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 南京 210096)
(东南大学计算机网络和信息集成教育部重点实验室 南京 210096)

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稀疏性问题,提高推荐系统的性能.因此,该文首先提出基于实值的状态玻尔兹曼机(Real-Valued Conditional Restricted Boltzmann Machine,R_CRBM)模型,此模型不需要将评分数据转化为一个K维的0-1向量,并且R_CRBM模型在训练过程中使用了训练数据中潜在的评分/未评分信息;同时该文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;最后,针对大数据环境下,普通平台很难完成R_CRBM模型训练的问题,该文提出基于Spark的并行化方案,较好地解决了该问题.

关键词 受限玻尔兹曼机;数据稀疏性;R_CRBM;社交网络;信任关系;大数据
中图法分类号 TP393 **DOI号** 10.11897/SP.J.1016.2016.00183

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

HE Jie-Yue MA Bei
(*School of Computer Science and Engineering, Southeast University, Nanjing* 210096)
(*MOE Key Laboratory of Computer Network and Information Integration, Southeast University, Nanjing* 210096)

Abstract Restricted Boltzmann Machine (RBM) for Collaborative filtering has become one of the significant researches. Currently, RBM model for Collaborative filtering only use users rating data. However, there are serious data sparsity in users 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 a K dimensional 0-1 vector unit. Meanwhile, the training process of R_CRBM model also uses rated/unrated information. Moreover, the nearest trusted relationships are applied to the R_CRBM model in the recommended process. The experimental results from Baidu and Epinions datasets show that the R_CRBM model and the nearest trusted relationships help to improve the prediction accuracy of the recommendation system. Finally, due to a common platform is very difficult to train R_CRBM model in big data. Therefore, a parallelization scheme based on Spark is also proposed in this paper. The experimental result shows that the parallelization method for R_CRBM is a good solution for this problem.

Keywords restricted Boltzmann machine; data sparsity; real-valued conditional restricted Boltzmann machine; social network; trust relationships; big data

收稿日期:2015-01-18;**在线出版日期**:2015-05-11.**本课题得到江苏省自然科学基金**(BK2012742)**和软件新技术与产业化协同创新中心部分资助**.**何洁月**,**女**,1964**年生**,**博士**,**教授**,**中国计算机学会**(CCF)**会员**,**主要研究领域为数据密集型计算、生物信息学、数据挖掘和机器学习等**. E-mail: jieyuehe@seu.edu.cn.**马 贝**,**男**,1990**年生**,**硕士**,**主要研究方向为深度学习、协同过滤推荐**.

1 引 言

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商品.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,评分矩阵就好像是在内存中一样.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的核心,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(a))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[3],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[2,4].

2007 年 Salakhutdinov 等人[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 0-1 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,CRBM)模型,CRBM 训练过程中利用潜藏在评分数据中的评分/未评分数息,更加突出评分数据的作用.但是 Salakhutdinov 等人[2]所提模型的缺陷是:需要将实值的评分数据转化为一个 K 维的 0-1 向量,可见层与隐藏层之间的连接权重变为 $M\times K\times S$ 维,维数变为了原来的 K 倍,从而导致参数过多、训练过程复杂、模型训练时间较长;而且该模型只能将整型的评分数据转化为一个 K 维的 0-1 向量,如果训练数据中的评分是 Double 型的就无法转化.2013 年 Georgiev 等人[4]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,虽然作者改进了 RBM 的训练过程,提高了模型的推荐效果,但是此模型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、爱好等社会媒体信息,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会媒体信息.目前基于社交网络的推荐系统中一种常用的社会媒体信息是社会信任网络.

Golbeck[5]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[6]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 等人[7-8]使用类似于 Golbeck[5]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[7-8]算法.

Ma 等人[9-10]提出了一种基于概率矩阵分解的因子分析方法,该方法利用用户的评分信息和社交网络信息,可以很好的解决推荐系统数据稀疏性和

预测精度低的问题. Huang 等人[11]研究了口碑推荐的后影响，发现口碑推荐可以提高用户对项目的后评价. 这些方法均很好地利用了社交网络信息，提高了推荐系统的预测效果.

本文借鉴 Georgiev 等人[4]提出的实值 RBM 的思想，对 Salakhutdinov 等人[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一个 K 维的 0-1 向量，而且对训练数据的类型没有要求，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，本文创新性地将最近信任好友的概念加入 R_CRBM 模型，提出了基于 MoleTrust 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.

本文第 2 节论述 R_CRBM 模型和基于 MoleTrust 推理的最近信任好友 R_CRBM 算法；第 3 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一个 K 维的 0-1 向量，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 推理的最近信任好友 R_CRBM 算法.

2.1　Real-Valued Conditional RBM(R_CRBM)模型

在此，我们基于文献[4]的思想提出 R_CRBM 模型，如图 1 所示. 此模型中可见单元可直接表示实值的评分数据.

