大数据分析实验大作业
推荐系统

目录

華中科技大學
计算机科学与技术学院
School of Computer Science &Technology，HUST

一 任务背景

一 任务背景

◆实验目的

- 了解推荐系统的多种推荐算法并理解其原理。
- 实现User-User的协同过滤算法并对用户进行推荐。
- 实现基于内容的推荐算法并对用户进行推荐。

華中科技大學
计算机科学与技术学院
School of Computer Science &Technology，HUST

二 基于用户的协同过滤

任务描述

◆实验内容

- 实现基于用户的推荐算法，实现对指定用户进行动漫推荐，并对该算法进行评估

◆文件说明

- 提供动漫评分文件，分为训练集train_set和测试集test_set两部分。
- 其中train_set和test_set的第一列、第二列和第三列分别是用户id，动漫id和用户对动漫的评分（1-10分）。

	A	B	C
1	user_id	anime_id	rating
2	1	30276	8
3	1	22535	9
4	1	22199	7
5	1	37779	9
6	1	31240	8
7	1	14749	7
8	1	35847	7
9	1	4437	7

train_set.csv截图

	A	B	C	D
1	user_id	anime_id	rating	
2	265	37210	8	
3	648	11577	10	
4	266	19363	9	
5	465	37920	8	
6	20	10110	9	
7	140	27989	8	
8	602	10620	7	
9	672	34262	6	

test_set.csv截图

任务描述

◆ 实验要求

- 对训练集中的评分数据构造用户-动漫的效用矩阵，使用pearson相似度计算方法计算用户之间的相似度，因此构造出相似度矩阵。对单个用户进行推荐时，找到与其最相似的k个用户，用这k个用户的评分情况对当前用户的所有未评分动漫进行评分预测，选取评分最高的n个动漫进行推荐。

- 在测试集中包含100条用户-动漫评分记录，用于计算推荐算法中预测评分的准确性，对测试集中的每个用户-动漫需要计算其预测评分，再和真实评分进行对比，误差计算使用SSE误差平方和。

