

利用社交关系的实值条件受限玻尔兹曼机协同过滤推荐算法

何洁月, 马贝

东南大学计算机科学与工程学院 东南大学计算机网络和信息集成教育部重点实验室 江苏南京 中国 210096

jieyuehe@seu.edu.cn, lihuating0910@126.com

摘要 利用受限玻尔兹曼机 (Restricted Boltzmann Machine, RBM) 解决推荐问题已成为一个很有意义的研究方向。目前用于推荐的 RBM 模型中使用的仅仅是用户评分数据, 但用户评分数据存在着严重的数据稀疏性问题。随着互联网对人们生活的不断渗透, 社交网络已经成为人们生活中不可缺少的一部分, 利用社交网络中的好友信任关系, 有助于缓解评分数据的数据稀疏性问题, 提高推荐系统的性能。因此, 本文提出基于实值的条件受限玻尔兹曼机 (Real-Valued Conditional Restricted Boltzmann Machine, R_CRBM) 模型, 此模型不需要将评分数据转化为向量单元, 将数据中潜在的评分/未评分信息应用到模型训练过程中; 并且将最近信任好友关系应用到 R_CRBM 模型推荐过程中。在百度数据集和 Epinions 数据集上的实验结果表明 R_CRBM 模型和引入的最近信任好友关系均有助于提高推荐系统的预测精度。

关键词 RBM; R_CRBM; 社交网络; 信任关系

Real-Valued Conditional Restricted Boltzmann Machine Based on Social Network for Collaborative Filtering

Jieyue He, Bei Ma

School of Computer Science and Engineering, MOE Key Laboratory of Computer Network and Information Integration

Southeast University, Nanjing, 210096, China

jieyuehe@seu.edu.cn, lihuating0910@126.com

Abstract Restricted Boltzmann Machine (RBM) for Collaborative filtering has become one of the significant researches. Currently, RBM model for Collaborative filtering only use user rating data. However, there are serious data sparseness problem in user rating data. As the Internet continues to penetrate on people's lives, social networks have become an indispensable part of life, friends trust relationships in social network can help alleviate the problem of sparse rating data and improve the performance of the recommendation system. Therefore, a Real-Valued Conditional Restricted Boltzmann Machine (R_CRBM) model is proposed in this paper. In the R_CRBM model, rating data does not need to be converted to vector data unit and rated/unrated

information is applied to the model training process. Moreover, nearest trusted relationships are applied to the R_CRBM model in recommended process. The experimental results on Baidu and Epinions datasets show that the R_CRBM model and nearest trusted relationships help to improve the prediction accuracy of the recommendation system.

Keywords RBM; R_CRBM; Social Network; Trust Relationships

1 绪论

随着互联网和信息技术的快速发展,微博、即时通讯、搜索引擎、电子商务、网络游戏等网络业务越来越普及,网络信息服务已经渗透到人们生活的各个方面,导致互联网用户的数量急剧增长。急剧增加的不仅仅是互联网用户的数量,还包括各种繁多的交易数据。

面对互联网上如此海量的商品,用户不得不浪费大量的时间来选择自己感兴趣的物品。基于此,推荐系统应运而生,出现了很多商用推荐系统,比如为用户推荐图书和其它商品的 Amazon,中国最大的电子商务平台淘宝网,电影推荐系统 MovieLens,文章推荐系统 GroupLens 等。

协同过滤推荐系统的应用最为广泛和成功。协同过滤算法分成两类^[1]:基于内存的协同过滤(Memory-based CF)、基于模型的协同过滤(Model-based CF)。基于内存的协同过滤,首先是计算用户(或项目)之间的相似度,然后是聚合最相似的若干用户(或项目)的评分进行预测。推荐过程中主要根据评分矩阵来进行,就好像是评分矩阵在内存中一样。基于模型的协同过滤是从已有的评分矩阵中学习出一个紧凑的模型,后续推荐中用这个模型进行预测。建立用户模型是该方法核心,目前常用的模型包括回归模型、贝叶斯模型、聚类模型、马尔可夫模型、隐语义模型、奇异值分解模型、受限玻尔兹曼机(Restricted Boltzmann Machine, RBM)模型^[2]等。其中 RBM 模型因其准确度较高近年来受到较大关注。

RBM(图3左)可以被视为一个无向图模型,它由两层二进制单元组成:一个可见层,表示数据,一个隐层,可视为特征提取器增加学习能力,并且层内无连接。RBM 模

型已经被实验证明是一种有效的解决推荐问题的方法^{[2][3]}。

2007 年 Salakhutdinov^[2]首次将 RBM 模型应用于解决推荐问题,作者对传统的 RBM 做了两点改变:首先,可见层用一个长度为 K 的向量单元表示评分数据;其次,用户可能只对若干个项目进行评分,对没有评分的项目使用一种特殊的(Missing)单元表示,这种单元不与任何隐层的单元连接。每一个用户都有一个单独的 RBM,这些 RBMs 对应一个共同的隐层。所有的 RBMs 之间的权重和偏置是共享的,所以如果两个用户对同一项目进行了评分,那么将会使用同一个权重。作者同时提出了条件受限玻尔兹曼机(Conditional RBM, 简记为 CRBM)模型,CRBM 训练过程中利用潜藏在评分数据中的评分/未评分信息,更加突出评分数据的作用。但是 Salakhutdinov^[2]所提模型的缺陷是需要将实值的评分数据转化为 0-1 数据,而且参数过多、训练过程复杂、模型训练时间较长。2013 年 KostadinGeorgiev^[3]提出了可直接处理实值评分数据的 RBM 模型并且改进了模型的训练过程,使 RBM 的可见单元可以直接表示实值,模型的训练和预测过程得到了简化,但是模型只利用评分数据,未能解决数据的稀疏性问题;此外,虽然作者改进的 RBM 的训练过程,提高了模型的推荐效果,但是使所有用户对同一项目的预测评分都相同,缺乏可解释性。

