Journal of Computer Applications
计算机应用,2015,35(9):2569 - 2573
ISSN 1001-9081
CODEN JYIIDU
2015-09-10
http://www.joca.cn

文章编号:1001-9081(2015)09-2569-05
doi:10.11772/j.issn.1001-9081.2015.09.2569

基于主题模型的个性化图书推荐算法

郑祥云,陈志刚*,黄 瑞,李 博
(中南大学 软件学院,长沙 410075)
(* 通信作者电子邮箱 czg@csu.edu.cn)

摘 要: 针对传统推荐算法精准度不高的问题,在潜在狄利克雷分布(LDA)主题挖掘模型的基础上提出了一种新的适用于图书推荐(BR)的数据挖掘模型——BR_LDA 模型。通过对目标借阅者的历史借阅数据与其他图书数据进行内容相似度分析,得到与目标借阅者历史借阅图书内容相似度较高的其他图书。通过对目标借阅者的历史借阅数据及其他借阅者的历史借阅数据进行相似性分析,得到最近邻借阅者的历史借阅数据。通过求解图书被推荐的概率,最终得到目标借阅者潜在感兴趣的图书。特别地,当推荐数量为 4 000 时,BR_LDA 模型比基于多特征方法和关联规则方法精准度分别提高了 6.2%、4.5%;当推荐数量为 500 时,BR_LDA 模型比协同过滤的近邻方法和矩阵分解方法分别提高了 2.1%、0.5%。实验表明本模型能够更准确地向目标借阅者推荐历史感兴趣类别的新图书及潜在感兴趣的新类别的图书。

关键词: 图书推荐;图书管理系统;数据挖掘;推荐算法
中图分类号: TP301.6 **文献标志码:** A

Personalized book recommendation algorithm based on topic model

ZHENG Xiangyun, CHEN Zhigang*, HUANG Rui, LI Bo
(*School of Software, Central South University, Changsha Hunan* 410075, *China*)

Abstract: Concerning the problem of high time complexity of traditional recommendation algorithms, a new recommendation model based on Latent Dirichlet Allocation (LDA) model was proposed. It was a data mining model applied to Book Recommendation (BR) in library management systems, named Book Recommendation_Latent Dirichlet Allocation (BR_LDA) model. Through the content similarity analysis of historical borrowing data of the target borrowers with other books, other books which had high content similarities with historical borrowing books of the target borrowers were gotten. Through the similarity analyses performed on the target borrowers' historical borrowing data and historical data from other borrowers, historical borrowing data of the nearest neighbors were gotten. Books which the target borrowers were interested in could be finally gotten by calculating the probabilities of the recommended books. In particular, when the number of recommended books is 4 000, the precision of BR_LDA model is 6.2% higher than multi-feature method and 4.5% higher than association rule method; when the recommended list has 500 items, the precision of BR_LDA model is 2.1% higher than collaborative filtering based on the nearest neighbors and 0.5% higher than collaborative filtering based on matrix decomposition. The experimental results show that this model can efficiently mine data of books, reasonably recommend new books which belong to historical interested categories and new books in potential interested categories to the target borrowers.

Key words: Book Recommendation (BR); library management system; data mining; recommendation algorithm

0 引言

图书管理系统中的图书推荐(Book Recommendation, BR)[1]是图书管理系统日趋人性化的一个重要体现,是实现图书管理系统自动地向借阅者进行个性化推荐[2]图书的过程。图书管理系统通过对目标借阅者及其最近邻借阅者的历史借阅记录进行数据分析,挖掘出有用的模式,合理、及时地向借阅者推荐潜在感兴趣图书,使得图书管理系统的个性化推荐服务受到了广大借阅者的好评。

图书推荐算法[3]主要以借阅者自己的历史借阅数据或者最近邻借阅者的历史借阅数据为前提,通过对数据进行分析,得到用户潜在感兴趣的图书。传统的图书推荐算法通常只考虑使用借阅者的历史借阅数据进行图书推荐,造成推荐的精准度偏低。在传统的图书推荐算法的基础上,有研究者将目标借阅者的邻近借阅者的借阅数据进行分析并应用于图书推荐中,但由于未考虑目标借阅者本身的历史借阅数据,推荐效果不佳。在一些其他的研究中,以基于图书之间的相似度[4]和借阅者之间的相似度[5]的方法在图书推荐中取得了一定的研究成果,但是由于数据分析仍存在一定的不合理性且使用了关联相似度计算,导致推荐精准度不高且时间复杂

