·人工智能及识别技术· 文章编号：1000-3428(2017)01-0247-06 文献标志码：A 中图分类号：TP311

# 基于标签主题的协同过滤推荐算法研究

文俊浩[1,2]，袁培雷[1,2]，曾 骏[1,2]，王喜宾[1,2]，周 魏[1,2]

（1. 信息物理社会可信服务计算教育部重点实验室，重庆 400030；2. 重庆大学 软件学院，重庆 401331）

**摘 要：** 传统基于标签的推荐算法仅考虑用户的评分信息，导致推荐准确度不高。为解决该问题，提出一种改进的协同过滤推荐算法。对用户-标签矩阵、资源-标签矩阵进行潜在 Dirichlet 分布建模，发掘推荐系统中的潜在语义主题，从语义层面计算用户对各资源的偏好概率，将计算出的偏好概率与协同过滤算法计算出的资源相似度相结合，预测用户偏好值，实现个性化推荐。在 Movielens 数据集上的实验结果表明，与传统基于标签的推荐算法相比，该算法能消除标签中存在的同义词、多义词等语义模糊问题，同时提高推荐准确度。

**关键词：** 标签主题；协同过滤；潜在 Dirichlet 分布模型；个性化推荐；相似度

**中文引用格式：** 文俊浩，袁培雷，曾 骏，等. 基于标签主题的协同过滤推荐算法研究[J]. 计算机工程，2017，43(1)：247-252，258.

**英文引用格式：** Wen Junhao，Yuan Peilei，Zeng Jun，et al. Research on Collaborative Filtering Recommendation Algorithm Based on Topic of Tags[J]. Computer Engineering，2017，43(1)：247-252，258.

# Research on Collaborative Filtering Recommendation Algorithm Based on Topic of Tags

WEN Junhao[1,2]，YUAN Peilei[1,2]，ZENG Jun[1,2]，WANG Xibin[1,2]，ZHOU Wei[1,2]

（1. Key Laboratory of Dependable Service Computing in Cyber Physical Society，Ministry of Education，Chongqing 400030，China；
2. School of Software Engineering，Chongqing University，Chongqing 401331，China）

**【Abstract】** Aiming at the problem that traditional tag-based recommendation algorithm only considering the user's rating information leads to the low accuracy of recommendation，this paper puts forward an improved collaborative filtering recommendation algorithm. It builds the Latent Dirichlet Allocation（LDA） model for user-tag matrixes and resource-tag matrixes and explores the latent semantic topics in the recommendation system. The probability of users choosing resources from the semantic level is computed. Then the probability is combined with resource similarity calculated through collaborative filtering algorithm to predict user preference values to realize personalized recommendation. Experimental results on the Movielens dataset show that the proposed algorithm can eliminate the semantic ambiguities such as synonyms and polysemes，and improve the accuracy of recommendation，compared with the traditional tag-based recommendation algorithm.

**【Key words】** topic of tag；collaborative filtering；Latent Dirichlet Allocation（LDA） model；personalized recommendation；similarity

**DOI：** 10.3969/j.issn.1000-3428.2017.01.043

## 0 概述

信息膨胀使用户很难获取信息中隐藏的价值，也使得用户发现其感兴趣的信息的速度急剧下降。个性化推荐系统通过分析用户历史行为，预测用户对信息的偏好值，将偏好值较高的信息推荐给用户，实现个性化推荐，有效地解决了信息过载问题。在 Web2.0 时代，用户可以通过社会标签组织、管理和共享资源，社会标签也因其易用性成为资源分类和索引的重要方式[1]。社会标签既能体现资源本身的特征，又能反映用户的兴趣偏好[2]。将社会标签与个性化推荐相结合，能够有效地提高个性化推荐的效率。

目前，个性化推荐主要包括基于内容的推荐、协

**基金项目：** 国家自然科学基金(61379158，61502062)；重庆市科技计划项目(cstc2014jcyjA40054)。

**作者简介：** 文俊浩(1969—)，男，教授、博士，主研方向为数据挖掘、服务计算；袁培雷，硕士；曾 骏，讲师、博士；王喜宾、周 魏，博士。

**收稿日期：** 2016-01-18 **修回日期：** 2016-02-19 **E-mail：** jhwen@cqu.edu.cn

同过滤推荐和混合推荐[3],其中协同过滤推荐的应用最广,此方法主要根据用户对资源的历史行为信息,寻找用户或者资源的近邻集合,以此来计算用户对资源的偏好值。基于社会标签的个性化推荐技术的最新研究进展主要包含以下 3 类方法[4]:基于图论的方法,基于张量的方法,基于主题的方法。基于图论的方法能解决数据的稀疏性问题[5],但将标签用二维向量表示,忽略了系统中的整体信息;基于张量的方法能解决多维数据的降维问题[6],但没有充分挖掘用户的偏好;基于主题的方法能产生更具解释性的推荐结果[7]。

