838	支持废除死刑

Table 13: 观点总结

用户 4 分析

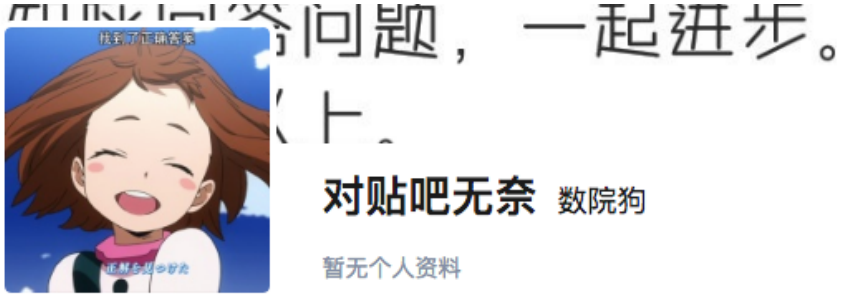


Figure 14: 用户 4 的资料

回答 id	观点总结
496608227	支持废除死刑，因为其不可逆性
551278375	支持废除死刑，但是目前时机不成熟
551431050	支持废除死刑，但是目前时机不成熟

Table 14: 观点总结

用户 5 分析

Figure 15: 用户 5 的资料

回答 id	观点总结
550893228	支持死刑
554688291	支持死刑，认为人类社会永远达不到废除死刑的发达程度
551431050	支持废除死刑，但是当前中国时机不成熟

Table 15: 观点总结

用户 6 分析

Figure 16: 用户 6 的资料

回答 id	观点总结
553613932	支持死刑
554688291	支持死刑
550121621	支持死刑

Table 16: 观点总结

用户 7 分析

Figure 17: 用户 7 的资料

回答 id	观点总结
473205072	支持死刑
554688291	支持死刑，因为人类的文明程度不可能发达到不需要死刑
550893228	支持死刑

Table 17: 观点总结

用户 8 分析

Figure 18: 用户 8 的资料

回答 id	观点总结
473205072	支持死刑
550121621	支持死刑
555055812	支持死刑
555998966	支持死刑
554688291	支持死刑
551527855	支持死刑

Table 18: 观点总结

用户 9 分析

Figure 19: 用户 9 的资料

回答 id	观点总结
554688291	支持死刑
550893228	支持死刑
550121621	支持死刑
541108664	支持死刑
391790520	支持死刑

Table 19: 观点总结

用户 10 分析

Figure 20: 用户 10 的资料

回答 id	观点总结
550121621	支持死刑
554688291	支持死刑
551527855	支持死刑

Table 20: 观点总结

总结

- 在这个问题下，呈现出明显的舆论对抗学界的倾向。绝大多数支持死刑的回答都是在嘲讽废除死刑的人，回答大多数属于宣泄情感。而法学界的主流观点则是支持废除死刑，大多数支持废除死刑的回答态度比较端正，条理比较清晰。
- 基于这一现象，应该考虑引入数据过滤清洗机制，将态度不严谨、不认真的回答清洗掉，以及点赞过多的 voter 清洗掉。

采样方法

- 与前两个问题的采样方法不同，在这个问题下，我采取直接找寻点赞数为 3 或 4 的 voter，分析他们点赞的回答。不再取定观点相反的两个基准回答。这是因为，在这个回答下面，认为“中医可靠”和“中医是伪科学”的大体各占一半，而不是前两个问题的压倒性舆论倾向，所以随机取样可以很容易找到两种观点的支持者。

用户 1 分析

Figure 21: 用户 1 的资料

回答 id	观点总结
35498669	根据定义，中医不属于科学
35733968	根据定义，中医不属于科学
103457580	认为中医属于巫术
79991241	举自身的从业经历的例子，认为中医是可信的

Table 21: 观点总结

用户 2 分析

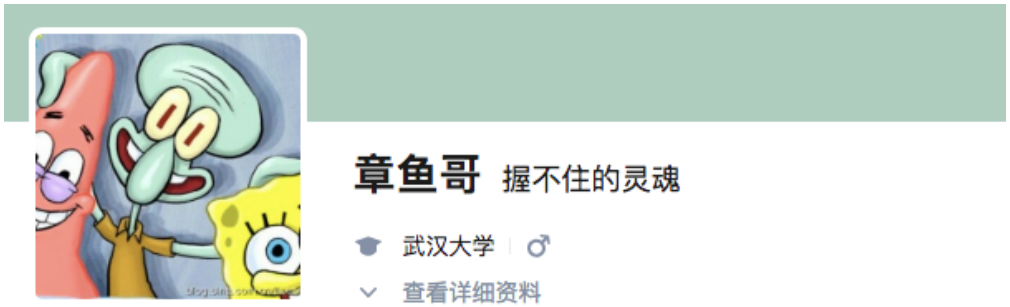


Figure 22: 用户 2 的资料

回答 id	观点总结
104484293	中医应该成为现代医学研究的对象，而不是现代医学研究的指导。
116678255	中医不是科学，而是经验学
35733968	根据定义，中医不是科学

Table 22: 观点总结

用户 3 分析

Figure 23: 用户 3 的资料

回答 id	观点总结
35613818	中医不是科学的
103457580	中医是巫术
95885753	中医不是科学，是巫术

Table 23: 观点总结

用户 4 分析

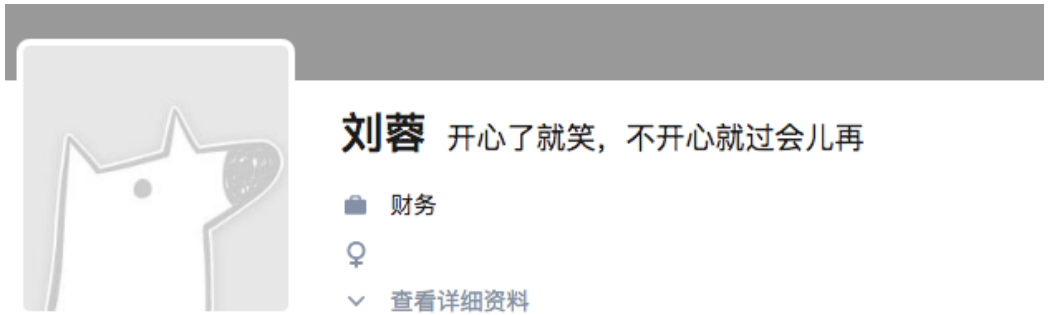


Figure 24: 用户 4 的资料

回答 id	观点总结
35656226	认为中医有效，中西医各有所长
35544843	举例，认为中医有效，中医经验是宝库
35535198	认为中医学是经验学，是可靠的
18414301	认为中医是可靠的，西医也有不可靠的药物

Table 24: 观点总结

用户 5 分析

Figure 25: 用户 5 的资料

回答 id	观点总结
546989545	认为中医有很多失传方法，中医是可靠的
104484293	中医应该成为现代医学研究的对象，而不是现代医学研究的指导。
116678255	中医不是一门科学，而是一门经验学。

Table 25: 观点总结

用户 6 分析

Figure 26: 用户 6 的资料

回答 id	观点总结
35498669	根据定义，中医不是科学
103457580	中医是巫术
35613818	中医不是科学

Table 26: 观点总结

用户 7 分析

Figure 27: 用户 7 的资料

回答 id	观点总结
35535198	用故事说明，中医是可靠的、受西医尊重的
35494773	认为这是政治斗争，西医想以科学之名打压中医
35544843	中医是可靠的，中医经验是宝库

Table 27: 观点总结

用户 8 分析

Figure 28: 用户 8 的资料

回答 id	观点总结
35733968	根据定义，中医不是科学
80062226	认为中医应该废医验药，用科学方法改造
35544843	认为中医可靠，中医的经验是宝库

Table 28: 观点总结

用户 9 分析

Figure 29: 用户 9 的资料

回答 id	观点总结
35544843	认为中医可靠，中医的经验是宝库
35535198	认为中医学是经验学，是可靠的
35529959	认为中医的大量经验是可靠的

Table 29: 观点总结

用户 10 分析

Figure 30: 用户 10 的资料

回答 id	观点总结
35498669	根据定义，中医不是科学
35733968	根据定义，中医不是科学
35751226	中医不是科学，应该消亡

Table 30: 观点总结

总结

- 大体情况来看，在中医话题下，绝大多数人符合“观点一致”假设，即一个 voter 点赞的回答观点总是一致的。
- 同时这也侧面证明，点赞数在 3-4 的 voter 是靠谱的。

