

基于群体兴趣的个性化推荐候选集构建

石宇¹, 查梦娟¹, 梁宇², 邱惠²

(1.武汉大学信息管理学院, 武汉 430072; 2.国网鄂州供电公司电力调度控制中心, 鄂州 436000)

摘要: 对个性化推荐候选集构建方案进行优化, 在保证对用户感兴趣项目覆盖率较高的前提下, 压缩候选集规模, 从而改善个性化推荐效果。采用融合用户群体兴趣与个性化特征的方法构建推荐候选集, 先以用户群体兴趣为基础构建初始候选集, 在此基础上基于用户个性化特征对其进行过滤, 生成推荐候选集。在模型构建的基础上, 以电影数据为例进行模型效果验证。实验结果显示, 该模型在覆盖率小幅降低的情况下, 可将候选集规模压缩为原来的3.4%; 该模型能显著提升个性化推荐的最终效果, 因此该策略是一种较理想的候选集构建策略。

关键词: 个性化推荐; 群体兴趣; 候选集

中图分类号: G35; TP391

DOI: 10.3772/j.issn.1673-2286.2017.1.008

1 引言

在个性化推荐中, 候选集的构建对最终效果具有重要影响, 一般来说, 如果候选集构建能过滤大量用户不感兴趣的资源, 用户对推荐结果的满意度将显著提升^[1]。在当前研究与实践中个性化推荐候选集构建方式主要包括四种: 第一, 基于项目相似度的推荐候选集构建, 即在计算系统待推荐项目与用户感兴趣项目的相似度基础上, 将相似度最高的N条结果或高于某个阈值的结果作为候选集^[2-4]; 第二, 基于项目与用户兴趣模型相似度的候选集构建, 即先通过向量空间或关键词列表的形式获得用户兴趣描述, 在此基础上计算待推荐项目与用户兴趣的相似度, 并将相似度最高的N条结果或高于某个阈值的结果作为候选集^[5-7]; 第三, 基于情境的候选集构建, 在基于情境的个性化推荐中, 通常将与情境信息相匹配的项目作为推荐候选集^[8-10]; 第四, 基于最近邻用户的候选集构建, 即先确定用户的最近邻集合, 在此基础上从最近邻感兴趣的项目集中选出一部分作为候选集(如热门项目或高评分项目^[11-14])。这些候选集构建方案虽然在研究与实践中得到一定应用, 但也存在一些问题。前三种均采用基于相似度计算的思路, 但相似度高的项目中存在大量用户不感兴趣的项目; 基于最近邻用户的候选集构建中当最近邻集较小

时, 候选集对用户兴趣的覆盖率较低, 当最近邻集较大时, 则候选集中会包含大量用户不感兴趣的项目^[15]。

为解决上述问题, 本文拟基于用户群体兴趣构建候选集。根据对豆瓣电影的预调研, 当用户群体规模较大时, 用户群体兴趣具有较强的稳定性。将笔者抓取的豆瓣电影用户(约83万人)看作一个群体, 对其2014年10月5日—12日的观影记录进行统计。2014年10月5日—9日, 调研的101 486部影片中用户共观看16 727部, 观影频次172 567次; 其中Top500热门影片的观影频次占总频次的55.7%, 10月5日—9日83.8%的Top500热门影片也出现在10月10日—12日的热门影片中, 前者观影总频次占10月10日—12日观影总频次的53.3%。基于此, 为优化个性化推荐项目候选集的构建策略, 本文拟以用户群体兴趣为基础, 并以用户的个人兴趣对其包含的项目集合进一步过滤, 从而生成推荐候选集。

2 研究设计

2.1 基于群体兴趣的个性化推荐候选集构建模型

基于群体兴趣的个性化推荐候选集构建, 本质上是一个两阶段的构建策略: 首先, 获取用户所在群体可

能感兴趣的集合,并将其作为初始候选集;其次,引入用户的个性化特征对初始候选集进一步过滤,从而得到最终候选集。

2.1.1 群体划分及群体兴趣识别

用户群体的划分可基于聚类和复杂网络分析技术进行实现,鉴于群体规模与用户群体兴趣的稳定性具有显著的正相关关系,因此,与最近邻集的构建不同,群体划分除需考虑用户间相似性外,还需保证具有较大的群体规模。在群体划分的基础上,对于群体感兴趣的项目识别可基于投票机制进行,即将项目按照群体中对其感兴趣的用户数量进行排序,在此基础上取TopN或者高于某个阈值的项目作为初始候选集。随着时间的推移,用户的群体兴趣同样会发生漂移,因此在群体兴趣识别时,需要关注时间范围的限定,根据资源自身的特质进行合理地时间范围限定。

