基于潜在主题的混合上下文推荐算法

李 平[①②] 张路遥*[①] 曹 霞[①] 胡检华[①]
[①](长沙理工大学计算机与通信工程学院 长沙 410114)
[②](智能交通大数据处理湖南省重点实验室 长沙 410114)

摘 要：针对单个环境上下文中项目访问记录稀疏的问题，推荐系统难以获取与当前环境上下文关联的用户偏好。该文设计了一种新的上下文关联性推荐(CTRR)算法。CTRR 算法通过 CTRR_LDA 模型求解推荐项目出现在特定环境上下文的概率，并结合上下文后过滤推荐算法，对用户进行推荐。CTRR_LDA 模型是在(LDA)模型的基础上，结合环境上下文和项目特征上下文，提出的项目与环境上下文的关联概率模型。该模型将环境上下文划分为多个环境上下文因子，每个环境上下文因子表示为 K 维的主题分布，挖掘环境上下文因子中项目出现的潜在主题特征。利用 LDOS-CoMoDa 网站上真实的电影数据集进行实验，验证了算法的可靠性。

关键词：推荐；关联概率；潜在主题；环境上下文；项目特征上下文
中图分类号： TP391 **文献标识码**： A **文章编号**：1009-5896(2018)04-0957-07
DOI: 10.11999/JEIT170623

Hybrid Context Recommendation Algorithm Based on Latent Topic

LI Ping[①②] ZHANG Luyao[①] CAO Xia[①] HU Jianhua[①]
[①](*School of Computer and Communication Engineering, Changsha University of Science and Technology, Changsha* 410114, *China*)
[②](*Hunan Provincial Key Laboratory of Intelligent Processing of Big Data on Transportation, Changsha* 410114, *China*)

Abstract: In the recommendation system, a critical challenge is that individual environment context log may not contain sufficient item access records for mining his/her environment context preferences. This paper designs a Contextual Topic-based Relevance Recommendation (CTRR) algorithm. The CTRR algorithm uses the CTRR_LDA model and a postfiltering strategy to recommend items to users in a specific environment context. CTRR_LDA is an improved LDA model, which combines environment contexts and item feature contexts to calculate the probability of the item appeared. In this model, the environment context is divided into multiple environment context factors. Each environment context factor can be expressed as a K-dimensional topic distribution. Then the CTRR_LDA model is used to mine the latent topic of the items in each environment context factor. According to the experimental results on the LDOS-CoMoDa datasets, the reliability of algorithm is validated in the context-aware recommendation scenario.

Key words: Recommendation; Relevance probability; Latent topic; Environment context; Item feature context

1 引言

近年来，随着移动设备感知能力不断提升，推荐系统能从移动智能终端的上下文日志中获取用户的偏好和丰富的上下文信息。上下文信息对推荐系统产生了不容小觑的影响，受到工业界和学术界的广泛关注。在工业界推荐方面，京东到家手机 APP 利用环境上下文如用户下单和收单的位置、时间以及商品特征上下文如商品的价位向用户提供生鲜及超市产品配送。在学术界推荐领域，上下文感知推荐通过在传统的推荐算法中融入上下文信息，得到满足用户偏好同时符合上下文要求的推荐列表[1−6]。

目前，上下文后过滤推荐[7,8]是上下文感知推荐系统中较为成熟的算法。基本思想是将项目与当前上下文关联的概率[4]和已预测评分进行加权。但传统的上下文后过滤推荐在项目与上下文的关联概率求解上仅仅考虑了项目在环境上下文中出现频次的特性，忽略了项目多元的特征上下文。例如在歌曲推荐系统中，两首歌在书店这个场景下出现频次相等，

收稿日期：2017-06-28；改回日期：2017-11-20；网络出版：2017-12-12
*通信作者：张路遥 465888329@qq.com
基金项目：湖南省教育厅资助重点项目(14A004)
Foundation Item: The Scientific Research Fund of Hunan Provincial Education Department (14A004)

但针对在书店特定场景下出现的音乐类型的分析，轻音乐类型的歌曲在未来收听的概率将大于摇滚乐类型的歌曲。因此，传统的算法不能真实地反映项目与上下文之间的关联概率，同时由于单个环境上下文中项目的访问记录稀疏，推荐准确率并不理想。

