文章编号: 1007-5321(2015)03-0034-05 DOI: 10.13190/j.jbupt.2015.03.004

基于标签和因子分析的协同推荐方法

蔡国永， 吕 瑞， 樊永显
(桂林电子科技大学 广西可信软件重点实验室，桂林 541004)

摘要: 根据在线社区中群体的历史行为进行物品(或信息)推荐是当前研究热点之一，传统推荐算法都面临数据稀疏性问题的挑战. 针对传统推荐算法知识表示的局限性进行了研究，提出了一种基于标签系统的用户行为知识表示法，把用户在物品上历史行为的统计，转化为对用户在物品标签上的统计，从而缓解数据稀疏的情况. 为了降低标签维度过高导致的计算复杂性问题，提出了采用因子分析法，抽取出潜在重要且稳定的特征因子向量来最终表示用户的历史行为，并据此度量用户行为在特征因子向量上的相似性. 最后采用协同过滤的思想给出了一种新的协同推荐方法. 通过在真实数据集上的大量对比实验，表明该方法在处理具有稀疏性的数据集时，总是能保持更高且更稳定的推荐准确率.

关 键 词: 推荐系统; 数据稀疏性; 标签系统; 因子分析; 评分预测
中图分类号: TP301.6 **文献标志码:** A

Collaborative Recommendation Method Based on Tags and Factor Analysis

CAI Guo-yong， LÜ Rui， FAN Yong-xian
(Guangxi Key Laboratory of Trusted Software, Guilin University of Electronic Technology, Guilin 541004, China)

Abstract: Item (or information) recommendation is one of hot research topics currently. However the issue of sparseness in dataset challenges all traditional recommendation algorithms. Limitations of knowledge representation in traditional recommendation algorithms were studied. The tag-system-based knowledge to represent information of each user's behavior was proposed. That it the account on user's behavior on items is transferred to an account on a user's behavior on tags. To decrease the computation complexity on high dimensional tag-based datasets, a factor analysis method was taken to extract those most important latent factors to represent users' behaviors. Based on each user's representing vector of latent factors, a new way was given to compute similarities among users. By incorporating this similarity measure, a new collaborative recommendation method with low sensitivity to sparseness was built to meet the need of practical and dynamic datasets. Experiments were carried on real-world datasets to compare the proposed method with other state-of-the-art collaborative filtering and matrix factorization based recommendation methods. It is shown the proposed method can achieve better prediction accuracy while keeps a lower sensitivity to sparseness.

Key words: recommendation system; dataset sparseness; tag system; factor analysis; rating prediction

基于用户和项的 User-Item 评分矩阵模型是推荐系统领域里最为经典和有效的数据模型. 众多的

收稿日期: 2015-01-01
基金项目: 国家自然科学基金项目(61462018)；广西高校高水平创新团队及卓越学者计划资助项目；广西可信软件重点实验室基金项目(kx201202)
作者简介: 蔡国永(1971—)，男，教授，博士，E-mail: ccgycai@guet.edu.cn.

现实推荐问题都可以通过转化为 User-Item 评分矩阵进行建模和处理. 然而,由于 User 和 Item 的数据规模通常处于一个较大的数量级,而 User-Item 矩阵中有效的评分数据往往很少(不足 5%),这就产生了评分数据集的稀疏性问题[1-2].

针对 User-Item 矩阵的稀疏性问题,学者们从不同方面进行了研究,提出了一些应对数据稀疏性的方法. 在 User-Item 矩阵外引入其他相关信息来弥补数据稀疏,如利用社会网络[3]、信任网络[4]、矩阵分解[5]、扩散[6]、内部一致[7]、转移相似性[8]等方法发掘用户和物品潜在的关联关系等. 这些方法同样存在着一些不足,如涉及的辅助信息太多、推荐算法时间复杂度较高以及处理的数据集往往是静态的等.

标签(tag)是一种对网络资源进行标记、分享及识别的工具,标签信息相对稳定. 通过将商品抽象为标签,建立用户对标签的偏爱或兴趣分布,一方面可以了解用户的喜好特征,另一方面可以减弱原始数据集稀疏程度对推荐算法的影响. 因子分析方法可以抽取决定性因子,有效降维和除噪. 因此笔者提出一种基于标签和因子分析方法的推荐方法.