值得指出的是,用户群体规模较大时,可将时间范围严格限制,从而能够在避免数据稀疏性的基础上及时反映用户最新的兴趣变化;用户群体规模较小时,可将时间范围限制放松,并对用户在不同时间范围内的兴趣进行加权处理,从而在保障群体兴趣稳定的前提下避免数据稀疏性问题。

2.1.2 基于用户个性化特征的过滤

为减少候选推荐集合中非相关项目的数量,需基于用户个性化特征对其进行进一步过滤。借鉴当前个性化推荐研究成果,用户的个性化特征既可以是情境特征、兴趣向量或词表,也可以是感兴趣的项目集合、最近邻用户集,或者多种个性化特征的融合;而过滤策略也可采用相似度计算、最近邻匹配等方式实现。

本文拟采用兴趣向量空间模型来表征用户的个性化特征,通过计算相似度进行非相关项目的过滤。

2.2 评价方法

为评价基于群体兴趣的个性化推荐候选集构建模型的效果,拟基于其在个性化推荐中的应用效果进行评估。在算法实现中,拟基于向量空间模型进行用户兴趣建模,并通过项目与用户兴趣相关度的大小对候选集中项目进行个性化推荐结果排序;而与之对比的基准

算法,则采用同样的兴趣建模与排序策略,唯一的区别是其将采用全部资源作为初始推荐候选集。

在评价指标上,由于用户更关心推荐列表中排序靠前的项目能否满足需求,因此拟选用P@N指标进行效果评价^[16],衡量对于排序在TopN范围内的推荐项目用户感兴趣的可能性。需要指出的是,由于在推荐中考虑了时间因素,因此用户登录推荐系统时,每次都会显示不同的推荐结果,从而在效果评价时需要考虑用户登录因素的影响。本文,假设用户在有观影记录的日期才登录系统,登录频率为1次/天。进而本文P@N衡量的是一段时期内平均每天的推荐准确率。计算方法如公式(1)所示。

$$P@N = \frac{\sum N_{ri}}{\sum N_r} \quad (1)$$

其中,P@N指一段时间内推荐的前 N_r 条项目中用户喜欢项目的平均占比, $\sum N_r$ 指一段时间内推荐列表的总长度,即每次推荐的列表长度与用户登录次数的乘积; $\sum N_{ri}$ 指一段时间内推荐给用户的项目中用户喜欢的总数量。

3 数据

本文拟以电影作为研究对象,电影是个性化推荐研究的常用对象,很多研究均基于MovieLens、Eachmovie等提供的电影数据集开展^[17-18],因此基于其开展研究能保证结论具有较强代表性。在平台选择上,本文选择国内知名的电影网站豆瓣电影作为数据源,因其用户规模较大且经过10多年的发展积累大量的用户数据,从而能够为研究提供较丰富的数据支撑。

3.1 数据采集

数据采集目标包括两个:一是获得能够判断用户兴趣的影片(本文将用户标记为“看过”的影片视为用户感兴趣的影片)及相关信息,二是获得推荐候选影片的相关信息。具体实现上,先通过豆瓣的明星用户列表获取前50位明星用户的粉丝ID列表,总计830 682位用户;在此基础上,2014年11月20日—12月15日,笔者借助工具采集这些用户的全部观影记录,采集字段包括用户ID、影片URL、观影时间、评分;从统计结果看,用户观看过的影视作品(包括电影、电视剧、动画片、短片等类型)共101 486部,并采集这101 486部影视作品的基本信息,包括片名、分类类型、导演、演员、编

