VCM

电影推荐系统

• spring, spring mvc, mybatis, android, 爬虫, 最近邻居法

http://vcm.vatcore.com

crawler包(爬虫)

DoubanMovieSubjectCollectionsCrawler
 https://movie.douban.com/subject/26718838/collections
 爬电影下的最近评价用户列表,每页20个用户,总共200个用户

DoubanMovieSubjectCrawler
 https://movie.douban.com/subject/26718838/
 爬电影信息及movie_cover

DoubanMovieUserCollectCrawler

controller.NewController.AbstractNewThread有两个方法

• boolean addZombieUserRatingAndMovieByDoubanId(int doubanId) //添加僵尸用户评分及电影(都是爬下来的), doubanId为豆瓣电影Id

```
userList = DoubanMovieSubjectCollectionsCrawler.crawer(doubanId); //爬某电影下最近评价的200个用户
while (userIterator.hasNext()) {
    userMapper.add(user); //添加用户到user表
    addRatingAndMovieByDoubanUsername(user, null); //添加用户评分及数据库缺失电影
}
```

• boolean addRatingAndMovieByDoubanUsername(User user, User doubanUser) //user为写入数据库的用户, doubanUser为爬的对象, 如果doubanUser为null, 则爬写对象都为user ratingList = DoubanMovieUserCollectCrawler.crawer(user | doubanUser); //爬用户看过的电影评分与短评

```
for (Rating rating : ratingList) {
    if (movieMapper.findMovieByDoubanId(movie) == null) { //通过doubanId查询此电影是否在数据库中存在
        if (movieService.addMovieUseCrawlerByDoubanId(movie) == null) continue; //爬下电影信息及movie_cover, 并写入到movie表, 如果非电影则continue
    }
    ratingService.changeRating(rating); //添加或修改评分
}
```

- 如果爬虫线程只有一个, 则每个页面爬下需sleep(>2000), 如果有多个爬虫线程, 则sleep(>2000 * 线程数)
- 默认构造方法会生成threadId,管理员可以通过threadId查看线程进度,如添加僵尸用户每完成一个用户进度+0.5%

有三个类继承AbstractNewThread

BatchNewThread

调用addZombieUserRatingAndMovieByDoubanId(int doubanId)

RandomBatchNewThread

从数据库中随机 (unhandle_rating<10)得到doubanld, 调用addZombieUserRatingAndMovieByDoubanld(int doubanld),不停止线程则不停的随机并调用

ImportDoubanUsernameThread

爬下指定豆瓣用户评分到指定用户(当前登录用户)下, 调用addRatingAndMovieByDoubanUsername(User user, User doubanUser)

VCM	nameZh imdbld doubanld	0	首页	推荐	导入	移动端	root▼		
批量新建(正在执行的任务有:0个)									
尽量单个线程 当前任务状态:									
	当前任务状态: 强行终止所有执行任务(并不会瞬间停止)								
强行停止									
输入地址,抓取此 doubanld	次最近看过的用户及用户所有看过电影 批量增加	影和评分							
随机批量增加									
输入douba	anId抓取新建								
doubanld	增加								
将doubanUsername(id)的评分导入自己账户下									
doubanUserna	me(id) 导入								

普通用户只能使用最下面两个功能 userld为1的用户可以使用所有

handler包(常驻线程)

• 当启动tomcat时, 由spring将 ***Handler类注入到Handler类, 读Config类的配置决定是否开启线程

更新电影的平均评分与评分用户总数线程

RatingsAndUsersHandler

movieService.calculationAndUpdateRatingsAndUsers(); //找到100部unhandle_ratings(未处理的评价总数, 处理过后归零) >0的电影遍历统计评价过的用户总数和评价均分, 将结果update到movie表

最近邻居线程(5个)

- GenresRatingHandler
- FavoriteGenresHandler与写配置文件线程
- NearestNeighborHandler与写配置文件线程