近年来,随着社交网络的流行,利用社交网络中的社交关系来提供推荐服务受到了越来越多学者的关注和研究。相对于传统的推荐系统,基于社交网络的推荐系统具有可靠性高、转化潜在需求为实际购买力强等特点。由于人们在社交网络中表达了很多隐含的兴趣、爱好等社交媒体信息,因此基于社交网络的推荐系统可以充分利用这些隐含的社交媒体信息,能够提高推荐系统的性能。

目前基于社交网络的推荐系统中一种常用的社交媒体信息是社会信任网络。

Golbeck^[4]假设用户精确提供了对社交网络中其他用户的信任评分，使用信任值取代相似性的查找，解决数据稀疏性问题。但用户提供对社交网络中所有用户的信任评分是不可能的，于是作者提出了一种 TidalTrust^[5]推测机制：以广度优先搜索方式推测与其他用户之间的间接信任值。Massa^{[6][7]}使用类似于 Golbeck^[4]的方法，但其推导间接用户间之间的信任值的主要思想是：考虑预先设定的距离范围内的所有用户，对所有到达用户的路径上的信任值进行加权和，该方法被称为 MoleTrust^{[6][7]}算法。

因此，本文借鉴 Georgiev^[3]提出的实值 RBM 的思想，对 Salakhutdinov^[2]提出的 CRBM 模型进行了改进，提出了 R_CRBM 模型。此模型不需要将评分数据转化为向量单元，避免了数据转化过程中信息的丢失，降低了模型的训练难度，训练过程中使用了潜藏的评分/未评分信息，以提高模型的推荐效果；其次，本文创新性的将最近信任好友的概念加入 R_CRBM 模型，提出了基于 MoleTrust^{[6][7]}推理的最近信任好友 R_CRBM 算法。在百度数据集和 Epinions 数据集上的实验结果表明 R_CRBM 模型以及基于 MoleTrust 推理的最近信任好友 R_CRBM 算法，均提高了推荐系统的推荐效果。

论文结构如下：第二节论述了 R_CRBM 模型和基于 MoleTrust 推理的最近信任好友 R_CRBM 算法，第三节给出了本模型和算法的实验结果及其分析，最后总结本文的工作并提出下一步的研究方向。

2 算法描述

本节首先论述本文提出的 R_CRBM 模型的原理以及模型中各参数的训练方法；然后论述了基于 MoleTrust 推理的最近信任好友 R_CRBM 算法

条件受限玻尔兹曼机 (CRBM^[2]) 模型，虽然能利用潜藏在评分数据中的评分/未评分数据信息，但实值的评分数据需转化为 0-1 数据，参数过多、训练过程复杂、模型训练

时间较长。为此，我们借鉴文献^[3]的思想，提出直接利用实值的 R_CRBM 模型，同时在模型中加入最近好友关系以进一步强化推荐的有效性。2.1 节和 2.2 节将分别介绍 R_CRBM 模型和基于 MoleTrust^{[6][7]}推理的最近信任好友 R_CRBM 算法。

2.1 Real-Valued Conditional RBM (R_CRBM)

在此，我们基于论文^[3]的思想提出 R_CRBM 模型，如图 1 所示。此模型将使其可见单元可直接表示实值的评分数据。

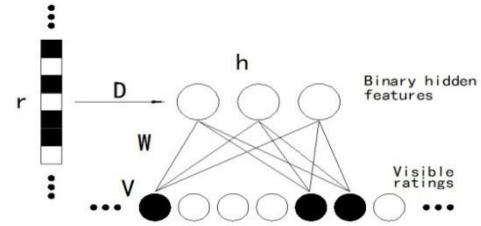


图 1: Real-Valued Conditional RBM。二进制向量 r 表示评分/未评分信息

R_CRBM 可以被视为一个无向图模型， v 为可见层，表示数据， h 为隐层，可视特征提取器， W 为可见层与隐层之间的连接权重， D 为 r 层和隐层之间的连接权重， c 表示可见单元的偏置， b 表示隐单元的偏置。在给定可见单元状态（输入数据）以及评分/未评分信息时，各隐单元的激活状态条件独立；反之，在给定隐单元状态时，可见层单元的激活状态亦条件独立，并且可通过 Gibbs 采样有效得到服从 R_CRBM 所表示分布的随机样本。

R_CRBM 考虑了评分/未评分这种潜藏信息，用 $r \in \{0,1\}^M$ 表示评分/未评分信息，其中 M 表示数据集中总的项目数，0 表示未对项目评分，1 表示已评过。由于将评分/未评分信息纳入考虑，因此向量 r 也将影响隐单元的状态（见图 1）。

