三部图张量分解标签推荐算法

廖志芳[1)] 李 玲[1)] 刘丽敏[2)] 李永周[3)]

[1)](中南大学软件学院 长沙 410075)

[2)](中南大学信息科学与工程学院 长沙 410075)

[3)](中南大学信息科学与工程学院计算机科学与技术博士后流动站 长沙 410083)

摘 要 三部图作为社会标签系统的表示方法,虽然可以简化标签系统元素间关系的表达,但也丢失了部分元素间的相关信息,而且不能有效处理标签系统中具有大量稀疏值和缺失值的数据.基于以上问题,文中提出了基于三部图的三维张量分解推荐算法(TTD 算法).首先分析三部图元素间可能丢失的信息,通过定义以三部图为基础的低阶张量分解模型,对高阶稀疏数据进行分析.该模型不仅包含三部图所表达的系统信息,同时还表达了三部图所丢失的元素间相互信息;在此基础上,利用缺失值处理,进行社会标签系统中的标签推荐预测.通过模型对比实验以及标签预测实验,表明 TTD 模型所揭示的社会标签系统中元素间的相互关系更加全面,同时在进行标签预测时,所得到的预测结果召回率和精确率得到了显著改善.

关键词 三部图;张量分解三部图模型(TTD);标签预测;社会标签系统

中图法分类号 TP391 **DOI 号**: 10.3724/SP.J.1016.2012.02625

A Tripartite Decomposition of Tensor for Social Tagging

LIAO Zhi-Fang[1)] LI Ling[1)] LIU Li-Min[2)] LI Yong-Zhou[3)]

[1)](*School of Software*, *Central South University*, *Changsha* 410075)

[2)](*School of Information Science & Engineering*, *Central South University*, *Changsha* 410075)

[3)](*School of Information Science & Engineering*, *Post-Doctor Site*, *Central South University*, *Changsha* 410083)

Abstract Although Tripartite Graph can reduce complexity among the relationships of social tagging system, it loses some information among the three elements, and it is also difficult to process the sparse data with missing values. In this paper, we present a Tripartite Tensor Decomposition (TTD) Algorithm to deal with these problems. We first analyzes the information may be lost in the Tripartite Graph, then propose a lower order tensor model based on tripartite graph to deal with the missing information and high-index sparse data. Comparing with tripartite graph model, TTD model reveals comprehensive relations in social tagging system, not only obtains the information between elements, but also gets the relation among three elements. The model is also applied in social tagging system for tagging recommendation by dealing with the missing value. The results of the model comparison experiment and social tagging predication experiment show that TTD model reveals the relations in social tagging system more comprehensive and the results show significant improvements in terms of effective measured through recall/precision when it is used for social tagging recommendation.

Keywords tripartite graph; Tripartite Tensor Decomposition model (TTD); tagging prediction; social tagging system

---

收稿日期:2011-05-17;最终修改稿收到日期:2012-05-31.本课题得到国家自然科学基金(61073105)、国家博士后科学基金(20100480950)资助.**廖志芳**,女,1968 年生,博士,副教授,研究方向为数据挖掘、推荐系统.E-mail:zfliao@csu.edu.cn.**李 玲**,女,1987 年生,硕士,研究方向为数据挖掘、推荐系统.**刘丽敏**,女,1976 年生,博士,讲师,研究方向为数据挖掘、计算机网络.**李永周**,男,1971 年生,博士,副教授,研究方向为模式识别、社会网络.

1 引 言

作为 Web2.0 的重要特征[1]，社会标签允许用户利用开放平台，对系统资源赋予个性化标签，为具有相同兴趣爱好的用户提供资源推荐及共享. 比较成功的社会标签系统有如音乐标签系统 last.fm、图片标签系统 flickr 以及 Web 网页标签系统 Del.icio.us 等，这些网站利用社会标签进行资源整合，搜索用户之间的联系，并通过推荐算法将用户感兴趣的标签推荐给使用同一资源的用户.

社会标签系统起源于传统的推荐系统，社会标签系统的主要元素包括{资源(item)，标签(tag)，用户(user)}，系统可定义为 $F:=(U,T,I,R)$，其中，U 为 user；T 为 tag；I 为 item；R 为 user、item 和 tag 之间的关系，其中 $R\in T\times U\times I$[2]. 而传统推荐系统包括{资源(item)，用户(user)}及两者之间的关系，系统定义为 $F:=(U,I,R)$，其中 $R\in U\times I$[3].

根据上述定义，在问题描述上传统推荐系统分析的是{item，user}之间的二维关系，而社会标签系统分析的是{item，user，tag}三维关系，因而传统推荐系统中的推荐算法不能直接应用于社会标签系统. 但是很多基于传统推荐系统中的推荐方法具有良好的性能和效果[3]，如协调过滤 CF 算法[4]、图模型推荐算法等，这些算法在推荐系统中都有良好的应用.

国内外很多学者对将推荐算法应用于社会标签推荐系统进行了研究，提出了很多算法. 这些算法可以分为两种，一种方法是将算法直接进行改进，将其设计成能处理三维关系的算法和模型，如 Zhang 与 Liu[5] 提出的超图模型，模型的结果能够很好地再现标签网络中的超度分布、超度-簇系数分布和节点距离排序，能够充分体现元素间的关系，但是在处理大量稀疏数据时算法复杂性较高. 另外一种就是将社会标签的三维关系转换为二维关系，直接应用传统推荐系统算法及模型，如扩展 PLSA (Probabilistic Latent Semantic Analysis)方法[6]，由 Cohn 和 Hofmann[7] 提出的基于内容和基于链接相结合的统一框架模型以及当前应用较为广泛的由 Mika[8] 提出的利用三部图的模型，这些模型的主要目的是将标签系统三维关系转换为二维关系.