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，可视为特征提取器；W 为可见层与隐层之间的连接权重矩阵；$\boldsymbol{D}$ 为 r 层和隐层之间的连接权重矩阵；c 表示可见单元的偏置；

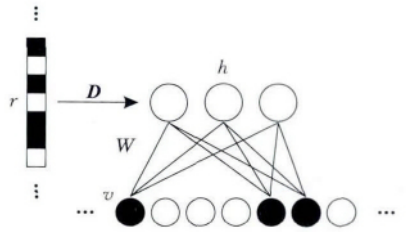


图 1　Real-Valued Conditional RBM
(二进制向量 r 表示评分/未评分信息)

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 表示已评过分. 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权重放到权重矩阵 $\boldsymbol{D}$ 中，因为权重 $\boldsymbol{D}$ 是 r 层和隐单元之间的连接权重(见图 1)而 r 表示评分/未评分信息，因此若用户对项目评分，那么从隐单元的偏置中减去的放到 $\boldsymbol{D}$ 中的权重将加回隐单元的偏置中(见式(1))，所以若用户对项目评分，那么从隐单元的偏置中减去的放到 $\boldsymbol{D}$ 中的权重不会对隐单元或可见单元产生任何影响. 但是如果评分缺失，那么从隐单元的偏置中减去的放到 $\boldsymbol{D}$ 中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 层间全连接、层内无连接的特殊结构可知：可见层和隐层之间是相互独立的. 当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

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

其中，$\sigma(x)=1/1+\exp(-x)$.

当给定隐单元的状态时，第 i 个可见单元的值为[12]

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

2002 年，Welling 和 Hinton[13]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

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

$$\Delta c_i = \varepsilon(\langle v_i \rangle_{\text{data}} - \langle v_i \rangle_{\text{recon}}) v_i > 0 \quad (4)$$

$$\Delta b_i = \varepsilon(\langle h_j \rangle_{\text{data}} - \langle h_j \rangle_{\text{recon}}) \quad (5)$$

$$\Delta D_{ij} = \varepsilon(\langle h_j \rangle_{\text{data}} - \langle h_j \rangle_{\text{recon}}) r_i \quad (6)$$

其中:$\langle \cdot \rangle$表示数学期望;$\langle \cdot \rangle_{\text{data}}$表示可见单元已知的情况下,隐层的概率分布;$\langle \cdot \rangle_{\text{recon}}$表示用 CD 算法重构后模型定义的分布;$\varepsilon$ 是学习率;$r \in \{0,1\}^M$是一个长度等于评分矩阵中项目数的 0-1 向量,用于指示项目是否被用户评分,即评分/未评分信息. 训练 R_CRBM 模型的伪代码见算法 1.

算法 1. 基于 R_CRBM 模型的协同过滤推荐算法(其中 CD 的步长为 1)的伪代码.

输入:训练数据集,评分/未评分数据

输出:训练好的 R_CRBM 模型

1. FOR $t=1$: $NumberEpochs$ DO:
2. FOR $n=1$: $NumberDataSamples$ DO:
3. Positive Phase:

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

4. Negative Phase:

①: $$P(v_i \mid h) = N\left(c_i + \sum_j h_j W_{ij}, 1\right)$$

或者②:可见单元的值等于该单元和所有隐单元的连接权重的和再加上该可见单元的偏置

5. Update Phase:

$$\Delta W_{ij} = \varepsilon(\langle v_i h_j \rangle_{\text{data}} - \langle v_i h_j \rangle_{\text{recon}}) v_i > 0$$

$$\Delta c_i = \varepsilon(\langle v_i \rangle_{\text{data}} - \langle v_i \rangle_{\text{recon}}) v_i > 0$$

$$\Delta b_i = \varepsilon(\langle h_j \rangle_{\text{data}} - \langle h_j \rangle_{\text{recon}})$$

$$\Delta D_{ij} = \varepsilon(\langle h_j \rangle_{\text{data}} - \langle h_j \rangle_{\text{recon}}) r_i$$

6. END FOR
7. END FOR

R_CRBM 模型的训练过程主要分为 3 个阶段:即 Positive Phase、Negative Phase 和 Update Phase.

第 1 行:$NumberEpochs$ 为训练数据集参与训练的次数;

第 2 行:$NumberDataSamples$ 表示训练数据集的行数(或者列数),即将训练数据集中的每行数据(或者每列数据)依次输入模型参与模型的训练;

第 3 行:Positive Phase 阶段,已知可见单元状态时(包括评分/未评分信息),求隐单元的激活概率;

第 4 行:Negative Phase 阶段,已知隐单元的状态时,求可见单元的值. 在该阶段求可见单元的值有两种方法:第 1 种是 $P(v_i \mid h) = N\left(c_i + \sum_j h_j W_{ij}, 1\right)$;第 2 种方法是可见单元的值等于该单元和所有隐单元的连接权重的和再加上该可见单元的偏置.

第 5 行:Update Phase 阶段,该阶段更新模型的所有参数. 值得注意的是计算 ΔW 和 Δc 时训练数据中该可见单元的值应大于 0(即用户对该项目评分).

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.

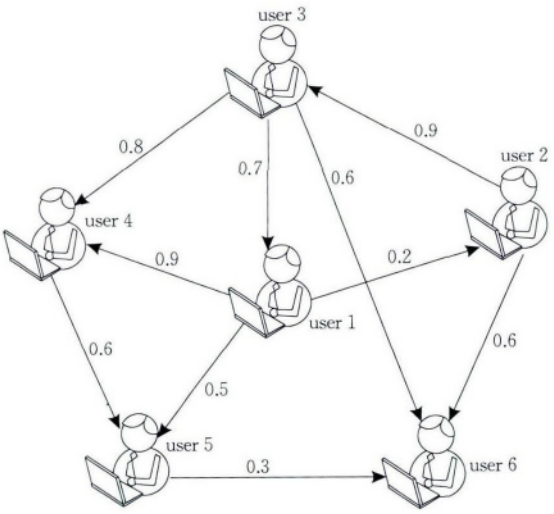


图 2 6 个用户组成的直接信任网络

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,如式(7)所示.