收稿日期:2015-04-23;修回日期:2015-06-16。 基金项目:国家自然科学基金资助项目(61379057,61309001,61272149,61103202);中南大学中央高校基本科研业务费专项资金资助项目(2015zzts228)。

作者简介:郑祥云(1992 -),男,湖南永州人,硕士研究生,主要研究方向:数据挖掘; 陈志刚(1964 -),男,湖南益阳人,教授,博士生导师,博士,CCF 会员,主要研究方向:无线网络、分布式计算; 黄瑞(1989 -),男,安徽安庆人,博士研究生,主要研究方向:社交网络; 李博(1988 -),男,河北衡水人,硕士研究生,主要研究方向:数据挖掘。

度较高。

通过对以上问题进行分析，本文提出了一种基于BR_LDA模型的图书推荐算法。该算法是将用于在社交网络中进行主题挖掘的三层概率模型[6] LDA进行适当改进之后所产生的，以目标借阅者及其最近邻借阅者的历史借阅数据为前提条件而建立的概率模型。该模型通过计算出图书被推荐的概率，能更加合理且高效地向借阅者推荐图书。

1 相关工作

在图书管理系统中图书推荐是一个完成向借阅者主动推荐图书的个性化的推荐过程，如何设计出有效的推荐方法是图书推荐系统的一个研究热点。

协同过滤方法[7]是通过对目标借阅者的历史借阅数据进行分析、计算出与其他借阅者之间的兴趣相似度，通过相似借阅者对图书的评价预测目标借阅者的兴趣度。协同过滤推荐方法主要有基于近邻模型、奇异值矩阵分解法等，对于有借阅者比较充分的偏好信息时，此方法可以很好地发现了用户的潜在的图书喜好。但随借阅者及图书数量增加，协同过滤方法面临着稀疏性及冷启动等问题，评分矩阵稀疏性问题及新用户问题导致了图书推荐精准度降低。

关联规则方法[8]是利用Apriori算法从事务集中产生项集，选出满足最小支持度的项集作为频繁集，然后从这些频繁集中找出同时满足最小支持度与最小置信度的频繁集产生强关联规则。在借阅者的历史借阅数据的基础上挖掘出与其他图书之间的关联度，并进行分析，把关联度较大的图书作为推荐对象。此方法可用于离线操作，但发现图书之间的关联规则是最为困难也最需要时间的一个过程，此方法面临的一个重大难题是当图书名称相同时，不能确保实现精准的推荐。

基于多特征的推荐方法[9]是根据通过构建中图分类树来计算图书之间的相似性，通过构建专业分类树来计算读者基于专业的相似性，最后通过综合图书与读者的综合特征相似性来预测读者感兴趣的图书并向读者推荐。此方法在图书资源较少时具有相对较高的查全率与查准率，由于提取内容的能力非常有限，对于高等院校数量庞大的图书资源很难准确且全面地进行挖掘，使得有很大的可能性导致推荐的图书与目标借阅者实际偏好存在误差，从而降低了图书推荐的精准度。

通过对相关工作进行分析与研究，本文设计出一种基于BR_LDA模型的图书推荐算法，相对更准确且效率更高地向借阅者推荐潜在感兴趣的图书。

2 基于主题模型的图书推荐算法

2.1 图书内容相似度

目标借阅者的历史借阅数据是分析目标借阅者的历史兴趣点并对其推荐感兴趣的图书的重要依据。对于目标借阅者经常借阅某一类别的图书，可以认为目标借阅者对此类别的图书比较感兴趣，从而此类别的图书被推荐的可能性就很大。设目标借阅者的历史借阅图书类别集合 $G=(g_1,g_2,\cdots,g_i,\cdots g_I)$ 及每一类别对应的关键词集合 $J=(j_1,j_2,\cdots,j_i,\cdots j_I)$，其中 $j_i=(m_1,m_2,\cdots,m_v)$（即类别 g_i 具有 v 个关键词）。对于一本其他的图书（非目标借阅者借阅过的图书）可以根据此图书对应的关键词集合 $N=(n_1,n_2,\cdots,n_k,\cdots n_u)$（即此图书对应有 u 个关键词）与目标借阅者历史借阅各类别图书的关键词的相似度进行分析，本文定义如下：

$$sim_1=\left(\sum_{i=1}^{I}\sum_{r=1}^{v_i*u}d_r\right)\Big/\left(\sum_{i=1}^{I}\sum_{r=1}^{v_i*u}d_0\right) \tag{1}$$