除了系统分析复杂性问题外，标签系统还存在数据极度稀疏的问题. 为解决这个问题，Symeonidis 等人[9] 首先提出将能完整地表示高维数据并且能维持高维空间数据的本征结构信息的 tensor(张量)应用于社会标签系统中，并且利用张量分解方法进行标签预测. 如果定义 $\boldsymbol{Y}$ 为张量，则 $\boldsymbol{Y}$ 分别表示 item、tag 和 user 的张量分解关系，不同的张量分解方法代表不同的 $\boldsymbol{Y}$. 其中典型张量分解方法有 Tucker 分解[10] 和 ParaFac[10] 以及由 Tucker 分解推广的高阶奇异值分解(High Order Singular Value Decomposition，HOSVD)[10]，但是高维张量导致了算法具有高阶的时间复杂度，同时在进行标签预测时，存在对标签预测结果的过拟合问题. 为改进此问题，Symeonidis[9] 采用三阶张量方法，定义稀疏张量 $\boldsymbol{Y}$，以值为 1 表示用户在标签上进行了标记，而 0 则表示没有用户没有标记. 该方法去除数据集中的部分噪音，在预测效果上高于传统张量分解方法. 与 Symeonidis 相似，Rendle 等人[10] 则是以正负来代表用户标签是否标记，如果用户对标签进行标记，则分类为“+”，如果没有标记，则分类为“−”，这种方法可以有效区分用户的标签分类，但是在标签预测过程中则不能为用户提供有效的标签序列. Rendle 等人[10-11] 在 2010 年提出了一个新的推荐算法：PITF (Pairwise Interaction Tensor Factorization)，它的主要思想就是在张量分解的时候，考虑 3 个二维关系两两之间的关联，PITF 分解算法是一种特殊的具有线性时间复杂度的 Tucker 分解算法，在推荐质量上也有了很好的提高.

通过以上研究分析，我们发现当前应用最为广泛地将三维关系转化为二维关系的为三部图模型，在三部图转换成二部图过程中存在的三元组信息丢失问题；同时由于社会标签系统数据稀疏性问题，三部图的处理方法会损失更多的三元组信息. 为解决上述两个问题，本文在分析张量分解方法的基础上，基于三部图的张量分解模型对标签系统中三元组{item，user，tag}的关系进行集成及描述，分析 3 个对象之间潜在的关系，并在分析其有效性的基础上利用张量分解三部图模型进行用户标签预测.

2 三部图基础

三部图将标签系统中 item、user、tag 定义为如图 1 的关系，其中，item 表示为 i，tag 表示为 j，而 user 表示为 k；U_{ij} 表示{item，tag}之间的关系，W_{ik} 表示{item，user}之间的关系，而 V_{jk} 表示{user，tag}之间的关系. 在进行标签系统推荐时，三部图通过计算 3 个关系对中的元素同时发生的次数，将三部图

转换成 3 个二部图的关系，即{item，user}，{item，tag}以及{user，tag}之间的关系来进行社会标签系统的描述如图 2 所示.

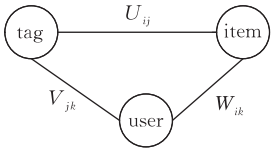


图 1　三部图基本结构

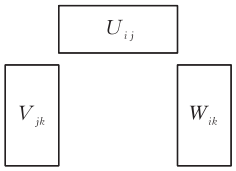


图 2　三部图分解为二部图

我们将同时发生的次数定义如下.

定义 1.　设 A,B 为{item}集中的两个标签 tag，则 A,B 同时发生的次数定义为

$$RC(A,B)=\frac{|A\cap B|}{|A\cup B|}.$$

从图 1 到图 2 的转换过程可以看出，这种方法简单明了，但是在标签系统中原始数据是以三元组出现，即$\{i_t,j_t,k_t\}$，说明用户 k_t在资源 i_t上进行了 j_t 标签标注，因此标签系统数据集为$\{i_t,j_t,k_t\}$，$t=1,\cdots,m$. 三部图在转换成二部图的过程中，如图 2 所示，因为只考虑了 U_{ij}、W_{ik} 和 V_{jk} 的关系，因而损失了部分$\{i_t,j_t,k_t\}$之间的信息. 以图 3 为例，我们可以看出，item 中 i_2与 tag 标签 j_2之间不仅仅存在着 $U_{i_2j_2}$ 的关系，同时 i_2与 j_2还可以沿着 item 中的 i_2 与 user 中的 k_2以及 user 中的 k_2到 tag 中的 j_2之间传递 i_2与 j_2之间的信息. 但是三部图到二部图的转换中，遗失了这部分信息.

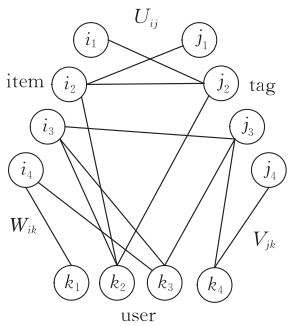


图 3　三部图三元组信息

因此三部图作为社会标签一个应用模型，虽然简化了标签系统的复杂关系，但是也损失了标签系统的部分信息，直接影响预测分析结果. 为解决这个问题，除了要考虑三元组中各个元素间直接联系外，还需要考虑元素间的间接联系. 基于此考虑，我们提出了本文的模型，简单地说，就是模型在考虑{item，tag}间的关系时，不仅要考虑{item，tag}间的直接联系，也需要考虑从{item，user}再到{user，tag}之间的间接关系.

3　基于三部图的张量分解方法

由于标签系统中的数据集为三元组，也为三阶张量分解方法进行三部图的转换提供了理论基础. 利用张量分解处理稀疏数据以及三部图简化处理方法的优越性，本文在三部图基础上建立张量分解三部图模型，定义如下

$$\boldsymbol{X}_{ijk}\approx\boldsymbol{Y}_{ijk}\tag{1}$$

其中不同的 Y 代表不同的张量分解方法.