R_CRBM 将潜藏在评分数据中的评分/未评分数据信息应用到模型的训练过程中，更加突出评分数据的作用。R_CRBM 的原理是：从隐单元的偏置中减去一部分放到权重矩阵 D 中，若用户对项目评分那么其不会对隐单元和可见单元产生任何影响。如果评分

缺失那么参数 D 中对应的缺失评分的权重将不会加到隐单元的偏置中，因此缺失评分将影响隐单元的特征提取。

根据 R_CRBM 的特殊结构层间全连接、层内无连接可知：可见层和隐层之间是相互独立的。给定可见单元状态时（包括评分/未评分信息），第 j 个隐单元的激活概率为：

$$P(h_j=1|v,r)=\sigma(b_j+\sum_i v_i W_{ij}+\sum_i r_i D_{ij}) \quad (1)$$

$$\text{其中, } \sigma(x)=\frac{1}{1+\exp(-x)}$$

当给定隐单元的状态时，第 i 个可见单元的激活概率为

$$P(v_i|h)=N(c_i+\sum_j h_j W_{ij},1) \quad (2)$$

2002 年, Hinton^[8]提出了 RBM 的快速学习算法，即对比散度（Contrastive Divergence, CD）算法。 R_CRBM 也采用该算法， R_CRBM 模型中的参数更新准则为

$$\Delta W_{ij}=\varepsilon(<v_i h_j>_{data}-<v_i h_j>_{recon})v_i>0 \quad (3)$$

$$\Delta c_i=\varepsilon(<v_i>_{data}-<v_i>_{recon})v_i>0 \quad (4)$$

$$\Delta b_i=\varepsilon(<h_j>_{data}-<h_j>_{recon}) \quad (5)$$

$$\Delta D_{ij}=\varepsilon(<h_j>_{data}-<h_j>_{recon})r_i \quad (6)$$

其中 $<.>$ 表示数学期望， $<.>_{data}$ 表示可见单元已知的情况下，隐层的概率分布， $<.>_{recon}$ 表示用 CD 算法重构后模型定义的分佈； ε 是学习率。

2.2 基于推理的最近信任好友 R_CRBM 算法

2.2.1 直接信任度计算

用户之间的直接信任关系可以用直接信任网络来表示。直接信任网络可以用一个有向图 $G=(U, E)$ 表示， U 是图中节点集合，每个节点代表一个用户， E 是网络中边的集合，每条边上的值表示朋友间的信任值，如图 2 所示。

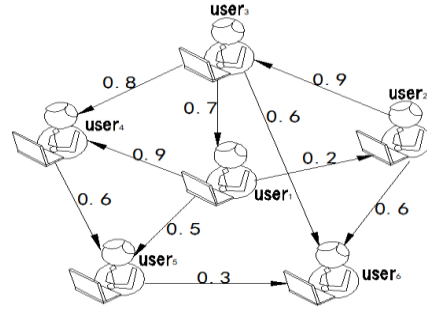


图 2: 6 个用户组成的直接信任网络

直接信任网络中的信任值一般可以表示成 $[0,1]$ 之间的实数，表达信任的程度，0 代表完全不信任，1 代表完全信任。然而，社交网络只是一个二值网络，0 代表不是好友，1 代表是好友。社交网络中好友关系大量存在，但是好友之间的信任值却无法获得。

本文的算法中使用 Pearson 系数计算社交网络中好友之间的直接信任值。Pearson 相关系数表示两个变量之间的关联性。用户 u 和用户 v 的 Pearson 相关系数，如公式 7 所示。

$$\text{sim}(u,v)=\frac{\sum_{c \in I_{u,v}} (R_{u,c}-\bar{R}_u)(R_{v,c}-\bar{R}_v)}{\sqrt{\sum_{c \in I_{u,v}} (R_{u,c}-\bar{R}_u)^2 * \sum_{c \in I_{u,v}} (R_{v,c}-\bar{R}_v)^2}} \quad (7)$$

其中， $I_{u,v}$ 表示用户 u 和 v 共同评分的项目集合， $R_{u,c}$ 表示用户 u 对项目 c 的评分，

$\bar{R}_u$ 和 $\bar{R}_v$ 分别表示训练集中用户 u 和 v 对项目的平均评分。利用 Pearson 系数计算社交网络中好友之间的直接信任值，基本思想是：首先用训练数据训练一个 R_CRBM 模型，模型训练好以后若两个用户是好友关系，那么对这两个用户的所有项目进行预测评分，然后使用 Pearson 系数计算两个用户的相似性，从而就获得了两个直接好友之间的信任值。

2.2.2 间接信任度计算

间接信任表示间接好友间的信任程度。在社交网络中，用户对其他所有用户提供信任评分是不可能的。 $MoleTrust^{[6][7]}$ 提供了一种推测机制，用好友之间的直接信任值推测间接好友的信任值，其主要思想是：考虑预先

设定的距离范围内的所有用户，对所有到达用户的路径上的信任值进行加权和。

2.2.3 基于 MoleTrust 推理的最近信任好友 R_CRBM 算法

所谓最近信任好友关系是指和用户是好友关系，同时两人之间的信任值大于阈值 0.6^[6]，并且该好友对用户要预测的项目已评分过，我们将这种好友关系称为最近信任好友关系（Nearest Trusted Friends, NTF）。