其中：v_i 为目标借阅者历史借阅图书类别 i 的关键词个数。

由式(1)可知，sim_1 值越大，则相似度越大，此图书被推荐的可能性就越大。式中，取 $d_0=1$，若 $n_k=m_i$（即此图书的关键词与目标借阅者的历史借阅某一类别图书的关键词相匹配）则 d_r 取值为1，否则取值为0。即 d_r 为：

$$d_r=\begin{cases}1, & n_k=m_i\\ 0, & n_k\neq m_i\end{cases} \tag{2}$$

2.2 最近邻借阅者

在给目标借阅者推荐图书时其他借阅者的历史借阅数据也是一项重要的不可忽视的参考依据，可能从中挖掘出目标借阅者新的感兴趣的图书。设有矩阵 $U(n,m)$ 表示 n 个目标借阅者与最近邻借阅者集合 $P=(p_1,p_2,\cdots,p_n)$ 及 m 个图书集合 $Q=(q_1,q_2,\cdots,q_m)$ 的评分矩阵，利用余弦相似性计算公计算与借阅者相似程度较高的其他借阅者作为目标借阅者的最近邻。计算方法如下：

$$sim_2(p_1,p_2)=\frac{\sum_{q\in Q_{1,2}}(U_{p_1,q}-\overline{U_{p_1}})(U_{p_2,q}-\overline{U_{p_2}})}{\sqrt{\sum_{q\in Q_1}(U_{p_1,q}-\overline{U_{p_1}})^2\sum_{q\in Q_2}(U_{p_2,q}-\overline{U_{p_2}})^2}} \tag{3}$$

式(3)中，$sim(p_1,p_2)$ 表示借阅者 p_1 与借阅者 p_2 的相似度，$Q_{1,2}$ 表示借阅者 p_1，p_2 具有共同评分的图书，Q_1 为借阅者 p_1 有过评分的图书，Q_2 为借阅者 p_2 有过评分的图书，$U_{p,q}$ 表示借阅者 p_1 对图书 q 的评分。$\overline{U_{p_1}}$ 与 $\overline{U_{p_2}}$ 分别表示借阅者 p_1 与 p_2 对图书评分的均值。其中，借阅者 p 对借阅图书 q 的评分 $U_{p,q}$ 计算方法如下：

$$U_{p,q}=\frac{t_q-T_{\min}}{T_{\max}-T_{\min}} \tag{4}$$

其中：t_q 为借阅者 p 借阅图书 q 所花时间（为一时间段），$T_{\min}$ 为统计开始时刻（本文假设开始时刻比任何借阅者开始借书时刻都早），$T_{\max}$ 为统计结束时刻。

2.3 LDA主题挖掘模型

潜在狄利克雷分布（Latent Dirichlet Allocation，LDA）模型[10-11]最先由Blei等[10]提出，是以概率生成为条件的包括文档、主题、词在内的3层结构模型，基于不存在相互关联的词语和文档，通过对文档进行建模，挖掘出潜在的主题。LDA 3层概率生成模型的生成过程如图1所示。

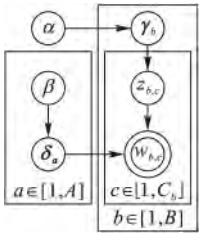

图1 LDA概率生成模型

图1中：α、β 均为参数；γ_b 表示第 b 篇文档下主题的分布；δ_a 表示第 a 个主题下词语的分布；$z_{b,c}$ 为在概率 γ_b 下所抽取的

主题,表示第 b 篇文档中的第 c 个词语的主题; 词语 $w_{b,c}$ 是根据概率分布抽取产生的,表示第 b 篇文档中的第 c 个词语。A 为主题的总个数,B 为文档总个数,C_b 为第 b 篇文档中的词语的总个数。