GenresRatingHandler

userList = userService.findHundredUsersIfRatingCountMoreThanXAndUnhandleRatingMoreThanY(); //找到100个rating_count(评价总数) >60并且unhandle_rating>30的用户 for (User user : userList) {

ratingService.selectAndAddGenresRatingListToGenresRatingByUserld(rating); //统计用户评价的每部电影类型下的平均评分和与总平均评分的相差, 比例, 类型占的比例, 类型总量, 大于用户电影平均评分的类型并且此类型电影评价总数大于10会写入genres_rating表中(图中绿色的行)

userService.makeUnhandleRatingsZeroByUserld(user); //unhandle_rating归零

}

ratingCount: 评价过的电影总计

avgRating(s): 平均电影评分

count: 用户评价过的此类型电影总数

ratio: 此类型占所有评价电影比例, 这个用户总共对416部电影评价了, 剧情类

型的评价了189部,则ratio = 189 / 416

avgRating(y): 类型的平均评分

avgDifference:类型的平均评分-电影的平均评分

avgRatio:类型的平均评分/电影的平均评分

weight: 类型权重, (ratio + avgRatio) * avgRating, 由FavoriteGenresHandler操

作

	ratingCount	avgRating(s)
	416	7.5962
avaPating(v) avaDifference(v-s) avaPatio(ar	My/e)weight	(/r ± ar\ * v1

genresid	nameZh	count	ratio(r)	avgRating(y)	avgDifference{y-s}	avgRatio(ar){y/s}	weight{(r + ar) * y}
1	剧情	189	0.45432693	7.8942	0.29799986	1.0392301	11.790437726224155
10	动作	155	0.37259614	7.4065	-0.18970013	0.97502697	9.981170377795479
13	喜剧	149	0.35817307	7.5839	-0.0123000145	0.9983808	10.287968728261689
6	冒险	112	0.26923078	7.75	0.15380001	1.020247	9.993452683091164
4	科幻	97	0.23317307	7.4433	-0.15290022	0.9798715	9.029054472010245
14	爱情	89	0.2139423	7.3708	-0.22539997	0.97032726	8.729014111954427
5	惊悚	76	0.1826923	7.3158	-0.2803998	0.96308684	8.382291306732732
16	奇幻	69	0.16586539	7.8261	0.22989988	1.0302651	9.361036631113379
3	犯罪	68	0.16346154	7.7353	0.13910007	1.0183117	9.141370890903033
24	动画	56	0.13461539	8.1429	0.5467005	1.0719702	9.825106574872677
12	家庭	43	0.103365384	7.7209	0.12470007	1.0164161	8.645720716267473
2	悬疑	37	0.088942304	7.4054	-0.19080019	0.9748821	7.87804522126266
8	传记	13	0.03125	7.2308	-0.36539984	0.951897	7.108939658419615
11	灾难	13	0.03125	7.2308	-0.36539984	0.951897	7.108939658419615
15	恐怖	11	0.026442308	6.7273	-0.8688998	0.88561386	6.135675602487153
9	战争	11	0.026442308	7.4545	-0.14169979	0.98134595	7.51255778062049
23	历史	9	0.021634616	7.7778	0.1816001	1.0239067	8.132011394588009
19	音乐	7	0.016826924	8.5714	0.9751997	1.12838	9.816025765523861
21	儿童	7	0.016826924	7.1429	-0.4533	0.94032544	6.836843603557265
20	歌舞	5	0.012019231	8.4	0.8037996	1.105816	9.389815569794445
26	古装	5	0.012019231	6.8	-0.7961998	0.89518446	6.168985253381837
22	短片	5	0.012019231	8.0	0.4038	1.0531582	8.521419160068035
7	同性	4	0.009615385	8.0	0.4038	1.0531582	8.502188391983509
25	武侠	3	0.0072115385	6.6667	-0.9295001	0.8776362	5.899014279179953
18	纪录片	3	0.0072115385	8.6667	1.0705004	1.1409258	9.950561985502478
27	运动	2	0.0048076925	8.0	0.4038	1.0531582	8.463726852089167
17	情色	2	0.0048076925	8.0	0.4038	1.0531582	8.463726852089167
28	西部	1	0.0024038462	8.0	0.4038	1.0531582	8.444496082141995