文献[8]提出一种基于标签的协同过滤推荐算法,能够较好地削弱数据稀疏性。随后很多研究者将标签与协同过滤算法相结合,提出多种推荐算法[9-10]。文献[11]使用二维向量来表示用户的兴趣,通过标签的 TF-IDF 权重向量来计算用户间相似度。文献[12]提出一种基于标签的电影推荐算法,根据用户对电影的标注记录,利用相应的算法计算用户与标签、电影与标签之间的关系值,然后计算用户与电影间的关系值,最后根据两者间的关系值排序,进行 Top-N 推荐。文献[13]先通过用户的历史行为求出用户之间的标签集,然后利用标签集的语义距离来度量两者之间的相似性,根据相似邻居做出推荐。文献[14-15]采用不同的标签相似度度量方法,分别提出了不同的推荐策略。但是这些相似度度量方法本质上是通过关键词完全匹配的方式来度量用户之间的相似度,不能解决标签的语义模糊问题。

基于以上分析,本文提出一种基于标签主题的协同过滤推荐算法(ColLDA)。该算法将潜在 Dirichlet 分布(Latent Dirichlet Allocation,LDA)模型引入到标签系统中,以此来发现用户与标签、资源与标签之间的潜在语义关系,并计算用户选择某个资源的条件概率,然后将计算出的概率与通过协同过滤算法计算出的资源相似度相结合,预测用户偏好值。

## 1 LDA 模型

LDA 模型[16]已成为主题建模的一个标准,目前已被广泛应用到信息检索、主题挖掘、社交网络分析等领域。LDA 是一个具有文本、主题和单词 3 层结构的贝叶斯概率生成模型。LDA 模型的基本思想:文本集中存在着 $Z$ 个互相独立的隐含主题,每个文本是这 $Z$ 个隐含主题上的多项式分布,每个主题是词汇表中所有单词上的多项式分布。

LDA 建模的图模型如图 1 所示,图中的白色圆形代表模型中的潜在变量;黑色圆形代表模型中的可观测变量;矩形代表模型的重复抽样过程,右下角的常数表示抽样次数;黑色箭头表示 2 个指定变量间的条件依赖关系。外层的矩形代表文本集中的一个文本对主题的抽样,里面的矩形代表同一个文本内主题对单词的重复抽样。其中,$w$ 表示文本集中某一文本中的一个单词;$z$ 表示该单词所描述的主题;$\theta$ 表示文本集中的一个文本在 $Z$ 个主题上的多项式分布,该分布符合超级参数为 $\alpha$ 的 Dirichlet 先验分布,即 $\theta \sim Dir(\alpha)$;$\phi$ 表示每个主题在词汇表中单词上的多项式分布,该分布符合超级参数为 $\beta$ 的 Dirichlet 先验分布,即 $\phi \sim Dir(\beta)$。

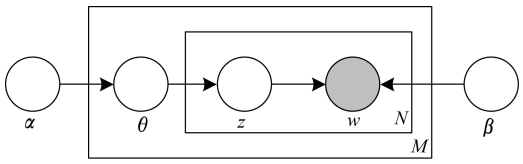


**图 1 LDA 模型**

在 LDA 建模的过程中,用条件概率 $p(z|d)$ 表示一个文本 $d$ 中的主题 $z$ 分布概率,用条件概率 $p(w|z)$ 表示每一个主题 $z$ 中一个单词 $w$ 的分布概率,则单词 $w$ 在文本 $d$ 中的分布概率公式为:

$$p(w|d)=\sum_{i=1}^{Z}p(w|z_i)p(z_i|d) \tag{1}$$

其中,$Z$ 表示隐含主题的个数,它的大小必须提前设定,不同的 $Z$ 值会对整个 LDA 的建模产生影响。如果 $Z$ 取值较大,则表示主题很多,通过模型训练出的每个主题所包含的范围就会比较小;如果 $Z$ 取值过小,则表示主题很少,通过模型训练出的每个主题所包含的范围就会比较广。一般通过 Gibbs 抽样构造 LDA 模型,对于文本 $d_i$ 中的每个单词 $w_i$ 循环抽样,估算由 $w_i$ 生成一个新的主题 $z_i=n$ 的概率 $p(z_i=n|w_i,d_i,z_{-i})$。

$$p(z_i=n|w_i,d_i,z_{-i})\propto\frac{C_{w_in}^{WZ}+\beta}{\sum_{w}C_{wn}^{WZ}+W\beta}\cdot\frac{C_{d_in}^{DZ}+\alpha}{\sum_{z}C_{d_iz}^{DZ}+Z\alpha} \tag{2}$$

其中,$\boldsymbol{C}^{DZ}$ 表示所有的文本-主题的原始分布矩阵;$\boldsymbol{C}^{WZ}$ 表示所有的主题-单词的原始分布矩阵;$C_{wn}^{WZ}$ 表示单词 $w$ 被赋予主题 $z_i=n$ 的次数;$C_{d_in}^{DZ}$ 表示 $d_i$ 文本里的单词被赋予主题 $z_i=n$ 的次数,需要注意,计算公式中不包括当前单词 $w_i$;$z_{-i}$ 表示除当前主题 $z_i=n$ 与当前单词 $w_i$ 以外的所有主题-单词分布以及文本-主题分布;参数 $\alpha$ 是文本-主题先验分布的超级参数;$\beta$ 是主题-单词先验分布的超级参数,在公式中作为平滑参数。

采用 Gibbs 重复抽样,当抽样次数达到一定值后,文本中隐含主题的概率会逐渐服从于 Dirichlet 先验分布,此时,LDA 模型中的参数 $\alpha$,$\beta$ 收敛。这