2.4　个性化图书推荐模型

LDA 模型主要适用于在社交网络中挖掘出主题,为研究者对微博等社交网络[12]的研究带来了便利,并逐渐成为了一种主要的研究手段。在图书管理系统中,为能比较准确地挖掘出学生潜在感兴趣的图书,通过对目标借阅者的历史借阅数据与其他图书进行内容相似度分析、找出最近邻借阅者并分析其历史借阅数据,然后结合 LDA 概率模型并对其进行适当的改进,本文提出了图书推荐_潜在狄利克雷分布(Book Recommendation_ Latent Dirichlet Allocation, BR_LDA) 模型,其生成过程如图 2 所示。

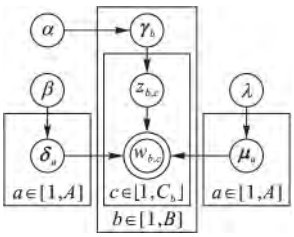

图 2　BR_LDA 图书推荐模型

假设在一所高校里有 B 个图书馆分馆,即有 B 个子图书管理系统,总共 A 种图书类别,根据借阅者的历史借阅数据最终得到借阅者对 D 种图书感兴趣,此时 BR_LDA 图书推荐模型的生成过程描述如下:

1) 新引入参数 λ 为狄利克雷[13]先验参数,从 λ 的狄利克雷先验分布中抽选最近邻借阅者在借阅类别 a 中图书的分布 $\boldsymbol{\mu}_a$,是 n 维向量(n 为图书总数) 。

2) 参数 α 为狄利克雷先验参数,从 α 的狄利克雷先验分布中抽选出第 b 个图书馆分馆中的类别的分布为 $\boldsymbol{\gamma}_b$,是 K 维向量(K 为类别总数) 。

3) $l_{b,c}$ 为在选取第 b 个图书馆分馆的前提下选取第 b 个分馆中第 c 个图书的类别。

4) 参数 β 是狄利克雷先验参数,首先随机选取的类别 a,从 β 的狄利克雷先验分布中抽选出类别 a 中图书与目标借阅者历史借阅图书的相似度的分布 $\boldsymbol{\delta}_a$,是 m 维向量(m 为图书总数) 。

5) $S_{b,c}$ 表示第 b 个图书馆分馆中第 c 个图书(图书总个数为 C_b) 。

参数 α、β、λ 均为根据实际经验给出的先验参数。对于上述的 V 维向量的概率分布 $\boldsymbol{\mu}_a$ 及 $\boldsymbol{\delta}_a$ 具体描述如下:

1) $\boldsymbol{\mu}_a = (u_1, u_2, \cdots, u_i, \cdots u_n)$

其中: n 表示在 a 类别里有 n 本图书,u_i 表示最近邻借阅者阅读第 i 书的概率,可表示为:

$$u_i = \frac{\text{借阅第 } i \text{ 本书的时长}}{\text{借阅图书的总时长}}$$

2) $\boldsymbol{\delta}_a = (r_1, r_2, \cdots, r_i, \cdots r_m)$

其中: m 表示在 a 类别目标借阅者借阅了 m 本图书,r_i 表示最近邻借阅者第 i 本图书与目标借阅者借阅图书相似度的概率,可表示为:

$$r_i = \frac{\text{近邻用户第 } i \text{ 本图书与目标借阅者图书的相似度}}{\text{目标借阅者与最近邻借阅者图书总相似度}}$$

3) 由于 $\boldsymbol{\gamma}_b$ 为 K 维向量可表示为:

$\boldsymbol{\gamma}_b = (y_1, y_2, \cdots, y_i, \cdots, y_K)$

其中概率 $y_i = 1/K$ 为抽出第 b 个分馆中各类别的概率。

2.5　BR_LDA 模型求解

根据 BR_LDA 模型,可以得出图 2 中全部变量的联合概率分布:

$$P(\boldsymbol{S}_c, \boldsymbol{l}_b, \boldsymbol{\gamma}_b, \underline{\delta}, \underline{\mu} \mid \alpha, \beta, \lambda) = \prod_{c=1}^{C_b} P(S_{b,c} \mid \boldsymbol{\delta}_{l_{b,c}}, \boldsymbol{\mu}_{l_{b,c}}) P(l_{b,c} \mid \boldsymbol{\gamma}_b) P(\boldsymbol{\gamma}_b \mid \alpha) P(\underline{\delta} \mid \beta) P(\underline{\mu} \mid \lambda) \tag{5}$$