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

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.

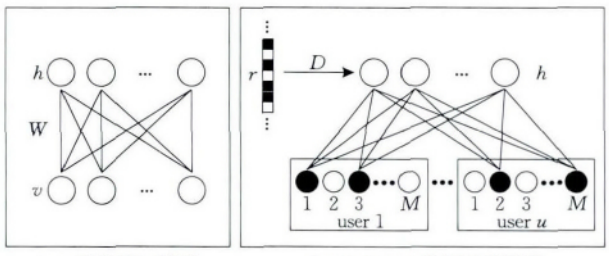
2.2.2　间接信任度计算

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 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(a)表示标准的 RBM 的模型图，图 3(b)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_CRBMs 对应一个共同的隐层，所有的 R_CRBMs 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.



(a) 标准RBM模型　　(b) R_CRBM协同过滤模型

图 3　模型结构图

Ziegler 和 Golbeck[14] 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(b)，假设要预测用户 1 对第 2 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 2 项目已评分过，因此，用户 u 是用户 1 关于第 2 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 2 个项目的预测评分来改善用户 1 对第 2 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 Based on 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 的预测评分如式(8)所示.

$$\hat{R}_{u,i}=mean_{ui}+\frac{\sum_{k=1}^{F}(friends_{ki}-mean_{ki})\times trust\_value_{uk}}{\sum_{k=1}^{F}trust\_value_{uk}} \tag{8}$$

其中

$$mean_{ui}=\frac{user\_mean_u+item\_mean_i}{2} \tag{9}$$

$$mean_{ki}=\frac{user\_mean_k+item\_mean_i}{2} \tag{10}$$

其中：$user\_mean_u$ 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；$friends_{ki}$ 表示最近信任好友 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[7]).

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. 其伪代码如算法 2 和算法 3 所示.

算法 2 的 R_CRBM_NTFMT 算法的输入是训练数据集、评分/未评分数据、好友社交关系网络和距离参数 $Distance$ 的值，输出是预测评分. 而算法 3 的 $ConstructTrustNet$ 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 $Trust\_net$.

算法 2. R_CRBM_NTFMT 算法的伪代码.

输入：训练数据集、评分/未评分数据、好友社交关系网络和距离参数 *Distance* 的值

输出：预测评分

//*Trust_net* 表示信任网络

//F 是用户 u 关于项目 i 的最近信任好友的数目

1. 训练一个 R_CRBM 模型
2. 调用 *ConstructTrustNet* 函数构建 *Trust_net*
3. FOR all ratings which need to be predicted DO:
4. 寻找用户 u 关于项目 i 的所有最近信任好友;
5. $mean_{ui}=\dfrac{user\_mean_u+item\_mean_i}{2}$
6. $mean_{ki}=\dfrac{user\_mean_k+item\_mean_i}{2}$
7. $\hat{R}_{u,i}=mean_{ui}+\dfrac{\sum_{k=1}^{F}(friends_{ki}-mean_{ki})\times trust\_value_{uk}}{\sum_{k=1}^{F}trust\_value_{uk}}$
8. END FOR

算法 3. *ConstructTrustNet* 函数的伪代码.

输入：R_CRBM 模型、好友社交关系网络和距离参数 *Distance* 的值

输出：*Trust_net*

//*Trust_net* 表示信任网络

//*Trust_net_direct* 表示直接信任网络

1. FOR all friendship(u_1, u_2) in Social_network DO:
2. 使用 R_CRBM 模型预测用户 u_1, u_2 对所有项目的评分;
3. 使用 Pearson 相关系数计算 u_1 和 u_2 之间的信任值;
4. 将 u_1 和 u_2 之间的信任值加入 *Trust_net_direct*
5. END FOR

//使用 *Trust_net_direct* 构建 *Trust_net*

6. 使用 MoleTrust 算法推理得到 *Trust_net*
7. END

从算法 2 可看出 R_CRBM_NTFMT 算法实现推荐的过程主要可划分为 3 步：

第 1 行：用训练数据训练一个 R_CRBM 模型；

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.

算法 3 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 算法和 RBM[4] 模型(包括文献[4]改进的训练过程)作为对比结果. 文献[4]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 Massa 等人[15] 从 Epinions.com 网站上收集得到，包含 40 163 个用户对 139 529 个项目的评分. 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[16-17])、根均方误差($RMSE$[18]).

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. 它的计算公式如式(11)所示.

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

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；$|R_{\text{test}}|$ 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. 它的计算公式如式(12)所示.

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

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；$|R_{\text{test}}|$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.