时公式中的先验概率可以通过式(3)和式(4)得到。

$$p(w_i|z_i=n)=\frac{C^{WZ}_{w_in}+\beta}{\sum_w C^{WZ}_{wn}+W\beta} \tag{3}$$

$$p(z_i=n|d_i)=\frac{C^{DZ}_{d_in}+\alpha}{\sum_z C^{DZ}_{d_iz}+Z\alpha} \tag{4}$$

## 2　基于标签主题的协同过滤推荐算法

从第 1 节介绍的 LDA 模型可以发现,在文本处理领域,该模型认为文本集中的每个文本隐含着若干独立的主题,可以通过 LDA 建模计算并发现文本集中的潜在主题。在标签系统中,用户可以对系统中的资源进行浏览并对感兴趣的资源添加相应的标签。标签能反映资源本身的特征以及用户的兴趣偏好,用户的这种行为可以将用户、资源与标签建立联系。类比于文本处理领域,将资源上用户添加的标签 $t\in T$ 类比为文本中的单词,用户 $u\in U$ 及资源 $r\in R$ 则可以类比为文本。

本文把 LDA 模型与标签相结合,将计算得到的信息应用到协同过滤算法中,提出一种基于标签主题的协同过滤算法,该推荐算法的基本框架如图 2 所示。该算法分为两部分,一部分根据用户的历史标注行为,构建用户-标签矩阵、资源-标签矩阵,对两者分别进行 LDA 建模,分别计算语义层的用户-标签概率矩阵以及语义层的资源-标签概率矩阵;最后,计算用户对各个资源喜好的条件概率,形成用户对资源的喜好概率矩阵。另一部分是根据用户的历史评分,统计用户评分矩阵,利用协同过滤思想依据 Pearson 相关系数计算资源间的相似度,根据相似度找到待推荐资源的相似资源集合。然后将两部分相结合计算用户对待推荐资源的偏好值,进行 Top-N 个性化推荐。

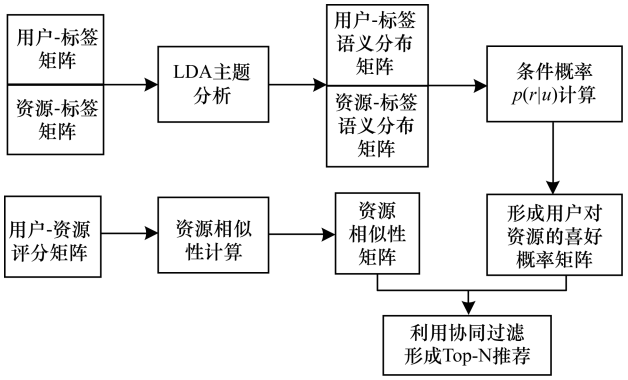


**图 2　推荐算法框架**

**定义 1**　用四元组 $F:=(U,T,R,A)$ 表示社会标签系统。表达式中用户集合表示为 $U=\{u_1,u_2,\cdots,u_i,\cdots,u_M\}$,$M$ 为用户总数,$i=1,2,\cdots,M$;资源集合表示为 $R=\{r_1,r_2,\cdots,r_j,\cdots,r_N\}$,$N$ 为资源总数,$j=1,2,\cdots,N$;标签集合表示为 $T=\{t_1,t_2,\cdots,t_k,\cdots,t_L\}$,$L$ 为标签总数,$k=1,2,\cdots,L$;$A$ 是一个包含用户、资源以及标签的三元关系集合,$A\in(U\times R\times T)$。若用户 $u$ 对资源 $r$ 打过标签 $t$,则 $(u,r,t)$ 是 $A$ 的一个元素,即 $(u,r,t)\in A$。根据系统中的历史标注行为,构造资源-标签的关系矩阵 $\boldsymbol{X}_{N\times L}$、用户-标签的关系矩阵 $\boldsymbol{Y}_{M\times L}$;$\boldsymbol{X}$ 中的元素 $x_{rt}$ 表示资源 $r$ 被打过标签 $t$ 的次数;$\boldsymbol{Y}$ 中的元素 $y_{ut}$ 表示用户 $u$ 被打过标签 $t$ 的次数。

### 2.1　语义层概率分布计算

首先将标签系统中收集得到的资源-标签矩阵 $\boldsymbol{X}$ 进行 LDA 潜在语义主题建模。设定系统中潜在主题个数为 $Z$,采用 Gibbs 抽样方法对资源-标签记录进行抽样:对资源 $r_i$ 上的每个标签 $t_i$ 进行多次重复抽样,估计由当前标签 $t_i$ 生成新的 Topic $z_i=n$ 的概率 $p(z_i=n|t_i,r_i,z_{-i})$。概率 $p(z_i=n|t_i,r_i,z_{-i})$ 同式(2),标签 $t_i$、资源 $r_i$ 分别对应公式中的 $w_i,d_i$。采样结束时,会得到多项式分布的超级参数 $\alpha$ 和 $\beta$。利用式(3)、式(4)得到资源上的主题分布概率 $p(z_i=n|r_i)$ 及主题中的标签分布概率 $p(t_i|z_i=n)$。