任务描述

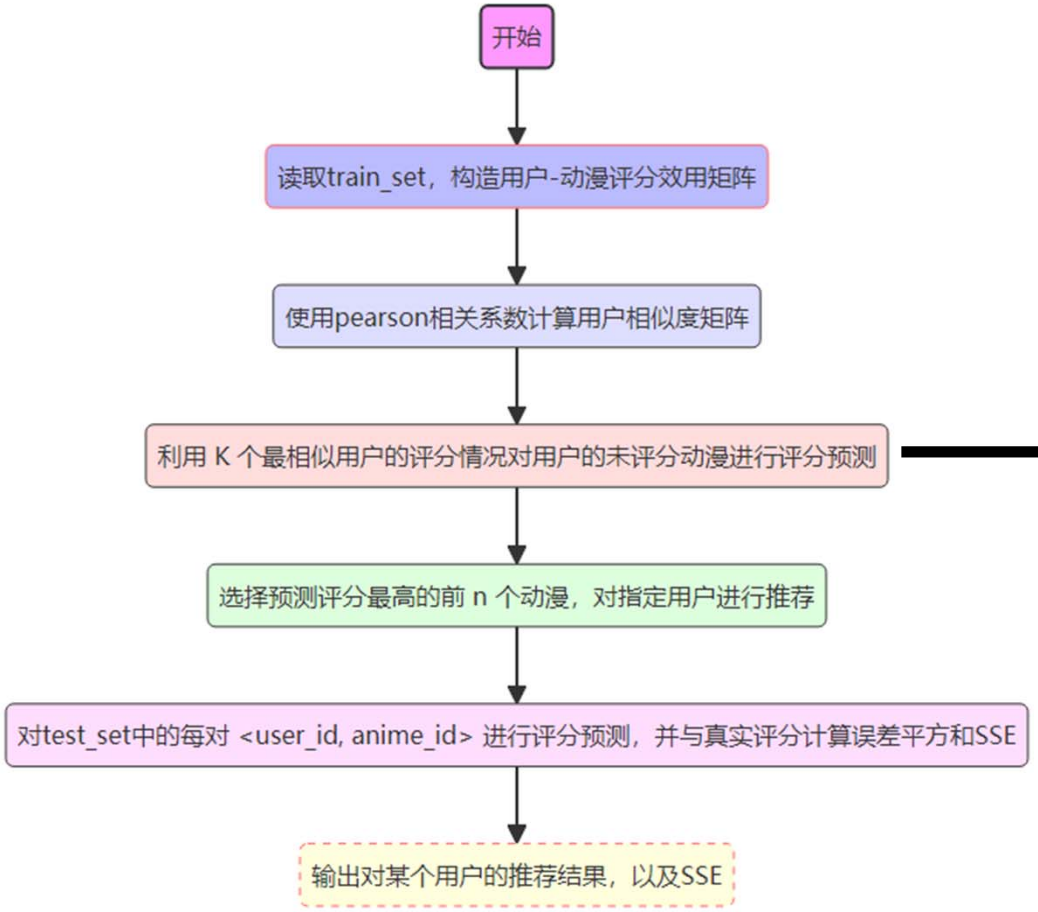
◆效用矩阵

横坐标是user_id

纵坐标是anime_id

	1	2	3	4	5	6	7	8
30276	8.00000	0.00000	8.00000	0.00000	0.00000	0.00000	0.00000	0.00000
22535	9.00000	8.00000	8.00000	0.00000	0.00000	0.00000	0.00000	0.00000
22199	7.00000	0.00000	0.00000	0.00000	0.00000	0.00000	8.00000	0.00000
37779	9.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
31240	8.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
14749	7.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
35847	7.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
4437	7.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
1575	10.00000	9.00000	8.00000	0.00000	0.00000	0.00000	0.00000	0.00000
6325	7.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
154	8.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
13667	8.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
13759	7.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
37965	9.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
9919	8.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
39940	7.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
37497	8.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
32188	10.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
33161	8.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000

算法流程

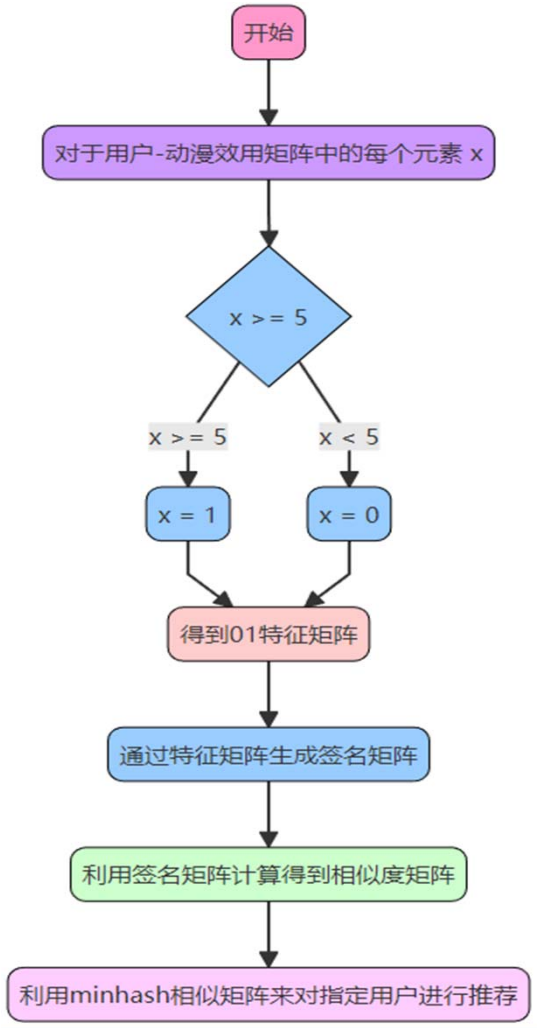


$$\text{predict\_rating} = \frac{\sum_{i=1}^{k} rating(i) * sim(i)}{\sum_{i=1}^{k} sim(i)}$$

进阶部分

◆进阶任务

- 使用minhash算法对效用矩阵进行降维处理，从而得到相似度矩阵



華中科技大學
计算机科学与技术学院
School of Computer Science &Technology，HUST

三 基于内容的推荐算法

任务描述

◆实验内容

- 实现基于内容的推荐算法，实现对指定用户进行推荐，并对该算法进行评估

◆文件说明

- 提供Anime数据集（anime.csv），包含用户对动漫评分、动漫标签等文件。
- 提供动漫评分文件，分为训练集train_set和测试集test_set两部分，其中train_set和test_set的第一列、第二列和第三列分别是用户id，动漫id和用户对动漫的评分。