1　推荐算法框架

推荐算法框架流程如图 1 所示. 图 1 中步骤①表示构建 User-Tag 矩阵. User-Tag 矩阵是一个以用户为行、标签为列的 2 维矩阵,它由原始的 User-Item 矩阵变换而来,矩阵中的每一行表示一个用户对所有标签的兴趣分布. 步骤②表示学习因子模型,通过对步骤①获得的 User-Tag 矩阵(一个样本)进行因子分析,挖掘出样本中不同兴趣分布背后隐藏的一些潜在因子并建立一种描述潜在因子与原标签关系的因子模型. 该因子模型是一个经过数据除噪、降维后的 Tag-Factor 矩阵. Tag-Factor 矩阵以标签为行、因子为列,每一列表示一个因子(Factor)与原始标签($\mathrm{Tag}_1 \sim \mathrm{Tag}_m$)的线性组合关系. 步骤③表示利用因子模型 Tag-Factor 矩阵为训练集中的所有用户建立新的 User-Factor 向量,这些向量的集合构成 User-Factor 矩阵. 步骤④计算不同用户向量之间的相似度. 基于获得的目标用户的"邻居",可以将"邻居"曾经打分较高且目标用户没有评分过的物品对目标用户进行推荐(步骤⑤),也可以定义一个基于目标用户"邻居"的评分预测方法来对目标用户未评分的物品进行评分预测(步骤⑥). 式(1)描述了一种常用的根据目标用户"邻居"预测目标用户未知评分的计算方法.

$$\widetilde{R}_{ui} = \bar{r}(u) + \frac{1}{k}\sum_{v\in N_k(u,i)} \mathrm{sim}(u,v)\,(r(v,i) - \bar{r}(v)) \tag{1}$$

其中: $\widetilde{R}_{ui}$ 为用户 u 对物品 i 的预测评分,$\bar{r}(u)$ 为用户 u 对其所评分过的所有物品的平均分,$r(v,i)$ 为用户 v 对物品 i 的评分,$N_k(u,i)$ 为目标用户"邻居"中对 i 有评分的用户集合,k 为 $N_k(u,i)$ 中元素的个数.

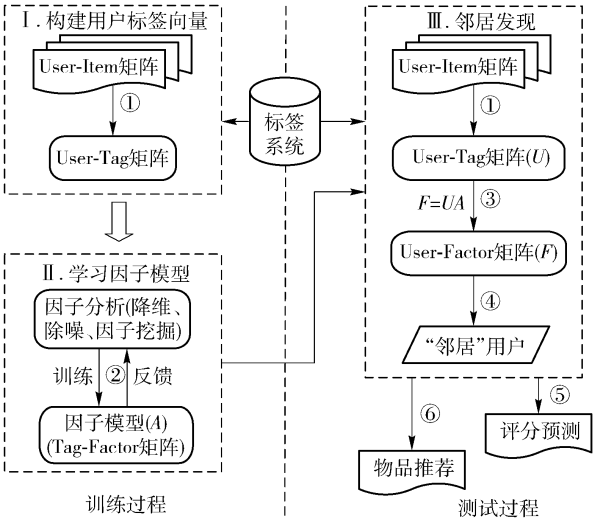

图 1　算法流程

2　潜在因子挖掘和邻居发现

2.1　User-Tag 矩阵构建

假设某标签系统为某类物品预先定义了 p 个标签. 用户 u 对物品 i 有过评分且物品 i 对应有 p 个标签中的 n 个标签,则物品 i 对应的每个标签将获得关注度 $1/n$. 由此用户 u 对标签 t 的关注度可以表示为

$$r_{ut} = \frac{\sum_{i\in D_k(u)} \frac{\mathrm{sgn}(u,i,t)}{N(i)}}{k} \tag{2}$$

其中: $\mathrm{sgn}(u,i,t) = \begin{cases} 1, & t \text{ 为物品 } i \text{ 的一个标签} \\ 0, & \text{其他} \end{cases}$, r_{ut} 为用户 u 对标签 t 的关注度,$N(i)$ 为物品 i 的标签个数,$D_k(u)$ 为用户 u 有过评分的物品的集合,k 为集合 $D_k(u)$ 中元素的个数. 用户 u 对所有标签的关注度可用 User-Tag 向量 $\boldsymbol{U}_u$ 表示:

$$\boldsymbol{U}_u = (r_{u1}, r_{u2}, \cdots, r_{up}) \tag{3}$$

多个 User-Tag 向量的集合构成了 User-Tag 矩阵.