根据图 3，考虑到三部图中三元组元素之间两两相关性，本文提出了如下张量分解模型.

$$\boldsymbol{Y}_{ijk}=\boldsymbol{U}_{ij}\boldsymbol{V}_{jk}+\boldsymbol{V}_{jk}\boldsymbol{W}_{ik}+\boldsymbol{U}_{ij}\boldsymbol{W}_{ik}\tag{2}$$

根据图 1 所示的三部图模型，item 节点 i 和 tag 节点 j 之间的关系为

$$\hat{\boldsymbol{U}}_{ij}=\sum_{k=1}^{N_k}\boldsymbol{X}_{ijk}\tag{3}$$

而根据式(2)所定义的张量分解模型，item 节点 i 和 tag 节点 j 之间的关系为

$$\tilde{\boldsymbol{U}}_{ij}=\sum_{k=1}^{N_k}\boldsymbol{Y}_{ijk}=\sum_{k=1}^{N_k}[\boldsymbol{U}_{ij}\boldsymbol{V}_{jk}+\boldsymbol{V}_{jk}\boldsymbol{W}_{ik}+\boldsymbol{U}_{ij}\boldsymbol{W}_{ik}]=\boldsymbol{U}_{ij}(\boldsymbol{V}_{j+}+\boldsymbol{W}_{i+})+(\boldsymbol{V}\boldsymbol{W}^{\mathrm{T}})_{ij}\tag{4}$$

其中，$\boldsymbol{U}_{ij}$是 $I\times J$ 矩阵，$\boldsymbol{V}_{jk}$是 $J\times K$ 矩阵，$\boldsymbol{W}_{ik}$是 $I\times K$ 矩阵，比较式(3)和式(4)，式(3)中三部图模型的 $\hat{\boldsymbol{U}}_{ij}$显示了{item，tag}之间的直接关系，而三维张量分解模型中的 $\tilde{\boldsymbol{U}}_{ij}$，根据式(4)，可以看到 $\tilde{\boldsymbol{U}}_{ij}$则不仅来自{item，tag}之间的直接关系 $\boldsymbol{U}_{ij}$，同样也来自{item，user}即 $\boldsymbol{V}_{jk}$、{user，tag}即 $\boldsymbol{W}_{ik}$以及 $\boldsymbol{V}_{jk}$和 $\boldsymbol{W}_{ik}$的传递关系. 同样也可以得到三部图模型中的 $\hat{\boldsymbol{V}}_{jk}$与张量模型 $\tilde{\boldsymbol{V}}_{jk}$以及三部图模型 $\hat{\boldsymbol{W}}_{ik}$与张量模型 $\tilde{\boldsymbol{W}}_{ik}$之间的比较结果.

由此可以得出，本文提出的基于三部图的张量分解模型所包含的信息$\tilde{\boldsymbol{U}}_{ij}$、$\tilde{\boldsymbol{V}}_{jk}$、$\tilde{\boldsymbol{W}}_{ik}$与三部图中的$\hat{\boldsymbol{U}}_{ij}$、$\hat{\boldsymbol{V}}_{jk}$、$\hat{\boldsymbol{W}}_{ik}$相比，不仅仅来自于模型中的二维直接关系，同时也来自于模型中三元组{item，user，tag}之间的关系，因此这个模型为标签系统提供比三部图更多的系统间的相关信息，解决了三部图在转换成二部图中信息丢失的问题.

该模型在进行标签预测时，为获得最优的标签

预测值，$\boldsymbol{X}_{ijk}$、$\boldsymbol{Y}_{ijk}$需要满足式(5)，即

$$\min_{\boldsymbol{U},\boldsymbol{V},\boldsymbol{W}} J = \min \sum_{ijk} (\boldsymbol{X}_{ijk} - \boldsymbol{Y}_{ijk})^2 + \alpha(\|\boldsymbol{U}\|^2 + \|\boldsymbol{V}\|^2 + \|\boldsymbol{W}\|^2) \quad (5)$$

其中，$\boldsymbol{X}_{ijk}$为标签系统实际值，$\boldsymbol{Y}_{ijk}$为标签预测值，其中α是一个模型参数，用于调整$\boldsymbol{U}$，$\boldsymbol{V}$，$\boldsymbol{W}$，使得$\boldsymbol{U}$、$\boldsymbol{V}$、$\boldsymbol{W}$的元素都有基本相同的数量级.

4 三部图张量分解算法

根据式(3)所定义的张量分解三部图模型进行社会标签的预测，本文将该算法命名为TTD(Tripartite tensor Decomposition). 表1列出了算法中的所有符号含义.

表1 算法中的符号

符号	定义	符号	定义
$\boldsymbol{X}$	初始张量，$\boldsymbol{X}=\boldsymbol{X}_\Omega+\boldsymbol{X}_m$	$\boldsymbol{U}^*$	$\boldsymbol{U}_{ij}$从二维到三维的扩展
$\boldsymbol{X}_\Omega$	$\boldsymbol{X}$中具有值的元素集合	$\boldsymbol{V}^*$	$\boldsymbol{V}_{ik}$从二维到三维的扩展
$\boldsymbol{X}_m$	$\boldsymbol{X}$中缺失值的元素集合	$\boldsymbol{W}^*$	$\boldsymbol{W}_{jk}$从二维到三维的扩展
N_i	资源项数量	$\boldsymbol{X}_{\boldsymbol{U}}^{i}{}^*$	$\sum_{i=1}^{N_i}(\boldsymbol{X}_{ijk}\boldsymbol{U}^*)$
N_j	标签数量	$\boldsymbol{X}_{\boldsymbol{U}}^{j}{}^*$	$\sum_{j=1}^{N_j}(\boldsymbol{X}_{ijk}\boldsymbol{U}^*)$
N_k	用户数量	$\boldsymbol{X}_{\boldsymbol{V}}^{i}{}^*$	$\sum_{i=1}^{N_i}(\boldsymbol{X}_{ijk}\boldsymbol{V}^*)$
$\|\boldsymbol{X}\|$	$\|\boldsymbol{X}\|^2=\sum_{i=1}^{N_i}\sum_{j=1}^{N_j}\sum_{k=1}^{N_k}\boldsymbol{X}_{ijk}^2$	$\boldsymbol{X}_{\boldsymbol{V}}^{k}{}^*$	$\sum_{k=1}^{N_k}(\boldsymbol{X}_{ijk}\boldsymbol{V}^*)$
$\boldsymbol{U}_{i+}$	$\sum_{j=1}^{N_j}\boldsymbol{U}_{ij}$，$\boldsymbol{U}_{+j}$、$\boldsymbol{V}_{i+}$、$\boldsymbol{V}_{+k}$、$\boldsymbol{W}_{j+}$、$\boldsymbol{W}_{+k}$定义类似	$\boldsymbol{X}_{\boldsymbol{W}}^{j}{}^*$	$\sum_{j=1}^{N_j}(\boldsymbol{X}_{ijk}\boldsymbol{W}^*)$
$\boldsymbol{e}_i=[1,1,\cdots,1]^{\mathrm{T}}$	长度为N_i的向量	$\boldsymbol{X}_{\boldsymbol{W}}^{k}{}^*$	$\sum_{k=1}^{N_k}(\boldsymbol{X}_{ijk}\boldsymbol{W}^*)$
$\boldsymbol{e}_j=[1,1,\cdots,1]^{\mathrm{T}}$	长度为N_j的向量	ε	收敛因子，$\varepsilon=0.001$
$\boldsymbol{e}_k=[1,1,\cdots,1]^{\mathrm{T}}$	长度为N_k的向量		