剧、上映时间、集数、片长等。

3.2 数据预处理

3.2.1 非电影数据剔除

鉴于本文的研究对象是电影,因此需先剔除抓取数据中剧集、短片、电视节目、戏曲等非电影类的影视作品。剔除主要依据影视作品基本信息中的集数、片长、片名三个字段进行。其中,通过是否包含集数字段可过滤剧集、多期类的电视节目;通过片长信息可过滤短片、单期类的电视节目、部分戏曲等其他非影视作品。为提高过滤召回率,将片名中包含特定关键词的影视作品过滤掉,如“演唱会”。过滤后,共得到75 694部电影。在此基础上,以这些影片的URL为特征,筛选出用户的电影观看记录信息。

3.2.2 影片类型规范化

用户在进行观影决策时,会考虑影片的类型、时间、演员、制片国家等多方面因素,其兴趣也可分为多个维度。其中影片类型为用户最重要、最常用的决策维度,为验证实验效果,本文将根据影片类型进行用户兴趣建模。参考互联网电影资料库(Internet Movie Database, IMDB)对电影类型的设定,将影片类型分为爱情、传记、动画、动作、犯罪、黑色、幻想、家庭、惊悚、剧情、科幻、恐怖、历史、冒险、西部、喜剧、悬疑、音乐、运动、战争20种^[19]。鉴于豆瓣电影平台对影片类型的划分与IMDB有一定差异,且存在一些英文表达、繁简字体等不规范情况,因此采用人工介入的方式对其进行规范处理。包括繁简字体转换、中英文转换,高频不规范类型的人工规范,以及与IMDB影片类型的映射。

3.2.3 样本抽取

在完成数据规范处理的基础上,随机抽取200位用户作为样本进行推荐实验。同时,考虑到避免观看影片数量过少对建模效果的影响,剔除观影数量在20部以下的36位用户,剩余164位用户构成最终研究样本,并以全量数据作为候选集构建依据进行实验。为便于实验开展,将用户观影记录划分为训练集和测试集,实验方法:将每个用户看过的影片按照时间排序,将前

60.0%部影片作为训练集,后40.0%部影片作为测试集用于检验效果。

4 实验

4.1 实验过程

在数据预处理的基础上,本文采用上述的方法进行实验,其主要环节包括基于用户群体兴趣的初始候选集构建、基于向量空间的用户兴趣建模、基于影片与用户兴趣向量相关度的候选影片过滤、基于相关度与影片热度的个性化推荐影片排序。

(1) 基于用户群体兴趣的初始候选集构建。从预调研数据可以看出,影片观看是一个低频需求,如果用户群体规模过小,容易导致群体兴趣不稳定,因此本文将全部830 582位用户作为一个群体。与此同时,发现以最近1天和超过1天的数据进行用户群体兴趣建模的预测能力相差不大。以2014年10月10日为目标日期为例,分别选取前1天、前3天及前5天全部用户观看电影记录并按观看频次排序,其中Top500热门电影数量在目标日期Top500热门电影中占比分别为74.6%、77.8%及79.2%($p>0.05$)。基于此,本文以最近1天的观影数据进行群体用户兴趣建模。在初始候选集生成方面,以目标日期前1天的热门影片作为候选影片,通过对50天数据分析发现,如果将影片按照观看频次进行排序,以观看频次超过总频次的80.0%作为热门影片的判断标准,其中有48天的热门影片阈值均为2,因此本文将观看频次大于或等于2的电影纳入实验组推荐候选集;而对照组的初始候选集则选择实验数据集的75 694部电影。

(2) 基于向量空间的用户兴趣建模。向量空间模型的基本思路:根据用户历史行为,统计用户感兴趣的特征词及频次,频次越高代表用户对相关特征越感兴趣^[20]。由于在相当长的一段时间内,用户的观影偏好较为稳定,因此以用户全部历史数据作为兴趣建模的数据。各影片类型的兴趣权重计算方法如公式(2)所示。

$$W_u(u_i, t_j) = \frac{F(u_i, t_j)}{\sqrt{\sum_{j=1}^{20} F(u_i, t_j)^2}} \quad (2)$$

其中, $W_u(u_i, t_j)$ 表示用户 u_i 的兴趣类型 t_j 的权重, $F(u_i, t_j)$ 表示类型 t_j 在用户 u_i 看过的影片中出现的频次。