那么将每一个图书初始化为时 t 的概率分布为:

$$P(S_{b,c} = t \mid \boldsymbol{\gamma}_b, \underline{\delta}, \underline{\mu}) = \sum_{a=1}^{A} P(S_{b,c} = t \mid \boldsymbol{\delta}_a, \boldsymbol{\mu}_a) P(l_{bc} = a \mid \boldsymbol{\gamma}_b) = \sum_{a=1}^{A} P(S_{b,c} = t \mid \boldsymbol{\delta}_a) P(S_{b,c} = t \mid \boldsymbol{\mu}_a) P(l_{bc} = a \mid \boldsymbol{\gamma}_b) \tag{6}$$

在图书子馆 b 中,$P(S_{b,c} = t \mid \boldsymbol{\delta}_a)$ 为在类别 a 下根据借阅者本身的历史借阅图书与图书 c 的内容相似度概率; $P(S_{b,c} = t \mid \boldsymbol{\mu}_a)$ 为在类别 a 下根据最近邻借阅者的借阅记录得到的最近邻借阅者对图书 c 感兴趣(借阅时间长短) 的概率; $P(l_{bc} = a \mid \boldsymbol{\gamma}_b)$ 为图书馆分馆 b 中抽选类别 a 的概率。考虑到以下两种特殊的情况下的图书都有可能成为推荐的对象,此时只取其中概率不为 0 的一项:

1) 一图书被最近邻借阅者借阅的概率较大但与目标借阅者的历史借阅图书的内容相似度为 0,即:

$P(S_{b,c} = t \mid \boldsymbol{\mu}_a) = 0$

2) 一图书与目标借阅者的历史借阅图书的内容相似度较大但其最近邻借阅者从未借阅过,即:

$P(S_{b,c} = t \mid \boldsymbol{\delta}_a) = 0$

从而可以得出整个图书馆分馆中的图书的似然函数为:

$$P(S \mid \underline{\gamma}, \underline{\delta}, \underline{\mu}) = \prod_{b=1}^{B} P(S \mid \boldsymbol{\gamma}_b, \underline{\delta}, \underline{\mu}) = \prod_{b=1}^{B} \prod_{c=1}^{C_b} P(S_{b,c} \mid \boldsymbol{\gamma}_b, \underline{\delta}, \underline{\mu}) \tag{7}$$

2.6　BR_LDA 模型算法设计

本文将 BR_LDA 模型应用在图书管理系统中来挖掘出学生潜在感兴趣的图书,通过上述对模型的分析与推导,根据 BR_LDA 模型生成过程,可设计出此模型的算法,可得其算法伪代码描述如下。

算法 1　BR_LDA 模型算法。

输入: 借阅者数据集、图书分馆数量 B、参数 λ、α 和 β。

输出: 被推荐的图书。

```
Sample a borrower r.
Calculate total number of books C_b of each sub-library b.
Calculate the total number of categories A .
Initialize an array G [n]  for storing books.
for a = 1 To A do
  calculate sim_1 and sample distribution δ_a ~ Dirichlet(β).
  calculate sim_2 and sample distribution μ_a ~ Dirichlet(λ).
end for
for b = 1 To B do
  sample distribution γ_b ~ Dirichlet(α).
  for c = 1 To C_b do
    sample category l_{b,c} according to γ_a.
    sample books S_{b,c} according to δ_{l_{b,c}} & μ_{l_{b,c}}.
```

14) put $s_{b,c}$ into G.
15) end for
16) end for
17) return G.

通过上述对算法的描述可知，该算法在执行时的一些变量由吉布斯抽样得到，进行概率计算，将推荐的图书存于数组 G [u] 里，最后返回 G。对算法分析知，第一部分 for 循环执行完成需要消耗时间为 $O(A)$，第二部分 for 循环嵌套执行完成需要消耗时间为 $O(B \cdot C_b) < O(B \cdot n)$，其中 $n = \max\{C_b \mid b \in [1,B]\}$。由于总类别数 A 远小于最大分馆图书总数 n 乘以分馆个数 B，故知本文提出的 BR_LDA 模型算法的时间复杂度为 $O(n \cdot B)$。

3 实验与分析

本文实验数据采用某高等院校图书管理系统数据。实验选取了 3 个图书管理子系统（即 3 个分馆）中的借阅者 1 245 位和借阅记录 11 988 条。实验分别将目标借阅者的历史借阅数据、最近邻借阅者历史借阅数据及内容相似度较高的图书进行了实验，并将其中的 70% 的数据作为训练数据，剩余的 30% 的数据用在实验后期检测模型预测效果。