如图 3 所示，在图 3 左边表示标准的 RBM 的模型图，图 3 中右边表示我们提出的 R_CRBM 模型的结构图。在利用 R_CRBM 模型进行推荐时，每一个用户都有一个单独的 R_CRBM，这些 R_CRBM 对应一个共同的隐层，所有的 R_CRBM 之间的权重和偏置是共享的，所以如果两个用户对同一项目进行了评分，那么将会使用同一个权重。

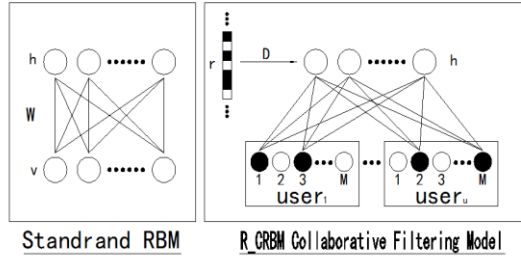


图 3： 左为标准 RBM，右为利用最近好友关系和 R_CRBM 模型进行推荐的模型，每个用户均有一个 R_CRBM 模型，共同评分的项目共享权重。

Ziegler 和 Golbeck^[9]研究了用户兴趣相似性与用户间信任的联系,结果表明两者之间存在着正相关性,即用户之间的信任度较高,则他们兴趣相似性也相对较高。如图 3(右),假设要预测用户 1 对第二个项目的评分,用户 u 是用户 1 的好友那么他们的兴趣也必然比较相似,恰好此时用户 u 对第二项目已评分过,因此,用户 u 是用户 1 关于第二个项目的最近信任好友。R_CRBM 模型训练好以后就可以利用用户 u 对第二个项目的预测评分来改善用户 1 对第二个项目的评分。

利用 MoleTrust 算法得到的信任网络,在信任网络中寻找最近信任好友,从而利用最近信任好友的预测评分来改善预测效果。基于此我们提出了基于 MoleTrust 推理的最近信任好友（Nearest Trusted Friends

MoleTrust, NTFMT）R_CRBM 算法，称其为 R_CRBM_NTFMT 算法。该算法在预测评分过程中考虑了均值因素的影响，因为即使两个用户的平均评分不同，但是利用 Pearson 系数求得的相似性也可能很高。R_CRBM_NTFMT 算法中用户 u 对项目 i 的预测评分如公式 8 所示。

$$\hat{R}_{u,i} = mean_{ui} + \frac{\sum_{k=1}^F (friends_{ki} - mean_{ui}) * trust_value_{uk}}{\sum_{k=1}^F trust_value_{uk}} \quad (8)$$

其中

$$mean_{ui} = \frac{user_mean_u + item_mean_i}{2} \quad (9)$$

$$mean_{ki} = \frac{user_mean_k + item_mean_i}{2} \quad (10)$$

$user_mean_u$ 表示训练集中用户 u 的平均评分； $item_mean_i$ 为训练集中项目 i 的平均评分； $mean_{ui}$ 表示用户 u 的平均评分和项目 i 的平均评分的平均，综合考虑用户平均评分和项目平均评分的影响； $mean_{ki}$ 表示用户 k 的平均评分和项目 i 的平均评分的平均；F 为最近信任好友的数目； $friends_{ki}$ 表示最新信任好友 k 的 R_CRBM 模型对项目 i 的预测评分； $trust_value_{uk}$ 表示用户 u 和 k 之间的信任值(大于阈值 0.6^[6])。

R_CRBM_NTFMT 算法的基本思想是：首先用训练数据训练一个 R_CRBM 模型，然后在用户的信任网络中寻找最近信任好友关系，最后利用最近信任好友的预测评分来改善用户对项目的预测评分。其伪代码如表 1 和表 2 所示。

R_CRBM_NTFMT 算法的输入是训练数据集、好友社交关系网络和距离参数 Distance 的值，输出是预测评分。而 ConstructTrustNet 函数的功能是用好友社交关系网络构建信任网络，其输入为训练好的 R_CRBM 模型、好友社交关系网络以及距离 Distance 的值，输出是信任网络 Trust_net。

表 1： R_CRBM_NTFMT 算法的伪代码

R_CRBM_NTFMT Algorithm

Input: Train dataset, Social_network, Distance

Output: Predict data

```

//Trust_net is trust network
// For user u ,he need to predict item i, F is the numbers
of his most trusted friends about item j
1.Train a R_CRBM model
2.Use ConstructTrustNetfunction to construct
   Trust_net
3.For all ratings which need to be predicted do:
4. Searching user u nearest trusted friends about item i;
5.  $mean_{ui} = \frac{user\_mean_u + item\_mean_i}{2}$ 
6.  $mean_{ki} = \frac{user\_mean_k + item\_mean_i}{2}$ 
7.  $\hat{R}_{u,i} = mean_{ui} + \frac{\sum_{k=1}^F (friends_{ki} - mean_{ki}) * trust\_value_{uk}}{\sum_{k=1}^F trust\_value_{uk}}$ 
8. End for

```

表 2: ConstructTrustNet 函数的伪代码

```

ConstructTrustNetFunction
Input: R_CRBM model,Social_network ,Distance
Output: Trust_net
//Trust_net is trust network
// Trust_net_direct is directly trust network
1.For all friendship(  $u_1, u_2$  ) in Social_network do:
2. predict scores for all items by R_CRBM model;
3. compute trustscore of (  $u_1, u_2$  ) by Pearson;
4. add the trustscore of (  $u_1, u_2$  ) into
   Trust_net_direct
5. End for
// Use Trust_net_direct to construct Trust_net
6.Use Moletrust[6][7]Algorithm to construct Trust_net of
   all users
7.End