(3) 基于影片与用户兴趣向量相关度的候选影片过滤。在技术实现上,首先,需要在固定类型顺序的基

础上将每部影片按照所属类型转换为类型向量;即如果电影属于某个类型,则将此类型下的值设为1,否则设为0。其次,计算用户观看过的每部影片在每个类型上的权重,其方法如公式(3)所示。

$$W_m(m_i, t_j) = \frac{V(m_i, t_j)}{\sqrt{\sum_{j=1}^{20} V(m_i, t_j)^2}} \quad (3)$$

其中, $W_m(m_i, t_j)$ 表示电影 m_i 的兴趣类型 t_j 的权重, $V(m_i, t_j)$ 表示电影 m_i 在类型 t_j 上的值。

最后,在获得用户兴趣向量与影片类型向量的基础上,计算用户兴趣向量与每一部影片类型的相关度,并将相关度大于0的影片作为该用户最终的候选推荐影片。相关度的计算方法如公式(4)所示。

$$Rel(u_i, m_j) = \sum_{k=1}^{20} W_u(u_i, t_k) \times W_m(m_j, t_k) \quad (4)$$

其中, $Rel(u_i, m_j)$ 表示用户 u_i 的兴趣与电影 m_j 类型的相关度, $W_u(u_i, t_k)$ 表示用户 u_i 的兴趣类型 t_k 的权重, $W_m(m_j, t_k)$ 表示电影 m_j 的兴趣类型 t_k 的权重。

试验组及对照组的相关度计算结果样例如表1和表2所示。

表1 试验组用户——电影相关度计算结果(局部)

用户编号	电影名	相关度	电影在前1天被观看频率/次
User 1	《大魔术师》	0.768	10
User 1	《两小无猜》	0.731	29
User 6	《毕业生》	0.862	7
User 12	《十二生肖》	0.382	22
User 86	《非法入侵》	0.677	3
User 55	《叶问2: 宗师传奇》	0.167	104
User 142	《大话西游之月光宝盒》	0.784	12
User 121	《怦然心动》	0.682	68
User 26	《肩上蝶》	0.913	5

表2 对照组用户——电影相关度计算结果(局部)

用户编号	电影名	相关度	电影被观看总频率/次
User 5	《小时代2: 青木时代》	0.915	28 689
User 11	《奥菲斯》	0.823	126
User 42	《春娇与志明》	0.839	57 552
User 6	《女儿红》	0.815	131

续表

用户编号	电影名	相关度	电影被观看总频率/次
User 88	《喜马拉雅星》	0.738	5 435
User 97	《平行理论》	0.669	602
User 116	《塔斯鲁波的手提箱2》	0.745	2
User 42	《绿鱼》	0.781	351
User 25	《星尘往事》	0.598	227

(4) 基于相关度与影片热度的个性化推荐影片排序。在个性化推荐排序中,首先考虑的因素是影片类型向量与用户兴趣向量间的相关度,但由于存在大量影片具有相同的类型向量,因此其在与用户兴趣向量的相关度上并无差异。同时,对群体用户而言,用户对影片的兴趣是一个缓慢衰减的过程,影片的热度越高则其第2天继续热门的概率也越高;对于相关度相同的影片,宜采用基于热度的方法进行排序,从而形成最终的推荐列表。对于对照组而言,同样存在类似的问题,因此也需引入热度数据对同相关度影片进行二次排序,需要指出的是,其热度用所有历史记录中用户观看电影的频次来衡量。

4.2 实验结果分析

根据评价指标设定与计算方式,分别统计试验组和对照组在不同长度推荐列表(N 值)下推荐的 $P@N$ 值,如表3所示。

表3 不同 N 值下试验组与对照组的 $P@N$ 值

$P@N$	试验组准确率	对照组准确率	p 值
$P@10$	0.29%	0.13%	0.000
$P@20$	0.23%	0.10%	0.000
$P@50$	0.14%	0.07%	0.000
$P@100$	0.13%	0.05%	0.000
$P@200$	0.10%	0.03%	0.000
$P@500$	0.09%	0.02%	0.000
$P@1\ 000$	0.07%	0.01%	0.000