针对以上问题，国内外学者提出多种改进思路。文献[8]采用图模型和上下文后过滤推荐结合的方式求解用户对电影的偏好。文献[9,10]提出利用公开关联数据库 DBpedia, OER 等获取项目特征语义的实体描述，进一步地丰富了推荐系统的项目特征上下文，并利用主题模型挖掘用户偏好的项目，缓解了数据稀疏性的问题，但没有考虑位置、时间等环境因素，推荐的准确率较差。文献[6]提出张量分解算法直接对用户、项目、各类别的上下文变量进行因式分解，来改善推荐的质量，但该算法通常采用梯度下降法求解核心张量和各个因子矩阵，迭代次数非常高。因此，算法的效率是目前该方法的一个瓶颈。文献[11,12]提出利用主题模型在大量用户的上下文日志中挖掘共同上下文感知偏好，并将每个用户的上下文感知偏好用共同上下文感知偏好表示。该算法缓解了数据稀疏对推荐系统的影响，但未考虑项目多元的特征信息。文献[13]提出基于位置感知的概率生成模型 LA-LDA。LA-LDA 模型由两个部分组成，将用户表示成一个文档，用 K 维的主题分布来表示用户的特征。同时考虑用户位置周边用户偏好对该用户偏好的影响和项目的位置主题特征，该方法在一定程度上提升了推荐的准确率。文献[14]将时间因素引入主题模型，通过分析社交媒体中用户的行为记录，提出一种动态的上下文感知混合模型，挖掘用户不断变化的兴趣。

为了获取与当前环境上下文关联的用户偏好，进一步地提升推荐准确率，本文在主题模型 LDA 的基础上，结合环境上下文和项目特征上下文，提出项目与环境上下文的关联概率模型：CTRR_LDA 模型。将每个环境上下文划分为多个环境上下文因子，挖掘环境上下文因子中项目出现的潜在主题特征，并且将每个环境上下文因子表示为 K 维的主题分布，主题的分布由项目分布和多元的项目特征上下文分布共同决定，利用潜在主题分布挖掘推荐项目出现在特定环境上下文的概率。同时，本文设计了一种上下文关联性推荐 CTRR 算法。CTRR 算法通过 CTRR_LDA 模型和上下文后过滤推荐算法相结合，对用户进行推荐。本文提出的基于主题模型的上下文关联性推荐算法缓解了数据稀疏性的问题，并结合上下文后过滤，考虑了用户的评分信息，提升了推荐结果的准确性。

2 上下文关联性推荐 CTRR 算法

2.1 CTRR 推荐算法框架

CTRR 算法总体可以分为两个部分，算法框架如图 1 所示。左半部分为本文提出的项目与环境上下文的关联概率模型 CTRR_LDA 模型。右半部分结合上下文后过滤推荐算法，考虑用户的评分信息，对用户进行推荐。

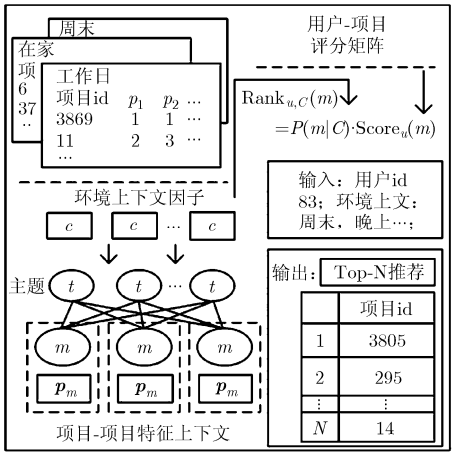

图 1 CTRR 算法框架

在上下文日志中一条环境上下文记录可以表示为{用户 id;环境上下文;项目 id;项目特征上下文}。其中，环境上下文是描述项目出现情景的信息，项目特征上下文是描述项目自身特征的信息。算法将 n 个环境上下文变量(时间，天气等)表示为一个环境上下文向量 $\boldsymbol{C}=\{C_1,C_2,\cdots,C_n\}$，一条记录中的环境上下文可以表示为 $\boldsymbol{c}=\{c_1,c_2,\cdots,c_n\}$。其中 c_i $(i=1,2,\cdots,n)$ 为环境上下文变量的具体内容，称之为环境上下文因子。例如 C_i 表示天气，c_i 的取值有晴天、下雨等。但单个环境上下文因子中项目的访问记录较为稀疏，不能很好地反映环境上下文和项目之间的关联关系。由此，本文提出挖掘每个环境上下文因子的 K 维潜在主题分布。同时考虑到项目多元的特征信息如电影类型、导演等，并将其构成项目 m 的 n 维项目特征上下文向量 $\boldsymbol{p}_m=\left\{p_1,p_2,\cdots,p_{N_m}\right\}$，综合潜在主题中项目特征上下文的分布，求解各个环境上下文因子与项目的关联概率，结合用户对项目的评分信息，最终返回用户在多个环境上下文因子条件下未评分项目推荐列表。

2.2 关联概率模型 CTRR_LDA 模型

本文在 LDA 模型[15]的基础上提出 CTRR_LDA 模型，构建环境上下文、主题、项目与项目特征上下文的 3 层结构。CTRR -LDA 模型如图 2 所示。