悬疑 37 0.088942304 7.3949685

恐怖 11 0.026442308 6.717824

11 0.026442308 7.4439993

0.021634616 7.766844

0.016826924 8.559326

0.016826924 7 1328382

0.012019231 8.388167

0.012019231 6.7904215

0.012019231 7.988731

0.009615385 7.988731

0 00721153856 657309

0.00721153858 654492

0 00480769257 988731

0.00480769257.988731

0.00240384627 988731

13 0.03125

平 無

武侠

经录片

people(p) | mine(m)

		ratingCountavgRa	ting(s)ratingCount	avgRating(s)	mpRatingRatio(mpr){[m]s/[p]s}
416 7.5962 193 7.5855 0.99859136	416 7.5962 193 7.5855 0.99859136	416 7.5962	193	7.5855	0.99859136

rating	count	rating	СО
2	1	2	14
4	10	4	15
6	118	6	36
8	230	8	60
10	57	10	68

[p]original | [m]original | [p]weighted

[p]y * r... 对[p]进行加权

genres	ld nameZI	coun	tratio(r)	avgRating(y	avgDifference{y-s}	avgRatio(ar){y/s	}weight{(r + ar) * y}	genresi	dnameZI	hcount	tratio(r)	avgRating(y)	avgDifference{y-s	avgRatio(ar){y/s	s}we
1	剧情	189	0.45432693	7.8942	0.29799986	1.0392301	11.790437726224155	1	剧情	122	0.63212436	8.1639	0.5784006	1.0762508	13
10	动作	155	0.37259614	7.4065	-0.18970013	0.97502697	9.981170377795479	14	爱情	70	0.3626943	7.7143	0.12880039	1.0169798	10.
13	喜剧	149	0.35817307	7.5839	-0.0123000145	0.9983808	10.287968728261689	13	喜剧	51	0.2642487	7.1765	-0.40899992	0.94608134	8.6
6	冒险	112	0.26923078	7.75	0.15380001	1.020247	9.993452683091164	10	动作	33	0.17098446	5.8788	-1.7066998	0.775005	5.5
4	科幻	97	0.23317307	7.4433	-0.15290022	0.9798715	9.029054472010245	6	冒险	23	0.119170986	7.0435	-0.5419998	0.9285479	7.3
14	爱情	89	0.2139423	7.3708	-0.22539997	0.97032726	8.729014111954427	16	奇幻	19	0.098445594	7.2632	-0.32229996	0.95751107	7.6
5	惊悚	76	0.1826923	7.3158	-0.2803998	0.96308684	8.382291306732732	4	科幻	18	0.09326425	5.8889	-1.6966	0.77633643	5.1
16	奇幻	69	0.16586539	7.8261	0.22989988	1.0302651	9.361036631113379	5	惊悚	16	0.08290155	6.625	-0.96049976	0.87337685	6.3
3	犯罪	68	0.16346154	7.7353	0.13910007	1.0183117	9.141370890903033	12	家庭	16	0.08290155	7.375	-0.21049976	0.97224975	7.7
24	动画	56	0.13461539	8.1429	0.5467005	1.0719702	9.825106574872677	3	犯罪	14	0.07253886	7.5714	-0.014099598	0.9981412	8.1
12	家庭	43	0.103365384	7.7209	0.12470007	1.0164161	8.645720716267473	23	历史	12	0.062176164	8.5	0.91450024	1.120559	10.
2	悬疑	37	0.088942304	7.4054	-0.19080019	0.9748821	7.87804522126266	9	战争	12	0.062176164	8.0	0.41450024	1.0546438	8.9
8	传记	13	0.03125	7.2308	-0.36539984	0.951897	7.108939658419615	24	动画	11	0.05699482	8.7273	1.1417999	1.150524	10.
11	灾难	13	0.03125	7.2308	-0.36539984	0.951897	7.108939658419615	8	传记	11	0.05699482	7.4545	-0.13099957	0.98273027	7.7
15	恐怖	11	0.026442308	6.7273	-0.8688998	0.88561386	6.135675602487153	20	歌舞	10	0.051813472	7.8	0.21450043	1.0282776	8.4
9	战争	11	0.026442308	7.4545	-0.14169979	0.98134595	7.51255778062049	2	悬疑	7	0.03626943	6.0	-1.5854998	0.79098284	4.9
23	历史	9	0.021634616	7.7778	0.1816001	1.0239067	8.132011394588009	27	运动	5	0.025906736	7.2	-0.38549995	0.94917935	7.0
19	音乐	7	0.016826924	8.5714	0.9751997	1.12838	9.816025765523861	26	古装	5	0.025906736	9.6	2.0145006	1.2655725	12.
21	儿童	7	0.016826924	7.1429	-0.4533	0.94032544	6.836843603557265	7	同性	4	0.020725388	6.0	-1.5854998	0.79098284	4.8
20	歌舞	5	0.012019231	8.4	0.8037996	1.105816	9.389815569794445	19	音乐	4	0.020725388	8.0	0.41450024	1.0546438	8.6
26	古装	5	0.012019231	6.8	-0.7961998	0.89518446	6.168985253381837	25	武侠	4	0.020725388	9.5	1.9145002	1.2523894	12.
22	短片	5	0.012019231	8.0	0.4038	1.0531582	8.521419160068035	17	情色	3	0.015544041	4.6667	-2.9187999	0.6152132	2.9
7	同性	4	0.009615385	8.0	0.4038	1.0531582	8.502188391983509	28	西部	1	0.005181347	6.0	-1.5854998	0.79098284	4.7
25	武侠	3	0.0072115385	56.6667	-0.9295001	0.8776362	5.899014279179953	genresi	dnameZi	hcount	ratio(r)	avgRating(y)avgDifference{y-	s}avgRatio(ar){y	/s}w
18	纪录片	3	0.0072115385	58.6667	1.0705004	1.1409258	9.950561985502478	1	剧情	189	0.45432693	7.88308	0.29758024	1.0392301	1
27	运动	2	0.004807692	58.0	0.4038	1.0531582	8.463726852089167	10	动作	155	0.37259614	7.3960667	-0.1894331	0.97502697	9
17	情色	2	0.004807692	58.0	0.4038	1.0531582	8.463726852089167	13	喜剧	149	0.35817307	7.573217	-0.012282848	0.9983808	1/
28	西部	1	0.002403846	28.0	0.4038	1.0531582	8.444496082141995	6	冒险	112	0.26923078	7.7390833	0.15358353	1.020247	9
								4	科幻	97	0.23317307	7.432815	-0.15268469	0.9798715	9
								14	爱情	89	0.2139423	7.3604174	-0.2250824	0.97032726	8
								5	惊悚	76	0.1826923	7.305495	-0.28000498	0.96308684	8
								16	奇幻	69	0.16586539	7.815076	0.22957611	1.0302651	9
								3	XP.STE	68	0.16346154		0.1389041	1.0183117	9

