文章编号：2096-1618(2021)03-0305-06

一种融合 PageRank 和 PersonalRank 的多层个性化推荐算法

欧如月， 陶宏才
(西南交通大学计算机与人工智能学院，四川 成都 611756)

摘要：传统的推荐系统只能实现一种类型的实体推荐，为解决一次性进行多种类型实体即多层推荐的问题，提出一种融合 PageRank 和 PersonalRank 的多层个性化推荐算法。利用图数据模型中的顶点描述实体，边描述实体间关联关系的这种特性，在图中将用户作为第一层实体即起始点，而将用户的历史行为(如评论过的电影)作为第二层实体，根据第二层实体依次给用户推荐第三层、第四层直到第 N 层的实体列表。通过爬虫爬取豆瓣网电影获取数据集，实验结果表明该模型具有多层推荐的效果，并较 PersonalRank 算法有更高的准确率和召回率。

关 键 词：推荐系统；多层推荐；PageRank；PersonalRank；图模型
中图分类号：TP311.5 **文献标志码**：A
doi：10.16836/j.cnki.jcuit.2021.03.011

0 引言

随着信息技术和互联网的发展，人们所面对的信息量每天都在不断地上升，而要从中获取自身所需要的信息越来越难。推荐系统可以帮助人们从海量信息里面获取所需信息时节省大量时间，而运用了个性化推荐算法的推荐系统，更能够提高用户从海量信息中获取所需信息时的效率。

传统的个性化推荐算法只能用于推荐一种类型的物品。当用户需要同时购买多种类型的物品时，需要推荐系统具有多层推荐的能力，即根据用户的历史行为，为用户一次性推荐多种类型的物品。具有多层推荐能力的推荐系统灵活多变，可以一次性满足用户的多种需求，节省了时间成本。传统的 PageRank 和 PersonalRank 算法都不能实现多层推荐，目前对个性化推荐算法的研究来看，多层个性化推荐研究极少。本文将传统的 PageRank 和 PersonalRank 算法进行融合，提出一种新的多层个性化推荐算法，既保障达到逐层推荐的效果，同时在推荐的准确率及召回率上有较好的提升。

1 相关算法

1.1 PageRank

基于链接的排名算法能够改善 Web 搜索，其中 PageRank 算法已在 Google 搜索引擎中成功使用，并引起广泛关注[1]。PageRank 算法最早由美国斯坦福大学研究生 Page 和 Brin 提出[2]，其核心思想是根据网页链接到其他网页和其他网页链接到本网页的引用关系来计算网页的得分，并且根据分值进行排序[3]。该算法用以评价网页的重要性，在学术领域也有较多应用，如期刊评价[4-5]、论文评价[6-7]、作者评价[8]以及机构评价[9]等。PageRank 算法运用于有向图[10]，图中把网页当作顶点，把网页之间的超链接跳转关系当作边。每个网页的初始访问概率为 $1/N$，N 代表顶点个数。当用户点击网页，通过超链接在各个网页之间不停地跳转，每个网页的重要性(用 PR 描述)最终会被收敛到一个稳定的 PR 值。

如图 1 所示，图中共有 4 网页，分别为 A、B、C 和 D，每个网页有出链和入链。例如，网页 A 有 3 条出链和 2 条入链。

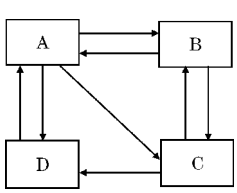

图 1 网页模型图

网页的重要性排名由所有指向它的入链的权值之和决定，每个网页的 PR 值为

$$PR(u)=\sum_{v\in Iu}\frac{PR(v)}{L(v)} \tag{1}$$

其中：Iu 为指向节点 u 的所有入链顶点的集合，$PR(v)$ 为网页 v 的 PR 值，$L(v)$ 为网页 v 的所有出链数。

收稿日期：2020-03-28
基金项目：国家自然科学基金资助项目(61806170)