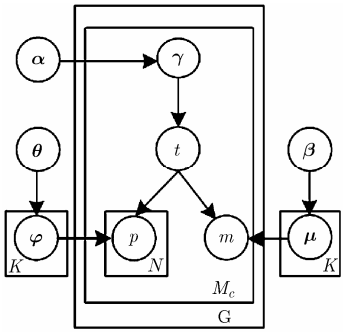

图 2　CTRR_LDA 模型

在 CTRR_LDA 模型中，将每个环境上下文因子 c 表示为一个文档，文档由该环境上下文因子 c 中出现过的所有项目 M_c 和各个项目的多元特征上下文 $\boldsymbol{p}_m$ 构成。每个环境上下文因子可以由 K 维的主题分布表示。每一维为环境上下文因子中一类主题出现的概率，主题的分布由项目分布和项目特征上下文分布共同决定。CTRR_LDA 模型通过环境上下文因子中潜在的主题分布挖掘环境上下文因子和项目的潜在关联关系。

CTRR -LDA 模型的生成过程描述如下：

(1)对于每一个主题 t_i，其中 $t_i \in \{t_1, t_2, \cdots, t_K\}$，项目的多项分布 $\boldsymbol{\mu}_{t_i}$ 满足 $\mathrm{Dir}\left(\boldsymbol{\mu}_{t_i} \middle| \boldsymbol{\beta}\right)$ 的先验分布，其中 $\boldsymbol{\beta}$ 为狄利克雷先验参数。

(2)对于每一个主题 t_i，其中 $t_i \in \{t_1, t_2, \cdots, t_K\}$，项目特征上下文的多项分布 $\boldsymbol{\varphi}_{t_i}$ 满足 $\mathrm{Dir}\left(\boldsymbol{\varphi}_{t_i} \middle| \boldsymbol{\theta}\right)$ 的先验分布，其中 $\boldsymbol{\theta}$ 为狄利克雷先验参数。

(3)对于环境上下文 G 中每一个环境上下文因子 c，其中 $c \in \mathrm{G}$。

(a)主题的多项分布 $\boldsymbol{\gamma}$ 满足 $\mathrm{Dir}(\boldsymbol{\gamma}|\boldsymbol{\alpha})$ 的先验分布，其中 $\boldsymbol{\alpha}$ 为狄利克雷先验参数；

(b)环境上下文因子 c 中出现的第 i 个项目 m_i；

(i)项目 m_i 所对应的主题 $t_i \sim \mathrm{Mult}(\boldsymbol{\gamma}_c)$；

(ii)项目 $m_i \sim \mathrm{Mult}\left(\boldsymbol{\mu}_{t_i}\right)$；

(iii)项目 m_i 中的一个特征上下文 $p_j \sim \mathrm{Mult}\left(\boldsymbol{\varphi}_{t_i}\right)$。

根据以上对 CTRR_LDA 模型生成过程的描述，在给定先验参数 $\boldsymbol{\alpha}, \boldsymbol{\theta}, \boldsymbol{\beta}$ 条件下，项目、项目特征上下文、主题和各个隐性变量的联合概率分布，如式(1)所示。

$$\begin{aligned} &P(\boldsymbol{\gamma}, \boldsymbol{\varphi}, \boldsymbol{\mu}, \boldsymbol{t}, \boldsymbol{m}, \boldsymbol{p} | \boldsymbol{\alpha}, \boldsymbol{\theta}, \boldsymbol{\beta}) \\ &= \prod_{c=1}^{\mathrm{G}} P(\boldsymbol{\gamma}_c|\boldsymbol{\alpha}) \prod_{k=1}^{K} P(\boldsymbol{\varphi}_k|\boldsymbol{\theta}) \prod_{k=1}^{K} P(\boldsymbol{\mu}_k|\boldsymbol{\beta}) \\ &\quad \cdot \prod_{i=1}^{M_c} \left((P(t_i|\boldsymbol{\gamma}_c) P(m_i|t_i, \boldsymbol{\mu}) \prod_{j=1}^{N} P\left(p_j|t_i, \boldsymbol{\varphi}\right) \right) \end{aligned} \quad (1)$$

在式(1)中，G 个环境上下文中主题的生成过程、K 个主题中项目的生成过程、K 个主题中项目特征上下文的生成过程是相互独立的，所以模型生成过程有(G+2K)个相互独立的 Dirichlet-Multinomial 共轭分布。其中，CTRR_LDA 模型中 $\boldsymbol{\gamma}_c$ 表示每个环境上下文中主题的分布，其反映了先前环境上下文对项目主题的偏好。$\boldsymbol{\varphi}$ 为主题 t_i 中项目特征上下文的分布，以 $P\left(p_j \middle| t_i, \boldsymbol{\varphi}\right)$ 的后验概率生成项目 m_i 中第 j 个特征上下文。