/people/{userId}

02061964301613

7 86694816047114

7 09892559954198

7 09892559954198

5 16029541702744

8 50941557741106

8 49021189848973

8 45180453692702

8.45180453692702

8 43260085614566

0.951897

-0.14150047

-0.4526615

2026676

-0.7950783

0.40323114

0.40323114

0.88561386

0.98134595

1 0239067

1 12838

0 94032544

1 105816

0.89518446

1.0531582

1 0531582

0.8776362

1 1409258

1 0531582

1.0531582

可实时查看用户与目标用户(第一张)的状态

第一, 二张都是原始数据, 第三张为第一张相对第二张加权(用户电影平均评分/目标用户电影平均评分) 得来

绿色行:会写入genres_rating表

根据weight推测目标用户喜欢:剧情, 冒险, 动画用户喜欢:剧情, 爱情, 动画

这两个用户并不会成为邻居,成为邻居的必要条件为喜欢的三个类型必须完全相同

成为邻居后,要算出离用户最近的三个邻居,用户的推荐电影从这三个用户中得出,统计电影出现的喜爱类型数量,从大到小排序,再将电影列表返回给用户

邻居坐标系x轴为类型平均分, y轴为类型所占比例

FavoriteGenresHandler

依赖 GenresRatingHandler处理后的结果, 处理genres_rating表, 写入结果至favorite_genres表, 内部有线程每10秒检查 nextGenresRatingIdINI与nextGenresRatingId(下一个处理的id), 两者不同则写入磁盘更新nextGenresRatingId.ini, .ini文件只有在线程启动时读入, 其余时间都在写入备份, 配置文件路径由Config类指定

int count = genresRatingService.findAndAddFavoriteGenresByGenresRatingId(genresRating); //算出类型的weight((ratio + avgRatio) * avgRating), 找出weight最大的三个类型(推测出的用户喜欢类型)写入到favorite_genres表, 返回处理genres_rating的行数

if (count == 0) { ... } //到表底了或卡住了(id不连续)...