经过多次迭代之后,所有网页的 *PR* 值都会被收敛到一个稳定的值,最终对所有 *PR* 值进行排序即可得到网页的重要性排名。

以图 1 中的网页 A 为例,由式(1)经计算可得网页 A 第一次迭代后的 *PR* 值为0.375,经过多次迭代后,网页 A 的 *PR* 值收敛到0.33。

1.2 PersonalRank

基于图的推荐算法,除了 PageRank 算法之外,还有 PersonalRank 算法[10]。PersonalRank 在 PageRank 算法基础之上作了一些改变,如图 2 所示,PersonalRank 不再使用图 1 的方式构建图模型,而是采用二分图的方式构建图,在图中只包含两种类型的顶点,分别是用户和物品。经过迭代计算,最终可以得到每个用户的物品推荐列表,而 PageRank 不能针对某个用户进行推荐,因此 PersonalRank 算法增加了用户的个性化,在度量节点间相似程度方面得到应用[11]。PersonalRank 算法是一种随机游走算法[12],可以求出图中所有物品节点的重要性排名。

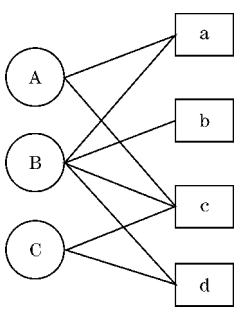

图 2 用户-物品模型图

图 2 中 A、B 和 C 分别表示 3 个不同的用户,a、b、c 和 d 分别表示 4 个不同的物品,用户和物品之间的边表示他们之间具有相关性。如果给节点 A 推荐物品,从节点 A 出发的游走过程为:首先由于节点 A 与节点 a 和 c 具有相关性,因此从节点 A 游走到节点 a 和节点 c。到达节点 a 和节点 c 之后,以概率 a 选择继续游走,游走时节点 a 和节点 c 的 PR 值按该节点的出度数均分之后继续游走[12],直至游走结束。

每个顶点的 *PR* 值计算方式为

$$PR(j)=\begin{cases} a\cdot\sum\limits_{i\in in(j)}\dfrac{PR(i)}{|\,out(i)\,|} & \text{if}(j\neq u)\\ (1-a)+a\cdot\sum\limits_{i\in in(j)}\dfrac{PR(i)}{|\,out(i)\,|} & \text{if}(j=u)\end{cases} \tag{2}$$

式中,$PR(j)$ 表示用户节点 j 被访问的概率,$PR(i)$ 表示物品节点 i 被访问的概率,$out(i)$ 表示物品节点 i 的出链个数,$in(j)$ 表示用户节点 j 的入链个数,a 表示继续访问的概率,一般取值为0.8。

经过多次迭代后,每个物品的 *PR* 值就会收敛到某个数,最终对所有 *PR* 值进行排序即可得到物品节点重要程度的排名。

以节点 a 为例,由式(2)经计算可得节点 a 第一次游走后的 *PR* 值为0.16,经过多次迭代后,节点 a 的 *PR* 值收敛到0.089。

2 多层个性化推荐算法

2.1 算法的思想

本文提出一种融合 PageRank 和 PersonalRank 算法的多层个性化推荐算法,命名为 PP-Rank 算法。其主要思想包括 3 个方面:一是从降低时间复杂度的角度出发,将错综复杂的图采用子图的方式进行存储,算法只在对应的子图中进行迭代;二是利用图中的多度关联关系形成层次丰富的结构,实现适用性广泛,灵活多变的多层次的个性化推荐;三是将被推荐者作为图的起始点,从起始点出发,局部迭代实现个性化推荐。