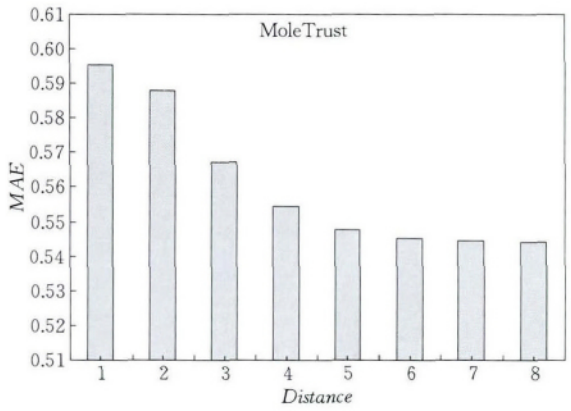


图 4　MoleTrust 算法中参数 *Distance* 值对 *MAE* 值的影响

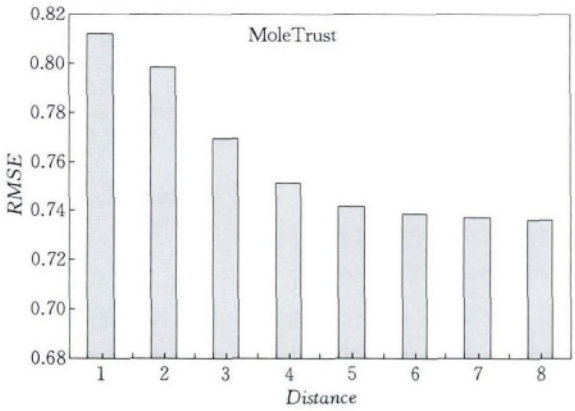


图 5　MoleTrust 算法中参数 *Distance* 值对 *RMSE* 值的影响

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[4] 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.

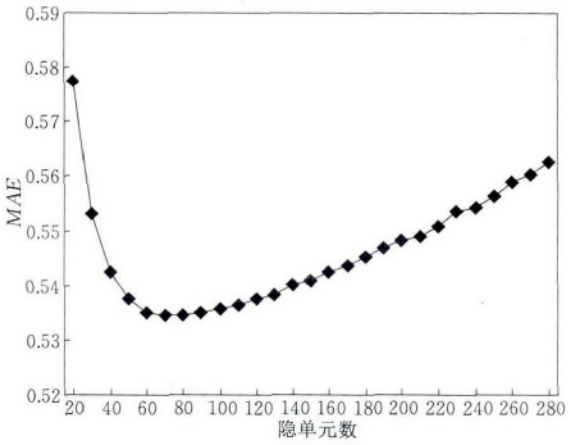


图 6　RBM 模型中隐单元数目对 *MAE* 值的影响

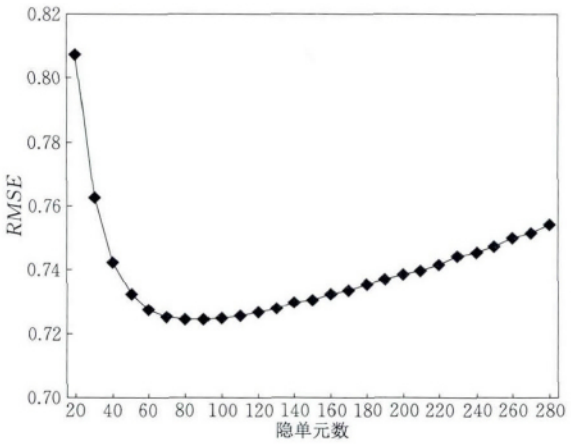


图 7　RBM 模型中隐单元数目对 *RMSE* 值的影响

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.

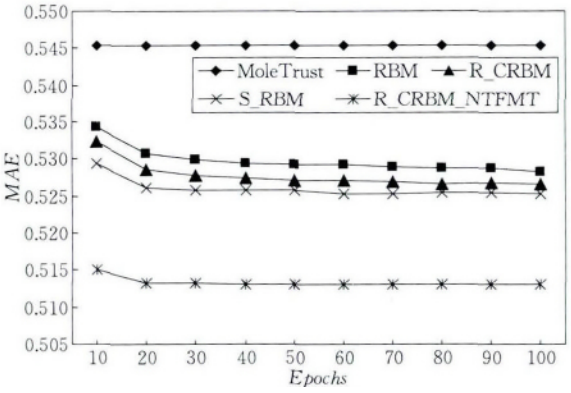


图 8 百度数据集上各方法的 *MAE* 值

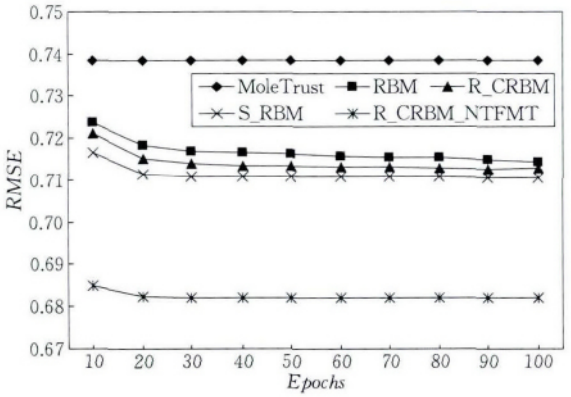


图 9 百度数据集上各方法的 *RMSE* 值

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 模型推荐效果比 RBM 模型有了一定程度的提高，但是其缺点是使所有用户对同一项目的预测评分均相同，这缺乏可解释性. 从两图中可看出我们提出的 R_CRBM_NTFMT 算法预测性能优于所有的算法，而且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，达到了较好的预测效果.

3.3.4 数据稀疏性实验

本实验通过改变训练数据集和测试数据集占评分数据的比例，观测数据在不同稀疏性的情况下对实验结果的影响. 本实验分别将评分数据中的 20%、40%、60%、80%作为训练数据集，相应地测试数据集占评分数据的比例分别为 80%、60%、40%、20%. 实验结果如图 10 和图 11 所示.