2.3 CTRR_LDA 模型推导

目前，有多种算法可以用来估计主题模型中的参数。比如 EM 算法和吉布斯采样算法。本文的 CTRR_LDA 模型中采用 Gibbs Sampling 算法对分布进行采样，由此对主题模型进行训练。在采样过程中，利用后验概率通过迭代不断更新项目的主题，直到项目的分布达到一个平稳的状态，也就是马尔科夫链收敛。在已知项目、项目特征上下文和隐性变量条件下项目主题的后验概率如式(2)所示。

$$\begin{aligned} P\left(\boldsymbol{t} \middle| \boldsymbol{m}, \boldsymbol{p}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\theta}\right) &= \frac{P\left(\boldsymbol{t}, \boldsymbol{m}, \boldsymbol{p} \middle| \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\theta}\right)}{P\left(\boldsymbol{m}, \boldsymbol{p} \middle| \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\theta}\right)} \\ &\propto P\left(\boldsymbol{t}, \boldsymbol{m}, \boldsymbol{p}, \middle| \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\theta}\right) \\ &\propto P(\boldsymbol{t}) P(\boldsymbol{m}|\boldsymbol{t}) \prod_{j=1}^{N} P\left(\boldsymbol{p}_j|\boldsymbol{t}\right) \end{aligned} \quad (2)$$

本文定义 $\boldsymbol{p}_{m_i} = \left\{p_1, p_2, \cdots, p_{N_{m_i}}\right\}$。模型中主题向量 $\boldsymbol{t}$ 是一个潜在变量，$-i$ 为去除第 c 个上下文中第 i 项目 m_i，根据上下文记录中其他的项目和项目特征上下文，结合贝叶斯法则求解该项目的主题，如式(3)所示。

$$\begin{aligned} &P\left(t_i{=}z \middle| m_i{=}x, \boldsymbol{p}_{m_i}{=}\boldsymbol{y}_{m_i}, \boldsymbol{t}_{-i}, \boldsymbol{m}_{-i}, \boldsymbol{p}_{-i}\right) \\ &\quad \propto \int P\left(t_i{=}z, m_i{=}x, \boldsymbol{P}_{m_i}{=}\boldsymbol{Y}, \right. \\ &\quad \left. \boldsymbol{\gamma}_c, \boldsymbol{\varphi}_z, \boldsymbol{\mu}_z \middle| \boldsymbol{t}_{-i}, \boldsymbol{m}_{-i}, \boldsymbol{P}_{-i}\right) \mathrm{d}\boldsymbol{\gamma}_c \boldsymbol{\varphi}_z \boldsymbol{\mu}_z \\ &= \int P\left(t_i{=}z \middle| \boldsymbol{\gamma}_c\right) \mathrm{Dir}\left(\boldsymbol{\gamma}_c \middle| \boldsymbol{n}_{c,-i}{+}\boldsymbol{\alpha}\right) \mathrm{d}\boldsymbol{\gamma}_c \\ &\quad \cdot \int P\left(m_i{=}x \middle| \boldsymbol{\mu}_z\right) \mathrm{Dir}\left(\middle| \boldsymbol{\mu}_z \middle| \boldsymbol{n}_{z,-i}{+}\boldsymbol{\beta}\right) \mathrm{d}\boldsymbol{\mu}_z \\ &\quad \cdot \prod_{j=1}^{N_{m_i}} \int P\left(p_j{=}y \middle| \boldsymbol{\varphi}_z\right) \mathrm{Dir}\left(\boldsymbol{\varphi}_z \middle| \boldsymbol{n}'_{z,-i}{+}\boldsymbol{\theta}\right) \mathrm{d}\boldsymbol{\varphi}_z \\ &= \frac{n^{(c)}_{z,-i}+\alpha_z}{\sum_z \left(n^{(c)}_{z,-i}+\alpha_z\right)} \cdot \frac{n^{(z)}_{x,-i}+\beta_x}{\sum_x \left(n^{(z)}_{x,-i}+\beta_x\right)} \\ &\quad \cdot \prod_{p \in \boldsymbol{P}_{m_i}} \frac{n^{(z)}_{y,-i}+\theta_y}{\sum_y \left(n^{(z)}_{y,-i}+\theta_y\right)} \end{aligned} \quad (3)$$