2.2 算法过程

2.2.1 构建多层的图模型

根据所爬取的数据集,分析出构建图模型所需的所有实体和实体间的关联关系。准备好所有的实体和实体间的关联关系后,开始构建多层的图模型。在图模型中,用顶点来描述实体,用边来描述实体之间的关联关系[13]。首先确定第一层顶点(一般为被推荐者)即图的起始点。通过第一层实体和第二层实体之间的关联关系,确定第二层顶点。通过第二层实体和第三层实体之间的关联关系,确定第三层顶点,以此类推,直至图模型构建完成。图 3 为多层的图结构,在图中第一层为用户,第二层为用户评价过的电影,第三层为候选推荐导演,第四层为候选推荐电影,第五层为候选推荐演员。

2.2.2 利用子图的方式存储图进行局部迭代

在多层个性化推荐的过程中,首先要对数据进行持久化存储,形成复杂的网络结构,推荐算法才能在图模型之上进行迭代计算实现多层个性化推荐。但是,图的存储方式会影响算法在图模型之上的迭代效率,如果改变图的存储结构,就可以改变算法的时间复杂度。而如果把图存储为一个一个的子图,子图与子图之间没有关联关系,当算法在图模型上迭代时,将由全局迭代转变为局部子图迭代,这样可以降低算法的时间复杂度。图 3 为使用子图前的图结构,图 4 为使用子图后用户 1 的子图结构,图 5 为使用子图后用户 2

的子图结构。

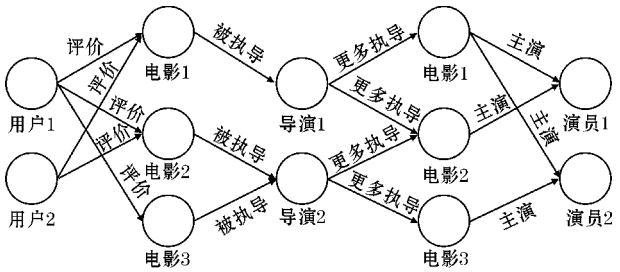

图 3　未使用子图的图结构

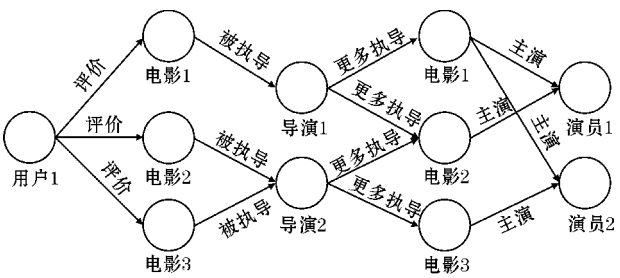

图 4　用户 1 的子图结构

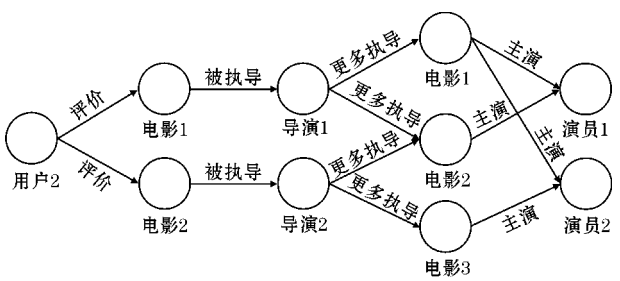

图 5　用户 2 的子图结构

2.2.3　从图的起始点出发进行局部迭代

当完成构建多层的子图之后,将在子图中进行局部迭代,具体方式是从图的起始点出发,根据层与层之间的关联关系到达下一层,以此类推,直到到达图的最后一层,至此完成多层推荐。例如,图 5 中起始点为用户,根据评价边到达第二层用户评价过的电影,从第二层开始通过被执导边到达第三层,从第三层开始通过更多执导边到达第四层,从第四层开始通过主演边到达第五层。在此迭代过程中,通过第三层开始可以给用户分别推荐他可能感兴趣的导演、电影和演员。

从算法的角度进行分析,当算法在子图上进行迭代时,首先第一层顶点的 Rank 值(被推荐的概率)为固定值 1。再根据第一层和第二层的关联关系迭代至第二层,计算第二层的所有顶点的 Rank 值,得到第二层所有顶点的 Rank 值列表。以此类推,当迭代完成时,可得第三层至最后一层的各层对应的 Rank 值列表。将每一个 Rank 值列表降序排列,在列表中排名前 N 的 Rank 值所对应的实体将被作为该层的推荐结果。