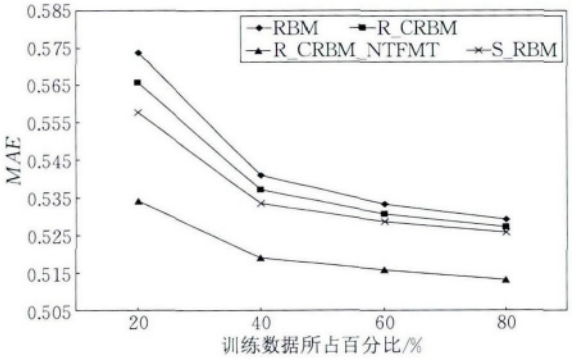


图 10 数据稀疏性比较_*MAE* 值

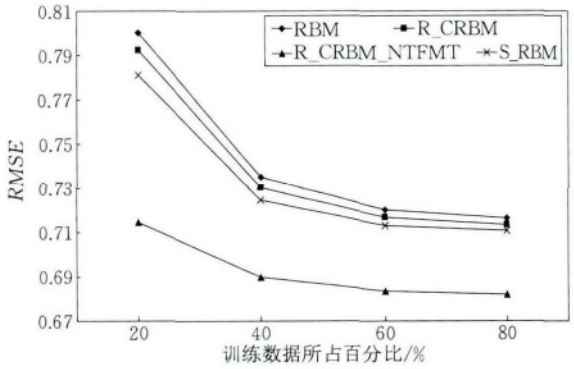


图 11 数据稀疏性比较_*RMSE* 值

图 10 中，横坐标表示不同稀疏度的训练数据集，纵坐标为评价指标 *MAE* 的值；图 11 坐标系中，横坐标表示不同稀疏度的训练数据集，纵坐标为评价指标 *RMSE* 的值. 从两图中可发现随着训练数据集占整个数据集比例的提高，也就是说随着数据稀疏性的减少，几种算法的 *MAE* 值和 *RMSE* 值均减小，也就是推荐效果不断变好. 两图中 R_CRBM 模型的推荐效果均优于 RBM 模型的结果，表明评分/未评分信息确实有利于提高推荐结果，在一定程度上解决了数据稀疏性问题；S_RBM 模型的效果较

RBM 模型也有了一定程度的提高；我们提出的 R_CRBM_NTFMT 算法的推荐结果远远优于 RBM 和 S_RBM 模型，表明我们利用的最近信任好友关系确实有利于提高推荐效果，而且数据越稀疏预测效果比 RBM 和 S_RBM 模型提高的越多. 当训练数据占 20% 时，我们的 R_CRBM_NTFMT 算法，*MAE* 比 RBM 和 S_RBM 模型分别提高 6.92% 和 4.25%，*RMSE* 比 RBM 和 S_RBM 模型分别提高 10.70% 和 8.55%；当训练数据占 80% 时，我们的 R_CRBM_NTFMT 算法，*MAE* 比 RBM 和 S_RBM 模型分别提高 3.06% 和 2.27%，*RMSE* 比 RBM 和 S_RBM 模型分别提高 4.76% 和 4.03%.

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.

3.4.1　训练数据参与训练次数 *Epochs* 对推荐结果的影响

本实验主要考查随着训练数据集参与训练次数 *Epochs* 的增加，各算法的推荐结果的变化情况. 实验结果如图 12 和图 13 所示.

图 12 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 13 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.

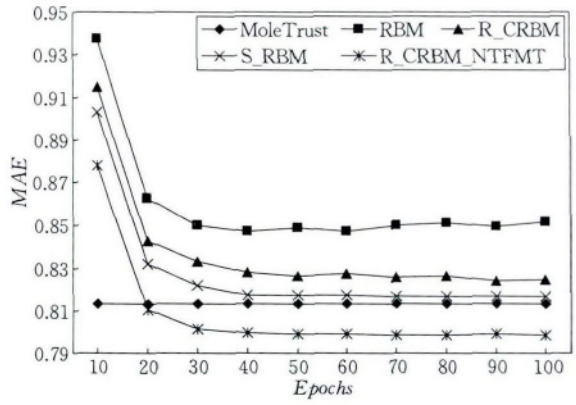


图 12　Epinions 数据集上各方法的 *MAE* 值

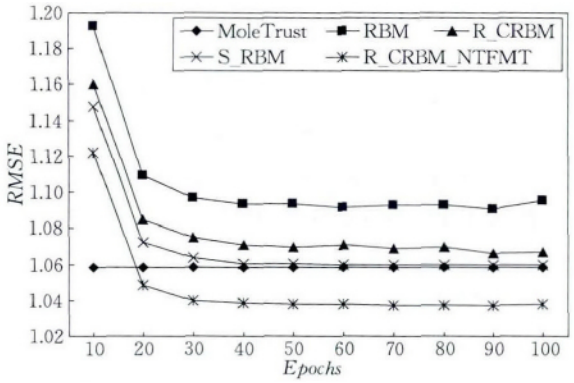


图 13　Epinions 数据集上各方法的 *RMSE* 值

3.4.2　数据稀疏性实验

本实验将评分数据中的 20%、40%、60%、80% 作为训练数据集，相应地测试数据集占评分数据的比例分别为 80%、60%、40%、20%. 实验结果如图 14 和图 15 所示.