当计算出资源上的主题分布概率及主题中的标签分布概率后,利用公式 $p_{\mathrm{LDA}}(t|r)=\sum_{i=1}^{Z}p(t|z_i)p(z_i|r)$ 得到标签在资源上基于主题层的概率值,计算结束后,可得到语义层的资源-标签概率矩阵 $\boldsymbol{X}'$,元素 $x'_{jk}$ 表示资源 $r_j$ 上出现标签 $t_k$ 的概率 $p(t_k|r_j)$,资源 $r_j$ 基于语义层面的标签分布概率对应行向量 $\boldsymbol{x}'_j$。

将标签系统中收集得到的用户-标签矩阵 $\boldsymbol{Y}$ 同样进行 LDA 潜在语义主题建模。计算用户上的主题分布概率 $p(z_i=n|u_i)$ 及主题上的标签分布概率 $p(t_i|z_i=n)$,利用公式 $p_{\mathrm{LDA}}(t|u)=\sum_{i=1}^{Z}p(t|z_i)p(z_i|u)$ 得到标签在用户上基于主题层的概率值。在计算结束后,可得到语义层的用户-标签概率矩阵 $\boldsymbol{Y}'$,元素 $y'_{ik}$ 表示用户 $u_i$ 打过标签 $t_k$ 的概率 $p(t_k|u_i)$,用户 $u_i$ 基于语义层面将标签的喜好度对应行向量 $\boldsymbol{y}_j'$。

通过以上方法可以得到资源-标签语义分布矩阵 $\boldsymbol{X}'$和用户-标签语义分布矩阵 $\boldsymbol{Y}'$。

### 2.2　语义层用户-资源偏好矩阵计算

根据 2.1 节计算得到的用户对标签的偏好概率 $p_{\mathrm{LDA}}(t|u)$ 以及标签在资源上的分布概率 $p_{\mathrm{LDA}}(t|r)$ 来计算一个用户 $u$ 对资源 $r$ 的偏好概率,其大小代表用户对资源的偏好值。应用贝叶斯定理的 $p(r|u)$ 计算公式如下:

$$\begin{aligned}p(r|u)&=\sum_{t\in T}p(r|t)p(t|u)\\&=\sum_{t\in T}\frac{p(t|r)p(r)}{p(t)}p(t|u)\\&=p(r)\sum_{t\in T}\frac{p(t|r)p(t|u)}{p(t)}\end{aligned} \tag{5}$$

在三元关系集合 $A$ 中，$(u,t,r)\in A$ 表示用户 $u$ 对资源 $r$ 打了一个标签 $t$，将标注记录中的用户 $u$ 及其标注的资源 $r$ 定义为一条书签(Bookmark)，用 $B:=\{(u,r)\mid\exists t:(u,t,r)\in A\}$ 来表示。式(5)中的概率 $p(r)$ 的计算公式为：

$$p(r)=\frac{N_r}{N_B} \tag{6}$$

其中，$N_B$ 表示标注记录中所有书签的总数目；$N_r$ 表示标注记录中包含资源 $r$ 的书签数目。

标签 $t$ 的概率 $p(t)$ 计算公式为：

$$p(t)=\frac{N_{t\in T_b}}{N_B} \tag{7}$$

其中，$T_b$ 表示在书签 $b$ 中，用户 $u$ 对资源 $r$ 打过的所有标签的集合；$N_{t\in T_b}$ 表示标签 $t$ 出现在 $T_b$ 中的书签数目。

式(5)中的概率 $p(t|u)$ 与 $p(t|r)$ 使用 2.1 节基于 LDA 模型得到的用户对标签的偏好概率 $p_{\mathrm{LDA}}(t|u)$ 以及标签在资源上的分布概率 $p_{\mathrm{LDA}}(t|r)$ 表示。条件概率 $p(r|u)$ 的计算公式为：

$$p(r|u)=p(r)\sum_{t\in T}\frac{p_{\mathrm{LDA}}(t|r)p_{\mathrm{LDA}}(t|u)}{p(t)} \tag{8}$$

用户 $u\in U$ 对资源 $r\in R$ 喜爱的概率为 $p(r|u)$。依据用户对资源的偏好概率构造用户-资源偏好概率矩阵，记为：

$$\boldsymbol{p}_{M\times N}=\begin{bmatrix} p_{u_1r_1} & p_{u_1r_2} & \cdots & p_{u_1r_j} & \cdots & p_{u_1r_N}\\ p_{u_2r_1} & p_{u_2r_2} & \cdots & p_{u_2r_j} & \cdots & p_{u_2r_N}\\ \vdots & \vdots & & \vdots & & \vdots\\ p_{u_ir_1} & p_{u_ir_2} & \cdots & p_{u_ir_j} & \cdots & p_{u_ir_N}\\ \vdots & \vdots & & \vdots & & \vdots\\ p_{u_Mr_1} & p_{u_Mr_2} & \cdots & p_{u_Mr_j} & \cdots & p_{u_Mr_N} \end{bmatrix} \tag{9}$$

其中，$u_i\in U,i=1,2,\cdots,M$；$r_j\in R,i=1,2,\cdots,N$。该矩阵记录了用户对资源的偏好值，能够在一定程度上较好地反映出用户对所有资源的喜爱程度。