2.3　算法的描述和实现

PP-Rank 算法适用于具有多层的有向子图,在图中将被推荐者作为图的起始点,根据层与层之间的关联关系链接到当前层的下一层。所以,在图中除了起始点的 Rank 值为固定值 1 之外,其余每一层的每一个节点将得到一个 Rank 值,这个值将作为在该层所有节点中被推荐的重要依据,Rank 值越大,越容易被推荐。根据式(3)可计算得到图中所有节点的 Rank 值。

$$PR(v)=\begin{cases}1 & v=source\\ a\sum_{v\in Iu}\dfrac{PR(v)}{L(v)} & v!=source\end{cases}\tag{3}$$

式中,$PR(v)$是节点 v 的 Rank 值,Iu 是所有链接到节点 u 的节点集合,节点 v 是属于集合 Iu 的一个实体,$L(v)$是实体 v 的对外链接数,a 是调节参数。

PP-Rank 算法伪代码描述如下:

输入:源顶点 id,路径规则 steps,继续游走的概率 alpha

输出:从源顶点开始的每一层的 rank 列表

```
(1)function PP-Rank(source, steps, alpha){
(2)ranks = [ ]
(3)sources = [source]
(4)rank = {source:1}
(5)for step in steps
(6)   for source in sources
(7)      nextLayerVertexs = getVertex(source,step.direction,step.labels,step.degree)
(8)for v in nextLayerVertexs
(9)         rank[v] = (rank[v]? rank[v]:0) + rank[source] * alpha/nextLayerVertexs.length
(10)      end for
(11)   end for
(12)   ranks.push(rank)
(13)   sources=nextLayerVertexs
(14)   rank=mixed(rank,sources)
(15)end for
(15)return ranks
(16)end function
```

2.4　算法的性能分析

由 PP-Rank 算法可知,在迭代过程中,首先遍历图模型的所有层,再遍历每层中的所有顶点。根据当前顶点所在层获取下一层的所有顶点,再遍历下一层的所有顶点,计算所有顶点的 Rank 值。假设图模型有 m 层和 n 个顶点,则算法的时间复杂度为 $O(mn^2)$。由

于该算法在迭代过程中,开辟了两个数组空间和两个对象空间,所以算法的空间复杂度为 $O(1)$。

3 实验与分析

3.1 数据集

本文的多层个性化推荐实验以导演-电影-演员多层推荐为例,所用实验数据集是使用爬虫在豆瓣网上通过爬取而获得。该数据集包括 4500 多个用户,30000 多部留有用户历史行为记录的电影,50000 多个候选推荐导演,30000 多部候选推荐电影,40000 多个候选推荐演员。每个用户至少对一部电影留下过历史行为记录。表 1 展示了数据集中作为图模型的第一层用户的部分字段信息,表 2 展示了数据集中作为图模型的第二层用户留下历史行为记录的电影的部分字段信息,表 3 展示了数据集中作为图模型的第三层候选推荐导演的部分字段信息,表 4 展示了数据集中作为图模型的第四层候选推荐电影的部分字段信息,表 5 展示了数据集中作为图模型的第五层候选推荐演员的部分字段信息。