```

从表 1 可看出 R_CRBM_NTFMT 算法实现推荐的过程主要可划分为三步:

第 1 行~第 1 行: 用训练数据训练一个 R_CRBM 模型;

第 2 行~第 2 行: 调用 ConstructTrustNet 函数, 获得信任网络;

第 3 行~结尾: 首先, 在信任网络中寻找用户 u 关于项目 i 的最近信任好友; 然后, 根据公式 (8) - (10) 计算 R_CRBM_NTFMT 算法对相应预测项目的预测评分, 即用最近信任好友的预测评分来改善用户对项目的预测评分。

表 2 是 ConstructTrustNet 函数的伪代码, 该函数的功能是用好友社交关系网络构建信任网络, 主要可划分为两步:

第 1 行~第 5 行: 若 (u_1, u_2) 是社交网络中的好友关系, 用训练好的 R_CRBM 模型预测 u_1, u_2 对所有项目的评分; 然后用 Pearson 系数计算 u_1, u_2 的信任值, 从而构建了好友之间的直接信任网络;

第 6 行~结尾: 用直接信任网络, 通过 Moletrust 算法推理得到用户之间的信任网络。

3 实验结果及分析

本节通过实验验证我们所提算法的性能, 实验数据采用百度推荐大赛数据集和 Epinions 数据集, 其中 80% 的数据作为训练数据 20% 的数据作为测试数据, MoleTrust 算法^{[6][7]}和 RBM^[3]模型(包括论文^[3]改进的训练过程)作为对比结果。论文^[3]改进了标准 CD 算法的训练过程, 改进思想是: 可见单元的值等于对应隐单元连接权重的和 (sum weight) 再加上偏置, 我们将此时的 RBM 模型我们称其为 S_RBM。

3.1 数据集

百度推荐大赛的数据集, 可从 <http://www.datatang.com/data/44268> 下载。首先对数据集进行了预处理, 将评分数据中没有好友关系的用户的数据剔除, 得到一个 3193 个用户对 7889 部电影的评分数据, 包括其好友关系网络。

Epinions 数据集由 Paolo Massa^[10]从 Epinions.com 网站上收集得到。我们从该数据集中抽取了一个 1963 名用户对 2436 个项目的评分数据集作为实验数据, 包括其好友关系。

3.2 评价指标

目前，绝大多数的推荐系统都使用预测准确度来评价推荐算法的性能，预测准确度是比较推荐算法的预测评分与用户实际评分的相似程度。预测准确度的常用度量方法有平均绝对误差(MAE^{[11][12]})、根均方误差(RMSE^[13])。

平均绝对误差计算预测评分与用户实际评分之间的平均绝对误差值。它的计算公式如公式 11 所示。

$$MAE = \frac{\sum_{(u,i) \in R_{test}} |R_{u,i} - \hat{R}_{u,i}|}{|R_{test}|} \quad (11)$$

其中， R_{test} 表示测试集数据， $R_{u,i}$ 表示

用户 u 对项目 i 的实际评分， $\hat{R}_{u,i}$ 是用户 u 对项目 i 的预测评分， $|R_{test}|$ 是测试集中数据的个数。推荐算法的准确度是所有用户预测评分与用户实际评分之差的平均。

根均方误差计算用户实际评分与预测评分之间的根均方误差。它的计算公式如公式 12 所示。

$$RMSE = \sqrt{\frac{\sum_{(u,i) \in R_{test}} (R_{u,i} - \hat{R}_{u,i})^2}{|R_{test}|}} \quad (12)$$

其中， R_{test} 表示测试集数据， $R_{u,i}$ 表示

用户 u 对项目 i 的实际评分， $\hat{R}_{u,i}$ 是用户 u 对项目 i 的预测评分， $|R_{test}|$ 是测试集中数据的个数。根均方误差在求和之前对系统预测评分与用户实际评分的误差进行平方，因此评分之间的误差越大，其对根均方误差的影响会比平均绝对误差更大。

3.3 百度数据集实验结果及分析

3.3.1 MoleTrust 算法中参数 Distance 值的确定

MoleTrust 算法考虑预先设定的距离 (Distance) 范围内的所有用户。为了计算用户 u 对用户 v 的信任值，前节点用户的信任值以信任边为权值进行加权和，即构建了一个以用户 u 为核心的距离 Distance 范围内的