出度为 2 的 voter 分析

回答 id	观点总结
35524824	认为中医与汉字一样，不可以废除
35544843	认为中医可靠

Table 31: 用户 025b5d3bc018295444d77267da1ef188

回答 id	观点总结
116678255	中医是经验学，不是科学，发展前景不如西医
35535198	中医经验学，是可靠的

Table 32: 用户 02713f7a9b1a748f43d5eac9607983ca

回答 id	观点总结
116678255	中医是经验学，不是科学，发展前景不如西医
35544843	中医是可靠的

Table 33: 用户 0271c7c27458e3b0183e5ffc025e3998

出度为 2 的 voter 分析

回答 id	观点总结
124137808	认为中医可靠
422514482	由于自身的患病历史，相信中医

Table 34: 用户 027b57d320c49964225ccbb26a10d202

回答 id	观点总结
161029977	中医不是科学，很多观点已经被证实为错误的，唯一的价值是进博物馆
141142248	中医不科学，反对中医

Table 35: 用户 027da839ec29c55a887533ac572821c3

回答 id	观点总结
35708808	认为中医是发展的，未来科学研究会逐渐验证中医的正确性
49054422	观点不明

Table 36: 用户 0280fb17cec9471eb0693773e7e7796a

定量分析方法

- 第一步，对于某个问题下的所有回答，进行人工标注。比如在“是否支持计划生育”话题下，根据答案观点，分为“反对一胎政策和其他计划生育手段”、“反对强制一胎政策，但是认为计划是有必要的，应该改进”和“支持强制一胎政策”共 3 种观点，对于观点不明的回答进行过滤。
- 第二步，对于出度大于等于 2 的 voter，统计其点赞答案的类别。
- 第三步，利用信息熵来衡量其点赞答案的多样性。

定量分析 1：计划生育问题

voter 出度	统计数量	平均信息熵	理论随机熵
2	218	0.174	0.421
3	63	0.267	0.608
4	28	0.146	0.712
5	17	0.344	0.777
6	7	0.230	0.819
7	4	0.535	0.849
8	5	0.548	0.871
10	2	0.148	0.900
11	2	0.277	0.910
14	1	0.234	0.931

Table 37: 根据 voter 出度划分的平均信息熵，为 0 代表观点非常一致，1 代表观点最不一致

理论随机熵

- 考虑，若 voter 点赞的答案与其所持观点无关，完全随机，那么其熵值为多少?
- 设 voter 的出度为 d, 这 d 个回答中，有 x 个是第 1 类回答，y 个是第 2 类回答，那么有 $d-x-y$ 个第 3 类回答。熵为 $e(x,y)=$ $\frac{1}{d}\left[d\log d-x\log x-y\log y-(d-x-y)\log(d-x-y)\right]$。
- 由凸函数的性质可知，$e(x,y)\leq 1$。
- 出现这种情形的概率为

$$p(x,y)=\frac{d!}{x!\,y!\,(d-x-y)!\,3^d}$$

，且有

$$\sum_{0\leq x+y\leq d} p(x,y)=1$$

理论随机熵的性质

上界

$$E(d) = \sum_{0 \leq x+y \leq d} p(x,y)\, e(x,y) \leq 1$$

(不证自明)

极限

$$\lim_{d \to \infty} E(d) = 1$$

(已证明)

极限证明

为方便化简，设

$$h(x) = \begin{cases} x \log x & , \quad x > 0 \\ 0 & , \quad x = 0 \end{cases} \tag{1}$$

接着化简原式，

$$E(d) = \log d - \frac{1}{d} \sum_{0 \le x+y \le d} \frac{d!}{x!\, y!\, (d-x-y)!\, 3^d} \left[h(x) + h(y) + h(d-x-y) \right] \tag{2}$$

$$= \log d - \frac{1}{d} \sum_{0 \le x+y \le d} \frac{d!}{x!\, y!\, (d-x-y)!\, 3^{d-1}} h(x) \tag{3}$$

$$= \log d - \sum_{\substack{x>0 \\ 0<x+y \le d}} \frac{(d-1)!}{(x-1)!\, y!\, (d-x-y)!\, 3^{d-1}} \log x \tag{4}$$

$$= \log d - \sum_{0 \le x+y \le d-1} \frac{(d-1)!}{x!\, y!\, (d-1-x-y)!\, 3^{d-1}} \log(x+1) \tag{5}$$

极限证明（续 1）

不难看出，我们得到了新的概率项 $\frac{(d-1)!}{x!\,y!\,(d-1-x-y)!\,3^{d-1}}$，即:

$$\sum_{0\le x+y\le d-1} \frac{(d-1)!}{x!\,y!\,(d-1-x-y)!\,3^{d-1}} = 1 \tag{6}$$

并且，概率项 $(x, y, d-1-x-y), (y, x, d-1-x-y), (d-1-x-y, x, y)$ 等具有对称性，易知概率项的个数总共为 $\frac{d(d+1)}{2}$，接下来我们讨论 d 是 3 的整数倍的情形，其他情况不难推广。由于 $d-1$ 不是 3 的倍数，固不论 x, y 取何值，等值概率项 $(x, y, d-1-x-y)$ 都出现了 3 的倍数次。我们将每 3 个相同的概率项合并为 1 个，将这些概率项构造成新集合 $A(d)$（有 $\frac{d(d+1)}{6}$ 个元素），那么

$$\sum_{(x,y,d-1-x-y)\in A(d)} \frac{(d-1)!}{x!\,y!\,(d-1-x-y)!\,3^{d-1}} = \frac{1}{3} \tag{7}$$

$$E(d) = \sum_{(x,y,d-1-x-y)\in A(d)} \frac{(d-1)!}{x!\,y!\,(d-1-x-y)!\,3^{d-1}} (\log\frac{d}{x+1} + \log\frac{d}{y+1} + \log\frac{d}{d-x-y}) \tag{8}$$

极限证明（续 2）

分析式8，

$$\log \frac{d}{x+1} + \log \frac{d}{y+1} + \log \frac{d}{d-x-y} = \log \frac{d^3}{(x+1)(y+1)(d-x-y)} \tag{9}$$

利用拉格朗日乘数法或均值不等式，可得当 $x = y = \frac{d}{3}$ 时分母取得最大值，即

$$\log \frac{d}{x+1} + \log \frac{d}{y+1} + \log \frac{d}{d-x-y} \geq \log \frac{d^3}{(\frac{d}{3}+1)(\frac{d}{3}+1)\frac{d}{3}} \tag{10}$$

$$= \log \frac{27d^3}{d^3+2d^2+d} \tag{11}$$

因此

$$E(d) \geq \sum_{(x,y,d-1-x-y)\in A(d)} \frac{(d-1)!}{x!\,y!\,(d-1-x-y)!\,3^{d-1}} \log \frac{27d^3}{d^3+2d^2+d} \tag{12}$$

$$= 1 - \frac{1}{3}\log\left(1+\frac{2}{d}+\frac{1}{d^2}\right) \tag{13}$$

极限证明（续 3）

由于

$$\lim_{d\to\infty} 1-\frac{1}{3}\log\left(1+\frac{2}{d}+\frac{1}{d^2}\right)=1$$

且 $E(d)<1$，由两边夹定理可知，

$$\lim_{d\to\infty} E(d)=1$$

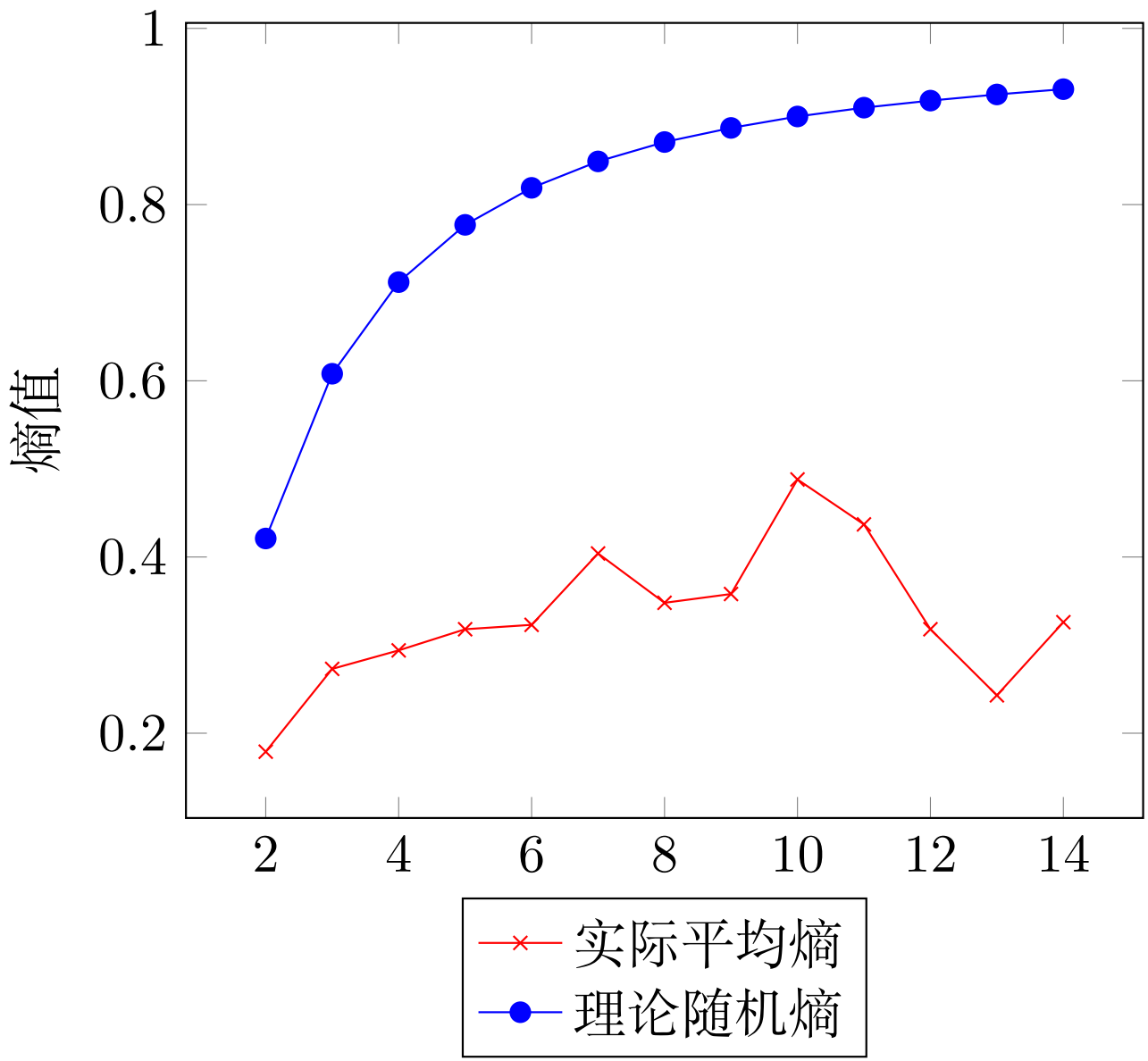
实际上，根据数值计算结果，$E(d)$ 应该还随 d 单调增加，不过严格证明暂时还没有想到。