### 2.3 资源相似性计算

设用户对资源的评分矩阵为 $\boldsymbol{H}_{M\times N}$，元素 $H_{u_ir_j}$ 表示用户 $u_i$ 对资源 $r_j$ 的评分值。在用户评分矩阵 $\boldsymbol{H}$ 上寻找目标资源的相似资源集合，资源相似度采 Pearson 相关系数来度量，计算公式如式(10)所示。

$$s_{mn}=\frac{\sum_{u\in U_{mn}}(H_{ur_m}-\overline{H_{r_m}})(H_{ur_n}-\overline{H_{r_n}})}{\sqrt{\sum_{u\in U_{mn}}(H_{ur_m}-\overline{H_{r_m}})^2\sum_{u\in U_{mn}}(H_{ur_n}-\overline{H_{r_n}})^2}} \tag{10}$$

其中，$U_{mn}$ 为对资源 $r_m$ 与资源 $r_n$ 都做过评价的用户集合的交集；$\overline{H_{r_m}}$ 为所有用户对资源 $r_m$ 的评分均值；$\overline{H_{r_n}}$ 为所有用户对资源 $r_n$ 的评分均值。

通过式(10)计算系统中任意资源间的 Pearson 相似度，由此可构成资源相似度矩阵 $\boldsymbol{S}_{N\times N}$：

$$\boldsymbol{S}_{N\times N}=\begin{bmatrix} 1 & s_{12} & \cdots & s_{1j} & \cdots & s_{1N}\\ s_{21} & s_{22} & \cdots & s_{2j} & \cdots & s_{2N}\\ \vdots & \vdots & & \vdots & & \vdots\\ s_{j1} & s_{j2} & \cdots & 1 & \cdots & s_{jN}\\ \vdots & \vdots & & \vdots & & \vdots\\ s_{N1} & s_{N2} & \cdots & s_{Nj} & \cdots & 1 \end{bmatrix} \tag{11}$$

其中，$j=1,2,\cdots,N$；$s_{mn}$ 表示资源 $r_m$ 与资源 $r_n$ 的相似度。

### 2.4 偏好值预测

根据用户标注记录计算出的基于语义层的用户偏好矩阵以及用户评分记录计算出的资源相似度，可计算用户 $u$ 对未使用资源 $r_m$ 预测偏好值 $prob_{ur_m}$，该值表示用户 $u$ 对特定资源 $r$ 的偏好程度。

$$prob_{ur_m}=\frac{\sum_{j=1}^{Q}p_{ur_j}\times s_{r_jr_m}}{Q} \tag{12}$$

其中，$Q$ 为资源 $r_m$ 的邻近资源集合的个数；$p_{ur_j}$ 表示用户 $u$ 对曾使用资源 $r_j$ 的偏好程度；$s_{r_jr_m}$ 表示资源 $r_j$ 和 $r_m$ 的相似度。

在计算预测偏好值时，充分利用了用户的历史标注记录和用户评分记录，标签隐含的主题信息以及资源间的相似度，从而提高推荐算法的准确度。

## 3 协同过滤推荐算法描述

设 $\boldsymbol{X}_{N\times L}$ 为资源-标签的关系矩阵、$\boldsymbol{Y}_{M\times L}$ 为用户-标签的关系矩阵、$\boldsymbol{H}_{M\times N}$ 为用户-资源的评分矩阵。本文所提出的基于标签主题的协同过滤算法具体描述如下：

**算法 1** 基于标签主题的协同过滤算法

**输入** 用户-标签-资源记录、用户-资源评分记录、推荐列表资源数 $N$

**输出** 目标用户 $u$ 的 Top-N 推荐列表

**步骤 1** 根据输入信息，统计资源-标签的关系矩阵 $\boldsymbol{X}_{N\times L}$、用户-标签的关系矩阵 $\boldsymbol{Y}_{M\times L}$ 以及用户-资源评分矩阵 $\boldsymbol{H}_{M\times N}$。

**步骤 2** 计算基于语义层的用户对资源的偏好矩阵。按照图 2 所示的流程将 LDA 模型用于矩阵 $\boldsymbol{X}_{N\times L}$ 和矩阵 $\boldsymbol{Y}_{M\times L}$，然后计算出用户对资源的偏好矩阵 $\boldsymbol{p}_{M\times N}$。

**步骤 3** 基于用户资源的评分矩阵 $\boldsymbol{H}_{M\times N}$，根据式(10)计算资源间的 Pearson 相关系数，由此构成资源间的相似度矩阵 $\boldsymbol{S}_{N\times N}$。

**步骤 4** 根据步骤 1～步骤 3 的结果，用式(12)计算用户 $u$ 对未标注资源的预测偏好值 $prob_{ur_m}$。

**步骤 5** 将步骤 4 中计算得到的偏好值降序

排列,并取前 $N$ 个资源组成 Top-N 推荐列表 $Top\_list = \{r_1, r_2, \cdots, r_N\}$ 并输出。

## 4　实验结果与分析

### 4.1　实验数据集

本文实验采用的数据集是 GroupLens 项目组开放的 Movielens 数据集。该数据集来源于 Minnesota 大学开发的电影推荐系统 Movielens,用户可以对电影评分,分值范围为 1 ~5,同时可以对自己看过的电影打标签。