信任网络。图 4 和图 5 验证 MoleTrust 算法中距离 Distance 对预测结果的影响。

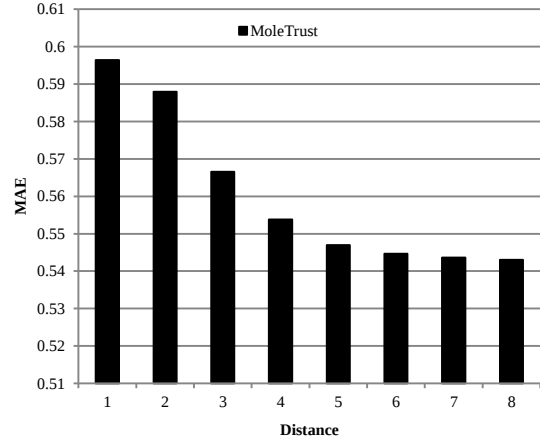


图 4: MoleTrust 算法中参数 Distance 值对 MAE 值的影响

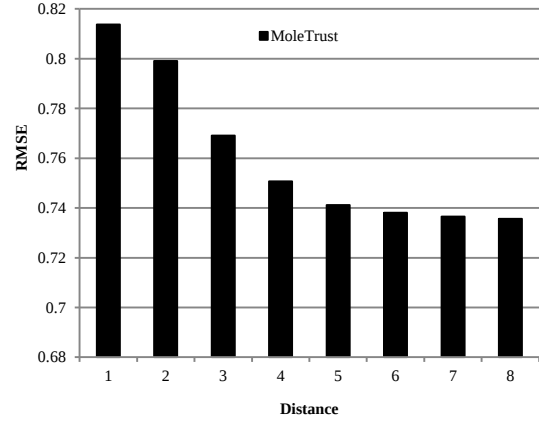


图 5: MoleTrust 算法中参数 Distance 值对 RMSE 值的影响

图 4 坐标系中，横坐标表示最大距离 Distance 的值，纵坐标为评价指标 MAE 的值。图 5 坐标系中，横坐标表示距离 Distance 的值，纵坐标为 RMSE 的值。从两张图中可明显看出当 Distance 达到 6 以后，MAE 和 RMSE 基本就不再减少。因此在下面的实验中我们使用 MoleTrust 算法中 Distance 为 6 所对应的信任网络，并且将 Distance 为 6 所对应 MoleTrust 算法的预测结果作为一个对比结果。

3.3.2 RBM 模型中隐单元数目的确定

本实验的目的是考察 RBM^[3]模型中隐单元数目对推荐结果的影响，确定隐单元的数目使 RBM 模型的推荐结果最优，实验是基于训练次数 Epochs 为 10 进行的，实验结

果如图 6 和图 7 所示。

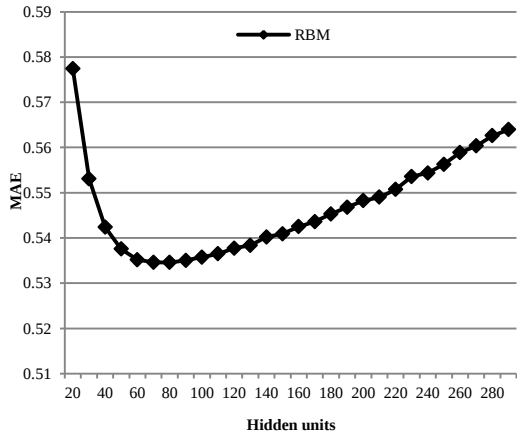


图 6: RBM 模型中隐单元数目对 MAE 值的影响

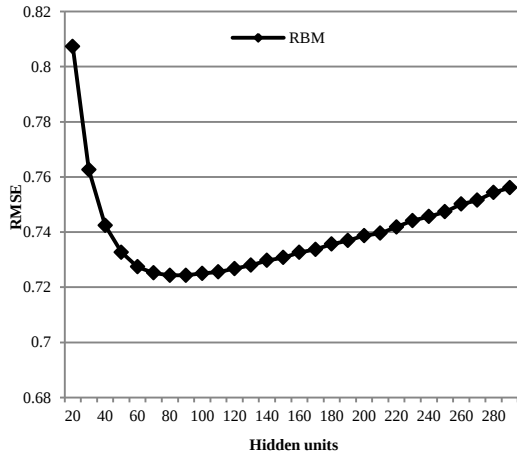


图 7: RBM 模型中隐单元数目对 RMSE 值的影响

图 6 坐标系中，横坐标表示 RBM 模型中隐单元数目，纵坐标为评价指标 MAE 的值。图 7 坐标系中，横坐标表示 RBM 模型中隐单元数目，纵坐标为 RMSE 的值。从图 6 和图 7 可发现随着隐单元数目的增加，MAE 与 RMSE 值先降后升，说明模型的效果呈现了先变好后变差的趋势。隐单元用于提取数据的特征，当模型提取的特征较少时这些特征不足以表达数据的特征，此时增加提取的特征数，模型的效果会逐渐变好，但是当模型提取的特征达到一定程度后，此时增加提取的特征会导致特征过多反而影响模型的效果。从图 6 中可发现隐单元数为 80 时对应的 MAE 值最小，图 7 中隐单元数为 90 时对应的 RMSE 值最小，综合考虑隐单元为 80 和 90 时的 MAE、RMSE 值，隐单元数为 80 时效果更好，因此下面的实验中 RBM、

S_RBM 和 R_CRBM 模型的隐单元数目都为 80。

3.3.3 训练数据参与训练次数 Epochs 对推荐结果的影响

本实验中使用 MoleTrust 算法中 Distance 为 6 所对应的信任网络，RBM、S_RBM 和 R_CRBM 模型的隐单元数目均为 80。实验结果如图 8 和图 9 所示。