定量分析 2：中医是否科学（分 3 类观点）

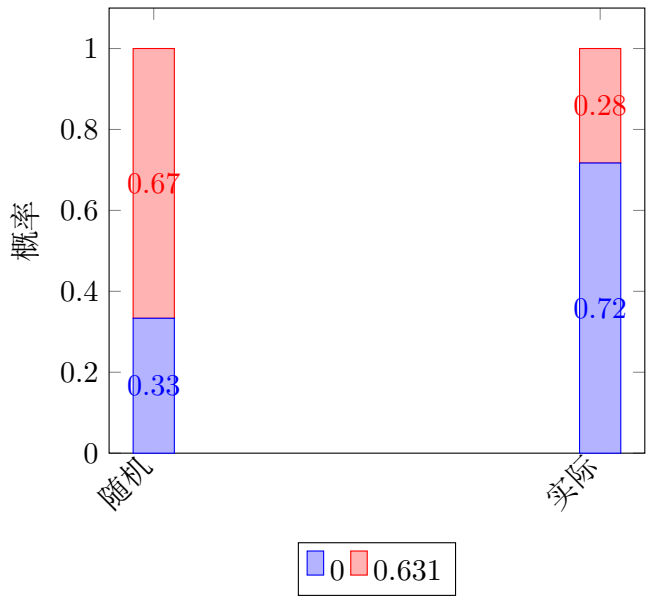
voter 出度	统计数量	平均信息熵	理论随机熵
2	4870	0.179	0.421
3	1392	0.273	0.608
4	584	0.294	0.712
5	284	0.318	0.777
6	121	0.323	0.819
7	80	0.404	0.849
8	46	0.348	0.871
9	41	0.358	0.887
10	34	0.488	0.900
11	11	0.437	0.910
12	13	0.318	0.918
13	8	0.243	0.925
14	5	0.326	0.931

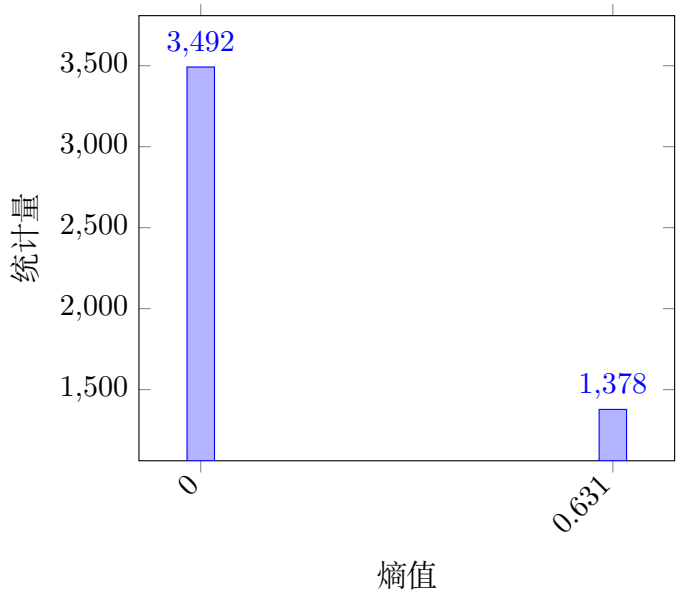
Table 38: 根据 voter 出度划分的平均信息熵，为 0 代表观点非常一致，1 代表观点最不一致

折线图 1

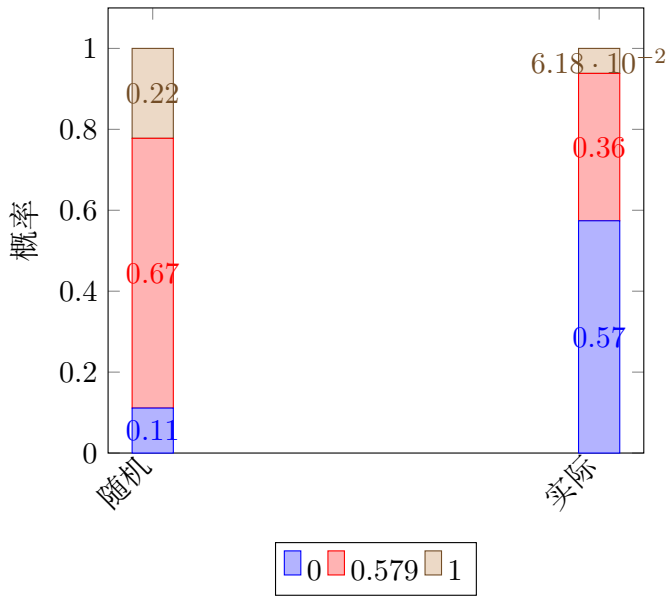


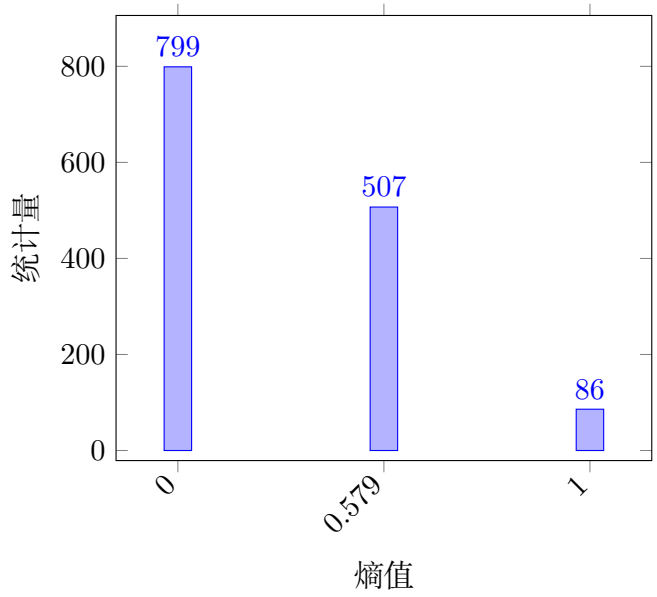
熵值统计图 $(d = 2)$


概率
1
0.8
0.6
0.4
0.2
0
0.67
0.33
0.28
0.72
随机
实际
0
0.631

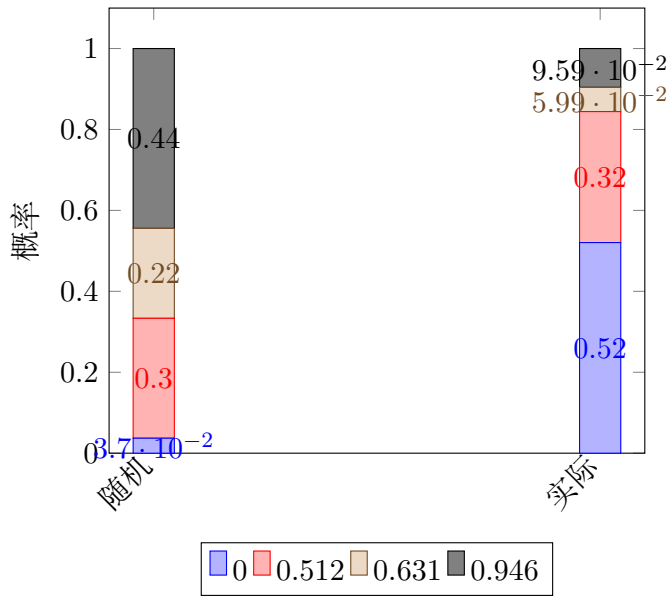

统计量
3,500
3,000
2,500
2,000
1,500
3,492
1,378
0
0.631
熵值

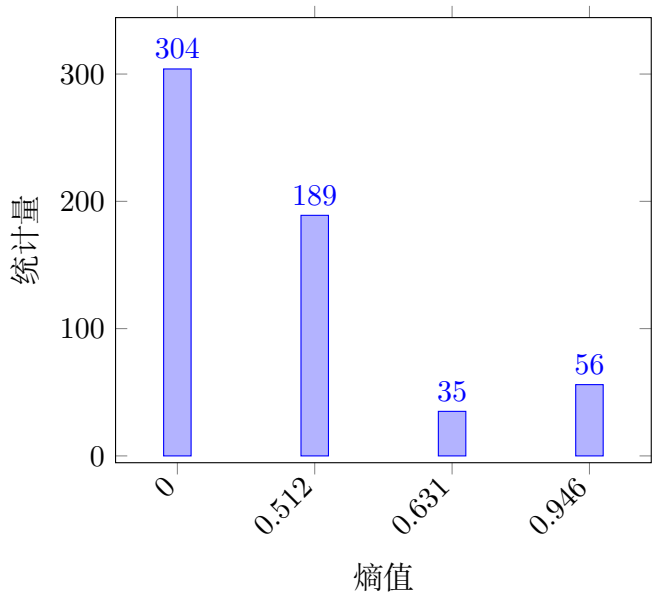
熵值统计图 $(d = 3)$


概率
1
0.8
0.6
0.4
0.2
0
0.22
0.67
0.11
$6.18 \cdot 10^{-2}$
0.36
0.57
随机
实际
0
0.579
1