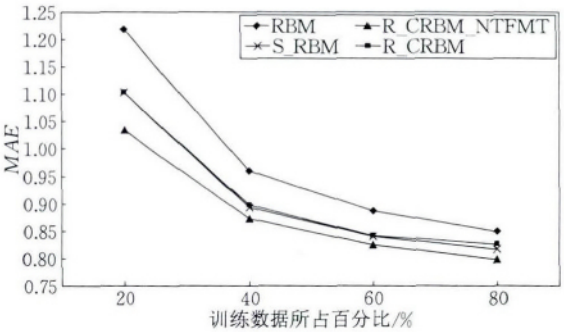


图 14　数据稀疏性比较_*MAE* 值

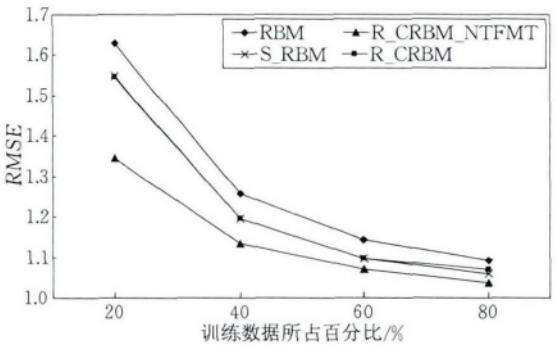


图 15　数据稀疏性比较_*RMSE* 值

图 14 中，横坐标表示不同稀疏度的训练数据集，纵坐标为评价指标 *MAE* 的值；图 15 坐标系中，横坐标表示不同稀疏度的训练数据集，纵坐标为评价指标 *RMSE* 的值. 从两图中可发现随着训练数据集占整个数据集比例的提高，几种算法推荐效果均不断变好. 两图中 R_CRBM 和 S_RBM 模型的效果

较为接近，R_CRBM 和 S_RBM 模型的推荐结果均优于 RBM 模型的结果，表明评分/未评分信息确实有利于提高推荐结果，在一定程度上解决了数据稀疏性问题；我们提出的 R_CRBM_NTFMT 算法的推荐结果远远优于 RBM 和 S_RBM 模型，表明我们利用的最近信任好友关系确实有利于提高推荐效果，而且数据越稀疏预测效果比 RBM 和 S_RBM 模型提高的越多，充分说明我们的算法能有效缓解数据稀疏性问题. 当训练数据占 20% 时，我们的 R_CRBM_NTFMT 算法，*MAE* 比 RBM 和 S_RBM 模型分别提高 15.11% 和 6.23%，*RMSE* 比 RBM 和S_RBM模型分别提高 17.45% 和 13.14%；当训练数据占 80%时，我们的 R_CRBM_NTFMT 算法，*MAE* 比 RBM 和 S_RBM 模型分别提高 5.89% 和 2.20%，*RMSE* 比 RBM 和 S_RBM 模型分别提高 5.06%和 2.95%.

综合对比两个数据集上的实验结果，可发现 R_CRBM 模型比 RBM 模型取得了更好的推荐效果，说明 R_CRBM 模型中利用的潜在的评分/未评分信息有助于提高推荐精度；我们提出的 R_CRBM_NTFMT 算法取得了最好的推荐效果，推荐效果比 RBM、R_CRBM 和 S_RBM 模型均有了较大程度地提高，说明我们算法中使用的最近信任好友关系是值得信赖的. 至此我们得出结论：我们的 R_CRBM 模型和 R_CRBM_NTFMT 算法中使用的评分/未评分信息和社交关系信息均有助于提高推荐效果，有效地解决了数据稀疏性问题.

3.5 基于 Spark 的大数据环境下的并行化实验

在大数据环境下，由于数据量巨大，普通平台无法处理大数据问题并且此时 R_CRBM 模型的参数数量将变得极其巨大，R_CRBM 模型的训练将面临巨大的挑战，因此，针对大数据下的 R_CRBM 模型，本文提出了基于 Spark 的并行化方案.

本实验的数据集采用完整的 Epinions 数据集，包含 40163 个用户对 139529 个项目的评分以及用户之间的关系数据.

本实验使用 IBM 高性能计算平台，使用其中 10 个计算节点，每个节点 8GB 内存.

3.5.1 基于 Spark 的 R_CRBM_NTFMT 算法

Spark 是一个基于内存计算的开源集群计算系统，其目的是更快速地进行数据分析. Spark 创新地提出了“弹性分布式数据集”(Resilient Distributed Datasets，RDD)的概念，RDD 是一种内存分布式数据集，它可以将中间结果缓存在内存中从而省去不必要的磁盘读写，提高运行速度.

图 16 为 R_CRBM 模型的并行化方案. 图中参数 $\theta=\{W,b,c,D\}$. 在并行化分解样本阶段将训练数据集切分到各个分片上；并行化阶段针对每个分片上的训练数据进行参数学习，得到各参数的更新值，即为式(3)～(6)；汇总阶段汇总每个分片上的参数得到平均后的参数. 本文的 R_CRBM_NTFMT 算法在 R_CRBM 模型训练好以后有一个寻找最近邻居的过程，其基本思想与 R_CRBM 模型的并行化方案类似.