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	Anime_id	Name	Genres	Ranked	Popularity	Score-10	Score-9	Score-8	Score-7	Score-6	Score-5	Score-4	Score-3	Score-2
2	1	Cowboy B	Action, Ac	39	1251960	182126	131625	62330	20688	8904	3184	1357	741	1580
3	5	Cowboy B	Action, Dr	159	518	30043	49201	49505	22632	5805	1877	577	221	109
4	6	Trigun	Action, Sc	201	558913	75651	86142	49432	15376	5838	1965	664	316	533

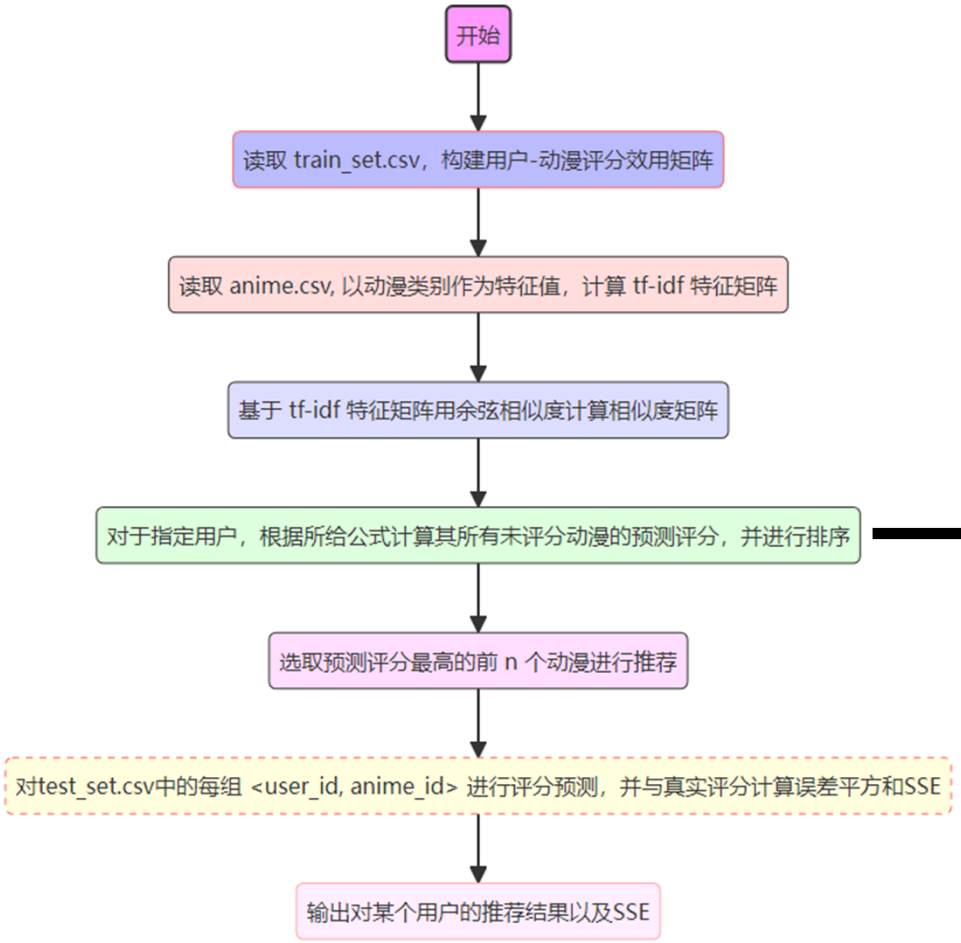
anime.csv截图

任务描述

◆实验要求

- 给定输入文件：train_set.csv, test_set.csv, anime.csv
- 利用余弦相似度的计算方法，得到相似度矩阵。
- 通过相似度矩阵完成：
- （1）对指定用户进行排名前k的动漫推荐，k值可根据需求更改。
- （2）利用误差平方和公式评估推荐算法的准确性。