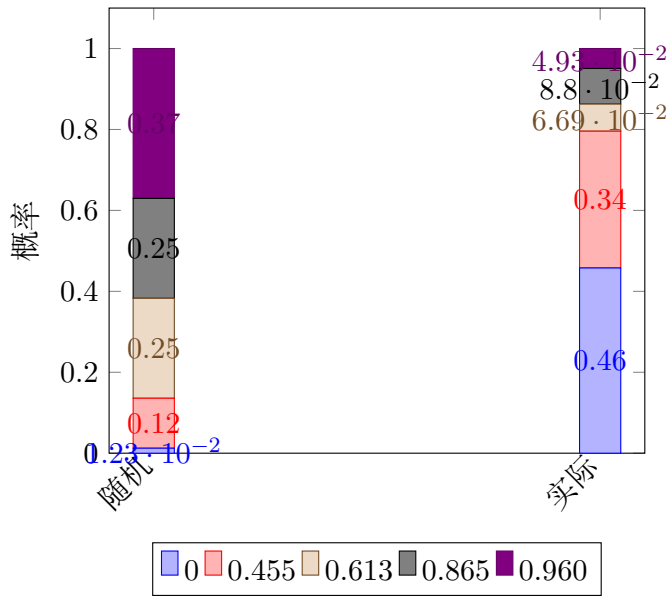

统计量
800
600
400
200
0
799
507
86
0
0.579
1
熵值

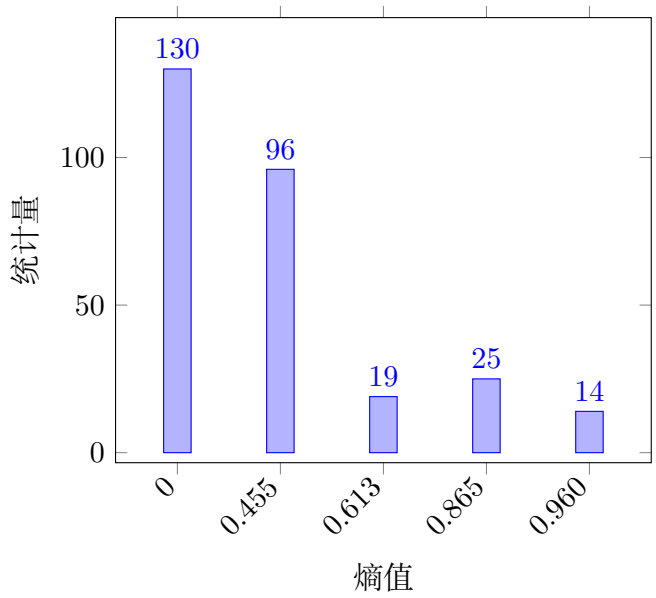
熵值统计图 $(d = 4)$


概率
1
0.8
0.6
0.4
0.2
0
0.44
0.22
0.3
$3.7 \cdot 10^{-2}$
$9.59 \cdot 10^{-2}$
$5.99 \cdot 10^{-2}$
0.32
0.52
随机
实际
0
0.512
0.631
0.946

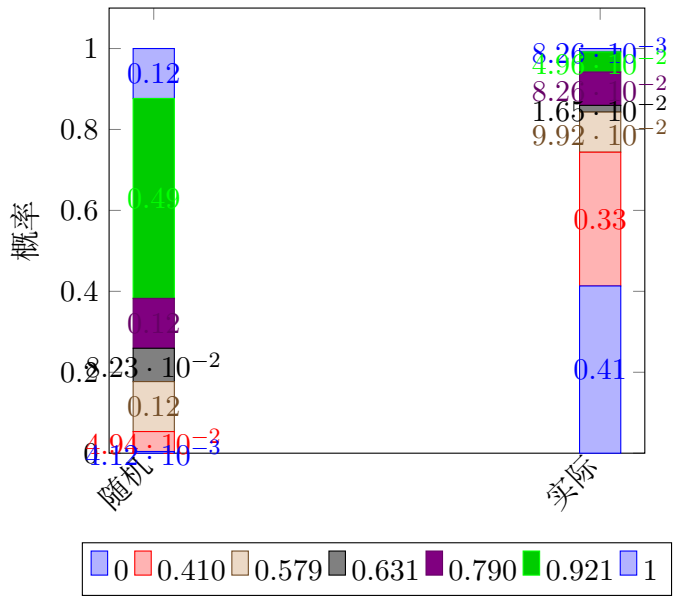

统计量
300
200
100
0
304
189
35
56
0
0.512
0.631
0.946
熵值

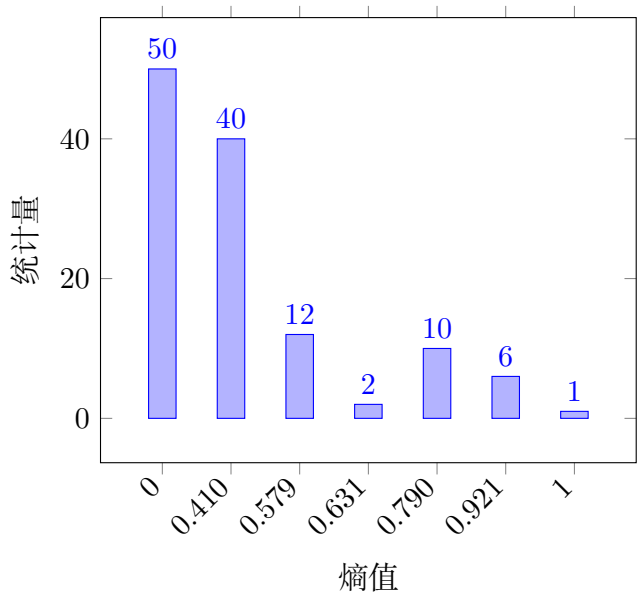
熵值统计图 ($d = 5$)


概率
1
0.8
0.6
0.4
0.2
0
0.37
0.25
0.25
0.12
$1.23 \cdot 10^{-2}$
$4.93 \cdot 10^{-2}$
$8.8 \cdot 10^{-2}$
$6.69 \cdot 10^{-2}$
0.34
0.46
随机
实际
0
0.455
0.613
0.865
0.960


统计量
130
96
19
25
14
100
50
0
0
0.455
0.613
0.865
0.960
熵值

熵值统计图 ($d = 6$)


概率
1
0.8
0.6
0.4
0.2
0.12
0.49
0.12
$8.23 \cdot 10^{-2}$
0.12
$4.94 \cdot 10^{-3}$
$4.12 \cdot 10^{-3}$
$8.26 \cdot 10^{-3}$
$4.96 \cdot 10^{-2}$
$8.26 \cdot 10^{-2}$
$1.65 \cdot 10^{-2}$
$9.92 \cdot 10^{-2}$
0.33
0.41
随机
实际
0
0.410
0.579
0.631
0.790
0.921
1


统计量
40
20
0
50
40
12
2
10
6
1
0
0.410
0.579
0.631
0.790
0.921
1
熵值

小结论

根据这些数据观察，我们得出结论：点赞者支持的答案大多数共享同一个观点，我们可以进行一些聚类算法的尝试。

衡量标准

为了衡量某种聚类方法的准确性，我们可以采用一些被广泛使用的 metrics，比如 ARI，NMI，V-Measure，FMI。前面这几种方法都是在预测类别和真实类别都已有的情况下的手段。还有几种聚类方法可以在仅有预测标签的情形下使用，比如 Silhouette Coefficient，Calinski-Harabaz Index，Davies-Bouldin Index 等。

第一个思路：相似度矩阵与聚类算法

- 第一条路线是使用传统的聚类算法，但这样必须先计算出来答案之间的相似度，所以整个算法流程分为两步，第一步是计算相似度矩阵，第二步是聚类算法。
- 已经调研过可用的算法有：AP 聚类、谱聚类、层次聚类、DBSCAN。

初期探索 (α_0 阶段)

- 用 Share_{ij} 表示第 i 和 j 个回答之间共享了多少个点赞者 (有多少个人同时赞了这两个回答)。
- 构造 S 矩阵：确定一个 ϵ，$\text{S}_{ij} = 1$ 当且仅当 Share_{ij} 在 $\{\text{Share}_i\}$ 中大小排名小于 ϵ 或 Share_{ji} 在 $\{\text{Share}_j\}$ 中大小排名小于 ϵ。
- 调参数 ϵ、以及聚类参数 (kmeans 或 discretize)，每次实验重复 5 次取均值。