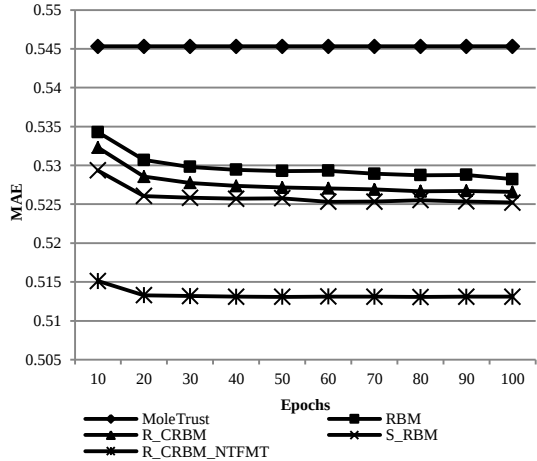


图 8: 百度数据集上各方法的 MAE 值

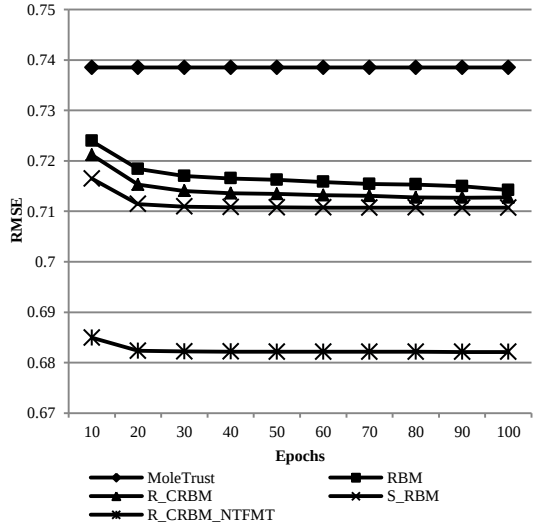


图 9: 百度数据集上各方法的 RMSE 值

图 8 坐标系中，横坐标表示训练集参与训练的次数 Epochs 的值，纵坐标为评价指标 MAE 的值；图 9 坐标系中，横坐标表示训练集参与训练的次数 Epochs 的值，纵坐标为评价指标 RMSE 的值。两图中 MoleTrust 表示 MoleTrust 算法中 Distance 为 6 所对应的预测结果。从两图中可发现 R_CRBM 模型的实验

结果均略优于 RBM 模型，说明潜藏的评分/未评分信息可以改善预测效果。S_RBM 推荐效果比 RBM 有了一定程度的提高，但是其缺点是缺乏可解释性。从两图中可看出我们提出的 R_CRBM_NTFMT 算法是一种比较稳定的算法，其 MAE 和 RMSE 基本是稳定的，受 Epochs 的影响不大，并且达到了较好的效果。

3.4 Epinions 数据集实验结果及分析

参数 Distance 和隐单元数目的确定与百度数据集类似，这里不再展示。实验中使用 MoleTrust 算法中 Distance 为 3 所对应的信任网络，并且将 Distance 为 3 所对应 MoleTrust 算法的预测结果作为一个对比结果，RBM、S_RBM 和 R_CRBM 模型的隐单元数目均为 80。实验结果如图 10 和图 11 所示。

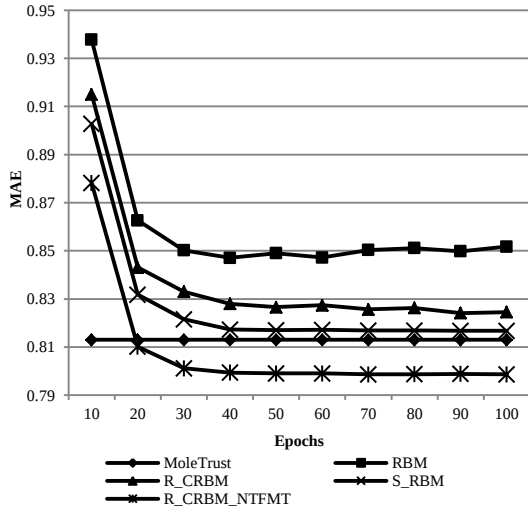


图 10: Epinions 数据集上各方法的 MAE 值

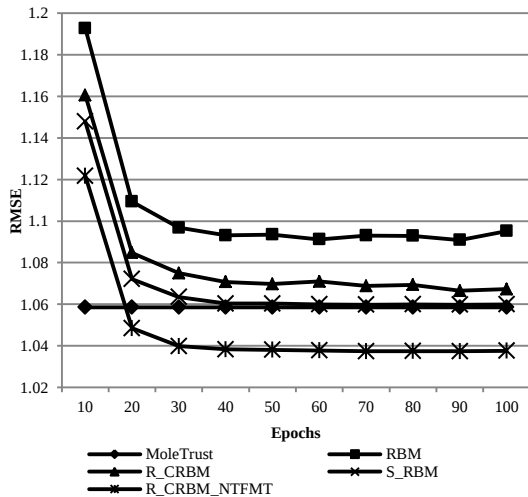


图 11: Epinions 数据集上各方法的 RMSE 值

图 10 坐标系中，横坐标表示训练集参与训练的次数 Epochs 的值，纵坐标为评价指标 MAE 的值；图 11 坐标系中，横坐标表示训练集参与训练的次数 Epochs 的值，纵坐标为评价指标 RMSE 的值。两图中 MoleTrust 表示 MoleTrust 算法中 Distance 为 3 所对应的预测结果，可发现 MoleTrust 算法在 Epinions 数据集上取得了较好的推荐效果。S_RBM、R_CRBM 模型的在此数据集上的实验结果仍然优于 RBM 模型。我们提出的 R_CRBM_NTFMT 算法取得了最好的预测效果。