4.1 TTD计算方法

为得到标签预测最优解，定义目标函数如式(3)所示，即

$$\min_{\boldsymbol{U},\boldsymbol{V},\boldsymbol{W}} J = \min \sum_{ijk} (\boldsymbol{X}_{ijk} - \boldsymbol{Y}_{ijk})^2 + \alpha(\|\boldsymbol{U}\|^2 + \|\boldsymbol{V}\|^2 + \|\boldsymbol{W}\|^2),$$

其中$\boldsymbol{Y}_{ijk}$为式(2)，即

$$\boldsymbol{Y}_{ijk} = \boldsymbol{U}_{ij}\boldsymbol{V}_{jk} + \boldsymbol{V}_{jk}\boldsymbol{W}_{ik} + \boldsymbol{U}_{ij}\boldsymbol{W}_{ik}.$$

为求解最优预测值，我们分别求解$\boldsymbol{U}_{ij}$、$\boldsymbol{V}_{jk}$以及$\boldsymbol{W}_{ik}$的最优值，通过迭代，求取式(5)的最优值.

为得到$\boldsymbol{U}_{ij}$，将$\boldsymbol{V}_{jk}$和$\boldsymbol{W}_{ik}$固定，计算式(6)

$$\min_{\boldsymbol{U}} J(\boldsymbol{U}) = \min \sum_{ijk} (\boldsymbol{X}_{ijk} - (\boldsymbol{U}_{ij}\boldsymbol{V}_{jk} + \boldsymbol{V}_{jk}\boldsymbol{W}_{ik} + \boldsymbol{U}_{ij}\boldsymbol{W}_{ik}))^2 + \alpha(\|\boldsymbol{U}\|^2 + \|\boldsymbol{V}\|^2 + \|\boldsymbol{W}\|^2) \quad (6)$$

令$\partial J/\partial \boldsymbol{U}=\boldsymbol{0}$，得到$\boldsymbol{U}$的最优解：

$$\boldsymbol{U} = (\boldsymbol{V}\boldsymbol{V}^{\mathrm{T}}\boldsymbol{e}_j^{\mathrm{T}} + \boldsymbol{e}_i\boldsymbol{W}\boldsymbol{W}^{\mathrm{T}} + 2\boldsymbol{V}\boldsymbol{W}^{\mathrm{T}} + \alpha\boldsymbol{I})^{-1} \cdot (\boldsymbol{X}_{\boldsymbol{V}^*}^{k} + \boldsymbol{X}_{\boldsymbol{W}^*}^{k} - (\boldsymbol{V}\cdot\boldsymbol{V})\boldsymbol{W}^{\mathrm{T}} - \boldsymbol{V}(\boldsymbol{W}\cdot\boldsymbol{W})) \quad (7)$$

按照同样的方法得到$\boldsymbol{V}$和$\boldsymbol{W}$的最优解，其最优解分别为

$$\boldsymbol{V} = (\boldsymbol{U}\boldsymbol{U}^{\mathrm{T}}\boldsymbol{e}_i\boldsymbol{e}_k^{\mathrm{T}} + \boldsymbol{e}_i\boldsymbol{e}_k^{\mathrm{T}}\boldsymbol{W}^{\mathrm{T}}\boldsymbol{W} + 2\boldsymbol{U}\boldsymbol{W} + \alpha\boldsymbol{I})^{-1} \cdot (\boldsymbol{X}_{\boldsymbol{U}^*}^{j} + \boldsymbol{X}_{\boldsymbol{W}^*}^{j} - (\boldsymbol{U}\cdot\boldsymbol{U})\boldsymbol{W} - \boldsymbol{U}(\boldsymbol{W}\cdot\boldsymbol{W})) \quad (8)$$

$$\boldsymbol{W} = (\boldsymbol{U}^{\mathrm{T}}\boldsymbol{U}\boldsymbol{e}_j\boldsymbol{e}_k^{\mathrm{T}} + \boldsymbol{e}_j\boldsymbol{e}_k^{\mathrm{T}}\boldsymbol{V}^{\mathrm{T}}\boldsymbol{V} + 2\boldsymbol{U}^{\mathrm{T}}\boldsymbol{V} + \alpha\boldsymbol{I})^{-1} \cdot (\boldsymbol{X}_{\boldsymbol{U}^*}^{i} + \boldsymbol{X}_{\boldsymbol{V}^*}^{i} - (\boldsymbol{U}^{\mathrm{T}}\cdot\boldsymbol{U}^{\mathrm{T}})\boldsymbol{V} - \boldsymbol{U}^{\mathrm{T}}(\boldsymbol{V}\cdot\boldsymbol{V})) \quad (9)$$

分别将$\boldsymbol{U}$，$\boldsymbol{V}$，$\boldsymbol{W}$最优值代入式(5)，通过不断迭代求解，最终得到式(3)的最优值，即$\boldsymbol{Y}_{ijk}$的最优解以用于标签推荐.

4.2 TTD算法收敛性分析

由式(3)可知，为取得最优预测值，须满足：

$$\min_{\boldsymbol{U},\boldsymbol{V},\boldsymbol{W}} J = \min \sum_{ijk} (\boldsymbol{X}_{ijk} - \boldsymbol{Y}_{ijk})^2 + \alpha(\|\boldsymbol{U}\|^2 + \|\boldsymbol{V}\|^2 + \|\boldsymbol{W}\|^2),$$

其中，$\boldsymbol{Y}_{ijk}=\boldsymbol{U}_{ij}\boldsymbol{V}_{jk}+\boldsymbol{V}_{jk}\boldsymbol{W}_{ik}+\boldsymbol{U}_{ij}\boldsymbol{W}_{ik}$. 即，求解

$$\min \sum [\boldsymbol{X}_{ijk} - (\boldsymbol{U}_{ij}\boldsymbol{V}_{jk} + \boldsymbol{V}_{jk}\boldsymbol{W}_{ik} + \boldsymbol{U}_{ij}\boldsymbol{W}_{ik})]^2 + \alpha(\|\boldsymbol{U}\|^2 + \|\boldsymbol{V}\|^2 + \|\boldsymbol{W}\|^2).$$

根据TTD计算方法，先对$\boldsymbol{U}_{ij}$进行求解，为求$\boldsymbol{U}_{ij}^n$最小值，则算法表示为：令$\partial J/\partial \boldsymbol{U}=\boldsymbol{B}+\boldsymbol{A}\boldsymbol{U}=\boldsymbol{0}$，得到$\boldsymbol{U}_{ij}^n$最小值. 接着将$\boldsymbol{U}_{ij}^n$代入，得到$\boldsymbol{V}_{jk}^n$，以此类推，得到$\boldsymbol{W}_{ik}^n$.