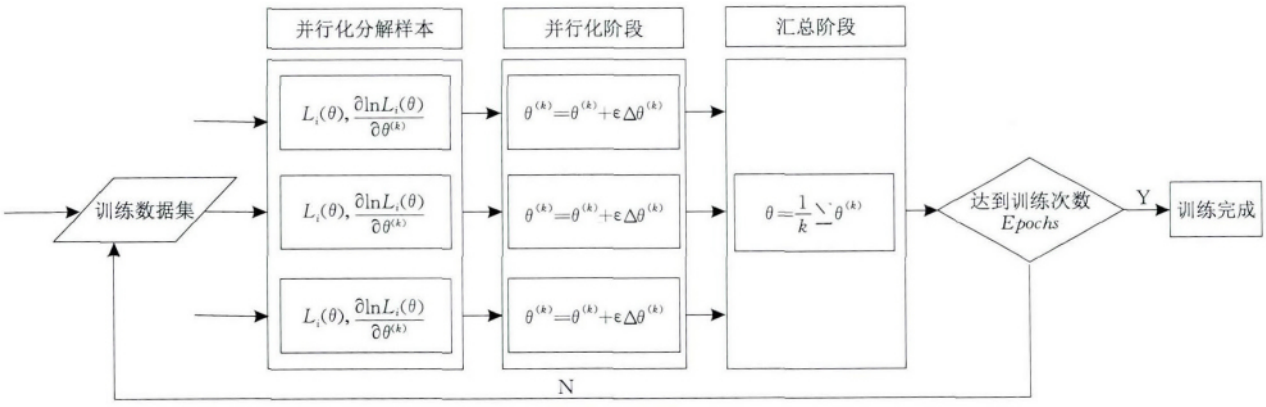


图 16 R_CRBM 模型并行化

3.5.2 可扩展性实验结果

本实验中，R_CRBM 模型的隐单元数量为 180，训练次数 *Epochs* 为 10，实验结果如图 17 和图 18 所示.

图 17 中，横坐标表示集群的节点数，纵坐标表示 R_CRBM_NTFMT 算法的运行时间，时间单位为分钟；图 18 中横坐标表示集群的节点数，总坐标表示加速比. 从两图中可发现随着集群节点数的增加，算法的运行时间越来越少，说明基于 Spark 的并行化方案是有效的；当集群从 1 个节点增加到 4 个

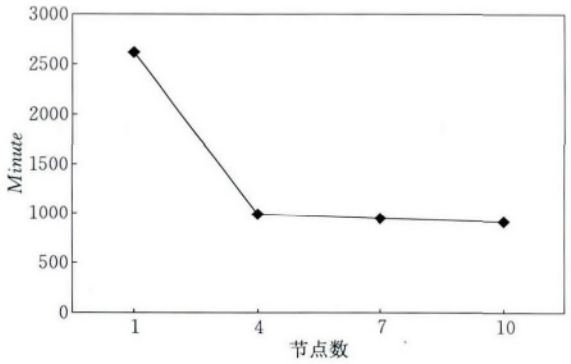


图 17　随集群节点变化算法运行时间变化

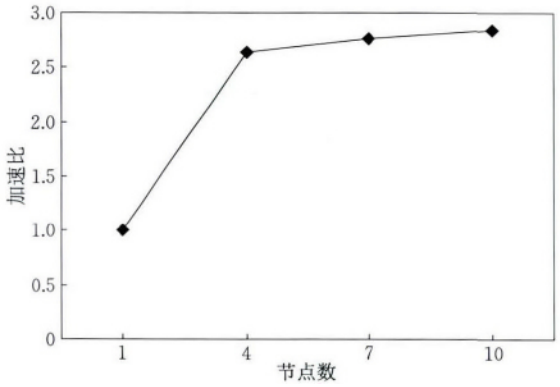


图 18　随集群节点变化算法加速比

节点时算法加速非常明显，加速比是 2.64，但是随着集群数量的继续增加，算法的加速情况变得较为缓慢，此时节点之间的通讯开销逐渐占据主导.

本实验中使用到 3 个数据（训练数据、测试数据、社交数据），3 个数据总的数据量约为 23 GB，使用 Spark 读取这 3 个数据仅需 20 s 左右，很好地解决了普通环境无法处理大数据的问题. 同时 10 个节点时基于 Spark 的并行化方案实现了 2.84 倍左右的加速，较好地解决了大数据环境下 R_CRBM 模型的训练问题.

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百度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，本文提出了 R_CRBM 模型并将 R_CRBM 模型和用户的社交关系相结合提出了 R_CRBM_NTFMT 算法. 实验结果表明我们的方法很好地解决了数据稀疏性问题，并且在数据越稀疏的情况下我们的 R_CRBM_NTFMT 算法的预测效果比 RBM 和 S_RBM 模型提高的越多. 最后，针对大数据环境下算法面临的挑战，本文提出了基于 Spark 的 R_CRBM_NTFMT 算法并行化方案，取得了一些初步的有效结果，未来将进一步优化其性能以适应在大数据下的推荐预测.

本论文中采用了文献[4]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[19] 提出了持续对比散度（Persistent Contrastive Divergence，PCD）算法；Tieleman 和 Hinton[20] 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 等人[21] 提出了 Parallel Tempering（PT）算法，通过交换相邻两个分布的状态，可以将低温下的状态传递到高温状态中，这样便可以从局部最优值中跳出，有更大的概率转移到距离较远的峰值中去；Ji 等人[22] 提出了 Parallel Tempering with Equi-Energy（PTEE）算法，用于解决 PT 算法中当相邻两个状态的能量差距很大时交换概率低的问题等等，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//Proceedings of the 14th Conference on Uncertainty in Artificial Intelligence. Madison, USA, 1998: 43-52

[2] Salakhutdinov R, Mnih A, Hinton G. Restricted Boltzmann machines for collaborative filtering//Proceedings of the 24th International Conference on Machine Learning. Corvallis, USA, 2007: 791-798

[3] Zhang Chun-Xia, Ji Nan-Nan, Wang Guan-Wei. Introduction of restricted Boltzmann machines. China Science Paper Online, 2013(in Chinese)
（张春霞，姬楠楠，王冠伟. 受限波尔兹曼机简介. 中国科技论文在线，2013）