由此，推导出 3 个隐性变量 $\hat{\gamma}_{cz}, \hat{\varphi}_{zx}, \hat{\mu}_{zy}$ 分别为

$$\hat{\gamma}_{cz}=\frac{n_{z,-i}^{(c)}+\alpha_z}{\sum_z\left(n_{z,-i}^{(c)}+\alpha_z\right)} \tag{4}$$

$$\hat{\varphi}_{zx}=\frac{n_{x,-i}^{(z)}+\beta_x}{\sum_x\left(n_{x,-i}^{(z)}+\beta_x\right)} \tag{5}$$

$$\hat{\mu}_{zy}=\frac{n_{y,-i}^{(z)}+\theta_y}{\sum_y\left(n_{y,-i}^{(z)}+\theta_y\right)} \tag{6}$$

其中，$\hat{\gamma}_{cz}$，$\hat{\varphi}_{zx}$，$\hat{\mu}_{zy}$ 分别为上下文 c 中主题 z 出现的概率，主题 z 中项目 x 出现的概率和主题 z 中项目特征上下文 y 出现的概率。$n_{z,-i}^{(c)}$，$n_{x,-i}^{(z)}$，$n_{y,-i}^{(z)}$ 分别为上下文 c 中主题 z 出现项目的个数，主题 z 中项目 x 出现的次数，主题 z 中项目特征上下文出现 y 的频次。于是，每个环境的上下文因子 c 的 K 维主题表示为 $\hat{\boldsymbol{\gamma}}_{\boldsymbol{c}}=(\hat{\gamma}_{c1},\hat{\gamma}_{c2},\cdots,\hat{\gamma}_{cK})$。

2.4 上下文后过滤推荐

在本节中，本文算法将结合上下文后过滤推荐算法，同时考虑用户的评分信息和项目与环境上下文的关联概率，生成用户在特定上下文条件下项目的推荐列表。

根据用户-项目 2 维评分矩阵和相似度计算方法得到用户间的相似度为 $S_{u,v}$，相似度最高的前 n 个用户作为用户 u 的近邻用户集 N_u。根据近邻用户集历史访问纪录预测用户对项目的评分 $\mathrm{Score}_u(m)$。如式(7)所示。

$$\mathrm{Score}_u(m)=\bar{r}_u+\frac{\sum_{v\in N_u}S_{u,v}\left(r_{v,m}-\bar{r}_v\right)}{\sum_{v\in N_u}S_{u,v}} \tag{7}$$

其中，$\bar{r}_u$ 为用户 u 对已评分项目的平均评分，$r_{v,m}$ 为用户 v 对项目 m 的评分。之后，根据 CTRR_LDA 模型得到项目 m 在特定的环境上下文 $C=\{c_1,\cdots,c_k,\cdots,c_n\}$ 下出现的概率，如式(8)所示。

$$\begin{aligned}P(m|C)&=\sum_{c_k\in C}P(m|c_k)P(c_k|C)\\&=\frac{1}{|C|}\sum_{c_k\in C}\sum_t\left(\mu_{tm}\gamma_{ct}\prod_{j=1}^{N}\varphi_{tp_j}\right)\end{aligned} \tag{8}$$

假设各个环境上下文因子之间是条件独立的，并在推荐算法中占同等权重，由此 $P(c_k|C)$ 可以简化为 $1/|C|$。其中 c_k 为一个环境上下文因子，γ_{ct}，μ_{tm}，φ_{tp_j} 可以由式(4)，式(5)，式(6)求得。上下文后过滤利用项目和上下文的关联概率对已预测评分加权调整，生成用户在特定上下文的项目推荐分数 $\mathrm{Rank}_{u,C}(m)$ 为

$$\mathrm{Rank}_{u,C}(m)=P(m|C)\cdot\mathrm{Score}_u(m) \tag{9}$$

推荐分数由高到低排列，选取前 n 个项目生成用户 u 在特定环境上下文 C 条件下的 Top-N 推荐列表。通过上述对算法的描述，可设计出算法的整体过程，其伪代码描述如表 1 所示。

表 1 CTRR 算法

上下文关联性推荐算法(CTRR)

Input: the dataset of LDOS-CoMoDa; parameters $\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\theta}$;

Output: item-context relevance probability $\mathrm{Rank}_{u,C}(m)$;

(1) for each topic $t\in K$ do

(2) draw a distribution over movies $\boldsymbol{\mu}_t\sim\mathrm{Dir}(\boldsymbol{\beta})$;

(3) draw a distribution over features $\boldsymbol{\varphi}_t\sim\mathrm{Dir}(\boldsymbol{\theta})$;

(4) end

(5) for each environmental context $c\in C$ do

(6) draw a distribution over topic $\boldsymbol{\gamma}_c\sim\mathrm{Dir}(\boldsymbol{\alpha})$;

(7) for each movie $m\in M_c$ do

(8) assign a topic t given the environment context $t\sim\mathrm{Mult}(\boldsymbol{\gamma}_c)$;

(9) draw a movie from the chosen topic $m_{c,t}\,|\,t\sim\mathrm{Mult}(\boldsymbol{\mu}_t)$;