假如$J^n - J^{n-1} > \varepsilon$，$\varepsilon$表示最小阈值，则需要继续计算$\boldsymbol{U}_{ij}^{n+1}$、$\boldsymbol{V}_{jk}^{n+1}$以及$\boldsymbol{W}_{ik}^{n+1}$.

根据以上算法过程，可以知道

$$J(\boldsymbol{U}^{n+1}) = \arg\min J(\boldsymbol{U}^n) \Rightarrow J(\boldsymbol{U}^{n+1}) < J(\boldsymbol{U}^n),$$

同理可推出，$J(\boldsymbol{V}^{n+1}) < J(\boldsymbol{V}^n)$，$J(\boldsymbol{W}^{n+1}) < J(\boldsymbol{W}^n)$，则

$$J(\boldsymbol{U}^{n+1},\boldsymbol{V}^{n+1},\boldsymbol{W}^{n+1}) < J(\boldsymbol{U}^n,\boldsymbol{V}^{n+1},\boldsymbol{W}^{n+1}) < J(\boldsymbol{U}^n,\boldsymbol{V}^n,\boldsymbol{W}^{n+1}) < J(\boldsymbol{U}^n,\boldsymbol{V}^n,\boldsymbol{W}^n).$$

即$J^{n+1} < J^n$，由此可以得到算法TTD是收敛的.

4.3 缺失值的处理

在标签数据集中，如果某用户没有用标签对某资源项进行标注，或者某标签与某资源项没有对应关系，则称该对应关系是缺失的，形式化描述就是三元组(i,j,k)无实际有效值，即为缺失值. 由于社会

标签数据极度稀疏，即社会标签数据集中非零值的相对数量少，所以标签系统缺失值是大量存在的，在缺失值处理中，本文采用与文献[12]相同的方法.

如果标签系统中的元素为非零值，则需要满足式(10)，其中 Ω 表示张量中所有被赋予了非零值的元素的集合.

$$\min_{\boldsymbol{Y}} \|\boldsymbol{X}-\boldsymbol{Y}\|_{\Omega}^{2}=\sum_{(ijk)\in\Omega}(\boldsymbol{X}_{ijk}-\boldsymbol{Y}_{ijk})^{2} \quad (10)$$

如果 $\boldsymbol{X}_m$ 表示标签系统 F 中缺失值元素组成的集合，因此对系统中的 $\boldsymbol{X}$，有 $\boldsymbol{X}=\boldsymbol{X}_{\Omega}+\boldsymbol{X}_m$，在进行缺失值处理时，常用方法主要包括将缺失值全部置 0 或者值平均值，并按式(3)张量分解方法来依次迭代计算缺失值 $\boldsymbol{X}_m^0,\boldsymbol{X}_m^1,\boldsymbol{X}_m^2,\cdots$，则数据应满足式(11)：

$$\min_{\boldsymbol{X}_m^{t+1}}\|(\boldsymbol{X}_{\Omega}+\boldsymbol{X}_m^t)-\boldsymbol{X}^{t+1}\|^2 \quad (11)$$

其中，$\boldsymbol{X}_{\Omega}$就是输入张量 $\boldsymbol{X}$ 中的非零值元素集合，$\boldsymbol{X}_m^t$ 是 $\boldsymbol{X}^t$ 中的缺失值部分第 t 次迭代的解.

$\boldsymbol{X}^{t+1}$是下一次迭代值. 当连续两次迭代差值达到了收敛要求时，即连续两次迭代后的 $\boldsymbol{X}^{t+1}$ 与 $\boldsymbol{X}^t$ 之间的变化已经很小时，迭代终止，此时得到的缺失值最近似于标签系统中的 $\boldsymbol{X}_{ijk}$. 而由于每次迭代前都把缺失值部分用前一次迭代的数据进行了填充，所以缺失值部分得到了逐次优化，最后就可以对缺失值部分进行较合理的预测. 详细的证明可参看文献[12].

4.4　TDD 算法

三部图张量分解算法(TDD)将社会标签数据集描述为张量 $\boldsymbol{X}_{ijk}$，迭代计算式(3)中的 $\boldsymbol{U}_{ij}$、$\boldsymbol{V}_{ik}$ 和 $\boldsymbol{W}_{jk}$ 的最优解并得到 $\boldsymbol{X}_{ijk}$ 的最优解，其算法表述如下.

这个算法包含 3 个输入参数：存在缺失值的张量数据、收敛因子 ε 以及最大迭代次数，算法输出是预测张量 $\boldsymbol{Y}_{ijk}$.

首先算法初始化缺失值部分，我们将每个缺失值元素都初始化为张量三维的平均值，经过比较，这种初始化方法比初始值全部赋 0 的效果更好些. 从第 6 行到第 12 行是求预测值迭代计算部分，当满足第 12 行中的收敛条件时算法停止迭代计算，第 14 行给出结果并返回.

算法 1.　三部图张量分解算法.