2.2 潜在因子挖掘

设 $\boldsymbol{X}=(X_1,X_2,\cdots,X_p)^{\mathrm{T}}$ 是可观测的 p 维随机变量，则因子模型可用一个二维矩阵 $\boldsymbol{A}$ 表示，$\boldsymbol{A}=(a_{ij})_{p\times m}$ 且 $\boldsymbol{A}$ 满足

$$\left.\begin{aligned} X_1-\mu_1&=a_{11}f_1+a_{12}f_2+\cdots+a_{1m}f_m+\varepsilon_1\\ X_2-\mu_2&=a_{21}f_1+a_{22}f_2+\cdots+a_{2m}f_m+\varepsilon_2\\ &\vdots\\ X_p-\mu_p&=a_{p1}f_1+a_{p2}f_2+\cdots+a_{pm}f_m+\varepsilon_p \end{aligned}\right\} \quad (4)$$

其中：$\mu_1,\mu_2,\cdots,\mu_p$ 为随机变量 $X_1,X_2,\cdots,X_p$ 的数学期望；$f_1,f_2,\cdots,f_m(m<p)$ 为公共因子，$\varepsilon_1,\varepsilon_2,\cdots,\varepsilon_p$ 为特殊因子，它们都是不可观测的随机变量. 式(4)也可以写成

$$\boldsymbol{X}=\boldsymbol{\mu}+\boldsymbol{AF}+\boldsymbol{\varepsilon}\approx\boldsymbol{\mu}+\boldsymbol{AF} \quad (5)$$

设 $\boldsymbol{X}_{(1)},\boldsymbol{X}_{(2)},\cdots,\boldsymbol{X}_{(n)}$ 是一组 p 维个体组成的样本，其中 $\boldsymbol{X}_{(i)}=(x_{i1},x_{i2},\cdots,x_{ip})^{\mathrm{T}}$，则样本协方差矩阵的估计量为

$$\boldsymbol{S}=\frac{1}{n-1}\sum_{i=1}^{n}(\boldsymbol{X}_{(i)}-\overline{\boldsymbol{X}})(\boldsymbol{X}_{(i)}-\overline{\boldsymbol{X}})^{\mathrm{T}} \quad (6)$$

其中 $\overline{\boldsymbol{X}}$ 为期望，即 $\overline{\boldsymbol{X}}=\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{X}_{(i)}$

由式(4)的性质[9]可得

$$\boldsymbol{S}\approx\boldsymbol{AA}^{\mathrm{T}} \quad (7)$$

设样本的协方差阵 $\boldsymbol{S}$ 的特征值为 $\lambda_1\geqslant\lambda_2\geqslant\cdots\geqslant\lambda_p\geqslant0$，相应的单位正交特征向量为 $\boldsymbol{l}_1,\boldsymbol{l}_2,\cdots,\boldsymbol{l}_p$，从而当前 m 个特征值的总和远大于最后 $(p-m)$ 个特征值的总和时有

$$\begin{aligned}\boldsymbol{S}&=\lambda_1\boldsymbol{l}_1\boldsymbol{l}_1^{\mathrm{T}}+\cdots+\lambda_m\boldsymbol{l}_m\boldsymbol{l}_m^{\mathrm{T}}+\lambda_{m+1}\boldsymbol{l}_{m+1}\boldsymbol{l}_{m+1}^{\mathrm{T}}+\cdots+\\ &\lambda_p\boldsymbol{l}_p\boldsymbol{l}_p^{\mathrm{T}}\approx\lambda_1\boldsymbol{l}_1\boldsymbol{l}_1^{\mathrm{T}}+\cdots+\lambda_m\boldsymbol{l}_m\boldsymbol{l}_m^{\mathrm{T}}=\boldsymbol{AA}^{\mathrm{T}}\end{aligned} \quad (8)$$

则 $\boldsymbol{A}=(a_{ij})_{p\times m}=(\sqrt{\lambda_1}\boldsymbol{l}_1,\sqrt{\lambda_2}\boldsymbol{l}_2,\cdots,\sqrt{\lambda_m}\boldsymbol{l}_m)$ 即为因子模型的解.