α_0 阶段部分调参结果

S 构造参数	聚类参数	RMA	FMI
随机	kmeans	48.37	42.86
真实关系矩阵	kmeans	100	100
share 矩阵	kmeans	54.68	72.28
eps=4	discretize	55.90	72.03
eps=4	kmeans	55.84	73.02

Table 39: 由表可知，初步阶段的方法的聚类结果显著好于随机结果，但是仍然有较大的改进空间。

α_1 阶段

- 在这一阶段，我尝试修改 S 矩阵的构造方法，将原先的对称化策略由或运算改为取平均值。
- 其他调参策略不变，发现结果有所改观，计算结果随 eps 变化不再像 α_0 阶段那样大，但是最好结果没有太大改变。

α_1 阶段部分调参结果

S 构造参数	聚类参数	RMA	FMI
随机	kmeans	48.37	42.86
真实关系矩阵	kmeans	100	100
share 矩阵	kmeans	54.68	72.28
eps=12	discretize	55.99	71.14
eps=4	kmeans	55.66	73.04

Table 40: 由表可知，初步阶段的方法的聚类结果显著好于随机结果，但是仍然有较大的改进空间。

总结与思考

- 实际上，我发现使用真实类别的关系矩阵进行聚类，聚类准确率可以高达 100%，这说明聚类算法本身是没有什么问题的，主要是相似度矩阵计算不够准确。
- 也就是说，相似度矩阵接近真实关系矩阵是聚类效果好的充分条件。
- 这个问题可以抽象成一个经典推荐算法的问题，使用 ItemCF 协同过滤算法优化相似度矩阵。
- 考虑到 VA 矩阵的稀疏性，我们可以使用矩阵填充算法将其先补全，然后再计算相似度矩阵。

协同过滤初期探索阶段 (β_0 阶段)

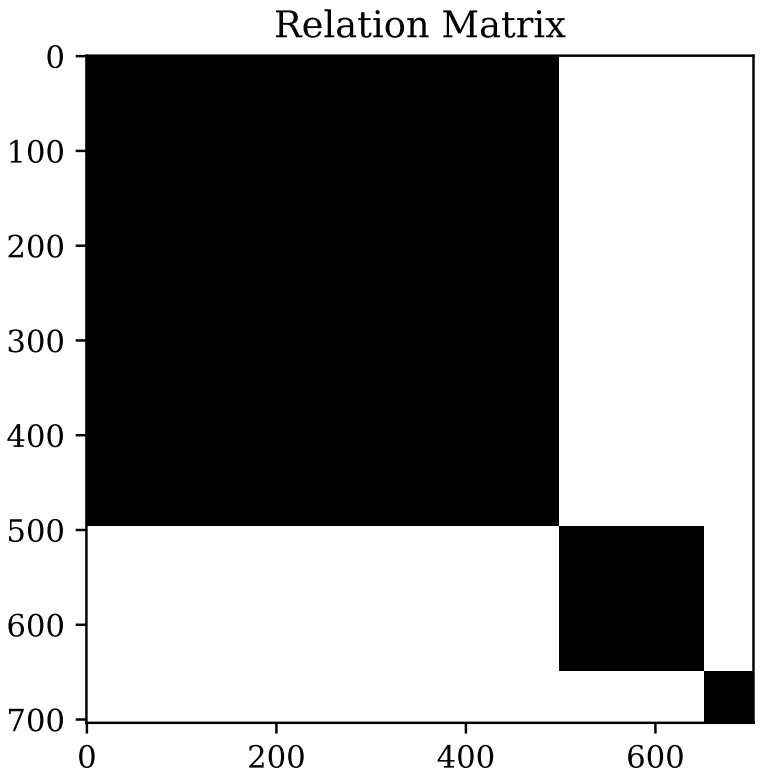
- 我借鉴了 ItemCF 协同过滤算法，计算出相似性矩阵，再使用谱聚类。在 ItemCF 中，有几种相似性度量手段，我都已经尝试过了。
- Jaccard 相似度:$J(A,B)=\frac{|A\cap B|}{|A\cup B|}$
- 改进 Jaccard 相似度:$J(A,B)=\frac{|A\cap B|}{\sqrt{|A|\,|B|}}$
- 欧式距离相似度: $E(X,\,Y)=\frac{1}{1+\|X-Y\|_2}$
- 余弦相似度: $C(X,\,Y)=\frac{X\,Y}{\|X\|_2\,\|Y\|_2}$
- Pearson 系数: $\rho(X,\,Y)=\frac{\mathrm{Cov}(X,Y)}{\sigma_X\,\sigma_Y}$

β_0 阶段实验结果

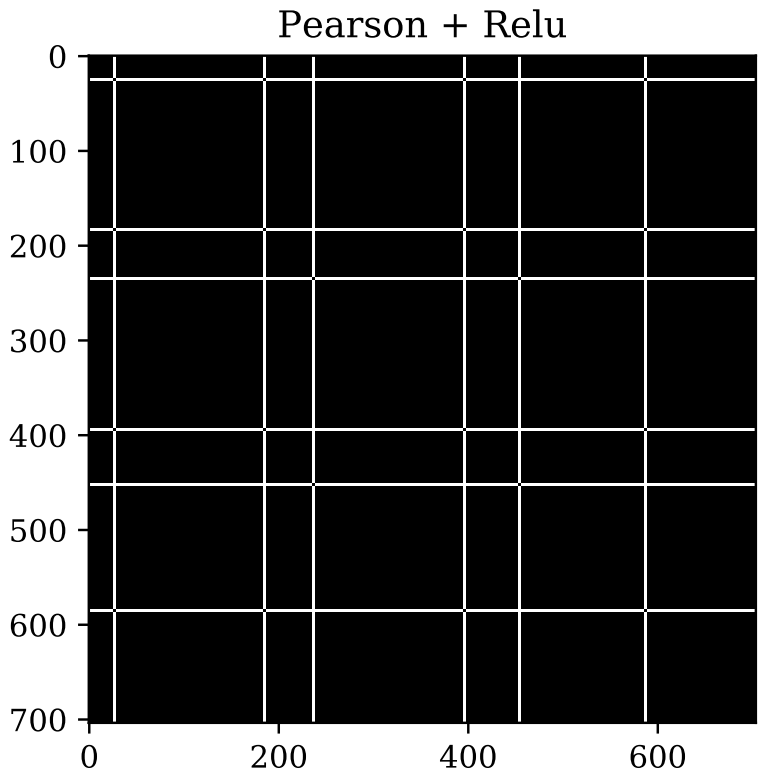
S 构造参数	聚类参数	RMA	FMI
随机	kmeans	48.37	42.86
真实关系矩阵	kmeans	100	100
Jaccard 相似度	discretize	55.06	71.90
改进的 Jaccard 相似度	discretize	55.24	72.11
欧式距离相似度	discretize	52.04	53.54
余弦相似度	discretize	54.82	71.65
Pearson 系数 +sigmoid	discretize	53.89	67.34
Pearson 系数 +relu	discretize	54.93	71.48