输入：

a：Item-Tag-User Tensor $\boldsymbol{X}$ with missing values；

b：ε for convergence test；

c：max-iteration

输出：

Predicated Tensor $\boldsymbol{Y}$

1. Initialize Missing Value part of $\boldsymbol{X}$ as $\boldsymbol{Y}_0$；
2. Do $t=0$ to max-iteration
3. $\quad\boldsymbol{Y}^t=\boldsymbol{X}_{\Omega}+\boldsymbol{Y}_m^t$；
4. update $\boldsymbol{U}$ by (7)；
5. update $\boldsymbol{V}$ by (8)；
6. update $\boldsymbol{W}$ by (9)；
7. input $\boldsymbol{U},\boldsymbol{V},\boldsymbol{W}$ to Compute (3) as J^{t+1}；
8. Do Step 6 to Step 7
9. $\quad$if $\|J^{t+1}-J^t\|<\varepsilon$, go to Line 11；
10. End Do；
11. Compute $\boldsymbol{Y}_{ijk}=\boldsymbol{U}_{ij}\boldsymbol{V}_{jk}+\boldsymbol{V}_{jk}\boldsymbol{W}_{ik}+\boldsymbol{U}_{ij}\boldsymbol{W}_{ik}$；
12. return $\boldsymbol{Y}_{ijk}$；

算法空间复杂度为 $O(N_iN_jN_k)$，时间复杂度为 $O(N_iN_jN_k\times m)$，其中 m 为迭代次数. 实验数据表明，$m\approx 20$ 左右，算法得到收敛.

5　实验结果及分析

本文实验包括两部分：第一部分实验是三部图张量分解模型与三部图之间的预测效率比较；第二部分实验中，我们使用三部图张量分解方法对社会标签进行预测. 实验所用数据集为 Last.fm，Bibsonomy 和 Movielens. 考虑到性能比较的一致性，本文主要选择当前常用的张量分解推荐算法以及作为 Benchmark 的 Folkrank 进行比较，其张量分解方法包括 Tucker[10]、ParaFac[10]、Pairwise[10]、Lower-2D (LOTD-2D)[12]算法.

5.1　数据集

本文对 3 个数据集分别选择了稍大以及稍小数据集进行比较，数据集列在表 2 中，可以看到各个实验数据集的详细信息. 每一个数据集可以看做一个大小为 $item\times tag\times user$ 的张量.

表 2　实验数据集

数据集	*item*	*tag*	*user*
Last.fm	100	157	280
Last.fm	100	157	560
Movielens	98	200	246
Movielens	98	200	592
Bibsonomy	160	101	204
Bibsonomy	160	101	2040

表 2 中 Last.fm 数据集来自 Last.fm 网站的网页数据，该网站为其用户提供个性化的多媒体服务，同时允许用户对多媒体文件添加标签，数据集转换为张量大小分别为 $100\times157\times280$，$100\times157\times560$；

Movielens 数据集来自 Movielens 网站，这是一个在线电影推荐系统，数据集张量分别为大小为 98×200×246，98×200×592；Bibsonomy 数据集下载自 bibsonom-y.org 网站，小数据集张量大小为 160×101×204，大数据集张量大小为 160×101×2040.

5.2 算法性能比较指标

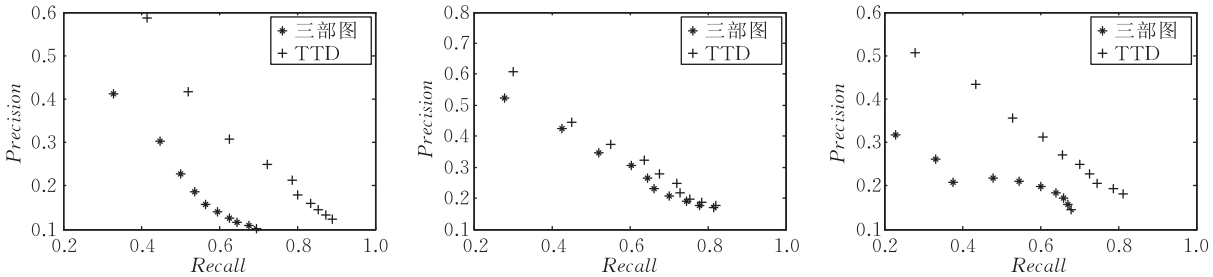
实验中采用传统的精度-召回率方法，依据前 N 个坐标值的曲线状态来衡量算法对标签的预测质量. 在实验中，对于每个用户，随机屏蔽一标签列 $Post(i,k)$，该 $Post(i,k)$ 数据作为缺失值，对每个屏蔽的标签列 $Post(i,k)$，我们将预测结果排序，N_j 个值对应 N_j 个 tags，实验中依次选取排序最靠前的 $N=1,2,\cdots,10$ 个值，并设定与这些选取值有关的标签的预测值为正值“positive”. 精度和召回率按如下方法计算：

$$Precision(T_{\text{test}},N)=\frac{|t\in Top(i,k,N)\cap t\in Post(i,k)|}{N\times|Post|} \tag{12}$$

$$Recall(T_{\text{test}},N)=\frac{|t\in Top(i,k,N)\cap t\in Post(i,k)|}{|t\in Post(i,k)|\times|Post|} \tag{13}$$

5.3 张量分解三部图模型与三部图有效性比较

本文在 3 个真实小数据集上对三部图张量分解模型（TTD）与三部图模型进行比较，实验以 Folkrank[11] 算法作为基准进行标签推荐，将三部图模型以及 TTD 模型作为 Folkrank 的矩阵输入，进行标签预测，并比较 TTD 与三部图预测结果，通过实验，α 分别取值 0.5，0.8，0.9 以及 1，其中以 0.8 的实验效果最好，因此本文实验结果均为 $\alpha=0.8$ 的结果. 实验结果如图 4.



(a) Movielens数据集性能比较 (b) Bibsonomy数据集性能比较 (c) Last.fm数据集性能比较

图 4 TTD 与三部图推荐效果比较

从图 4 中可以看到三部图张量分解模型（TTD）的预测效果比三部图更好. 这说明在同样的时间和空间消耗下，三部图张量分解模型能更准确的描述社会标签系统中的相互关系.