2.3 “最近邻居”用户计算

假设数据集的因子模型为 $\boldsymbol{A}$，为 User-Item 评分矩阵建立的 User-Tag 矩阵为 $\boldsymbol{U}$，则 User-Factor 矩阵 $\boldsymbol{F}$ 可表示为

$$\boldsymbol{F}=\boldsymbol{UA}=\begin{bmatrix}\sigma_{11}&\sigma_{12}&\cdots&\sigma_{1m}\\ \sigma_{21}&\sigma_{22}&\cdots&\sigma_{2m}\\ \vdots&\vdots&&\vdots\\ \sigma_{n1}&\sigma_{n2}&\cdots&\sigma_{nm}\end{bmatrix} \quad (9)$$

其中：$\sigma_{ij}=\sum_{k=1}^{p}r_{ik}a_{kj}$. 矩阵 $\boldsymbol{F}$ 中每一行表示一个 User-Factor 向量. 针对 User-Factor 矩阵 $\boldsymbol{F}$，可以利用不同的方法来计算任意 2 个行向量的距离(用户的相似度)，如 CityBlock 距离、Pearson 距离等. 根据到不同用户间的距离则可获得和目标用户最相似的 N 个邻居用户.

3 实验结果与分析

为验证所提出方法在评分预测方面的准确性以及对稀疏数据集的低敏感性，选取了推荐领域 MovieLens 1 M(http://www.datatang.com/data/44521)数据集进行实验(注：在其他数据集上得到的实验结果类似，不再重复结果). 该数据集包含 2000 年 3 900 个匿名用户对 6 040 部电影的 1 000 209 个评分(稀疏度为 4.2%)，标签的种类共有 18 种，每部电影由 1 个或多个标签构成. 数据集分为 2 个部分：训练集包含 3 000 个用户，测试集包含剩余的 3 040 个用户. 训练集用来学习因子模型，然后用训练集获得的因子模型对测试集中的用户进行 User-Factor 向量生成，并基于生成的 User-Factor 向量计算用户的 Top-N 个邻居及利用 Top-N 邻居对目标用户进行评分预测. 为方便起见，下面将提出的方法记为 Native 方法.

3.1 实验一：因子个数选择

在建立因子模型的过程中，为了确定应当从多少个样本用户中抽取多少个因子来建立因子模型，进行了以下实验. 实验结果如图 2 所示.

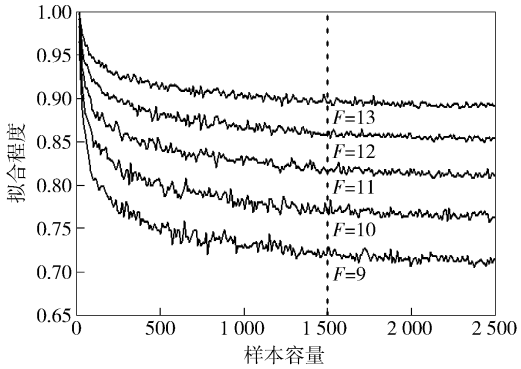

图 2 拟合程度和样本数量关系

图 2 中横轴代表从训练集中随机选择的用户数，纵轴表示得到的因子模型对训练集中选择的样本数据的拟合程度，K 表示指定的因子个数. 从图 2 中可知，拟合程度随着样本数的增多而下降，其原因是当因子个数一定时，随着数据量的增大，拟合所有数据的难度随之增加. 当样本数增加到 1 500 个时，因子的拟合程度基本收敛. 一般而言，为了防止过拟合，通常选择拟合程度为 80%. 因此使用的训练

样本为从 3 000 个用户中随机选择 1 500 个，挖掘的因子个数选定为 $K=11$.