表 1 图模型第一层部分数据展示

user_id	昵称	性别	年龄	省份	学历	职业	签名
Bearish	爱新觉罗	男	36 岁	陕西	专科	古筝手	细水长流

表 2 图模型第二层部分数据展示

mv_id	电影名	导演	主演	类型	语言	标签	评分	推荐
1291585	风之谷	宫崎骏	岛本须美,松田洋治	动画	日语	动画-日本	8.9	45

表 3 图模型第三层部分数据展示

director_id	导演	职业	代表作
1291585	风之谷	宫崎骏	岛本须美,松田洋治

表 4 图模型第四层部分数据展示

tj_mv_id	电影	导演	类型	地区	评分
38131	浪客剑心	古桥一浩,古田部胜义	剧情	日本	8.9

表 5 图模型第五层部分数据展示

actor_id	演员	职业	代表作
1024606	宫内幸平	演员	风之谷/聪明的一休/七龙珠

在以上数据集中的所有表中,将数据集随机拆分成训练集和测试集,两者比例为 8∶2。

3.2 数据处理

爬取的数据集可能会存在数据重复、数据缺失以及数据乱码等问题。为减少实验中产生的误差,实验研究之前应该对数据做处理,即对数据做清洗,整理的操作。

3.2.1 去除重复数据

数据集中可能存在重复的数据(如用户、电影、导演、演员等信息均相同),只保留一条即可。具体方法是,使用一个列表保留所有的样本数据,遍历列表进行去重操作。

3.2.2 去除空值数据

数据集中可能存在必要的字段为空值的情况,这种缺乏必要字段的数据项会对实验结果造成比较大的影响,应该删除。具体方法是,使用一个列表保留所有的样本数据,遍历列表判断必要字段是否为空。如为空就删除此数据项;不为空则不对该数据项做操作。

3.2.3 乱码数据处理

数据集中可能存在无法识别的繁体字、字符以及表情符号。对于繁体字可以采取插件库的方式将其转换为中文简体字,而无法识别的字符和表情符号需要通过正则表达式的方式将其进行匹配,去除无法识别的字符和表情符号。

3.3 数据导入

将数据集按照实体类型进行分层,再建立层与层之间的关联关系。准备好数据后,需要对数据进行持久化存储,即存到 MySQL 数据库中,这样可以提高操作数据的效率,节省时间。

3.4 实验评价指标

本文采用准确率和召回率[14]作为实验评价指标。召回率和准确率计算方式如表 6 所示。

表 6 召回率和准确率计算方式

	相关	不相关
检索到	A	B
未检索到	C	D

准确率=$A(A+B)^{-1}$,召回率=$A(A+C)^{-1}$。

3.5　实验对比

为验证 PP-Rank 算法的推荐效果,将 PP-Rank 和 PersonalRank 算法进行对比,实验完成,两个算法所运行的时间以及消耗的内存如表 7 所示,两个算法在所爬取的数据集中所求得的召回率和准确率如表 8 所示。

表 7　不同算法的运行时间和消耗内存

算法	运行时间/s	消耗内存/MB
PP-Rank	179.26	497
PersonalRank	43.71	341

表 8　不同算法的召回率和准确率

算法	召回率	准确率
PP-Rank	0.79	0.46
PersonalRank	0.57	0.23

从表 7 可知,PP-Rank 和 PersonalRank 相比,PP-Rank 算法时间性能更好。这是因为 PP-Rank 算法只需要在子图中进行迭代,可以节省大量时间,而 PersonalRank 需要进行全局迭代。PP-Rank 算法内存消耗高于 PersonalRank 算法,这是因为 PP-Rank 算法需要另外的内存来保存层之间的关联关系。

从表 8 可知,相较于传统个性化推荐算法 PersonalRank,PP-Rank 算法在召回率和准确率上都有一定的提升。

3.6　多层推荐实验结果

在多层的图模型中,由于第一层是用户,第二层是用户的历史行为,所以 PP-Rank 算法从第三层开始作推荐。图模型的第三层用来给用户推荐导演,第四层用来推荐电影,第五层用来推荐演员。表 9 为第三层、第四层和第五层推荐的召回率和准确率。

