# 融合初始资源与协同过滤的二部图推荐算法

赵 红 郑 骏
(华东师范大学计算中心 上海 200062)

**摘 要** 推荐系统的产生主要是为了解决信息过载的问题。基于二部图网络与基于协同过滤的推荐算法是目前应用比较广泛的算法,二者都取得了一定的推荐效果。基于加权二部图网络的算法忽略对初始资源的配置,基于物品的协同过滤算法在推荐时也产生数据稀疏等问题。组合推荐算法融合初始资源配置以及基于物品的协同过滤算法来解决相关的问题,可以达到更好的推荐效果。算法实验在 MovieLens 数据集上实施,结果表明,与传统的推荐算法以及最近的组合推荐算法相比,该方法有更好的推荐效果。

**关键词** 推荐算法 二部图网络 协同过滤 初始资源配置

**中图分类号** TP391 **文献标识码** A DOI:10.3969/j.issn.1000-386x.2019.01.051

## BIPARTITE GRAPH RECOMMENDATION ALGORITHM BASED ON INITIAL RESOURCES AND COLLABORATIVE FILTERING

Zhao Hong Zheng Jun
(*Computing Center, East China Normal University, Shanghai 200062, China*)

**Abstract** The recommendation system aims to solve the problem of information overload. The recommendation algorithms based on bipartite graph network and collaborative filtering are widely used at present. Both of them have achieved certain recommendation results. The algorithm based on the weighted bipartite graph network ignores the configuration of initial resources, and the item-based collaborative filtering algorithm has problems of data sparsity and other issues. The combined recommendation algorithm which combined initial resource configuration with item-based collaborative filtering algorithm could achieve a better recommendation effect to solve the problems. The experiment was implemented on the MovieLens data set. The results show that the method achieves a better recommendation effect compared with the traditional and current combined recommendation algorithms.

**Keywords** Recommendation algorithm Bipartite graph network Collaborative filtering Initial resource configuration

## 0 引 言

信息技术和互联网的迅速发展给人们带来了巨大的便利,与此同时,网络中的信息量也飞速增长,信息过载问题随着信息时代的发展也逐渐浮现出来。由于信息过载,用户不能轻易地从大量信息中获取自己所需要的部分。为了解决这个问题,推荐系统应运而生。

推荐算法是个性化推荐系统的核心部分,它的本质是通过一定的方式将用户和物品联系起来,使得用户可以快速找到自己感兴趣的物品。采用不同的方式进行推荐可以产生不同的推荐系统,现有的推荐算法主要有如下几种:协同过滤推荐算法、基于内容的推荐算法、组合推荐算法以及基于网络结构的推荐算法等。

协同过滤推荐算法[1]是推荐系统中最基础的算法,因此也有许多关于该算法的研究。该算法主要是利用收集到的用户信息来计算它们之间的相似度,筛选出目标用户的邻居用户对物品的评价,以此来预测目标用户对物品的喜好程度。该算法不受数据格式的影响,可以处理复杂数据进行有效快速的推荐,同时也存在数据稀疏、冷启动之类的问题。

基于内容的推荐算法[3]也是常用的算法,该算法通过整理用户过去喜欢的物品信息,找出与之相似度高的物品对目标用户进行推荐。该算法在计算物品相似度

收稿日期:2018-05-22。**赵红**,硕士生,主研领域:Web 应用技术。**郑骏**,教授。

的基础上分析用户和物品的内部信息,通过用户的喜好和物品的属性来进行推荐。该算法在考虑用户喜好的同时也达到了推荐结果直观且便于理解的作用。但是该算法处理非文本信息难度比较高,比如音乐、图像等。

组合推荐算法[5-6]是通过结合多种推荐算法来实现推荐。协同过滤推荐和基于内容的推荐都有一定的局限性,在实际应用中,通常将多种推荐算法结合起来使用,这样可以达到更好的效果。例如将以上两种算法组合使用就可以取长补短,组合推荐算法往往比单一的推荐算法有更高的准确率,但是时间以及空间开销也因此增加。