(10) draw a feature from the chosen topic $p_{c,t}|t\sim\mathrm{Mult}(\boldsymbol{\varphi}_t)$;

(11) end

(12) end

(13) compute $P(m|C)=\frac{1}{|C|}\sum_{c_k\in C}\sum_t\left(\mu_{tm}\gamma_{ct}\prod_{j=1}^{N}\varphi_{tp_j}\right)$;

(14) compute $\mathrm{Rank}_{u,C}(m)=P(m|C)\cdot\mathrm{Score}_u(m)$;

(15) return $\mathrm{Rank}_{u,C}(m)$

在 CTRR 算法中，CTRR_LDA 模型采用 Gibbs Sampling 算法对分布进行采样，通过迭代不断更新项目的主题。对于每次迭代，更新项目主题的时间复杂度为 $O(\mathrm{G}\times M_c)$，G, M_c 分别为环境上下文因子的个数和环境上下文因子 c 中出现项目的个数。结合上下文后过滤推荐最终时间复杂度为 $O(N\times\mathrm{G}\times M_c+U\times U)$，其中 N 为 CTRR_LDA 模型的迭代次数，U 为用户的个数。因此，CTRR 算法相比主题模型只增加了额外的时间复杂度 $O(U\times U)$，算法的时间复杂度是可以接受的。

3 实验及结果分析

3.1 数据集及实验环境

本文实验使用具有丰富上下文信息的 LDOS-CoMoDa 电影数据集[7]，数据集的统计信息如表 2 所示，数据集中包含 121 个用户对 1232 个项目的 2296 次评分数据和相应的上下文信息。本文选取环境上下文为时间、日期类型和地点，项目特征上下

表 2 数据集统计信息

LDOS-CoMoDa 数据集：	
用户	121
电影	1232
评分	2296
评分等级	1~5
环境上下文	12
项目特征上下文	7
稀疏度	98.46%

文为电影地区和电影的上映时间，根据文献[16]认为这 5 项是与用户选择项目强相关的上下文变量。表 2 为数据集的统计信息，从表中可以看到数据的稀疏度为 98.46%，在测试过程中，本文只考虑项目内容非空的上下文记录。

3.2 基准算法与评价指标

本文采取 5 折交叉验证方法进行实验。由以下 3 个评测指标：准确率 Precision@N、召回率 Recall@N 和 F1 值来检测算法的推荐性能。衡量指标为

$$\text{Precision@N}=\frac{\left|\{\text{relevant items}\}\cap\{\text{Top-N items}\}\right|}{N} \tag{10}$$

$$\text{Recall@N}=\frac{\left|\{\text{relevant items}\}\cap\{\text{Top-N items}\}\right|}{\left|\{\text{relevant items}\}\right|} \tag{11}$$

$$\text{F1}=\frac{2\times\text{Precision@N}\times\text{Recall@N}}{\text{Precision@N}+\text{Recall@N}} \tag{12}$$

式(10)和式(11)中相关项目是指测试集中测试用户在特定上下文组合中已评分，且评分高于 3 分的电影，Top-N 指推荐列表中前 N 个项目。本文评测指标是测试集用户的准确率均值和召回率均值。F1 值综合准确率和召回率，用于衡量推荐算法的整体性能。

3.3 实验结果及分析

本节中，根据 2.3 节 CTRR_LDA 模型推导对参数进行设置和调整。首先将主题的数目设为 25，参数 α, β, θ 的初始值分别设置为 $2, 0.2, 0.2$，迭代次数设置为 200 次。本文实验采用的对比算法如表 3 所示，为方便描述，实验结果图中相应算法用缩写表示。

表 3 本文算法和对比实验算法

缩写	全称
Weight PoF	基于线性加权的上下文后过滤推荐算法
CGR-IA	融合项目特征上下文的图模型推荐算法
LPM	基于上下文的潜在偏好推荐算法
CTRR	基于主题模型的上下文关联性推荐算法

图 3、图 4 分别为 4 种算法准确率和召回率的实验结果对比。如图所示，随着推荐项目数量的增加，各个算法准确率整体呈现下降的趋势，召回率呈现不断上升的趋势。4 种算法中，Weight PoF 算法利用统计方法求解环境上下文和项目之间的关联关系，在稀疏的数据集上劣势明显，准确率和召回率最低。CGR-IA 算法和 LPM 算法的曲线基本一致。本文提出算法在稀疏数据集上有明显的优势，在推荐项目个数为 11 时，准确率较文献[8]提出的 LPM 算法提升 2%，随着推荐项目的增加，准确率逐渐平稳在 0.017 左右。并且推荐项目每增加 10，召回率会有 0.05 左右的提升。