综合对比两个数据集上的实验结果，可发现我们的方法均比改进前取得了更好的推荐效果。至此我们得出结论，我们算法中使用的最近信任好友关系是值得信赖的，并且利用除评分数据外的潜在信息有助于改善推荐效果。

4 总结及展望

当前，以 RBM 为基本模块的深度置信网模型被认为是最有效的深度学习算法，也使其在深度学习领域中占据着核心位置，RBM 目前已被应用于多种机器学习问题，协同过滤便是其中之一。目前的研究中 RBM 使用的还仅仅是用户的评分数据，众所周知推荐领域中存在严重的数据稀疏性问题，以我们实验所采用的 baidu 推荐大赛的数据集为例其超过 97.4% 的数据是缺失的，这将大大影响 RBM 模型的训练效果。基于此，本文首先提出了 R_CRBM 模型，该模型直接利用评分数据避免数据转化过程中信息的丢失，利用评分数据中潜藏的评分/未评分信息，缓解数据稀疏性问题；其次，利用社交关系中的好友关系，使模型综合了评分数据的数据特征和社交关系特征来进行推荐，增强了模型的表达能力。实验结果验证了这些潜在信息均有助于改善系统的推荐效果。

本论文中采用了论文^[3]中的训练方法（包括其所作的改进），本质上都是 CD 算法，但在深度学习领域，一些研究者在 CD 算法的基础上，已经对其作了一系列的改进，例

如，Tieleman^[14]提出了持续对比散度（Persistent Contrastive Divergence, PCD）算法，Tieleman 和 Hinton^[15]进一步改进了 PCD 算法，引入一组辅助参数以加快马氏链的混合率，提出了快速持续对比散度（Fast Persistent Contrastive Divergence, FPCD）算法，Desjardins^[16]等人提出了 Parallel Tempering（PT）算法等等，但是这些方法的应用领域都是传统的 0-1 数据，因此如何改进这些方法使其适用于像推荐这种数据是实值并且数据大量缺失的应用领域，将是我们未来的研究工作。

参考文献

[1]Breese J S, Heckerman D, Kadie C. Empirical analysis of predictive algorithms for collaborative filtering[C]//Proceedings of the Fourteenth conference on Uncertainty in artificial intelligence. Morgan Kaufmann Publishers Inc., 1998: 43-52.

[2] Salakhutdinov R, Mnih A, Hinton G. Restricted Boltzmann machines for collaborative filtering[C]//Proceedings of the 24th international conference on Machine learning. ACM, 2007: 791-798.

[3]Georgiev K, Nakov P. A non-iid framework for collaborative filtering with restricted Boltzmann machines[C]//Proceedings of the 30th International Conference on Machine Learning (ICML-13). 2013: 1148-1156.

[4]Golbeck J. Generating predictive movie recommendations from trust in social networks[M]. Springer Berlin Heidelberg, 2006.

[5]Golbeck J, Hendler J. Filmtrust: Movie recommendations using trust in web-based social networks[C]//Proceedings of the IEEE Consumer communications and networking conference. University of Maryland, 2006, 96.

[6] Massa P, Avesani P. Controversial users demand local trust metrics: An experimental study on epinions. com community[C]//Proceedings of the National Conference on artificial Intelligence. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999, 2005, 20(1): 121.

[7]Massa P, Avesani P. Trust metrics on controversial users: Balancing between tyranny of the majority[J]. International Journal on Semantic Web and Information Systems (IJSWIS), 2007, 3(1): 39-64.

[8]Welling M, Hinton G E. A new learning algorithm for mean field Boltzmann machines[M]//Artificial Neural

Networks—ICANN 2002. Springer Berlin Heidelberg, 2002: 351-357.

[9]Ziegler C N. Investigating correlations of trust and interest similarity-do birds of a feather really flock together?[J]. submitted to the Journal of Artificial Intelligence Research, 2005.

[10]Massa P, Avesani P. Trust-aware bootstrapping of recommender systems[C]//ECAI Workshop on Recommender Systems. 2006: 29-33.

[11]Breese J S, Heckerman D, Kadie C. Empirical analysis of predictive algorithms for collaborative filtering[C]//Proceedings of the Fourteenth conference on Uncertainty in artificial intelligence. Morgan Kaufmann Publishers Inc., 1998: 43-52.

[12]Herlocker J L, Konstan J A, Borchers A, et al. An algorithmic framework for performing collaborative filtering[C]//Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 1999: 230-237.

[13]Shardanand U, Maes P. Social information filtering: algorithms for automating “word of mouth”[C]//Proceedings of the SIGCHI conference on Human factors in computing systems. ACM Press/Addison-Wesley Publishing Co., 1995: 210-217.

[14]Tieleman T. Training restricted Boltzmann machines using approximations to the likelihood gradient[C]//Proceedings of the 25th international conference on Machine learning. ACM, 2008: 1064-1071.

[15]Tieleman T, Hinton G. Using fast weights to improve persistent contrastive divergence[C]//Proceedings of the 26th Annual International Conference on Machine Learning. ACM, 2009: 1033-1040.

[16] Desjardins G, Courville A C, Bengio Y, et al. Tempered Markov chain Monte Carlo for training of restricted Boltzmann machines[C]//International Conference on Artificial Intelligence and Statistics. 2010: 145-152.