表 9　各层推荐的召回率和准确率

层级	召回率	准确率
第三层	0.81	0.39
第四层	0.78	0.45
第五层	0.77	0.53

由表 9 可知,在图模型中距离第一层越远的层级,推荐的准确率越高,但召回率越低。距离第一层越近的层级,推荐的准确率越低,但召回率越高。

4　结束语

在个性化推荐领域,以往的研究几乎很少涉及多层推荐,本文提出的新算法可以较好地实现此目标。同时,经过对比实验,也可以看出多层推荐算法在召回率和准确率上都有所提高。当然本文算法也有不足之处:其一,算法只适用于老用户的多层个性化推荐情况,而不适用于新用户的多层个性化推荐;其二,因多层推荐算法中层级越丰富推荐效果会越好,故要求所用数据集须构建丰富的图模型层级。

参考文献:

[1] Xuanhui Wang, Tao Tao, Jiantao Sun, et al. DirichletRank:Solving the Zero-One Gap Problem of PageRank[J]. ACM Transactions on Information Systems,2008,26(2):1-66.

[2] 库珊,刘钊. 基于 PageRank 与 HITS 的改进算法的网页排名优化[J]. 武汉科技大学学报,2019,42(2):155-160.

[3] 戴炳荣,姜胜明,李顿伟,等. 基于改进 PageRank 算法的跨链公证人机制评价模型[J/OL]. 计算机工程,https://doi. org/10. 19678/j. issn. 1000-3428. 0056460,2020-05-24.

[4] Bollen J,Rodriquez M A,Van De Sompel H. Journal status[J]. Scientometrics,2006,69(3):669-687.

[5] 马凤. 基于 PageRank 算法的期刊影响力研究[J]. 情报杂志,2014,33(12):103-108.

[6] Chen P,Xie H,Maslov S,Redner S. Finding scientific gems with Google's PageRank algorithm[J]. Journal of Informetrics,2007,1(1):8-15.

[7] 苏成, Hee-Sop KIM. 基于 PageRank, HITS 和 SALSA 算法的学术论文评价[J]. 情报杂志,2015,34(6):48-54.

[8] Ding Y,Yan E,Frazho A,Caverlee J. PageRank for ranking authors in co-citation netw orks[J]. Journal of the American Society for Information Science and Technology,2009,60(11):2229-2243.

[9] Fiala D. Suborganizations of institutions in Library and Information Science Journals[J]. Information,2013,4(4):351-366.

[10] 邱苓芸,王铭,赵卫东. PageRank 算法改进研究[J]. 软件导刊,2017,16(2):74-76.

[11] 刘清,王帆,冯亮,等. 高效图推荐算法应用研

究[J]. 软件导刊,2019,18(8):49-51.
[12] 吴迪,周利娟,林鸿飞. 基于随机游走的就业推荐系统研究与实现[J]. 广西师范大学学报(自然科学版),2011(1):179-185.
[13] 杨华. 基于图的商品推荐算法研究[D]. 南昌:江西师范大学,2017.
[14] 滕传志,赵月旭. 基于随机森林-马尔可夫用户冷启动推荐系统[J]. 计算机工程与设计,2020,41(11):3094-3098.

A Multi-layer Personalized Recommendation Algorithm Combining PageRank and PersonalRank

OU Ruyue, TAO Hongcai
(School of Computing & Artificial Intelligence, Southwest Jiaotong University, Chengdu 611756, China)

Abstract: The traditional recommendation system can only realize the recommendation of one type of entity. In order to solve the problem of multiple types of entities at one time, that is, multi-layer recommendation, a multi-layer personalized recommendation algorithm combining PageRank and PersonalRank algorithms is proposed. Firstly, it utilizes the characteristics of using vertices to describe entities in the graph data model, and describes the relationship among entities by edges. Then, it takes users as the first-level entities in the graph, i. e. ,the starting point, and the historical behaviors(e. g. , reviewed movies) left by users as the second-level entities. Further, according to the second layer the third layer is recommended to the user in turn, and the fourth layer up to the Nth layer of the entity list is also recommended. The experiment on the data set obtained by crawling Douban movies shows that the model has a multi-layer recommendation effect, and has a higher accuracy and recall rate than the original PersonalRank algorithm.

Keywords: recommendation system; multi-tier recommendation; PageRank; PersonalRank; graph model