基于网络结构的推荐算法[7-8]不考虑用户和物品的信息,而是将用户与物品抽象为网络中的节点,用户与物品的关系隐藏在网络的连接中,该算法利用网络结构构成的信息进行推荐。Zhou 等[9]提出基于二部图网络的推荐算法,该方法在复杂网络物质扩散和热传导思想上得到启发,用资源分配的方法来计算用户间的相似性,取得了比其他算法更好的效果。在此基础上,将用户对项目的评分作为边权,提出了基于加权的二部图推荐算法[10-11],算法得到了进一步的优化。

本文考虑到各个算法的优缺点,将基于物品的协同过滤算法与结合初始资源配置的加权二部图网络推荐算法融合到一起,提高了推荐的效果。

# 1 基于网络结构的推荐算法

## 1.1 基于二部图的推荐算法

Zhou 等[9]提出了一种基于资源分配的方法。假设一种资源最初位于项目上,每个项目将它的资源平均分配给所有邻居用户,然后每个用户将收到的资源重新分配给它所选择过的所有项目。

算法思想是一种加权方法,图 1 表明了资源在二部图中的流动过程。图中上面三个节点是项目节点,下面四个节点是用户节点。

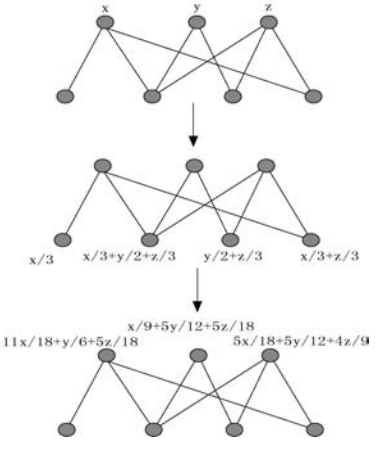


图 1 二部图的资源流动

设上面三个是节点 $x$,下面四个是节点 $y$。整个资源流动过程包括两步:资源首先从 $x$ 流向 $y$,然后重新流回 $x$。经过两部之后,$x$ 节点上资源由 $x$、$y$、$z$ 变成 $x'$、$y'$、$z'$。通过下式计算得到:

$$\begin{pmatrix} x' \\ y' \\ z' \end{pmatrix} = \begin{pmatrix} \frac{11}{18} & \frac{1}{6} & \frac{5}{18} \\ \frac{1}{9} & \frac{5}{12} & \frac{5}{18} \\ \frac{5}{18} & \frac{5}{12} & \frac{4}{9} \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} \tag{1}$$

考虑一般的二部图 $G(X,Y,E)$,其中 $E$ 是边的集合。$X$ 和 $Y$ 节点分别记为 $x_1,x_2,\cdots,x_n$ 和 $y_1,y_2,\cdots,y_m$。文中定义 $f(x_i)$ 为对象的初始资源,经过第一步,所有资源从 $X$ 流向 $Y$,$Y$ 中各个节点分配到的资源量为:

$$f(y_l) = \sum_{i=1}^{n} \frac{a_{il} f(x_i)}{k(x_i)} \tag{2}$$

式中:$k(x_i)$ 为 $x_i$ 的度,$a_{il}$ 为 $n \times m$ 的邻接矩阵,如下所示:

$$a_{il} = \begin{cases} 1 & x_i y_l \in E \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

接下来是第二步,所有的资源流回 $X$,最后集合 $X$ 中的节点的资源分配量为:

$$f'(x_i) = \sum_{j=1}^{n} w_{ij} f(x_j) \tag{4}$$

$$w_{ij} = \frac{1}{k(x_j)} \sum_{l=1}^{m} \frac{a_{il} a_{jl}}{k(y_l)} \tag{5}$$

对于任意一个用户,对他所有未选择的商品 $x_i$ 按照上述计算的 $f'(x_i)$ 进行降序排列,最后将资源量最大的那些商品推荐给目标用户,这种方法就是基于二部图的推荐算法(NBI)。

## 1.2 基于加权二部图的推荐算法

基于加权二部图的推荐算法(WNBI)[10-11]是在二部图推荐算法进行的改进算法,该方法将用户对项目的评分作为二部图中用户和项目的边权。资源扩散时,将资源按照边权占总边权的比例进行不均等分配。如图 2 所示,其中:

$$a_{il} = \begin{cases} x_i y_l & x_i y_l \in E \\ 0 & \text{otherwise} \end{cases} \tag{6}$$

