

基于改进 URP 模型和 K 近邻的推荐研究

夏利民 赵业东 彭东亮 张 伟

(中南大学信息科学与工程学院 长沙 410075)

摘 要 针对传统推荐系统中存在的冷开始和准确性等问题,提出了一种基于改进 URP 模型和 K 近邻的推荐方法。该方法利用改进的 URP 模型对用户和项目进行建模,可以有效地解决新用户的问题;并通过推荐项目的 K 近邻对预测等级进行优化,可以显著提高对新项目预测的准确性。实验结果表明,该方法可以有效地解决冷开始问题,并显著提高推荐结果的准确性。

关键词 URP 模型, K 近邻, 产生过程, Gibbs 抽样

中图法分类号 TP391 **文献标识码** A

Recommendation Research Based on Improved URP Model and K Nearest Neighbors

XIA Li-min ZHAO Ye-dong PENG Dong-liang ZHANG Wei

(School of Information Science and Engineering, Central South University, Changsha 410075, China)

Abstract The methods used to recommend products suffer from the problems such as cold starting and accurate. To address these problems, a new recommendation method based on improved URP model and K nearest neighbors was proposed. Users and items are modeled by improved URP model, and this model can solve the new user problem effectively. The rates predicted are optimized by K nearest neighbors to solve the new item problem. The experimental results show that the new method has good quality for recommendation.

Keywords URP model, K nearest neighbors, Proceed progress, Gibbs sampling

1 引言

当今信息量的日益增长促使了自动推荐系统的出现,这些推荐系统根据用户短期或长期的需求为用户提供相关信息,并且在日常生活中得到了广泛的应用,国内外专家学者在此方面做了很多研究工作。

协同过滤技术是研究最多且应用最成功的推荐技术,该技术可以分为基于内存的协同过滤和基于模型的协同过滤。在基于内存的协同过滤技术方面,国内外专家学者的主要研究集中在基于用户的协同过滤和基于项目的协同过滤上。例如,Heung-Nam Kim 等提出了基于协同标签的协同过滤方法来增强推荐的质量^[1];Jong-Seok Lee 等提出了两种方法共同预测的协同过滤技术来提高推荐的准确性并且该方法对稀疏的数据具有鲁棒性^[2]。在基于模型的协同过滤方面,国内外专家学者的主要研究集中于贝叶斯、奇异值分解(SVD)、潜在语义分析(LDA)、概率潜在语义分析(PLSA)等模型之上。例如,Kein Wang 和 Ying Tan 提出了一种新的基于朴素贝叶斯方法的协同过滤算法^[3];Hofman 提出了利用高斯 PLSA 模型作为协同过滤的方法^[4];Benjamin Marlin 则提出了基于 URP 模型的推荐方法^[5]。

本文针对推荐系统中的新项目、新用户和准确性问题提出了一种基于改进 URP 模型和 K 近邻的推荐算法。该方法

利用改进的 URP 模型对用户和项目进行建模,可以有效地解决新用户的问题;并通过推荐项目的 K 近邻对预测等级进行优化,可以显著提高对新项目预测的准确性。实验结果表明,该方法可以有效地解决冷开始问题,并显著提高推荐结果的准确性。

2 URP 模型

URP 模型是由 Benjamin Marlin 于 2004 年提出的^[5],如图 1 所示。该模型中 η 为条件参数 α 下的狄利克雷随机变量,且由 η 决定用户在隐含层 z (表示用户组)的概率分布; β 为用户组对项目评分的概率分布, α_β 使 β 满足狄利克雷分布;灰色椭圆表示可观察值;矩形框表示重复试验,其右下角的值表示重复试验的次数(N_U 为用户数量, N_D 为项目数量, K_U 为用户组数量)。