3.2　实验二：准确性评估

为了对比 Native 方法和其他方法在未知评分预测方面的差异，采用了均方根误差（RMSE，root-mean-square error）评估方法. RMSE 的表达式为

$$\sigma=\sqrt{\frac{\sum_{1}^{n}(r_i-\hat{r}_i)^2}{n}}$$

其中：r_i 为物品的真实评分，$\hat{r}_i$ 为对物品的预测评分，n 为实验中预测的物品的样本数.

实验中把 Native 方法与目前最好的一些改进方法进行了对比，对比的方法包括协同过滤方法（CityBlock 距离相似度、Pearson 距离相似度）、矩阵分解方法（SGD 方法、SVD ++ 方法）及为目标用户随机产生邻居的方法（记为 Random）. 其中，传统的协同过滤方法通过计算用户的 User-Item 向量的距离来寻找邻居；矩阵分解方法则通过建立拟合已有评分数据的矩阵模型来预测未知的评分；Random 方法提供了一种基于邻居用户进行评分预测的方法的基础对照. 本实验中用到的其他参数和方法设置：近邻数 $N=5$，相似度量采用 CityBlockSimilarity，评分预测采用式（1）.

实验中采用的其他方法的实现均来自 Apache Mahout（http://www.mahout.apache.org/），其中协同过滤方法以 Random 方法参考的邻居用户数和 Native 方法相同，矩阵分解方法 SGD 和 SVD ++ 的迭代次数为5 000. 每种方法重复计算 10 次并取平均值作为最后的结果. 不同算法的实验结果如图 3 所示.

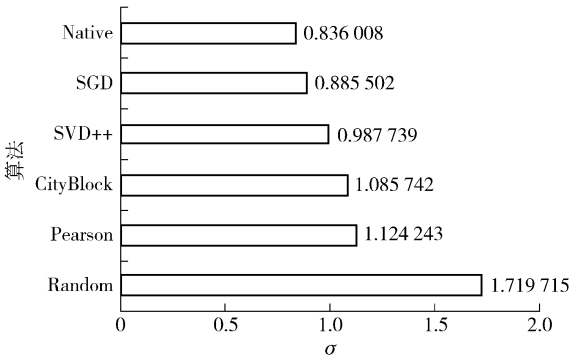

图 3　不同算法下的 RMSE 对比

由图 3 的实验结果可知，Random 方法效果最差，矩阵分解方法 SGD、SVD ++ 要优于采用 CityBlock、Pearson 相似性度量的传统协同过滤方法，而 Native 方法则具有最低的 RMSE 值，相比其他几种方法具有更高的算法准确率.

3.3　实验三：敏感性评估

算法的敏感性主要是指数据集的稀疏程度对算法评分预测准确性的影响. 好的推荐算法应当能适应不同稀疏程度的数据集. 为了验证 Native 方法在处理稀疏程度不断增加的真实数据集时的优越性，进行了以下对比实验. 首先对原始数据集中的评分数据进行随机抽样，使原始数据集的稀疏程度分别降低到 0.5%、1.0%、1.5%、2.0%、2.5%、3.0%、3.5%、4.0%；然后对比不同算法在不同数据稀疏度情况下准确率的变化情况. 实验结果如图 4 所示.