为验证在不同 N 值下,试验组和对照组在推荐准确率方面是否存在显著差异,本文选取卡方检验进行

两组策略的差异性检验,其依据是卡方检验适用于检验两个或两个以上变量是否有一致性。结果显示,在不同 N 值下试验组与对照组在推荐准确率方面均存在显著差异($p<0.001$),证明试验组和对照组存在显著性差异。从表3可以看出,在不同 N 值下,试验组准确率均比对照组高,随着 N 值的扩大,试验组和对照组的准确率都下降,但试验组的准确率下降趋势明显比对照组缓慢。因此,试验组推荐效果明显好于对照组。

显然,试验组效果明显好于对照组的主要原因是其候选集构建方法更合理。对每个用户的候选集规模统计发现,试验组每个用户的候选集平均规模为2 106部影片,而对照组的平均候选集为62 576部影片;同时,试验组对用户观看过的影片覆盖率83.9%。因此,试验组的候选集构建方案能够在仅牺牲覆盖率16.1%的情况下,将候选集规模缩减到对照组的3.4%,从而大幅减少误推荐的可能性。究其根本,具体到特定的日期,只有很少影片会被用户观看,即使对热门影片也是如此。通过统计2014年10月6日—10日所有用户的观影记录可知,在Top5 000历史热门影片中,平均有37.1%的影片并未出现在当天的观影记录中。

尽管试验组的推荐准确率明显好于对照组,但其效果仍较差($P@10$ 仅为0.29%),主要原因可能包括三个方面。(1)在按类型表示电影时,由于一部电影可能拥有不同的类型划分。本文认为一部电影在其所属所有类型上的比重是相同的,计算一部电影在各类型上的权重值也是相同的;但在实际中,一部电影在各类型下的权重是不同的,如《碟中谍》的类型划分为动作、惊悚和冒险,但其动作和冒险的成分明显大于惊悚,因此,将其类型权重设为相同不能展现电影的全部特征。

(2)尽管用户的观影兴趣相对稳定,但仍有部分用户的观影兴趣会随时间发生迁移,而使用的训练集数据中也有部分距离现在较为久远,这也可能导致兴趣分析产生偏差,进而影响推荐的准确率。(3)用户的观影决策会受多种因素的影响(如影片质量、类型、演员、上映时间等)。因此,仅基于电影类型进行用户兴趣模型构建不够全面,也存在少部分用户误操作产生噪声影响用户兴趣建模,从而影响个性化推荐的准确度。

5 结论

个性化推荐中,候选集构建对最终的推荐效果具有重要影响,然而基于当前策略构建的候选集中通常

包含大量用户不感兴趣的资源。为解决该问题,本文提出基于用户群体兴趣与个性化特征相融合的候选集构建策略,并以电影数据为例对模型效果进行实验验证。结果显示,本文提出的候选集构建策略能在小幅降低用户感兴趣项目覆盖率的情况下,大幅缩小候选集规模,从而改善个性化推荐的效果。

当然,本文也有一些局限性:(1)将超过83万位用户看成一个群体,未进行群体细分;(2)本文选取电影所属的所有类型作为电影特征向量,并赋予相同的权重,这会造成电影资源的准确表示产生偏差,并影响用户兴趣表示;(3)在进行用户兴趣建模时仅采用影片类型一个维度进行建模,并未建立用户兴趣的综合模型,因而其结果可能未全面反映两种推荐候选集构建策略的效果。后续研究中,拟针对这三方面进行优化:

(1)扩大样本用户规模,并在此基础上研究用户群体划分策略,从而建立更完备的基于用户群体兴趣的个性化候选集构建模型;(2)基于用户认知对影片的类型权重进行差异化赋权,从而更合理地建立用户兴趣模型;(3)将主题外其他影响用户观影的因素纳入兴趣模型,从而更全面地进行两种候选集构建策略的效果比较。

参考文献

- [1] 顾立平.用户行为模型驱动个性化服务研究综述[J].现代图书情报技术,2010(10):1-9.
- [2] SUN J,WANG G,CHENG X,et al.Mining affective text to improve social media item recommendation[J].Information Processing & Management, 2014, 51(4):444-457.
- [3] 姜书浩,潘旭华,薛福亮.一种基于项目聚类的自主推荐多样性优化算法[J].现代图书情报技术,2015(5):34-41.
- [4] DOOMS S,AUDENAERT P,FOSTIER J,et al.In-memory,distributed content-based recommender system[J].Journal of Intelligent Information Systems,2014,42(3):645-669.
- [5] 李树青.基于三词共现分析的学者主要研究兴趣识别及个性化外文推荐服务的实现[J].情报学报,2013,32(6):629-639.
- [6] 李志隆,王道平,关忠兴.基于领域本体的用户兴趣模型构建方法研究[J].情报科学,2015,32(11):69-73.
- [7] LI L,ZHENG L,YANG F,et al. Modeling and broadening temporal user interest in personalized news recommendation[J].Expert Systems with Applications,2014,41(7):3168-3177.
- [8] ALHAMID M F,RAWASHDEH M,DONG H,et al.RecAm:a collaborative