Table 41: 似乎使用协同过滤技术的效果不如一开始的方法

聚类预测的关系矩阵可视化

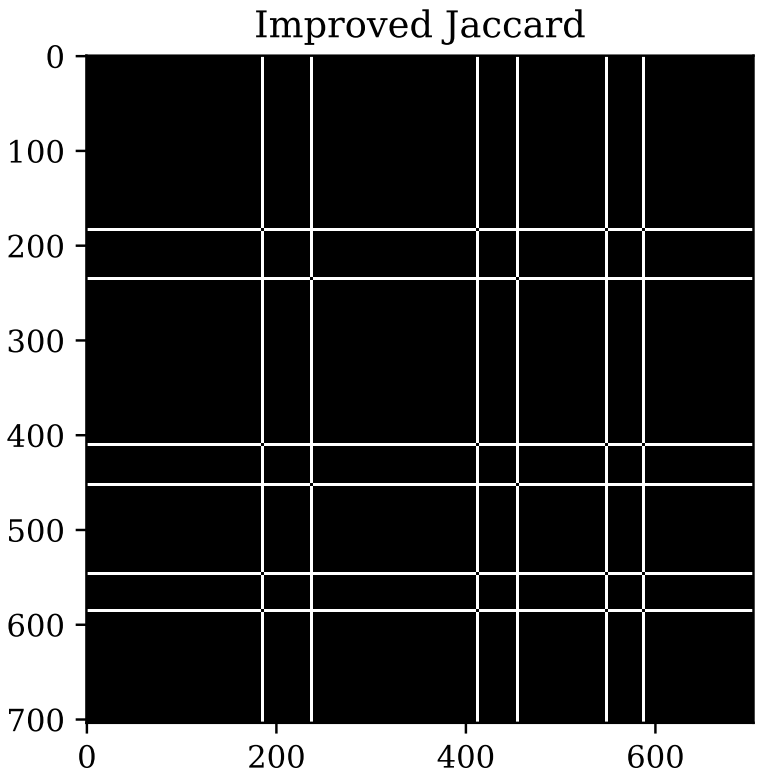


(a) 真实关系矩阵

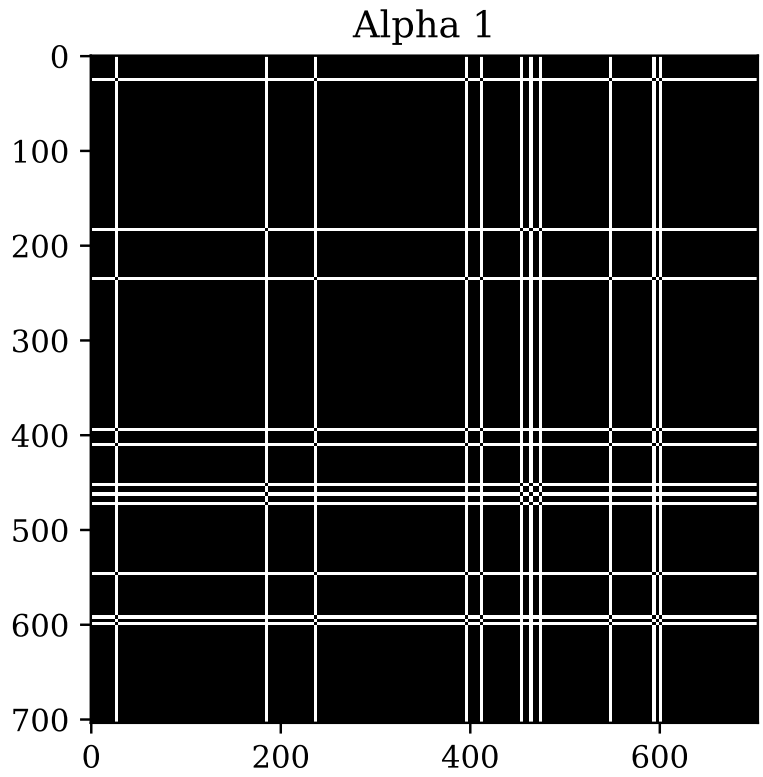


(b) ItemCF(Pearson + Relu)

聚类预测的关系矩阵可视化（续）



(a) ItemCF(Improved Jaccard)

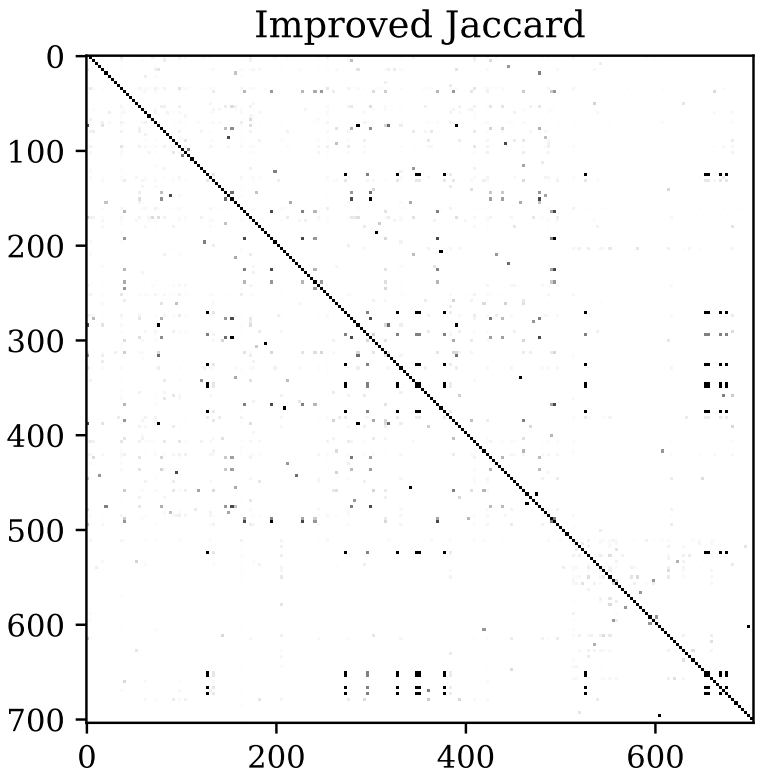


(b) $\alpha_1, \epsilon = 12$

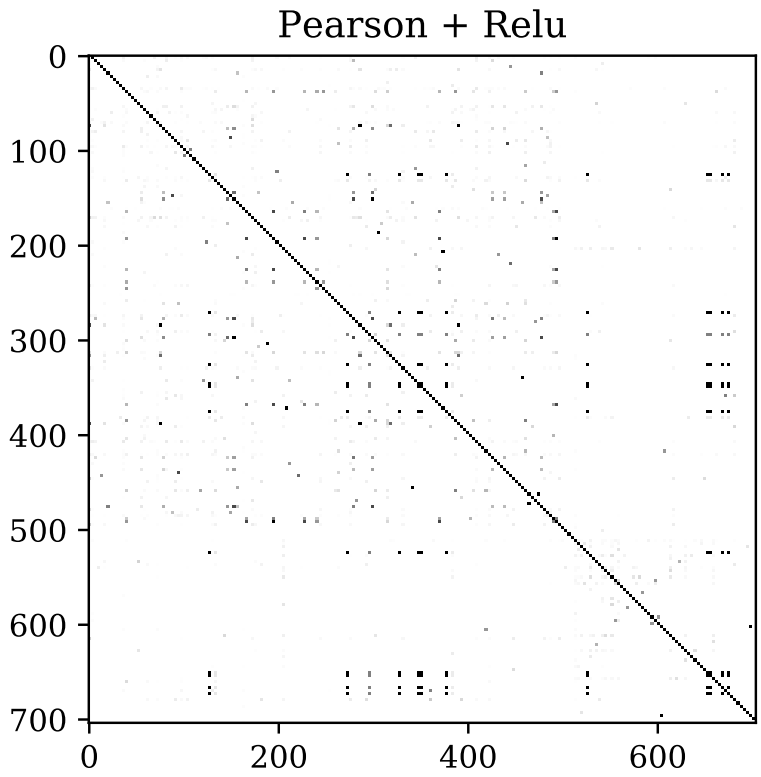
小结论

由此可以看出，预测结果效果还是很差的，有很大的提高空间。

S 矩阵可视化

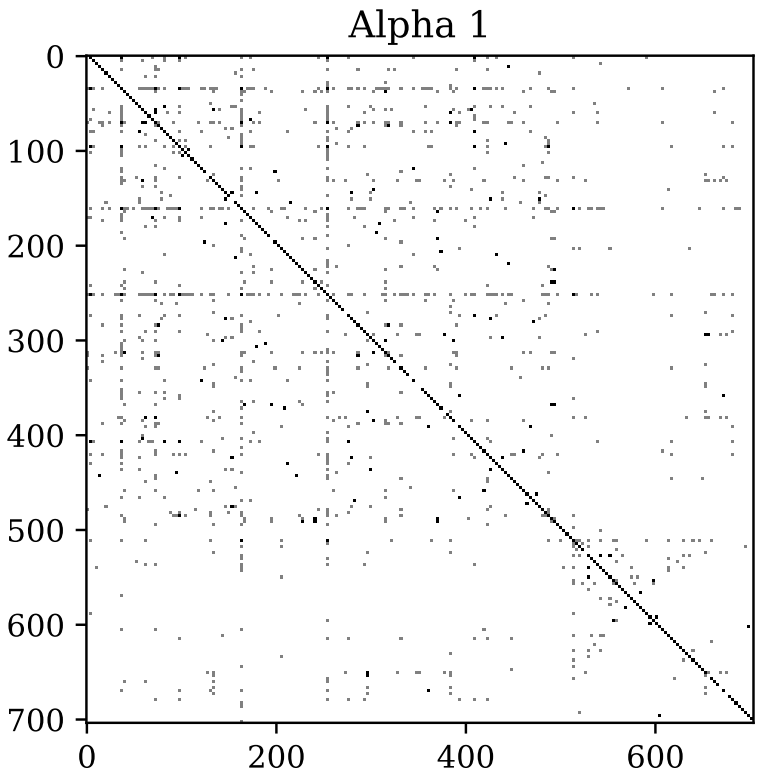


(a) ItemCF(Improved Jaccard)

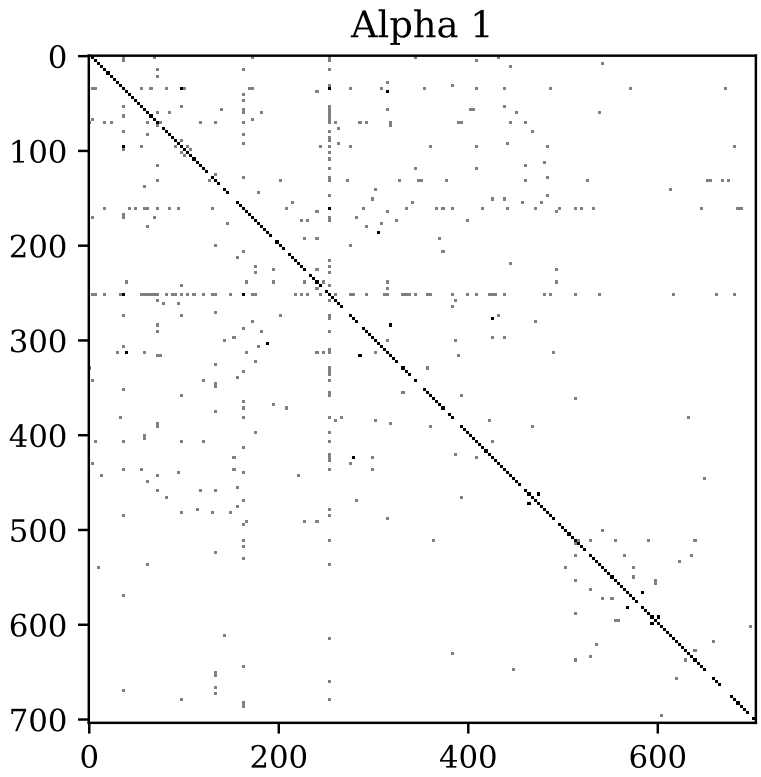


(b) ItemCF(Pearson + Relu)