[4] Georgiev K, Nakov P. A non-IID framework for collaborative filtering with restricted Boltzmann machines//Proceedings of the 30th International Conference on Machine Learning. Atlanta, USA, 2013: 1148-1156

[5] Golbeck J. Generating Predictive Movie Recommendations from Trust in Social Networks. Berlin Heidelberg: Springer, 2006

[6] Golbeck J, Hendler J. FilmTrust: Movie recommendations using trust in web-based social networks//Proceedings of the IEEE Consumer Communications and Networking Conference. Las Vegas, USA, 2006: 282-286

[7] Massa P, Avesani P. Controversial users demand local trust metrics: An experimental study on epinions. com community //Proceedings of the 20th National Conference on Artificial Intelligence. Menlo Park, USA, 2005: 121-126

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

[9] Ma Hao, et al. SoRec: Social recommendation using probabilistic matrix factorization//Proceedings of the 17th ACM Conference on Information and Knowledge Management. Napa Valley, USA, 2008: 931-940

[10] Ma Hao, et al. Recommender systems with social regularization //Proceedings of the 4th ACM International Conference on Web Search and Data Mining. Hong Kong, China, 2011: 287-296

[11] Huang Jun-Ming, et al. Exploring social influence via posterior effect of word-of-mouth recommendations//Proceedings of the 5th ACM International Conference on Web Search and Data Mining. Seattle, USA, 2012: 573-582

[12] Ji N, Zhang J, Zhang C, et al. Enhancing performance of restricted Boltzmann machines via log-sum regularization. Knowledge-Based Systems, 2014, 63(3): 82-96

[13] Welling M, Hinton G E. A new learning algorithm for mean field Boltzmann machines//Proceedings of the International Conference on Artificial Neural Networks (ICANN 2002). Madrid, Spain, 2002: 351-357

[14] Ziegler C N, Golbeck J. Investigating correlations of trust and interest similarity—Do birds of a feather really flock together. Decision Support Systems, 2005, 2005, 43(2): 1-34

[15] Massa P, Avesani P. Trust-aware bootstrapping of recommender systems//Proceedings of the ECAI Workshop on Recommender Systems. Riva del Garda, Italy, 2006: 29-33

[16] Breese J S, Heckerman D, Kadie C. Empirical analysis of predictive algorithms for collaborative filtering//Proceedings of the 14th Conference on Uncertainty in Artificial Intelligence. Madison, USA, 1998: 43-52

[17] Herlocker J L, Konstan J A, Borchers A, et al. An algorithmic framework for performing collaborative filtering//Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Berkeley, USA, 1999: 230-237

[18] Shardanand U, Maes P. Social information filtering: Algorithms for automating "word of mouth"//Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. Denver, USA, 1995: 210-217

[19] Tieleman T. Training restricted Boltzmann machines using approximations to the likelihood gradient//Proceedings of the 25th International Conference on Machine Learning. Helsinki, Finland, 2008: 1064-1071

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

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

[22] Ji N, Zhang J. Parallel tempering with equi-energy moves for training of restricted Boltzmann machines//Proceedings of the 2014 International Joint Conference on Neural Networks. Beijing, China, 2014: 120-127

HE Jie-Yue, born in 1964, Ph. D., professor. Her current research interests include data intensive computing, bioinformatics, data mining, machine learning.

MA Bei, born in 1990, M. S. His current research interests include deep learning and collaborative filtering.

Background

Collaborative filtering has become one of the most used approaches for recommender system. However, data sparsity is one of the most crucial challenges for the collaborative filtering algorithm or recommender system. Data sparsity will lead to the low prediction accuracy of collaborative filtering algorithms, which will seriously reduce the user

experience. In recent years, with the popularity of social network, social network relationship is becoming more and more important in people's daily life. Friends' opinion or view in social network tends to influence our decision. Therefore, the use of social relationship in social network can solve the problem of data sparsity.

Nowadays, deep learning has made significant breakthrough in many fields. Restricted Boltzmann machine model occupies a central position in the field of deep learning, which can be used to solve the recommendation problem. In current research, Restricted Boltzmann machine model for Collaborative filtering only uses users rating data. However, there are serious data sparseness problem in user rating data. Therefore, a Real-Valued Conditional Restricted Boltzmann Machine model (R_CRBM) is proposed in this paper. In R_CRBM model rating data does not need to be converted to a K dimensional 0-1vector unit. Meanwhile, the model training process uses rated/unrated information. Moreover, we also proposed a method R_CRBM_NTFMT which combines Real-Valued Conditional Restricted Boltzmann Machine model with Nearest Trusted Friends Based on MoleTrust in social network information. The experimental results on Baidu and Epinions datasets show that the R_CRBM model and R_CRBM_NTFMT algorithm help to improve the prediction accuracy of the recommendation system, the R_CRBM model and R_CRBM_NTFMT algorithm are good solutions to solve the problem of data sparsity. Moreover, the experimental results show that the train data more sparse the prediction accuracy of R_CRBM_NTFMT algorithm is much better than RBM and S_RBM model. Finally, due to a common platform is very difficult to train R_CRBM model in big data. Therefore, a parallelization scheme based on Spark is proposed in this paper, the experimental result show that the parallelization method for R_CRBM_NTFMT algorithm has good scalability.

This work was supported in part by the Natural Science Foundation of Jiangsu Province under Grant No. BK2012742 and Collaborative Innovation Center for Novel Software Technology and Industry of Jiangsu Province, Nanjing, 210046.