经过两次资源流动:

$$f'(x_i) = \sum_{j=1}^{n} w_{ij} f(x_i) \tag{7}$$

$$w_{ij} = \frac{1}{k(x_j)} \sum_{l=1}^{m} \frac{a_{il} a_{jl}}{k(y_l)} \tag{8}$$

# 2 基于物品的协同过滤推荐算法

基于物品的协同过滤算法(I_CF)[12-13]是目前应

用比较广泛的算法。该算法根据用户选择过的物品信息给目标用户推荐与这些物品类似的物品。该算法与其他推荐算法的不同之处在于:在计算相似度时,它只分析用户的历史行为,而不用考虑项目的特征,在计算出相似度的同时也可以用历史行为对推荐结果进行合理的解释。

具体方法为将两个项目作为在 $m$ 维空间的两个向量,它们之间的相似性通过计算这两个向量之间角度的余弦来测量。在 $m \times n$ 的评分矩阵中,项目 $i$ 和 $j$ 之间的相似度通过 $S(i,j)$ 来度量:

$$S_{ij} = \frac{\boldsymbol{i} \cdot \boldsymbol{j}}{\| i \| \ \| j \|} \tag{9}$$

## 3　基于初始资源配置与协同过滤的二部图推荐算法

### 3.1　初始资源的配置

经过上一节的介绍,加权的二部图推荐算法本质是资源的流动,因此如何对初始资源进行配置是值得研究的问题。

考虑位于对象 $x_i$ 上的初始资源作为其分配的推荐能力。在整个推荐过程中,给予 $x_i$ 的总资源是 $p_i = \sum_j f_i^j$,其中 $j$ 包括所有用户 $u_j$。在统一初始资源的配置下,$x_i$ 的总资源是 $p_i = \sum_j f_i^j = \sum_j a_{ij} = k(x_i)$。这也就是说,分配给对象的总推荐资源与其度数成正比,因此度数高的对象(如流行度高的电影)资源量得到增强。虽然基于加权二部图的推荐算法已经取得了比较好的推荐效果,但是这种统一的配置还是过于简化,以适当的方式抑制高度对象的影响可以进一步提高推荐效果。

基于上述讨论,文献[14]提出一个更复杂的初始资源分配方式来代替:

$$f_i^j = a_{ij} k^{\beta}(x_j) \tag{10}$$

式中:$\beta$ 是一个可调参数,以往统一配置的情况下,$\beta = 0$;$\beta > 0$ 时强化了高度数对象的影响;$\beta < 0$ 时弱化了高度数对象的影响。经过文献中的论证,在算法复杂性没有增加的情况下,使用适当的 $\beta$,抑制高度节点的影响,从而使推荐效果在准确率、流行度、个性化三个方面均优于以前加权的情况。

资源流动仍与加权的二部图推荐算法类似,主要包括两个步骤,从项目节点流向用户节点,再从用户节点流回项目节点。资源的流动分配如图 2 所示,对于每个项目,所有用户对该项目的评分之和为初始资源的度:

$$k(x_i) = \sum_{j=1}^{m} a_{ij} \tag{11}$$

在加入可调参数 $\beta$ 后,$k(x_i)$ 变为 $k^{\beta}(x_i)$,同时对于每个项目,评分 $a_{ij}$ 也随之变化:

$$a'_{ij} = a_{ij} k^{\beta-1}(x_i) \tag{12}$$

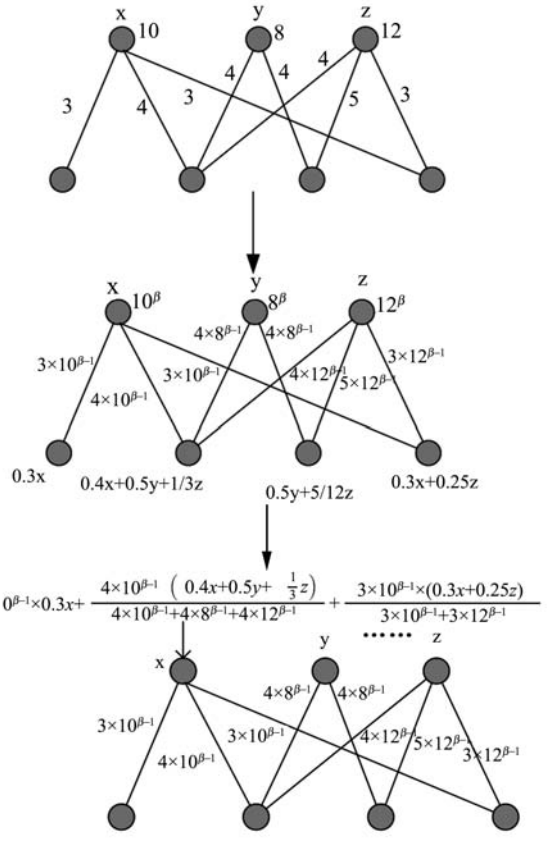


图 2　加入 $\beta$ 后的加权二部图资源流动

### 3.2　融合初始资源配置与协同过滤的二部图推荐算法

组合推荐算法近年来也有很多的应用,通过融合多种推荐算法往往可以达到取长补短的作用。基于物品的协同过滤推荐算法与加权二部图算法的融合[5](WNBI + CF)也在一定程度上提高了算法的效果。

根据以上内容的介绍,当项目与其他项目的总和相似度很大时,这意味着该项目在整个系统中非常活跃。所以当它被推荐给其他项目时,该项目应该具有更大的权重。因此,它被视为系统中的初始资源,同时考虑到数据的稀疏性等局限,我们可以将初始资源配置与基于物品的协同过滤结合到一起来进行推荐以达到比传统的加权二部图推荐更好的效果。

二者融合之后,按照二部图的流动分配资源,权重矩阵的计算变为:

$$w_{ij} = \frac{1}{k'(x_j)} \sum_{l=1}^{m} \frac{a'_{il} a'_{jl} S_j}{k'(y_l)} = \frac{1}{\sum_{l=1}^{m} k^{\beta-1}(x_j) a_{jl}} \sum_{l=1}^{m} \frac{k^{\beta-1}(x_i) k^{\beta-1}(x_j) a_{il} a_{jl} S_j}{\sum_{p=1}^{n} k^{\beta-1}(x_p) a_{pl}} \quad (13)$$

式中：

$$S_j = \sum_{k=1}^{m} S_{kj} \quad (14)$$

经过两次资源流动后，每个项目对应的资源量为：

$$f'(x_i) = \sum_{j=1}^{n} w_{ij} f(x_j) \quad (15)$$

对于任意一个用户，对他所有未选择的项目 $x_i$ 按照上述计算的 $f'(x_i)$ 进行降序排列，最后将资源量最大的那些商品推荐给用户。

# 4 实验分析

## 4.1 数据集

实验采用开源的 MovieLens 数据集，来自于网站 https://grouplens.org。该数据集中有 1 682 部电影，943 个用户，其中共有 100 000 条用户对电影的评分，评分区间为 1～5 分，得分高低反映了用户对该电影的偏爱程度。在本实验中该数据集被随机分为两部分，选取其中 90% 作为训练集，剩余 10% 作为测试集，测试集用于验证实验结果。

## 4.2 评价指标

本文采用准确率、召回率、F1 指标、覆盖率、新颖度等来对推荐算法进行评价。

### 4.2.1 准确率与召回率

准确率表示在得到的推荐列表中发生过的用户-物品评分记录的比例大小，召回率则表示用户-物品评分记录包含在推荐列表中的比例大小。准确率和召回率的高低代表着推荐效果的好坏。

$$P_{\text{recision}} = \frac{\sum_{u} | R(u) | \cap T(u)}{\sum_{u} | R(u) |} \quad (16)$$

$$R_{\text{ecall}} = \frac{\sum_{u} | R(u) \cap R(u) |}{\sum_{u} | T(u) |} \quad (17)$$