S 矩阵可视化（续）



(a) $\alpha_1, \epsilon = 12$



(b) $\alpha_1, \epsilon = 4$

尝试推广 α_1 阶段的相似性矩阵

- 根据矩阵可视化的结果，我们可以看出来，关系比较紧密的 answer 之间的相似度矩阵比较稀疏，可以尝试使用一些手段使其稠密一些。
- 一种方法是直接推广 share 矩阵：Share_{ij} 代表的是从节点 i 到达 j 经过恰好 1 个 voter 节点的路径数量。那么我们可以由此计算出恰好经过 2 个、3 个 $\cdots n$ 个节点的矩阵，将其加权作为新的 share 矩阵。根据图论相关定理，直接使用矩阵乘法即可完成。
- 第二种方法是直接推广相似度矩阵，类比上条，直接用相似度矩阵相乘得到新的矩阵。
- 由于第二种方法实验起来比较简单，称之为 α_2 阶段，而第一种方法为 α_3 阶段。

α_2 阶段矩阵构造方法

在这个阶段，我们采用的 S 矩阵构造方法为：

$$
\begin{aligned}
&\mathrm{S}_0 \leftarrow \mathrm{RankConstructFunc}(\mathrm{Share}, \epsilon_0) \\
&\mathrm{S}_1^* \leftarrow \mathrm{S}_0\,\mathrm{S}_0 \\
&\mathrm{S}_1 \leftarrow \mathrm{RankConstructFunc}(\mathrm{S}_1^*, \epsilon_1) \\
&\mathrm{S}_2^* \leftarrow \mathrm{S}_1\,\mathrm{S}_1 \\
&\cdots
\end{aligned}
$$

使用 S_2^* 作为相似性矩阵的部分调参结果

S 构造参数	聚类参数	RMA	FMI
(4, 2)	discretize	58.49	58.83
(8, 4)	discretize	56.96	71.38
(7, 4)	discretize	56.62	71.41
(8, 3)	discretize	56.69	68.27
(7, 3)	discretize	56.55	69.88

Table 42: 比 α_1 的结果好一些了，但是仍然没有显著改进，只是提升了 1 个点左右。

使用 S_1 作为相似性矩阵的部分调参结果

S 构造参数	聚类参数	RMA	FMI
(11, 4)	discretize	56.39	71.07
(15, 4)	discretize	56.33	71.40
(16, 4)	discretize	56.23	70.97
(3, 10)	discretize	56.22	72.48
(7, 4)	discretize	56.18	71.21
(8, 3)	discretize	56.15	68.30

Table 43: 比 α_1 的结果好一些了，但是仍然没有显著改进，只是提升了 1 个点左右。

使用 S_3^* 作为相似性矩阵的部分调参结果

S 构造参数	聚类参数	RMA	FMI
(7, 3, 10)	discretize	57.51	70.24
(7, 4, 11)	discretize	57.04	69.75
(8, 4, 10)	discretize	56.72	70.40
(4, 2, 6)	discretize	58.35	59.26
(4, 2, 7)	discretize	58.25	59.70
(4, 2, 8)	discretize	57.47	57.40

Table 44: 由此可见，eps 的前两项在低阶情况下效果好，往往意味着高阶情况下也不错。

使用 S_4^* 作为相似性矩阵的部分调参结果

S 构造参数	聚类参数	RMA	FMI
(7, 3, 10, 11)	discretize	57.10	70.42
(7, 4, 11, 11)	discretize	56.60	71.04
(8, 4, 9, 11)	discretize	57.31	71.44
(4, 2, 6, 9)	discretize	60.07	60.44
(4, 2, 7, 12)	discretize	59.70	60.22
(4, 2, 8, 6)	discretize	59.12	59.18

Table 45: 由此可见，eps 的前两项在低阶情况下效果好，往往意味着高阶情况下也不错。

将稠密化方法移植到 ItemCF 上

S 构造参数	聚类参数	RMA	FMI
Improved Jaccard (eps=3,12,7)	discretize	57.15	71.71
Improved Jaccard (eps=3,7,11)	discretize	57.01	70.99
Improved Jaccard (eps=3,11,13)	discretize	56.96	71.13
Improved Jaccard (eps=3,7,13)	discretize	56.95	71.16
Improved Jaccard (eps=3,11,7)	discretize	56.94	70.85
Improved Jaccard (eps=3,14,11)	discretize	56.93	70.59

Table 46: 由此可见，α_2 稠密化手段是改进相似度矩阵的有效方法。

α_3 实验：采用高阶的关联矩阵

在这个阶段的实验中，我计算出回答与回答之间的经过更多 voter 节点的路径数。在这里，我使用 Share_1 和 Share_2 来代表经过经过 1 个节点、2 各节点的路径数矩阵。然后将其分别使用 RankConstructFunc 计算相似度，最后取平均数。

S 构造参数	聚类参数	RMA	FMI
5, 13	discretize	55.68	72.38
3, 4	discretize	55.49	72.41
5, 5	discretize	55.42	72.30
2, 8	discretize	55.39	72.20
10, 8	discretize	55.36	72.08

F-SVD 矩阵分解算法

F-SVD 算法是基于隐状态假设的，即：假设每一个用户和回答都可以表示为一个 k 维的向量（隐状态），而用户对回答的赞同程度可以由它们的内积表示：$r_{ij} = p_i^T q_j$。回答与回答之间可以用向量差的范数来衡量其相似度。F-SVD 算法的训练过程是，将一部分评分作为训练集，剩余部分作为评测集。对训练集优化如下的损失函数：

$$\mathfrak{L} = \sum_{r_{ij} \in S_{training}} \|p_i^T q_j + b_i - \mu - r_{ij}\| + \lambda(\|P\| + \|Q\| + \|b\|)$$

其中的 b 是偏置项，用来避免用户本身存在的属性（如某个用户打分习惯较高）干扰训练，μ 是全局打分的平均值。在这里，如果用户给某回答点赞了，评分就是 1，否则为 unknown。

γ_0 实验：采用 F-SVD 矩阵分解算法

使用著名的 Funk SVD 矩阵分解算法来获得物品的隐状态表示（向量），然后计算向量之间的相似度。

S 构造参数	聚类参数	RMA	FMI
qi 特征向量，欧氏距离相似度	discretize	48.35	42.83
qi 特征向量，余弦相似度 +01 整流	discretize	48.54	43.01
qi 特征向量，Pearson+Relu	discretize	48.38	42.89

先填充再计算相似度

- 先前的方法是，用得到的回答隐状态表示 Q 直接计算相似度，但是效果很差，因此应该尝试先填充评分矩阵再计算相似度的方法。
- 在计算欧式距离相似度时，原先回答的距离为 $\|q_i - q_j\| = (q_i - q_j)^T (q_i - q_j)$，而在填充矩阵之后，回答之间的距离变为 $(q_i - q_j)^T P^T P (q_i - q_j) = (P(q_i - q_j))^T (P(q_i - q_j)) = \|P\,q_i - P\,q_j\|$。
- 由此看出，原先的回答的隐状态表示实际转换为了 $P\,q_i$。

γ_0 实验：采用 F-SVD 矩阵分解算法

使用著名的 Funk SVD 矩阵分解算法来获得物品的隐状态表示（向量），然后填充原矩阵，再利用原矩阵计算相似度。

S 构造参数	聚类参数	RMA	FMI
$P\,q_i$ 特征向量，欧氏距离相似度	discretize	50.03	48.34
$P\,q_i$ 特征向量，余弦相似度 +01 整流	discretize	48.43	42.94
$P\,q_i$ 特征向量，Pearson+Relu	discretize	48.29	42.70

γ_0 实验：采用 F-SVD 矩阵分解算法

- 可以看出来，这个方法的效果依然并不理想。
- 但是本身 F-SVD 是经典的推荐系统算法，而且本身效果在其他的推荐系统上已经得到了验证。所以在后续的工作中，可能需要解释为何在本课题中 F-SVD 的表现为何如此之差。
- 目前猜测的一个可能原因是，由于我们只能知道某个用户点赞的回答（评分为 1），而无法获得其负面评价（不赞同某个回答），因而评分矩阵的项要么为 1，要么为 unknown。矩阵的结构与电影评分系统的 scale(1-5，unknown) 大不相同。在知乎公开不赞同数据之前，这一点是无法改变的。