NearestNeighborHandler

依赖 FavoriteGenresHandler处理后的结果, 处理favorite_genres表, 写入结果至nearest_neighbor表, 写配置文件与处理流程与FavoriteGenresHandler 类似

int count = favoriteGenresService.findAndAddNearestNeighbor(favoriteGenres); //根据推测喜欢的三个类型找出相同用户(这些用户都为邻居), 算出用户与邻居(经过加权后)的距离,邻居坐标系x轴为类型平均分, y轴为类型所占比例, 用户与邻居对应的三个类型的距离依次算出相加就是距离, 找出离用户最短距离的三个邻居, 将结果写入nearest neighbor表

RecommenderController | RecommendMobileController

goodMovieList = nearestNeighborService.findNearestNeighborGoodMovieListByUserId(nearestNeighbor); //找出三个最近邻居喜爱类型的电影列表, 对电影 出现的类型次数进行计数, 每有一个喜欢类型, count++, 对map按值从大到小排序, 返回list VCM nameZh | imdbid | doubanid O 製図 数字 等入 移动線 面织・ 1994-09-10(多伦多电影节) 1994-10-14(美国: 1991-02-02(香港)

2011-05-28(中国大陆) 2011-05-26(美国

1992-01-22(法国) 1992-05-19(英国) 2013-11-01(中国大陆) 2012-09-08(多化多电影节 2000-09-05(威尼斯电影节) 2001-05-25(美国)

1989-06-02(多位多篇時) 1989-06-09(美国)

2004-06-25(中国大陆) 2004-02-13(美国

2012-06-08(美国/中国大陆) 2012-06-08(法国)

2017-02-14(中国大路) 2016-08-31(威尼斯电影节) 2016-12-25(美国) 126 2011-04-25(辛贝卡电影节) 97

2013-08-30(中国大陆) 2013-05-01(纽约首時) 2013-05-10(美国) 143

2000-10-13(中国大路) 2000-05-16(夏纳电影节) 2000-07-07(台湾)

2015-12-24(中国大陆) 2015-09-09(城尼斯电影节 2016-08-11(中国大陆/数据)

1993-01-01(香港)

2007-11-22 2013-12-03(中国大師)

1996-12-06(美国) 1997-03-14(英国)

2016-03-04(中国大陆/美国

2002-05-24(臺纳由銀竹) 2002-09-25(法国)

2015-09-30(中国大陆) 2010-04-16(中国大陆) 2007-10-11(西班牙) 1994-07-14(香港)

2003-12-10(多位多書時) 2004-01-09(美国)

1972-03-15((日约首時) 1972-03-24(美国

2000-10-27/89-3-80

2016-09-23(中国大路美国)

2013-12-13(美国) 2004-02-15 2009-12-26(英国) 2014-09-26((日的电影节) 2014-10-03(美国)

2010-06-16(法国) 2010-12-25(美国)

2012-11-22(中国大陆) 2012-09-28(田台電影节) 2012-11-21(美国

3578925 8 0/7)

1291548 8.7(20)

1309172 8.0(14) 1292757 8.5(4) 259340148.2581(31) 5322596 9.25(16)

3178770 7.6364(11)

3908423 8.0(7) 1781126 7.5(8) 1291549 8.9375(32)

1292720 9.2778(36) 6786002 9.2(25) 2054933 7.1429(7) 1291561 9.4146(41) 2365260 7.625(16) 3804891 7.8(20) 1292233 8.9(20)

1291853 8.2(10)

1292270 9.0(12) 1291879 9.0(12) 1296736 8.25(16)

1929463 9.0476(42)

256623298.9048(42)

1292402 8 8571/281

1292925 9.0(8)

数次 西西里的美丽 Maléna

社 校会學歷的秘 Mindenki

四四至3美術 Maléna 传说 違见你之前 Me Before You 接火车 Trainspotting 罗密称与朱丽 Roméo & Juliette D

原代符 L'illusionniste 近大郎座 Great Expectations 花田笛季 走起野歌幹 Hacksaw Ridge 少年計代 Boyhood 巴勢至母親 Notre Dame de Parts

haine à famour

1295467 9.0(2)