MovieLens 数据集包括用户评分数据集和用户-标签-电影数据集。该数据集包括 71 567 个用户对 10 681 部电影的 100 000 000 个评分记录和 100 000 个标签记录。该数据集中存在大量的多义和同义标签,如“play”既可能表示游戏,也可能表示戏剧;“rap”,“music”,“rock”均表示音乐类的电影。为减少实验数据集的稀疏度,对数据集进行预处理,仅选取每个标签被使用过 10 次以上以及每个用户对资源进行 10 次以上标注行为的记录作为实验数据集,数据集按 9∶1 的比例被随机分割为训练集和测试集。LDA 建模使用开放的 lingpipe 工具包,采用 Gibbs 抽样方法进行抽样,LDA 模型中的平滑参数 $\alpha$、$\beta$ 分别取 0.1,0.01,潜在主题数设定为 15,Gibbs 重复抽样的次数设定为 1 000。

### 4.2　评测指标

评价个性化推荐算法的最主要的评价指标为准确率(*Precision*)和召回率(*Recall*)。准确率是指利用推荐算法所产生的推荐列表中用户感兴趣的资源个数占整个推荐列表的比例,表示用户对系统推荐资源感兴趣的概率。召回率是指利用推荐算法所产生的推荐列表中用户感兴趣的资源的个数与系统中用户感兴趣的全部资源的比值,表示一个用户喜欢的资源被推荐的概率。

对于用户 $u \in U$,令 $P(u)$ 为给用户 $u$ 长度为 $N$ 的推荐列表,里面包含认为用户会打标签的物品。令 $D(u)$ 表示测试集中用户 $u$ 实际上打过标签的物品集合。推荐结果的准确率定义为:

$$Precision = \frac{\sum_{u \in U} P(U) \cap D(U)}{\sum_{u \in U} P(U)} \tag{13}$$

推荐结果的召回率定义为:

$$Recall = \frac{\sum_{u \in U} P(U) \cap D(U)}{\sum_{u \in U} D(U)} \tag{14}$$

准确率和召回率在一定程度上是相互制约的,为平衡两者之间的关系,引入综合衡量指标 $F$ 度量值(*F-measure*):

$$F\text{-}measure = \frac{2Precision \times Recall}{Precision + Recall} \tag{15}$$

### 4.3　结果分析

为了验证本文所提出的算法的推荐效果,设计 3 组实验进行对比。第 1 组 Collative 算法[8]是基于标签的协同过滤算法,利用标签计算资源间的相似度,采用协同过滤算法产生推荐结果;第 2 组 LDAProb 算法是根据 LDA 模型计算出的用户对资源的偏好概率,即式(8)求出的概率,将概率靠前的资源直接推荐给用户;第 3 组 ColLDA 算法是本文提出的基于标签主题的协同过滤算法。将上面的 3 组实验分别进行 5 次,每次选择不同的测试集和训练集,将所有实验的平均值作为实验结果。

在实验中,式(12)中 $Q$ 值的大小会对推荐结果产生影响,$Q$ 表示某个待推荐资源的邻近资源集合的个数。当进行 Top-N 推荐($N$ 值为 30)时,$F$ 度量值随 $Q$ 值的变化如图 3 所示。由图 3 可以看出,一开始随着 $Q$ 值的增大,$F$ 度量值逐渐增大,表明推荐结果逐渐变好;当 $Q$ 值为 200 时,$F$ 度量值最大,推荐效果达到最优;之后随着 $Q$ 值的逐渐增大,$F$ 度量值逐渐变小,推荐效果逐渐变差。实验结果表明,当 $Q$ 值为 200 时,以 $F$ 度量值为评价指标的推荐效果达到最佳,这说明 $Q$ 值的大小会对用户的偏好值产生影响。

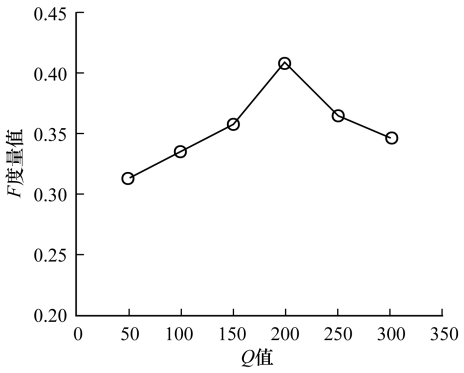


**图 3　$F$ 度量值随 $Q$ 值的变化情况**

当 $Q$ 值取 200 时,以推荐结果的准确率作为评价指标,3 种算法的对比实验结果如图 4 所示。

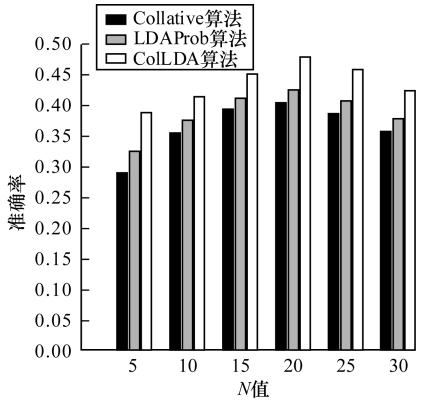


**图 4　3 种算法的准确率对比结果**

随着推荐资源个数 $N$ 的增大,3 种算法的推荐准确率先是不断增大,后逐渐缩小;当推荐资源个数 $N$ 为 20 时,3 种算法的准确率达到最优值。而且由图 4 可知,当推荐资源个数 $N$ 一定时,本文所提出的 ColLDA 算法的准确率高于 Collative 算法和 LDAProb 算法。