本文采用综合评测指标 F1 值，综合考虑推荐的准确率和召回率，反映推荐算法的总体性能。从表 4 中可以看到，CTRR 推荐算法在推荐项目个数为 11 时，综合推荐性能最优。CTRR 推荐算法和传统的 Weight PoF 算法相比，改善效果明显，F1 值有大幅度的提升，平均提高 73.602%, CTRR 推荐算法和 LPM 推荐算法相比，F1 值也有很大的改善，平均提高 42.321%。

以上利用 3 种评测指标对本文提出的算法和先前的算法进行对比实验，在实验中根据先验知识将

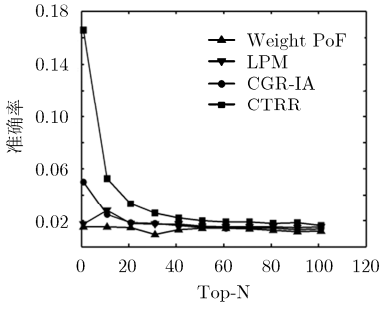

图 3　4 种算法准确率对比

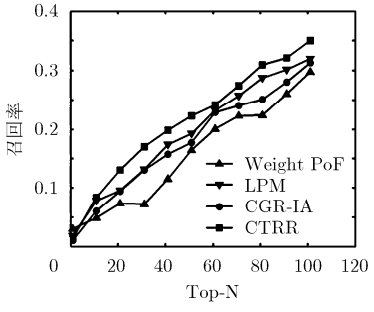

图 4　4 种算法召回率对比

表 4 4 种算法的 F1 值

名称	F@11	F@31	F@51	F@71	F@91
Weight PoF	0.0232	0.0165	0.0257	0.0255	0.0215
LPM	0.0407	0.0299	0.0285	0.0282	0.0279
CGR-IA	0.0346	0.0312	0.0266	0.0260	0.0246
CTRR	**0.0645**	**0.0446**	**0.0362**	**0.0348**	**0.0347**

CTRR-LDA 模型中的主题数目固定为 25。由于主题的数目是主题模型中的一个重要的预设参数，接下来，本文通过实验来分析主题数目对返回推荐列表推荐质量的影响。

图 5、图 6 分别为不同主题个数对应的准确率和召回率变化曲线。CTRR-LDA 模型中主题的个数为自变量，取值为 5 到 50，步长为 5。图中两条曲线分别表示项目推荐个数为 100 和 50 的情况，这两种情况在主题个数少于 10 时，准确率低于 0.012，召回率低于 0.05。这是因为在选取的主题个数很少的时候，大多数项目对每一个主题都有很强的相关性，这样在实验中实际所有的项目都并为一类，产生大量的噪音，导致推荐质量很低。在主题个数较大时，准确率和召回率曲线趋于平稳，并在主题个数为 25，Recall@100 和 Recall@50 召回率达到峰值 0.3501 和 0.2225，于是我们将其峰值的主题个数选作实验部分主题的预设参数。取值不同的推荐项目个数，主题个数对应准确率和召回率都具有相似的曲线。

4 结束语

本文提出 CTRR 推荐算法，能充分利用项目特征上下文信息和环境上下文信息。同时创新性地提出利用主题模型来求解项目出现在特定环境上下文中的概率，之后结合上下文后过滤得到的用户的项目推荐分数，返回 Top-N 推荐列表。通过在真实数据集 LDOS-CoMoDa 上的实验结果，表明该算法很好地缓解了传统推荐算法中数据稀疏性的问题，近一步地提升了 Top-N 推荐的准确率。本文未能考虑到用户社交上下文，如用户之间相似关系、信任关系等。用户社交上下文、项目特征上下文和环境上下文的融合是未来的工作重点，期望能达到更好的推荐效果。

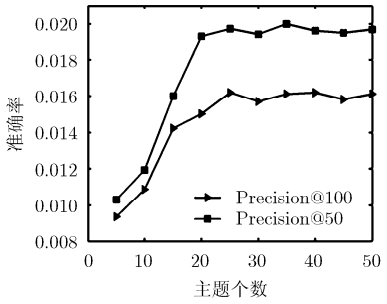

图 5 不同主题个数对应的准确率变化曲线

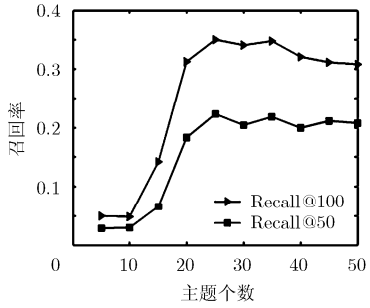

图 6 不同主题个数对应的召回率变化曲线

参 考 文 献

[1] ABOWD G D, DEY A K, BROWN P J, *et al*. Towards a better understanding of context and context-awareness[C]. International Symposium on Handheld and Ubiquitous Computing, Karlsruhe, Germany, 1999: 304–307.