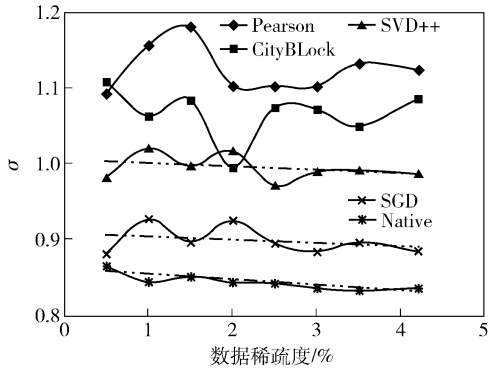

图 4　数据敏感性对比

图 4 中横轴表示数据集的稀疏程度（从左到右稀疏程度逐渐减弱），纵轴表示在相应稀疏度条件下算法的 σ 值，3 条虚线分别表示对 SVD ++、SGD 和 Native 方法结果的线性拟合. 从图 4 中可以看出，相比协同过滤方法和矩阵分解方法 Native 方法在不同稀疏度条件下均具有更低的 RMSE 值，即更好的准确率. 矩阵分解方法随着数据稀疏性的增加算法的准确率逐渐下降（由 SVD ++ 和 SGD 方法的拟合曲线可以看出），说明矩阵分解方法是一种对数据集稀疏程度敏感的方法. 而在实际应用中，随着用户和商品数量的不断增加，User-Item 矩阵中有效评分数增长的速度远小于矩阵所能表示的评分的增长速度，从而 User-Item 评分矩阵会变得越来越稀疏（如淘宝数据集的稀疏度已达百万分之一），但矩阵中大部分用户的评分数则会随着时间的延长而有所增加. 基于文中构建用户关于标签兴趣的方法可知，用户关注的物品越多计算用户关于标签的兴趣分布与用户的真实兴趣分布也越接近，因而利用因子分析方法计算的邻居用户也越精确. 因此，算法

的RMSE值也应当越来越低(当用户关于标签的兴趣分布未达到真实分布时)或者保持基本稳定(当用户关于标签的兴趣分布已逼近真实的兴趣分布时). 针对上面的分析,实验中Native方法在用户具有较多的评分情况下(对应于图4中数据稀疏度较低的情况,如4.0%,而在实际动态增长的数据集中则对应于稀疏度较高的情况)具有较低的RMSE值验证了这一点. 所以在实际动态增长的数据集中,Native方法将会随着数据集的稀疏程度的增加,准确率略有提高或者保持基本稳定.

4 结束语

针对实际应用中数据集稀疏度动态增加的特点,进行了深入的分析和研究,提出了一种基于标签和因子分解的协同推荐新方法,并通过在大量真实世界数据集上的实验,证实了所提出的方法在具有较好算法准确率的同时,对数据集的稀疏性具有较低的敏感性. 然而,所提出的方法也存在一定的局限性,当数据集涉及的时间跨度较大时,标签系统也可能会有较大变化,基于标签的兴趣建模方法的准确性也会逐渐降低,从而影响算法的准确性. 通过对标签系统进行及时跟踪,及时更新用户的标签兴趣变化将有助于保持预测的准确性.

参考文献:

[1] Tang Xiangyu, Zhou Jie. Dynamic personalized recommendation on sparse data [J]. IEEE Trans. on Knowledge and Data Engineering, 2013, 25(12): 2895-2899.

[2] Chadha H, Jain A, Singh A, et al. Recommendation system for highly sparse datasets: a hybrid approach [C] // International Conference on Technology and Business Management March. Dubai: American University in the Emirates, 2014: 24-26.

[3] Yang Xiwang, Guo Yang, Liu Yong. Bayesian-inference-based recommendation in online social networks [J]. Parallel and Distributed Systems, IEEE Transactions on, 2013, 24(4): 642-651.

[4] 周超, 李博. 一种基于用户信任网络的推荐方法 [J]. 北京邮电大学学报, 2014, 37(4): 98-102.
Zhou Chao, Li Bo. A recommendation method based on user trust network [J]. Journal of Beijing University of Posts and Telecommunications, 2014, 37(4): 98-102.

[5] Lin Chiajen, Kuo Tsungting, Lin Shoude. A content-based matrix factorization model for recipe recommendation [C] // Advances in Knowledge Discovery and Data Mining. Switzerland: Springer International Publishing, 2014: 560-571.

[6] Pan Ye, Cong Feng, Chen Kailong, et al. Diffusion-aware personalized social update recommendation [C] // Proceedings of the 7th ACM Conference on Recommender Systems. New York: Associate for Computing Machinery, 2013: 69-76.

[7] Ren Jie, Zhou Tao, Zhang Yicheng. Information filtering via self-consistent refinement [J]. Europhysics Letters, 2008, 82(5): 58007-58012.

[8] Sun Duo, Zhou Tao, Liu Runran, et al. Information filtering based on transferring similarity [J]. Physical Review E, 2009, 80(1): 017101-017104.

[9] 薛毅, 陈立萍. 统计建模与R语言 [M]. 北京: 清华大学出版社, 2007.
Xue Yi, Chen Liping. Statistical modeling and R software [M]. Beijing: Tsinghua University Press, 2007.