图 5 显示了本文所提出的 ColLDA 算法与 Collative 算法、LDAProb 算法在算法召回率上的对比结果,实验中 $Q$ 值取 200。实验结果表明,随着推荐资源个数 $N$ 的增大,3 种算法的召回率均逐渐增大,这说明随着推荐资源个数 $N$ 的增大,推荐列表中用户感兴趣的资源个数逐渐增多。在 $N$ 值固定时,本文所提出的 ColLDA 算法较 Collative 算法、LDAProb 算法具有更高的召回率。

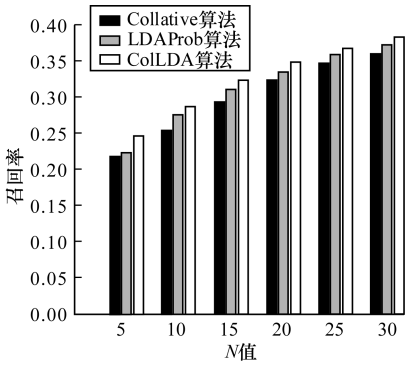


**图 5 3 种算法的召回率对比结果**

由于准确率和召回率在一定程度上是一对相互矛盾的指标,因此采用综合衡量指标 $F$ 度量值对3 种算法进行对比实验,实验中 $Q$ 值取 200。图 6 显示了 3 种算法在 $F$ 度量值上的对比结果,在不同的 $N$ 值上,本文所提出的 ColLDA 算法的 $F$ 度量值均高于 Collative 算法和 LDAProb 算法,并且在 $N$ 值为 25 时,3 种算法的 $F$ 度量值达到最优,说明当 $N$ 值为 25 时,3 种推荐算法在准确率和召回率之间达到了最佳平衡,这是由于 $N$ 值的大小会对准确率和召回率产生影响造成,$N$ 值过大或过小均不能取得最优解。

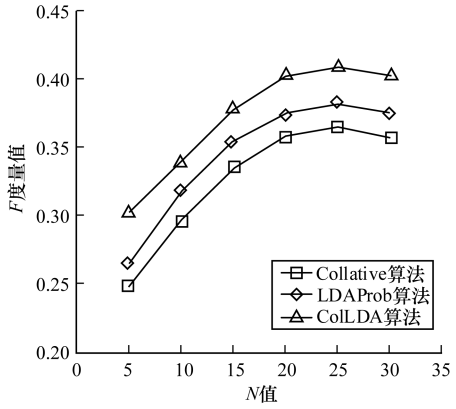


**图 6 3 种算法的 $F$ 度量值对比结果**

本文在 MovieLens 数据集上对 3 种算法的准确率、召回率和 $F$ 度量值评价指标进行对比实验。实验结果显示,在同义标签及多义标签存在的情况下,本文所提出的基于标签主题的协同过滤算法(ColLDA)较 Collative 算法和 LDAProb 算法能有效提高推荐结果的质量。

## 5 结束语

本文提出一种基于标签主题的协同过滤推荐算法,通过分析标签的特征,利用 LDA 模型分别对用户-标签矩阵、资源-标签矩阵进行建模,以此计算用户-资源偏好矩阵。在用户评分矩阵中利用协同过滤思想计算资源间的相似度,依据计算出的用户-资源偏好矩阵和资源间的相似度,预测用户偏好值,以此进行个性化推荐。最后在 Movielens 数据集上对推荐算法进行对比实验,结果证明,与传统基于标签的推荐算法相比,本文算法能够提高推荐准确度。下一步将利用上下文信息寻找提高标签质量的方法,并通过标签之间的层次关系提高推荐的准确度。

### 参考文献

[1] Kohi A,Ebrahimi S J,Jalali M. Improving the Accuracy and Efficiency of Tag Recommendation System by Applying Hybrid Methods[C]//Proceedings of the 1st International Conference on Computer and Knowledge Engineering. Washington D. C.,USA:IEEE Press,2011:242-248.

[2] 张 斌,张 引,高克宁,等. 融合关系与内容分析的社会标签推荐[J]. 软件学报,2012,23(3):476-488.

[3] 刘建国,周 涛,郭 强,等. 个性化推荐系统评价方法综述[J]. 复杂系统与复杂性科学,2009,6(3):1-10.

[4] Zheng N,Li Q. A Recommender System Based on Tag and Time Information for Social Tagging Systems[J]. Expert Systems with Applications,2011,38(4):4575-4587.

[5] Vinko Z,Gourab G,Guido C. Hypergraph Topological Quantities for Tagged Social Networks[J]. Physical Review E,2009,80(3):1-7.

[6] 李 贵,王 爽,李征宇,等. 基于张量分解的个性化标签推荐算法[J]. 计算机科学,2015,42(2):267-273.

[7] 李 敬,印 鉴,刘少鹏,等. 基于话题标签的微博主题挖掘[J]. 计算机工程,2015,41(4):30-35.

[8] Kim H N,Ji A T,Ha I,et al. Collaborative Filtering Based on Collaborative Tagging for Enhancing the Quality of Recommendation[J]. Electronic Commerce Research & Applications,2010,9(1):73-83.

[9] 王卫平,王金辉. 基于 Tag 和协同过滤的混合推荐方法[J]. 计算机工程,2011,37(14):34-35.