1291546 9.5294(34) 南干町間

議員 議員報社 Hotel Transylvania 比区辞料 Memeratio 形に特社 Dead Poets Society 百中之謀 意の東の庭 原房的女 可り動物型 Rose of the Guardar 逐大場 Babel

直の策のm A Good Woman 高秋型 Rec of the Guar 多大塔 Babel 人生選投幕 CHCK 印からの次 50 F

延期提升 Séraphine

网页与android端的推荐页(userId=269)

VCM

13:43

标记量必须大于60并且符合系统规则	
肖申克的救赎	9.6(35)
东邪西毒	9.0588(17)
阿飞正传	8.8889(18)
杀手没有假期	8.5(4)
纵横四海	8.8(15)
起风了	8.2857(7)
喜剧之王	8.6(30)
云中行走	7.3333(9)
功夫熊猫3	8.0(19)
功夫熊猫2	8.2222(18)
我家买了动物园	8.0(7)
情人	9.0(8)
精灵旅社	7.625(16)
记忆碎片	8.1481(27)
死亡诗社	8.7(20)
言叶之庭	8.2857(14)
痴男怨女	8.0(2)
守护者联盟	7.75(8)
通天塔	8.4444(9)
人生遥控器	7.3333(12)
初恋50次	8.0(14)
木偶奇遇记	8.5(4)
爱乐之城	8.2581(31)
超脱	9.25(16)
马达加斯加3	7.6364(11)
花落花开	8.0(2)
1 - /\	/ \

推荐

搜索

14.6K/s **令 訓** 联通 4G **訓** 移动 69%

我

Android

Task接口有三个默认方法(java8特性)用来与服务端通讯, ***Task类都实现这个接口, 使用默认方法的版本必须>=android7.0(api24)

控制器在controller.mobile下,一般与类名对应,除显示movie_cover外,服务端都用json给客户端传信息

Activity

StartPageActivity

起始页, 每次启动程序检查是否之前登录过并向服务端验证用户, 控制器为VerifyMobileController, 通过跳转MainActivity, 不通过清除之前的用户信息, 跳转到LoginAndRegisterActivity, 从没登录过的同样跳转. 跳转都会finish();

LoginAndRegisterActivity

有两个Fragment: LoginFragment, RegisterFragment. 可左右滑动切换

MainActivity

有三个Fragment: SearchFragment, RecommendFragment, MineFragment. 可左右滑动切换

MovieActivity

有一个或多个MovieFragment, 数量等于接受List<Movie>的size(), 可左右滑动切换

Fragment

MineFragment

显示nickname, 登出按钮点下跳转LoginAndRegisterActivity并finish();

LoginFragment

登录页, 登录成功跳转MainActivity

RegisterFragment

注册页

SearchFragment

搜索页, 从服务端接受List<Movie>, 点进一项跳转到MovieActivity并显示那项的具体信息

RecommendFragment

推荐页,与SearchFragment类似

MovieFragment

电影信息与评价页, 显示电影信息(不用从服务端获取, 在List<Movie>里), movie_cover(用doubanld向服务端获取), 用户的评分与短评. 并能修改评分与短评, 控制器为RatingMobileController





记忆碎片 Memento 评分:8.1481(27)

上映时间:2000-09-05(威尼斯电影节) 2001-05-25(美国)

时长:113分钟 imdb:tt209144

douban:1304447

简介:

当记忆支离破碎后,你该怎样面对生

活?本片的主人公就遭遇了这样的事情。

你的评分:

当前评分:



东邪西毒 東邪西毒 评分:9.0588(17)

上映时间:1994-09-17(中国大陆)

时长:100分钟 imdb:tt109688 douban:1292328

简介:

在恋人(张曼玉)负气嫁给兄长的

当晚,欧阳锋(张国荣)黯然离开白陀山,走进沙漠某个 小镇,成为一名杀手中介人。他的朋友、风流剑客黄药师 (梁家辉)每年都来小镇与他畅饮,酒话里有关白陀山的 消息令他忘却往事的念想渐次演变成自欺行为。

你的评分:10 当前评分:10

神一样的剧本

你的短评

提交

电影页面(userId=269)