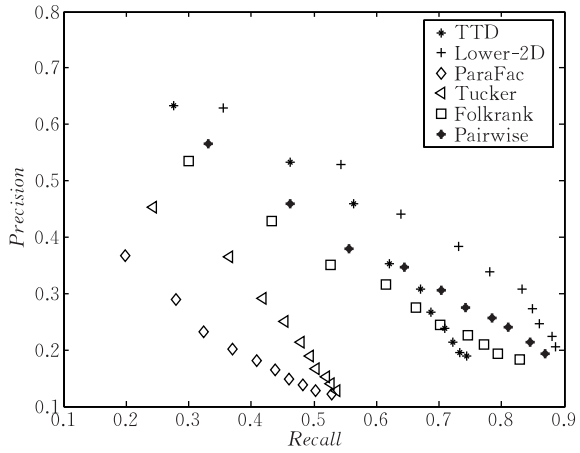
5.4 TTD 标签预测算法性能比较

在本节中，我们比较三部图张量分解算法和一些经典张量分解推荐算法的有效性，这些算法包括 Folkrank、Tucker、ParaFac、Pairwise 和 Lowor-2D，同样，$\alpha=0.8$，实验结果如图 5(a)～(f)所示.

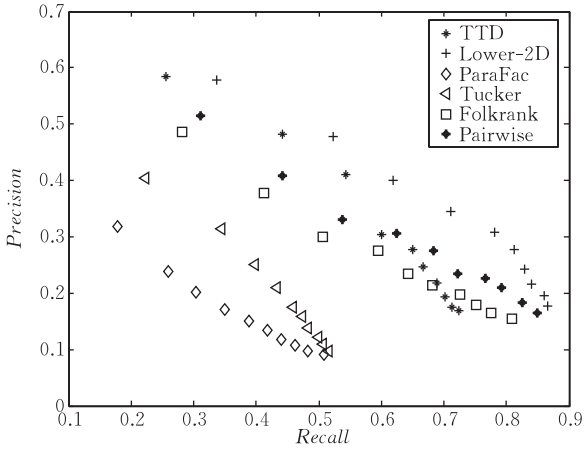
图 5(a)、(b)分别表示 Last. fm 在小数据集以及较大数据集上的实验结果；图 5(c)、(d)表示 Movielens 在小数据集以及较大数据集上的实验结果；图 5(e)、(f)表示 Bibsonomy 在小数据集以及较大数据集上的实验结果. 实验结果表明：在 Bibsonomy 和 Movielens 数据集上，三部图张量分解算法具有更好的 *Precision-Recall* 曲线. 在 Last. fm 数据集上，预测的前 1～3 标签有效性比 Lowor-2D 高，但是 4～10 中其它的数据比 Lowor-2D 低；与 Folkrank 相比，前 1～7 相对较高，而 8～10 较低；与其它传统算法相比则性能更好，其主要原因是用户、标签、资源三者分布关系对不同算法的影响所引起的，这部分内容在后续论文中进行论证及说明. 在图 5 较大数据集和较小数据集的实验中，我们将 Last. fm 和 Movielens 较大数据集用户数扩大一倍，在 Bibsonomy 中我们将用户数据量增大一个数量级，从实验结果可以看到，较大数据集的预测精度不如小数据集预测精度，而 Bibsonomy 的大数据集预测精度衰减更大，其主要原因是由于用户数目增大，使得数据稀疏性增大，因此实验结果不如小数据量结果，如图 5(b)、(d)、(f)所示.

从结果来看，而传统的张量分解方法 Truck 和 ParaFac 方法，这两种算法的准确性比其他算法低，但是 Folkrank 总能达到较高的精度.

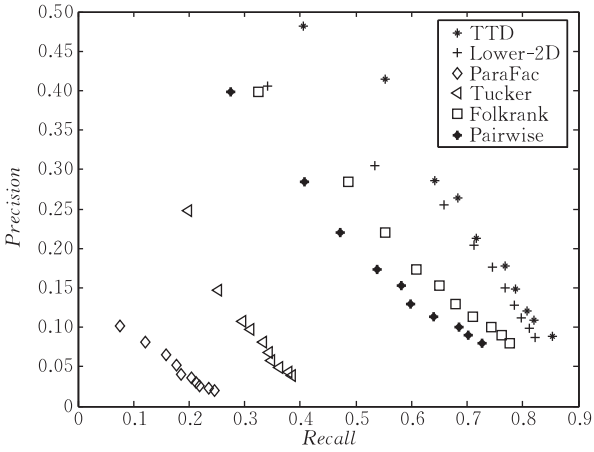
在表 3 中，我们比较了每个算法的时间，结果表明三部图分解模型（TTD）获得了高效率，但相比 Lower2D 等算法具有更高的算法复杂性.



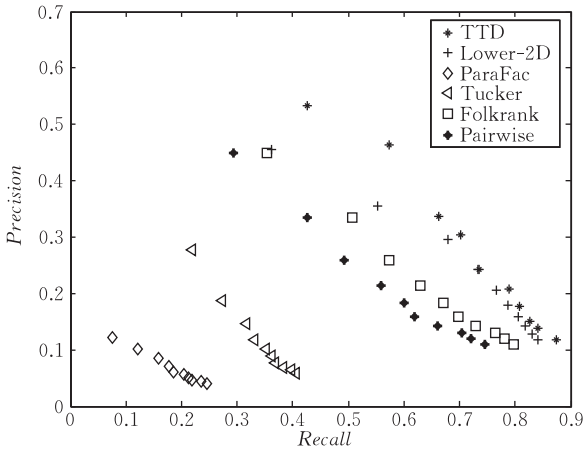
(a) Last.fm与其它算法性能比较(小数据集)



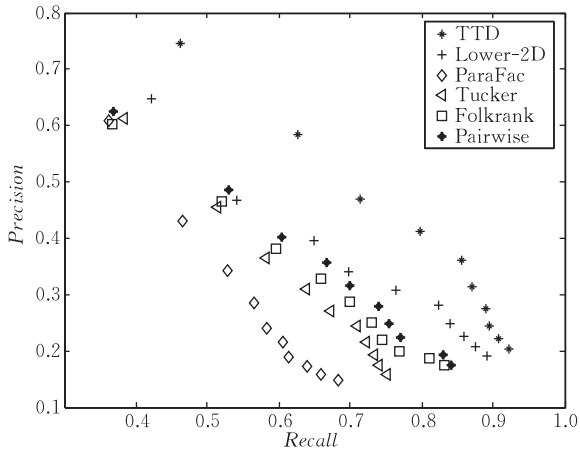
(b) Last.fm与其它算法性能比较(较大数据集)