### 4.2.2 F1 指标

F1 指标同时考虑准确率和召回率，可以比较全面地评价推荐算法的效果。

$$F_1(L) = \frac{2P(L)R(L)}{P(L) + R(L)} \quad (18)$$

式中：$P(L)$ 为准确率，$P(L)$ 为召回率。

### 4.2.3 覆盖率与新颖度

除了对推荐算法的精度进行评价，还可以计算算法的覆盖率与新颖度来进行评测。覆盖率主要代表推荐算法对长尾的发掘能力，它的高低在另一个方面评价推荐结果的好坏，下式表示在得到的推荐列表中包含物品的比例大小。

$$C_{\text{overage}} = \frac{| U_{u \in U} R(u) |}{| I |} \quad (19)$$

同时，新颖度是通过评估推荐列表中物品的平均流行度来评测算法。如果推荐列表中的物品都很热门，那表示新颖度比较低，反则说明推荐列表的物品比较新颖，推荐效果好。

## 4.3 实验结果

根据文献[14]的内容，$\beta$ 值取 0.8 时可以达到最好的效果，所以在本文中，取 $\beta$ 来进行实验。本实验设置推荐的 Top N 的值分别为 5、10、20、30、50、70、100。实验结果的表格和图中提到算法的中英文名称对照说明如表 1 所示，比较五种算法的推荐效果，实验结果平均值如表 2 所示。

**表 1 表格和图中提到算法的中英文名称对照说明**

| | |
|---|---|
| I_CF | 基于物品的协同过滤算法 |
| NBI | 二部图推荐算法 |
| WNBI | 加权二部图推荐算法 |
| WNBI_β + CF | 融合初始资源配置以及协同过滤的二部图加权算法 |

**表 2 四种算法的实验结果平均值**

| 算法 | 评价指标 | | | | |
|---|---|---|---|---|---|
| | Precision | Recall | F1 | Coverage | Popularity |
| I_CF | 10.17% | 30.15% | 12.19% | 30.83% | 6.01 |
| NBI | 10.29% | 16.03% | 11.23% | 19.48% | 6.18 |
| WNBI | 13.23% | 19.71% | 14.12% | 14.54% | 6.37 |
| WNBI_β + CF | 14.52% | 21.52% | 15.09% | 43.77% | 6.43 |

从表 2 中可以看出，WNBI_β + CF 算法的准确率（14.52%）和 F1 值（15.09%）以及覆盖率（43.77%）三个标准都相对较高，其中覆盖率有明显的提高，说明此方法有效降低了热门物品的影响；由于准确率得到提升，召回率方面相应有所降低；流行度方面四种算法的效果基本差别不大。