实验环境为 Intel P41.8 GHz 的 CPU，4 GB 的内存，500 GB的硬盘，操作系统为 Windows 7，采用 Matlab 作为实验仿真工具。实验将本文提出的 BR_LDA 模型图书推荐算法与协同过滤图书推荐方法、关联规则图书推荐方法和基于多特征的图书推荐方法的图书推荐精准度和算法时间复杂度分别进行了详细对比。根据参考文献 [14]，可以将本仿真实验的参数设置如表 1 所示。

表 1 实验参数设置

参数	值	参数	值
α	0.5 ~ 1	β	0.1 ~ 0.5
λ	0.1 ~ 0.5	迭代次数	100 ~ 135

在本实验中 α 是从 0.5 ~ 1 中的一个随机取值，$\lambda = \beta$ 是从 0.1 ~ 0.5 中的一个随机取值，由经验知当迭代次数 *Iteration* 大于 100 次的时候模型逐渐收敛，所以本文从 100 ~ 135 随机取一个自然数作为迭代次数。

3.1 图书推荐精准度

实验首先对各推荐算法的图书推荐精准度进行了对比分析，其中精准度本文定义如下：

$$precision = \left(\sum_{h=1}^{H} m_h \right) \Big/ \left(\sum_{r=1}^{R} s_r \right) \tag{13}$$

其中：m_h 为使用算法后所推荐的图书与后期检验预测图书匹配，R 为匹配总个数。s_r 为预留后期检验预测图书，H 为总个数，精准度取值范围：$precision \in [0,1]$。实验结果如图 3 所示。

由图 3 可以看出，图书数量对推荐的精准度有着一定的影响。当图书数量较少（500 本左右）时，四种算法的推荐精准度在 0.54 左右。随着图书数量的增加，精准度也随之提高，当图书数量达 1 000 本级别以上，很明显可以看出 BR_LDA 算法的精准度比其他算法要高。本文提出的算法的图书推荐精准度高是由于 BR_LDA 模型同时考虑了：

1) 目标借阅者及其最邻近借阅者的历史借阅数据。最近邻的数据可以用于给目标借阅者推荐潜在感兴趣种类的图书。

2) 内容相似度较高的其他未阅读的图书数据。给目标借阅者推荐已经熟知的种类的图书。

3) 图书及类别概率分布数据。根据实际情况将各个图书馆的藏书种类和数量的差异考虑在内。

通过多方面的概率分布约束，BR_LDA 模型比较准确地计算出被推荐图书的概率，提高了模型的推荐精准度。

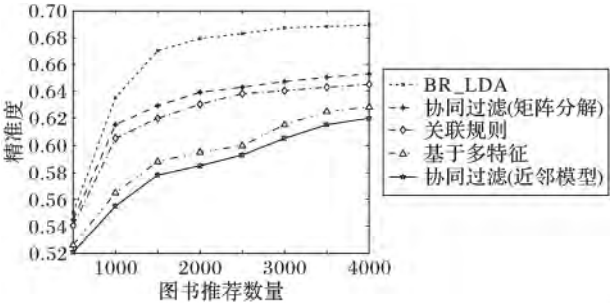

图 3 各算法图书推荐精准度比较

3.2 算法运行效率

在分析了图书推荐精准度之后，实验还记录了各算法对部分实验数据的训练时间，对各算法的运行效率进行了分析比较。算法的运行效率如表 2 所示。

表 2 各算法运行效率比较

算法名称	时间复杂度	训练时间/s
BR_LDA	$O(n \cdot B)$	16.5
关联规则	$O(n^2 \cdot k)$	26.7
基于多特征	$O(n \cdot m)$	17.8
协同过滤（矩阵分解）	$O(n_1 \cdot x + n_2 \cdot x + n_3 \cdot x)$	14.1
协同过滤（近邻模型）	$O(n^2 \cdot m)$	24.9

在表 2 中，n 为图书的个数，B 为图书馆分馆的个数，m 为最近邻借阅者个数，k 为频繁集的个数，n_1、n_2、n_3 为分解的评分矩阵非零元素的个数，x 为特征向量的维数[15]。