详解评价指标

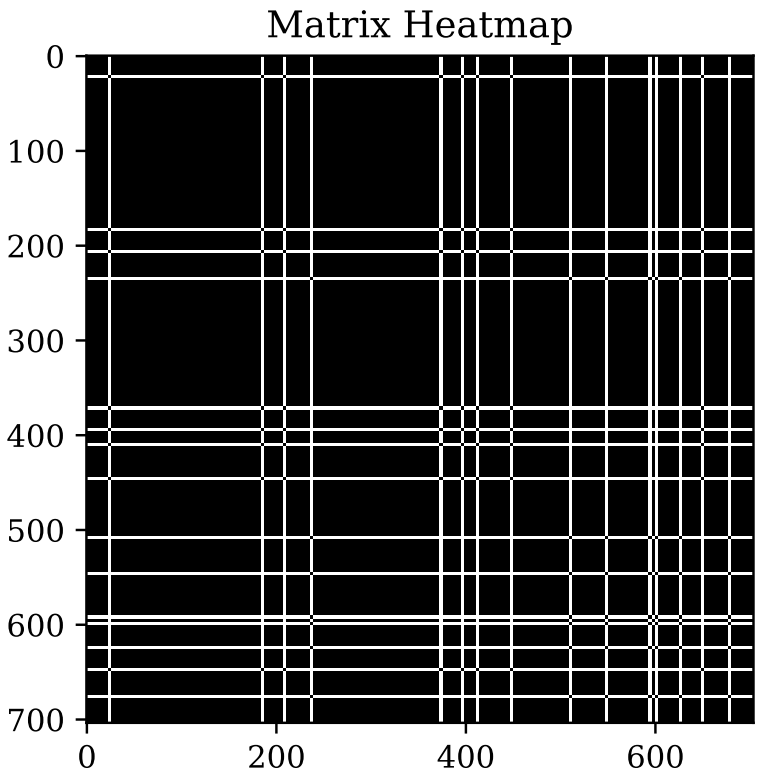
- 通过将预测标签和真实标签转化为关系矩阵，我们将评价问题转换为了二分类的评价。那么也就可以将这些关系分为 4 类：TP（真阳性）、TN（真阴性）、FP（假阳性）、FN（假阴性）。其中 TP 和 TN 表示与真实的关系相同，而 FP、FN 则是错误的。几乎所有主流的判别标准都是使用这 4 个指标进行排列组合、运算。
- Rand 指标：$\frac{TP+TN}{TP+TN+FP+FN}$
- Jaccard 指标：$\frac{TP}{TP+FP+FN}$
- Dice 指标：$\frac{2TP}{2TP+FP+FN}$
- Fowlkes–Mallows 指标：$\sqrt{\frac{TP}{TP+FP}} \cdot \sqrt{\frac{TP}{TP+FN}}$

详解评价指标

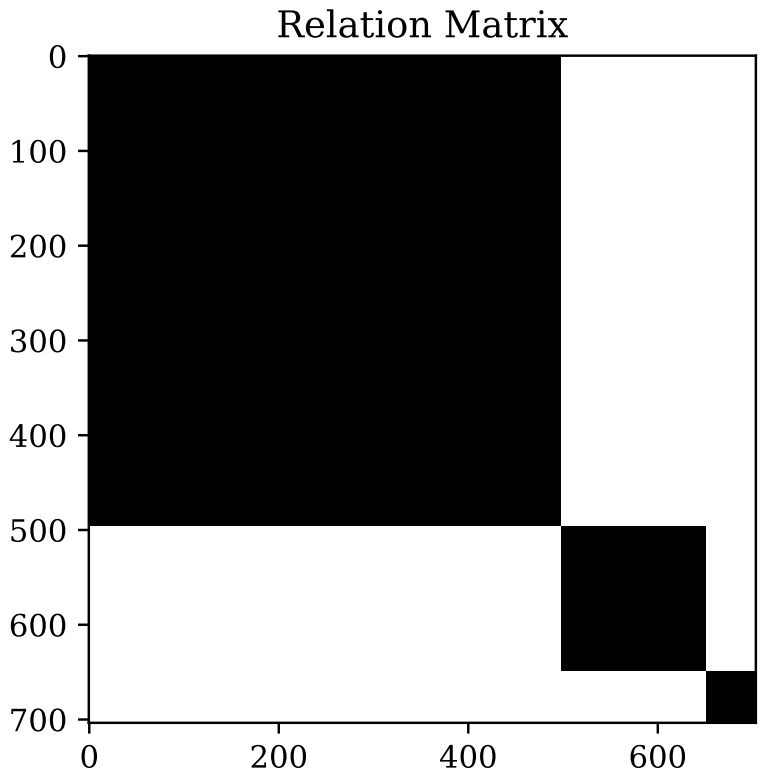
- 可以看出来，除了 Rand 指标之外，其他的指标都没有将 TN（真阴性）纳入考虑。
- 在我们记录过的实验结果中，出现了一些 RMA 和 FMI 不成正比的情况，主要是因为 TN 的数量异常导致的。

S 构造参数	TP	TN	FP	FN	RMA	FMI
(7, 4, 11, 11)	241840	36388	185894	31494	56.14	70.73
(4, 2, 6, 9)	145862	149122	73160	127472	59.52	59.61

详解评价指标

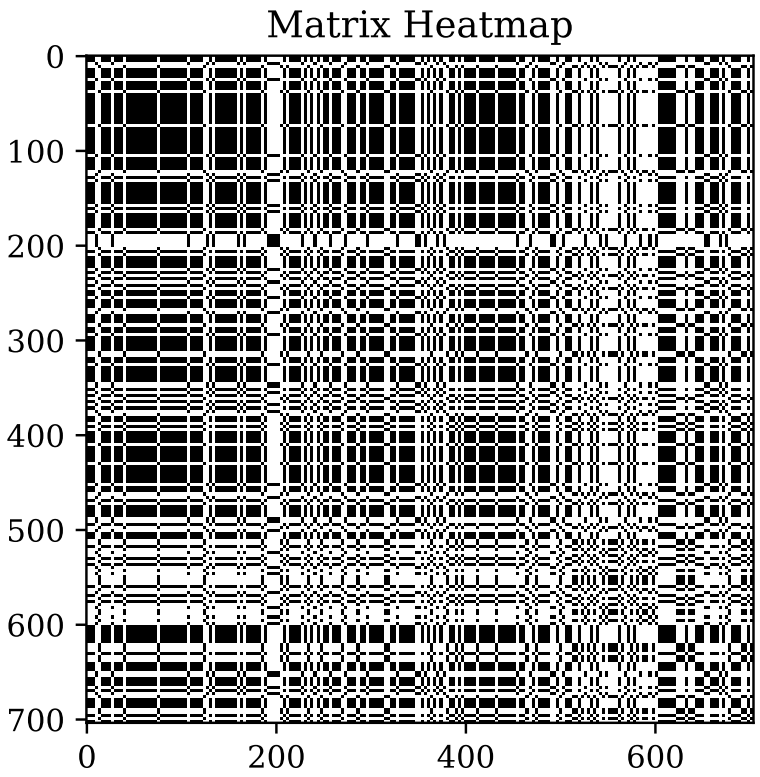


(a) 第一行结果矩阵可视化

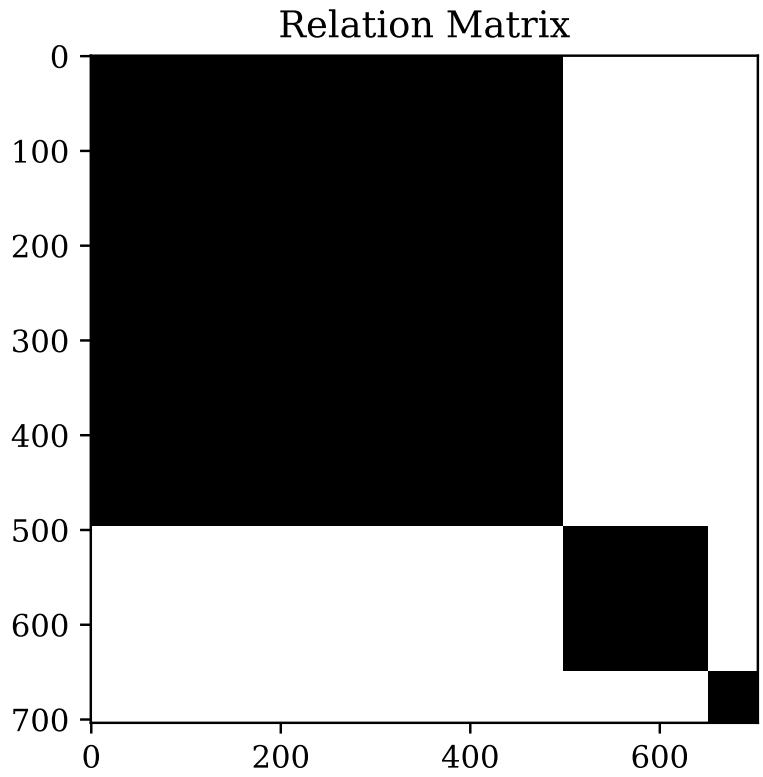


(b) 真实结果矩阵可视化

详解评价指标



(a) 第二行结果矩阵可视化



(b) 真实结果矩阵可视化

详解评价指标

- 可以看出，由于 FMI 指标没有将 TN 纳入考虑，高 FMI 的聚类结果往往存在的现象是，有大片本应为白色的区域预测结果为黑色，这也导致某一类别的预测数量过多。
- 因此，我们在论文中，将首要考虑 RMA 结果，同时兼顾 FMI 结果。

Case Study

为了研究聚类错误的回答出现错误的原因，我对其进行了一些case study，主要操作是取出某个错误的回答，观察它和其他回答的共享点赞数和相似度矩阵值。