此外，图 3 给出了四种算法在不同推荐列表个数的情况下 F1 值的变化，图 4 给出了五种算法在不同推荐列表个数的情况下覆盖率的变化。

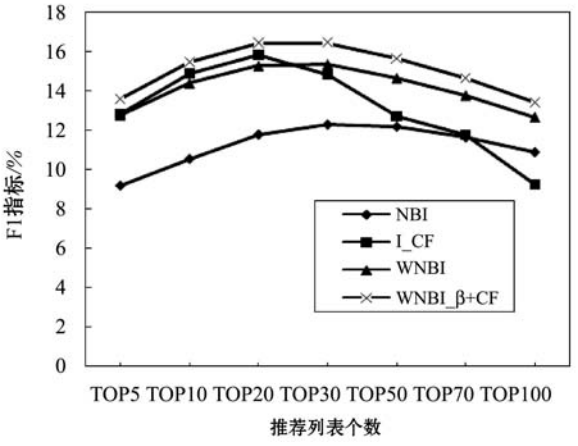


图 3　四种算法在不同推荐列表个数下的值

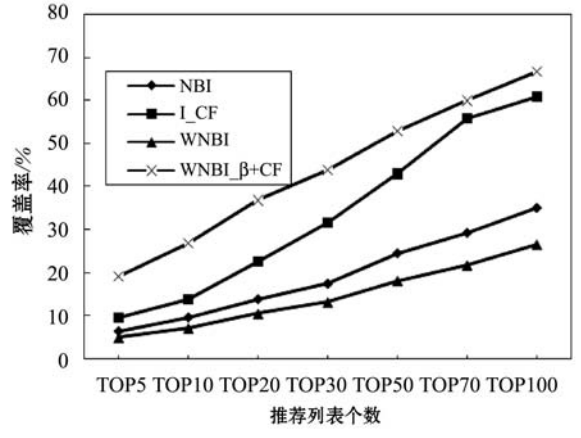


图 4　四种算法在不同推荐列表个数下的覆盖率

从表 2 以及图 3、图 4 中可以看出,融合初始资源配置以及协同过滤的二部图加权算法与其他三种算法相比,有更好的实验结果,表现出了比较好的推荐效果。

## 5　结　语

本文主要提出了一种组合推荐算法,在加权二部图的基础上加入可调参数 $\beta$ 和协同过滤算法进行融合,实现了一种融合初始资源配置以及基于物品的协同过滤的加权二部图算法。该方法考虑了初始资源的合理分配,降低了高度节点的影响,同时通过结合协同过滤算法与二部图算法,将二者的优缺点进行融合。通过实验表明,与一些传统的推荐算法相比,本算法具有较好的性能,提高了准确率、值、覆盖率等,其中覆盖率得到了明显提升。接下来,我们将进一步考虑用户这个因素的影响,通过考虑用户信任机制对推荐等的影响,以达到更好的推荐效果。

## 参　考　文　献

[1] Sarwar B, Karypis G, Konstan J, et al. Item-based collaborative filtering recommendation algorithms[C]//International Conference on World Wide Web. ACM, 2001:285-295.

[2] Wu Y, Xie M, Xie M, et al. CCCF: Improving Collaborative Filtering via Scalable User-Item Co-Clustering[C]//ACM International Conference on Web Search and Data Mining. ACM, 2016:73-82.

[3] Marko B. Balabanovic and Yoav Shoham 1997. Fab: Content-based, collaborative recommendation[C]//Communications of the ACM,2010.

[4] Bhagavatula C, Feldman S, Power R, et al. Content-Based Citation Recommendation[C]//The 16th Annual Conference of the North American Chapter of the Association for Computational(NAACL 2018), 2018.

[5] Hu X, Mai Z, Zhang H, et al. A Hybrid Recommendation Model Based on Weighted Bipartite Graph and Collaborative Filtering[C]//Ieee/wic/acm International Conference on Web Intelligence Workshops. IEEE, 2017:119-122.

[6] Aslanian E, Radmanesh M, Jalili M. Hybrid Recommender Systems based on Content Feature Relationship[J]. IEEE Transactions on Industrial Informatics, 2016(99):1.

[7] Wang X, Liu Y, Zhang G, et al. Mixed Similarity Diffusion for Recommendation on Bipartite Networks[J]. IEEE Access, 2017, 5(99):21029-21038.

[8] Zhang F, Liu Y, Xiong Q. A Novel Preferential Diffusion Recommendation Algorithm Based on User's Nearest Neighbors[J]. International Journal of Digital Multimedia Broadcasting,2017(4):1-7.

[9] Zhou T, Ren J, Medo M, et al. Bipartite network projection and personal recommendation[J]. Physical Review E Statistical Nonlinear & Soft Matter Physics, 2007, 76(4 Pt 2):046115.

[10] Liu J, Shang M, Chen D. Personal Recommendation Based on Weighted Bipartite Networks[C]//International Conference on Fuzzy Systems and Knowledge Discovery. IEEE, 2009:134-137.

[11] 张新猛, 蒋盛益. 基于加权二部图的个性化推荐算法[J]. 计算机应用, 2012, 32(3):654-657.

[12] Li D, Chen C, Lv Q, et al. An algorithm for efficient privacy-preserving item-based collaborative filtering[J]. Future Generation Computer Systems, 2016, 55(C):311-320.

[13] Shih T Y, Hou T C, Jiang J D, et al. Dynamically Integrating Item Exposure with Rating Prediction in Collaborative Filtering[C]//International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM, 2016:813-816.

[14] Zhou T, Jiang L L, Su R Q, et al. Effect of initial configuration on network-based recommendation[J]. Physics, 2008, 81(5):58004.