(下转第 258 页)

对象的尺寸较大时,差分背景探测器需要更多的时间来检测物体。虽然,文献[14]在数据集 Pet 2001 上实现了12 frame/s的处理速度,调整图像到原始大小的 1/4,即 384 像素 ×288 像素。然而,在该数据集中,本文使用原始尺寸的图像,即 768 像素 ×576 像素,达到了 6 frame/s 的速度。由于 Matlab 是解释性语言,虽然在处理矩阵运算方面有一定优势,但处理速度依然没有达到理想效果,使用 C + + 和并行处理可以实现高效处理,这也是本文未来的研究工作。

## 4 结束语

本文采用一种简单的背景差分方法,不需要任何离线训练操作,只使用背景图像来检测目标,与其目标形状或类型无关。利用背景图像信息以及外观模型与 IGM-PHD 跟踪器结合,可补偿漏检或背景差分探测器不准确检测的效果。实验结果表明,本文方法在检测和跟踪视觉目标上的有效性。未来将对外观模型进行深入研究,优秀的外观模型检测对 IGM-PHD 滤波的输入具有很重要的影响,在众多的方法中,弱监督方法的外观更新是研究热点。此外,如何设计并行程序以提高运行速度也是本文未来的研究工作。

### 参考文献


[1] 王 健. 基于目标跟踪的交通违停事件检测的研究[D]. 合肥:中国科学技术大学,2011.

[2] 孙新领,马绍惠,徐平平. 基于 SIR 粒子滤波和局部搜索算法的运动目标跟踪方案[J]. 湘潭大学自然科学学报,2016,38(2):84-88.

[3] 喻旭勇,王直杰. 基于灰度触发的 Mean Shift 自动跟踪算法[J]. 计算机工程,2014,40(1):228-231.

[4] Ning Jifeng,Zhang Lei,David Z,et al. Robust Mean-shift Tracking with Corrected Background-weighted Histogram[J]. IET Computer Vision,2012,17(1):62-69.

[5] Li Wenhui,Lin Yifeng,Fu Bo,et al. Cascade Classifier Using Combination of Histograms of Oriented Gradients for Rapid Pedestrian Detection[J]. Journal of Software,2013,8(1):1532-1539.

[6] 魏 运. 道路移动视觉环境感知中的多目标识别与跟踪方法研究[D]. 南京:东南大学,2013.

[7] Kalal Z,Mikolajczyk K,Matas J. Tracking-Learning-Detection[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2012,34(7):1409-1422.

[8] 连 峰. 基于随机有限集的多目标跟踪方法研究[D]. 西安:西安交通大学,2009.

[9] Kyriazis N,Argyros A. Scalable 3D Tracking of Multiple Interacting Objects[C]//Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Washington D. C.,USA:IEEE Press,2014:3430-3437.

[10] 周良毅,王 智,王营冠. 基于动态遮挡阈值的多视角多目标协作追踪[J]. 计算机研究与发展,2014,51(4):813-823.

[11] Mahler R. Multitarget Bayes Filtering via First-order Multitarget Moments[J]. IEEE Transactions on Aerospace and Electronic System,2003,39(4):1152-1178.

[12] 王 晓,韩崇昭,连 峰. 基于随机有限集的目标跟踪方法研究及最新进展[J]. 工程数学学报,2012,27(4):567-578.

[13] Ba-Ngu V N,Ma Wing-Kin. The Gaussian Mixture Probability Hypothesis Density Filter[J]. IEEE Transactions on Signal Processing,2006,54(11):4091-4104.

[14] Zhou Xiaolong,Li Youfu,He Bingwei,et al. GM-PHD-based Multi-target Visual Tracking Using Entropy Distribution and Game Theory[J]. IEEE Transactions on Industrial Informatics,2014,10(2):1064-1076.

[15] Bernardin K,Stiefelhagen R. Evaluating Multiple Object Tracking Performance:The CLEAR MOT Metrics[J]. Image Video Process,2008,23(1):1-10.


编辑 刘 冰

(上接第 252 页)


[10] 蔡 强,韩东梅,李海生,等. 基于标签和协同过滤的个性化资源推荐[J]. 计算机科学,2014,41(1):69-71.

[11] Zeng D,Li H. How Useful Are Tags ——An Empirical Analysis of Collaborative Tagging for Web Page Recommendation[C]//Proceedings of IEEE ISI 2008 PAISI,PACCF,and SOCO International Workshops on Intelligence and Security Informatics. Berlin,Germany:Springer,2008:320-330.

[12] Sen S,Vig J,Riedl J. Tagommenders:Connecting Users to Items Through Tags[C]//Proceedings of the 18th International Conference on World Wide Web. New York,USA:ACM Press,2009:671-680.

[13] Zhao S,Du N,Nauerz A,et al. Improved Recommendation Based on Collaborative Tagging Behaviors[C]//Proceedings of the 13th International Conference on Intelligent User Interfaces. New York,USA:ACM Press,2008:413-416.

[14] Givon S,Lavrenko V. Predicting Social-tags for Cold Start Book Recommendations[C]//Proceedings of Conference on Recommender Systems. New York,USA:ACM Press,2009:333-336.

[15] 陈毅波,揭志忠,吴产乐. 基于同义标签分组的协同推荐[J]. 湖南大学学报(自然科学版),2011,38(5):83-88.

[16] Blei D M,Ng A Y,Jordan M I. Latent Dirichlet Allocation[J]. Journal of Machine Learning Research,2003,3:993-1022.


编辑 陆燕菲