[2] BALTRUNA L, LUDWIG B, PEER S, *et al*. Context relevance assessment and exploitation in mobile recommender systems[J]. *Personal and Ubiquitous Computing*, 2012, 16(5): 507–526. doi: 10.1007/s00779-011-0417.

[3] 徐风苓, 孟祥武, 王立才. 基于移动用户上下文相似度的协同过滤推荐算法[J]. 电子与信息学报, 2011, 33(11): 2785–2789. doi: 10.3724/SP.J.1146.2011.00384.
XU Fengling, MENG Xiangwu, and WANG Licai. A collaborative filtering recommendation algorithm based on context similarity for mobile users[J]. *Journal of Electronics & Information Technology*, 2011, 33(11): 2785–2789. doi: 10.3724/SP.J.1146.2011.00384.

[4] 王立才, 孟祥武, 张玉洁. 上下文感知推荐系统研究[J].软件学报, 2012, 23(1): 1–20. doi: 10.3724/SP.J.1001.2012.04100.
WANG Licai, MENG Xiangwu, and ZHANG Yujie. Context-aware recommender system: A survey of the state-of-art and possible extensions[J]. *Chinese Journal of Software*, 2012, 23(1): 1–20. doi: 10.3724/SP.J.1001.2012.04100.

[5] WU W, ZHAO J, ZHANG C, *et al*. Improving performance of tensor-based context-aware recommenders using bias tensor factorization with context feature auto-encoding[J]. *Knowledge-Based Systems*, 2017, 128(C): 71–77. doi: 10.1016/j.knosys.2017.04.011.

[6] ZOU B, LI C, TAN L, *et al*. GPUTENSOR: Efficient tensor factorization for context-aware recommendations[J]. *Information Sciences*, 2015, 299(11): 159–177. doi: 10.1016/

j.ins.2014.12.004.

[7] WU H, YUE K, LIU X, *et al*. Context-aware recommendation via graph-based contextual modeling and postfiltering[J]. *International Journal of Distributed Sensor Networks*, 2015, 11(3): 16–26. doi: 10.1155/2015/613612.

[8] ALHAMID M F, RAWASHDEH M, DONG H, *et al*. Exploring latent preferences for context-aware personalized recommendation systems[J]. *IEEE Transactions on Human-Machine Systems*, 2016, 46(4): 615–623. doi: 10.1109/THMS.2015.2509965.

[9] ALHAHYARI M and KOCHUT K. Semantic context-aware recommendation via topic models leveraging linked open data[C]. International Conference of Web Information Systems Engineering, Lugano, Switzerland, 2016: 263–277.

[10] AIMUDENA R I, GUILLERMO J D, and MERCEDES G A. A Semantically enriched context-aware OER recommendation strategy and its application to a computer Science OER repository[J]. *IEEE Transactions on Education*, 2014, 57(4): 255–260. doi: 10.1109/TE.2014.2309554.

[11] YU K, ZHANG B, ZHU H, *et al*. Towards personalized context-aware recommendation by mining context logs through topic models[C]. Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining, Kuala Lumpur, Malaysia, 2012: 431–443.

[12] ZHU H, CHEN E, XIONG H, *et al*. Mining mobile user preferences for personalized context-aware recommendation [J]. *ACM Transactions on Intelligent Systems & Technology*, 2014, 5(4): 1–27. doi: 10.1145/2532515.

[13] YIN H, CUI B, CHEN L, *et al*. Modeling location-based user rating profiles for personalized recommendation[J]. *ACM Transactions on Knowledge Discovery from Data*, 2015, 9(3): 1–41. doi: 10.1145/2663356.

[14] YIN H, CUI B, CHEN L, *et al*. Dynamic user modeling in social media systems[J]. *ACM Transactions on Information Systems*, 2015, 33(3): 1–44. doi: 10.1145/2699670.

[15] BLEI D M, NG A Y, and JORDAN M I. Latent dirichlet allocation[J]. *Journal of Machine Learning Research*, 2003, 3(3): 993–1022.

[16] ODIC A, TKALCIC M, KOSIR A, *et al*. Predicting and detecting the relevant contextual information in a movie-recommender system[J]. *Interacting with Computers*, 2013, 25(1): 74–90. doi: 10.1093/iwc/iws003.

李　平：男，1972 年生，教授，主要研究方向为数据挖掘与大数据处理等.

张路遥：女，1993 年生，硕士生，研究方向为机器学习与数据挖掘.

曹　霞：女，1992 年生，硕士生，研究方向为推荐系统.

胡检华：男，1991 年生，硕士生，研究方向为机器学习和数据挖掘.