算法流程

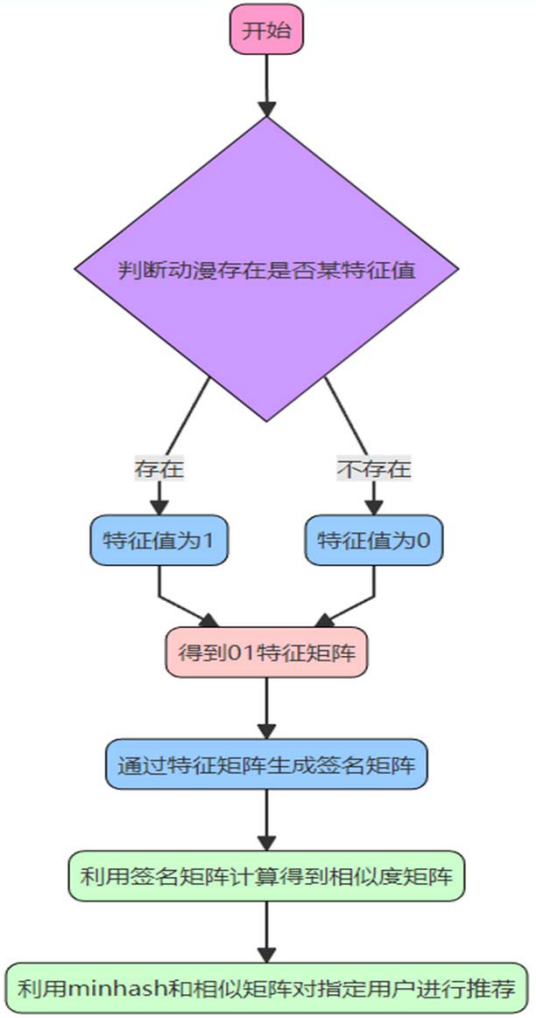


$$\text{score} = \frac{\sum_{i=1}^{n} score'(i) * sim(i)}{\sum_{i=1}^{n} sim(i)}$$

进阶部分

◆进阶任务

- 使用minhash算法对tf-idf特征矩阵进行降维处理，从而得到相似度矩阵




華中科技大學
计算机科学与技术学院
School of Computer Science &Technology，HUST

四 验收流程

基于用户的协同过滤

基础版对用户 629 推荐如下动漫：

Anime	Score
5114	9.405
9253	9.354
4181	9.320
11061	9.211
245	9.040
10030	9.000
12365	9.000
32281	8.984
2904	8.969
28851	8.969
38524	8.964
34599	8.931
30276	8.922
35839	8.909
7311	8.906
17074	8.900
2001	8.873
457	8.857
40591	8.848
1575	8.847

基础版 SSE = 167.81560664170405

总时间: 6.164 s.

进阶版对用户 629 推荐如下电影：

Movie	Score
9253	9.358
5114	9.342
245	9.138
1	9.102
11061	9.049
263	9.048
4181	9.047
32281	9.020
28851	8.977
5	8.935
2001	8.909
164	8.878
10030	8.870
38524	8.795
16067	8.773
17074	8.759
1210	8.758
1575	8.744
34599	8.734
457	8.710

总时间: 0.825 s.

MinHash 版 SSE = 238.48676372015785

基于内容的推荐算法

进阶版对用户 629 推荐如下电影:

Movie	Score
9253	9.358
5114	9.342
245	9.138
1	9.102
11061	9.049
263	9.048
4181	9.047
32281	9.020
28851	8.977
5	8.935
2001	8.909
164	8.878
10030	8.870
38524	8.795
16067	8.773
17074	8.759
1210	8.758
1575	8.744
34599	8.734
457	8.710

总时间: 0.954 s.

MinHash 版 SSE = 238.48676372015785

基础版对用户 629 推荐如下动漫:

Anime	Score
5114	9.405
9253	9.354
4181	9.320
11061	9.211
245	9.040
10030	9.000
12365	9.000
32281	8.984
2904	8.969
28851	8.969
38524	8.964
34599	8.931
30276	8.922
35839	8.909
7311	8.906
17074	8.900
2001	8.873
457	8.857
40591	8.848
1575	8.847

基础版 SSE = 167.81560664170405

总时间: 6.696 s.