在本实验中采取部分数据作为训练数据，由表 2 可知，基于矩阵分解的协同过滤方法训练时间相对较少，本文提出的 BR_LDA 模型算法稍次之。在算法运行的计算时间上，可以看出基于矩阵分解的协同过滤方法的在线计算时间取决于 x 的取值，当 x 取值很小（如 $x = 1$）时，此方法的最优；当用户和图书的数量较大即特征向量的维数 x 增大到几十或者几百时，BR_LDA 方法在线计算时间消耗最小，算法运行时间大小顺序为：关联规则 > 协同过滤（近邻模型） > 基于多特征 > 协同过滤（矩阵分解） > BR_LDA。

通过以上分析可以得知，BR_LDA 算法运行效率总体上是较优的，能相对有效地给借阅者推荐潜在感兴趣的图书。

4 结语

本文通过对图书馆里系统中的图书推荐的特点和三层概率模型 LDA 的分析，提出了基于 BR_LDA 模型的图书推荐算法。将目标借阅者及其最邻近借阅者的历史借阅数据结合起来，通过设定先验参数 α、λ 和 β，采用经典的吉布斯抽样方法对模型进行推导，计算出借阅者对特定类别下的图书感兴趣的概率。通过与协同过滤方法、关联规则方法和基于多特征的推荐方法比较，实验结果表明：BR_LDA 模型在推荐图书时

具有较高的精准度; 算法运行时相对较低的时间复杂度,能够高效率地给用户合理地推荐一些潜在感兴趣的图书,是图书管理系统中的一种较好的个性化推荐方法。

参考文献:

[1] LIU S. Research on the key issues for the recommender system [D]. Hefei: University of Science and Technology of China, 2014. (刘士琛. 面向推荐系统的关键问题研究及应用[D]. 合肥: 中国科学技术大学,2014.)

[2] ZHANG F. Survey of online social network based on personalized recommendation [J]. Journal of Chinese Computer Systems, 2014, 35(7): 1470 - 1476. (张富国. 基于社交网络的个性化推荐技术[J]. 小型微型计算机系统,2014,35(7): 1470 - 1476.)

[3] KONG Y. Recommendation algorithms in the big data era [D]. Xiamen: Xiamen University,2014. (孔远帅. 基于大数据的推荐算法研究[D]. 厦门: 厦门大学,2014.)

[4] WANG Z, HE M, DU Y. Text similarity computing based on topic model LDA [J]. Computer Science, 2013,40(2): 229 - 232. (王振振,何明,杜永萍. 基于 LDA 主题模型的文本相似度计算[J]. 计算机科学,2013,40(2): 229 - 232.)

[5] ZHU W. Research on user similarity function of recommendation system [D]. Chongqing: Chongqing University, 2014. (朱文奇. 推荐系统用户相似度计算方法研究[D]. 重庆: 重庆大学,2014.)

[6] BLEI D M. Introduction to probabilistic topic models [EB/OL]. [2015-01-11]. http://www.cs.princeton.edu/~blei/papers/Blei2011.pdf?origin=publication_detail.

[7] BOBADILLA J, ORTEGA F, HEMANDO A, *et al*. Improving collaborative filtering recommender system results and performance using genetic algorithms [J]. Knowledge-Based Systems, 2011, 24(8): 1310 - 1316.

[8] YANG Y, XIE K, ZHU Y, *et al*. Implementation of association rules recommendation model in recommendation system of e-commence Web [J]. Computer Engineering, 2004,30(19): 57 - 59. (杨引霞,谢康林,朱扬勇,等. 电子商务网站推荐系统中关联规则推荐模型的实现[J]. 计算机工程,2004,30(19): 57 - 59.)

[9] LI K, LIANG Z. Personalized book recommendation algorithm based on multi-feature [J]. Computer Engineering, 2012,38(11): 34 - 37. (李克潮,梁正友. 基于多特征的个性化图书推荐算法[J]. 计算机工程,2012,38(11): 34 - 37.)

[10] BLEI D M, ANDREW Y N, JORDAN M I. Latent Dirichlet allocation [J]. Journal of Machine Learning Research, 2003,3(1): 993 - 1022.