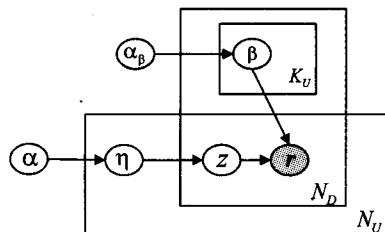


图 1 URP 模型的图解描述

到稿日期:2012-08-15 返修日期:2012-11-26 本文受国家自然科学基金(50808025),国家教育部博士点基金(20090162110057)资助。

夏利民(1963—),男,教授,博士生导师,主要研究方向为图像处理与模式识别;赵业东(1987—),男,硕士生,主要研究方向为图像处理, E-mail: zhaoyedong520@163.com;彭东亮(1989—),男,硕士生,主要研究方向为模式识别。

URP 模型的产生过程如下:

(1) 根据狄利克雷分布 $dir(a)$ 对每一位用户 u 产生一个多项式参数向量 $\eta(u)$ 。其中参数向量 $\eta(u)$ 包含了用户对不同态度 z 的概率,即用户属于不同用户组 z 的概率。

(2) 对于用户 u ,通过多项式分布 $multinomial(\eta(u))$ 得到其属于不同用户组 z 的概率。

(3) 对于每一用户组和项目对 (z, d) ,可以通过狄利克雷分布 $dir(\alpha_\beta(z, d))$ 得到一个伯努利参数向量 $\beta(z, d)$ 。参数 $\beta(z, d)$ 定义了用户组与项目相关的概率。

(4) 对于每一对 (z, d) ,通过二项式分布 $binomial(\beta(z, d))$ 得到一个二进制相关值 r 。

URP 模型可以有效地解决新用户的问题^[6],但其在建模过程中并未考虑隐含层中各用户组之间的相关关系^[7],而是假设各用户组之间相互独立,这与现实不符;且对新项目预测准确性差^[6],因此本文提出基于改进 URP 模型和 K 近邻的推荐算法。

3 基于改进 URP 模型和 K 近邻的推荐

本文推荐算法通过改进的 URP 模型来描述各用户组之间的相关关系,通过推荐项目的 K 近邻来解决新项目预测准确性问题。下面首先介绍改进的 URP 模型,然后说明对预测等级的 K 近邻处理,最后给出本文推荐算法的算法流程。

3.1 改进的 URP 模型

3.1.1 改进的 URP 模型及其产生过程

改进的 URP 模型如图 2 所示。相对于 URP 模型,其改进之处在于多项式参数向量 η 由两个参数 μ 和 Σ 共同决定。 η 服从以 μ 和 Σ 为参数的对数正态分布^[7]。对数正态分布是一个定义在单形上的分布,其通过转化一个由多变量高斯分布描述的随机变量获得,且 η 的产生过程如下^[8]:

(1) 根据高斯正态分布产生一个参数向量 θ ,其中 $\theta \sim N(\mu, \Sigma)$ 。

(2) 将 θ 转化为 $\eta: \eta_i = \frac{\exp(\theta_i)}{\sum_{i=1}^T \exp(\theta_i)}$ 。

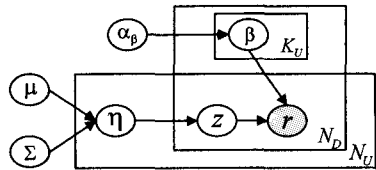


图 2 改进 URP 模型

改进 URP 模型的产生过程:

(1) 根据以 μ 和 Σ 为参数的对数正态分布对每一位用户 u 产生一个多项式参数向量 $\eta(u)$,产生过程如上所述。其中参数向量 $\eta(u)$ 包含了用户对不同态度 z 的概率,即用户属于不同用户组 z 的概率。

(2) 对于每一用户 u ,通过多项式分布 $multinomial(\eta(u))$ 得到其属于不同用户组 z 的概率。

(3) 对于每一用户组和项目对 (z, d) ,通过狄利克雷分布 $dir(\alpha_\beta(z, d))$ 得到一个多项式参数向量 $\beta(z, d)$ 。参数 $\beta(z, d)$ 定义了用户组对项目评分等级的概率。