- context-aware framework for multimedia recommendations in an ambient intelligence environment[J].Multimedia Systems, 2016,22(5):1-15.
- [9] ALHAMIDM F,RAWASHDEH M.Towards context-aware media recommendation based on social tagging[J].Journal of Intelligent Information Systems,2016,46(3):499-516.
- [10] 周朴雄,陶梦莹.移动网络环境下情景敏感的个性化信息推荐系统研究[J].图书情报工作,2012,56(19):80-84.
- [11] BELLOGiN A,CASTELLS P,CANTADOR I. Neighbor selection and weighting in user-based collaborative filtering:a performance prediction approach[J].Acm Transactions on the Web,2014,8(2):307-323.
- [12] KALELI C.An entropy-based neighbor selection approach for collaborative filtering[J].Knowledge-Based Systems,2014,56(C):273-280.
- [13] 奉国和梁晓婷协同过滤推荐研究综述[J].图书情报工作,2011,55(16):126-130.
- [14] 王晓耘,钱璐,黄时友.基于粗糙用户聚类的协同过滤推荐模型[J].现代图书情报技术, 2015(1):45-51.
- [15] 林耀进,胡学钢,李慧宗.基于用户群体影响的协同过滤推荐算法[J].情报学报,2013,32(3):299-305.
- [16] 胡潜,林鑫.社会化标注系统中基于标签和项目的兴趣建模比较研究[J].情报学报,2015,34(12):1296-1303.
- [17] 许海玲,吴潇,李晓东,等.互联网推荐系统比较研究[J].软件学报, 2009,20(2):350-362.
- [18] GUAN Y,ZHAO D,ZENG A, et al.Preference of online users and personalized recommendations[J].Physica A: Statistical Mechanics and its Applications,2013,392(16):3417-3423.
- [19] IMDB.Most popular by genre[EB/OL].[2016-06-11].http://www.imdb.com/genre/?ref_=n-v_ch_gr_5.
- [20] 周彩兰,王鹏.基于空间向量模型的用户建模算法改进[J].计算机与数字工程,2010,38(2):15-17,135.

作者简介

石宇,女,1993年生,硕士研究生,研究方向:数字化信息服务与用户,E-mail: 1084174531@qq.com。

查梦娟,女,1993年生,硕士研究生,研究方向:数字化信息服务与用户。

梁宇,女,硕士,中级工程师,研究方向:电力信息分析。

邱惠,女,硕士,高级工程师,研究方向:电力信息分析。

Construction of Candidate Item Sets in Personalized Recommendation Based on Group Interest

SHI Yu¹, ZHA MengJuan¹, LIANG Yu², QIU Hui²

(1.School of Information Management, Wuhan University, Wuhan 430072, China;

2.Center for Electric Operations Control of Ezhou Power Supply Company, Ezhou 436000, China)

Abstract: The purpose of this paper is to optimize the construction method of candidate item sets. We improve the effect of personalized recommendation by reducing the number of candidate item sets without sacrificing the coverage of items that users interested. The construction of candidate item sets in the personalized recommendation is based on the method of fusing group interest and the personalized features. First of all, this paper constructs the initial candidate item sets based on the group interest, then filter the candidate item sets based on user's personalized features. This paper verifies the effectiveness of the method by taking movie data as an example. The results show that the number of candidate item sets reduce to 3.4% of the original when the coverage of items decrease slightly. Moreover, this method can improve the effect of personalized recommendation significantly. Thus, it can be concluded that it is an ideal tactic to construct the candidate item sets.

Keywords: Personalized Recommendation; Group Interest; Candidate Item Sets

(收稿日期: 2016-10-30)