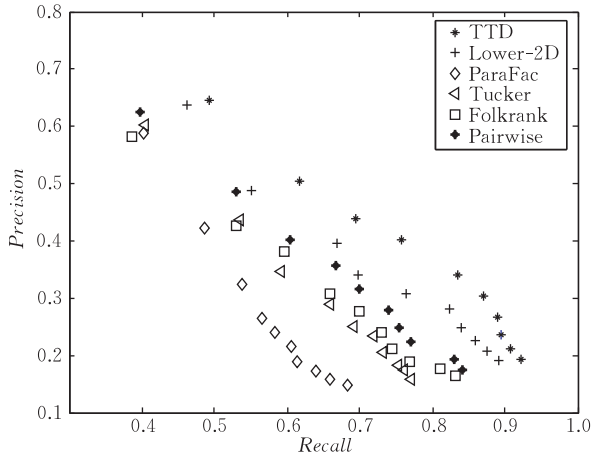
(c) Movielens与其它算法性能比较(小数据集)



(d) Movielens与其它算法性能比较(较大数据集)



(e) Bibsonomy数据集和其它算法性能比较(小数据集)



(f) Bibsonomy较大数据集和其它算法性能比较(较大数据集)

图 5　TTD 算法在不同数据集上的性能比较

表 3　运算时间比较(Movielens)

算法	时间/ms
TTD	87.6841
Lower-2D	8.9909
ParaFac	339.4411
Tucker	55.8750
Pairwise	1.59E+004
Folkrank	1.2401E+003

6　结　论

社会标签系统中的数据是非常稀疏的,并且存在缺失值.虽然三部图可以表示社会标签系统中的关系,但是处理社会标签系统中的稀疏和带缺失值的数据精确度不高,而张量分解是一种用低阶表示原始、稀疏数据的方法.

在本文中,我们提出基于三部图三维张量分解模型,获取社会标签系统中 item,tag 和 user 之间的潜在关系.在 3 个数据集上,我们以 Folkrank 为基准,比较了三部图和 TTD 模型的有效性,同时通过和当前流行的社会标签预测算法进行了比较,实验结果表明 TTD 模型在 recall/precision 精度上比其他算法的预测性能更好.

参 考 文 献

[1] Heymann P, Koutrika G, Garcia-Molina H. Can social bookmarking improve Web search//Proceedings of the International Conference on Web Search and Web Data Mining (WSDM'08). Standford, CA, USA, 2008: 195-206

[2] Symeonidis P, Nanopoulos A, Manolopoulos Y. A unified framework for providing recommendations in social tagging systems based on ternary semantic analysis. IEEE Transactions on Knowledge and Data Engineering, 2009, 22(2): 179-192

[3] Xu Ling, Wu Xiao, Li Dong, Yan Ping, Xiao Baohai. Comparison study of Internet recommendation system. Journal of Software, 2009, 20(2): 350-362. (in Chinese)
(许玲, 吴潇, 李东, 阎平, 晓保海. 互联网推荐系统比较研究. 软件学报, 2009, 20(2): 350-362)

[4] Liu Kai-Peng, Fang Bin-Xing. A novel page ranking algorithm based on social annotations. Chinese Journal of Computers, 2010, 33(6): 1014-1023(in Chinese)
(刘凯鹏, 方滨兴. 一种基于社会性标注的网页排序算法. 计算机学报, 2010, 33(6): 1014-1023)

[5] Zhang Zi-Ke, Liu Chuang. Hypergraph model of social tagging networks. Journal of Statistical Mechanics, 2010, P100005

[6] Hofmann T. A similarity-based probability model for latent semantic indexing. ACM Transactions on Information Systems, 2004, 22(1): 89-115

[7] Cohn D, Hofmann T. The missing link—A probabilistic model of document content and hypertext connectivity//In: Leen TK, Dietterich TG, Tresp V eds. Neural Information Processing Systems Conference (NIPS'00). MIT Press, 2000: 430-436

[8] Mika P. Ontologies are us: A unified model of social networks and semantics. Web Semantics: Science, Services and Agents on the World Wide Web Archive, 2007, 5(1): 5-15

[9] Symeonidis P, Nanopoulos A, Manolopoulos Y. Tag recommendations based on tensor dimensionality reduction//Proceedings of the 2008 ACM Conference on Recommender Systems. Lausanne, Switzerland, 2008: 43-50

[10] Rendle S, Schmidt-Thieme L. Pairwise interaction tensor factorization for personalized tag recommendation//Proceedings of the 3rd ACM International Conference on Web Search and Data Mining (WSDM10). New York, USA, 2010: 81-90

[11] Rendle S, Marinho B, Nanopoulos A, Thieme L. Learning optimal ranking with tensor factorization for tag recommendation//Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Paris, France, 2009: 727-736

[12] Cai Yuan-Zhe, Zhang Miao, Ding Chris, Chakravarthy Sharma. Low-order tensor decompositions for social tagging recommendation//Proceedings of the 4th ACM International Conference on Web Search and Data Mining (WSDM'11). Hong Kong, China, 2011: 695-704

LIAO Zhi-Fang, born in 1968, Ph. D., associate professor. Her research interests include dada mining, recommendation system.

LI Ling, born in 1987, M. S.. Her research interests include dada mining, recommendation system.

LIU Li-Min, born in 1976, Ph. D, lecturer. Her research interests include dada mining, network.

LI Yong-Zhou, born in 1971, Ph. D., associate professor. His research interests include pattern recognition, society network.

Background

We study the social tagging system predication with tensor decomposition based on tripartite model.

There are three elements{user, tag, item} in social tagging system, users use tags to present their interests in the system, such as use cartoon to tag the film he likes. But the data is quite sparse, and three dimension relationships among social tagging system cannot be fully described. The current work focuses on the methods or models to transfer complex three-dimension into two-dimension relationship so that efficient traditional models can be used in social tagging systems. The common used model is tripartite model, but unfortunately this model missing some useful values among three-dimension.

The paper presents the model based on tripartite model to meet the requirement of missing values among three objects in social tagging system, on the other hand, to solve the problem in data sparsity, the paper presents the tensor decomposition method. Combined with the tripartite model and tensor decomposition method, the Tensor Decomposition Tripartite Model is used to do the social tagging predication.

The research team focuses on the research area on data mining for several years, and now focuses on recommendation system with data mining technologies.

This work is supported by the National Natural Science Foundation of China under Grant No. 61073105 and Specialized Research Fund for the Doctoral Program of Higher Education of China under Grant No. 20100480950.