[11] ZHANG Z, MIAO D, GAO C. Short text classification using latent Dirichlet allocation [J]. Journal of Computer Applications, 2013,33(6): 1587 - 1590. (张志飞,苗夺谦,高灿. 基于 LDA 主题模型的短文本分类方法[J]. 计算机应用,2013,33(6): 1587 - 1590.)

[12] YAGER R R, YAGER R L. Social networks: querying and sharing mined information [C]// Proceedings of the 2014 47th Hawaii International Conference on System Sciences. Washington, DC: IEEE Computer Society, 2013: 1435 - 1442.

[13] MA Z. Bayesian estimation of the Dirichlet distribution with expectation propagation [C]// Proceedings of the 20th European Signal Processing Conference. Piscataway: IEEE, 2012: 689 - 693.

[14] GRIFFITHS T, STEYYERS M. Probabilistic topic models [J]. Handbook of Latent Semantic Analysis, 2007,427(7): 424 - 440.

[15] TU D, SHU C, YU H. Using unified probabilistic matrix factorization for contextual advertisement recommendation [J]. Journal of Software, 2013,24(3): 454 - 464. (涂丹丹,舒承椿,余海燕. 基于联合概率矩阵分解的上下文广告推荐算法. 软件学报,2013,24(3): 454 - 464.)

(上接第 2568 页)

[5] JIN H, WANG S, LI C. Community detection in complex networks by density-based clustering [J]. Physica A: Statistical Mechanics and its Applications, 2013,392(19): 4606 - 4618.

[6] XIA Z, BU Z. Community detection based on a semantic network [J]. Knowledge-Based Systems, 2012,26: 30 - 39.

[7] BARBIERI N, BONCHI F, MANCO G. Cascade-based community detection [C]// Proceedings of the 6th ACM International Conference on Web Search and Data Mining. New York: ACM, 2013: 33 - 42.

[8] DEV H, ALI M E, HASHEM T. User interaction based community detection in online social networks [M]// BHOWMICK S S, DYRESON C, JENSEN C S, *et al*. Database Systems for Advanced Applications, LNCS 8422. Berlin: Springer, 2014: 296 - 310.

[9] GREGORY S. Finding overlapping communities in networks by label propagation [EB/OL]. [2015-01-08]. http://iopscience.iop.org/1367-2630/12/10/103018/pdf/1367-2630_12_10_103018.pdf.

[10] XIE J, SZYMANSKI B K. Towards linear time overlapping community detection in social networks [M]// TAN P-N, CHAWLA S, HO C K, *et al*. Advances in Knowledge Discovery and Data Mining, LNCS 7302. Berlin: Springer, 2012: 25 - 36.

[11] LIU X, MURATA T. Advanced modularity-specialized label propagation algorithm for detecting communities in networks [J]. Physica A: Statistical Mechanics and its Applications, 2010,389(7): 1493 - 1500.

[12] XIE J, SZYMANSKI B K. Community detection using a neighborhood strength driven label propagation algorithm [C]// Proceedings of the 2011 IEEE Network Science Workshop. Piscataway: IEEE, 2011: 188 - 195.

[13] CORDASCO G, GARGANO L. Community detection via semi-synchronous label propagation algorithms [C]// Proceedings of the 2010 IEEE International Workshop on Business Applications of Social Network Analysis. Piscataway: IEEE, 2010,1 - 8.

[14] LOU H, LI S, ZHAO Y. Detecting community structure using label propagation with weighted coherent neighborhood propinquity [J]. Physica A, 2013,392(14): 3095 - 3105.

[15] UGANDER J, BACKSTROM L. Balanced label propagation for partitioning massive graph [C]// Proceedings of the 6th ACM International Conference on Web Search and Data Mining. New York: ACM, 2013: 507 - 516.

[16] KUZMIN K, SHAH S Y, SZYMANSKI B K. Parallel overlapping community detection with SLPA [C]// Proceedings of the 2013 International Conference on Social Computing. Washington, DC: IEEE Computer Society, 2013: 204 - 212.

[17] LI C, TANG Y, LIN H, *et al*. A parallelize overlapping community detection algorithm in complex networks based on label propagation [EB/OL]. [2014-12-28]. http://info.scichina.com/sciF/CN/10.1360/N112014-00258. (李春英,汤庸,林海,等. 基于标签传播的可并行复杂网络重叠社区发现算法[EB/OL]. [2014-12-28]. http://info.scichina.com/sciF/CN/10.1360/N112014-00258.)