(4) 对于每一对 (z, d) ,评分值 r 根据多项式分布 $multinomial(\beta(z, d))$ 得到。

3.1.2 改进 URP 模型的预测过程

本文对于后验概率分布 $P(r|u, d, D)$ 的估计采用 Gibbs 抽样法,其中 D 为观测数据组 (u, d, r) 的集合。记观测集合 D 中的第 n 组观测数据为 (u_n, d_n, r_n) , ψ 代表所有用户的多项式参数向量 η 以及所有用户组和多项目的多项式参数向量 $\beta(z, d)$ 。

则改进 URP 模型的似然函数为:

$$P(D|\psi) = \prod_n P(r_n|u_n, d_n, \psi) = \prod_n \sum_z P(z|u_n, \eta) P(r_n|z, d_n, \beta) \quad (1)$$

其中, $\begin{cases} P(z|u) \sim \text{multinomial}(\eta(u)) \\ P(r|z, d) \sim \text{multinomial}(\beta(z, d)) \end{cases}$

则改进 URP 模型的后验概率为:

$$P(\psi|D, \text{priors}) = P(\eta, \beta(z, d)|D, \mu, \Sigma, \alpha_\beta) \propto P(\eta|\mu, \Sigma) P(\beta(z, d)|\alpha_\beta) P(D|\psi) \quad (2)$$

其中, priors 为 μ, Σ 和 α_β , 且 $\begin{cases} P(\eta(u)) \sim \text{lognormal}(\mu, \Sigma) \\ P(\beta(z, d)) \sim \text{dirichlet}(\alpha_R) \end{cases}$ 。

采用 Gibbs 抽样的方法^[8,9]依次对用户组、多项式参数向量 η 和 $\beta(z, d)$ 抽样 M 次。对于第 m 次的抽样,得到的参数样本记为 $\psi^{(m)}$,且满足 $\psi^{(m)} \sim P(\psi|D, \text{priors})$ 。参数的每一次抽样产生一个概率矩阵 $P(r, u, d|\psi^{(m)})$,而 $P(r|u, d, \psi^{(m)})$ 可以依据此概率矩阵通过贝叶斯规则来计算。最终,本文采用 M 次 Gibbs 迭代预测的均值作为预测结果,即:

$$P(r|u, d, D) = E_{P(\psi|D, \text{priors})} [P(r|u, d, \psi)] \approx \frac{1}{M} \sum_{m=1}^M P(r|u, d, \psi^{(m)}) \quad (3)$$

由式(3)可得,对用户 u 推荐项目 d 的预测等级为:

$$R(u, d) = \arg \max_r P(r|u, d, D) \quad (4)$$

3.2 预测等级的 K 近邻处理

由于改进 URP 模型对新项目预测具有较大偏差,因此本文采用 K 近邻的方法对预测结果进行处理。对于一个项目的近邻,本文从项目属性和评分数据两个方面进行考虑,即两项目的相似性由其评分相似性和属性相似性两部分组成。

对于评分相似性,本文采用协同过滤技术中常用的相关性相似性进行度量^[10],即:

$$\text{sim}_r(i, j) = \frac{\sum_{u \in U(i, j)} (R_{u, i} - \bar{R}_i)(R_{u, j} - \bar{R}_j)}{\sqrt{\sum_{u \in U(i, j)} (R_{u, i} - \bar{R}_i)^2} \sqrt{\sum_{u \in U(i, j)} (R_{u, j} - \bar{R}_j)^2}} \quad (5)$$

式中, $U(i, j)$ 为对项目 i 和 j 均评分的用户集合, $R_{u, i}$ 为用户 u 对项目 i 的评分, $\bar{R}_i$ 为用户对项目 i 评分的均值, $R_{u, j}$ 为用户 u 对项目 j 的评分, $\bar{R}_j$ 为用户对项目 j 评分的均值。

对于属性相似性,本文通过比较两项目共有属性所占两项目所有属性的比例来获得,即:

$$\text{sim}_a(i, j) = \frac{\#(A_i \cap A_j)}{\#(A_i \cup A_j)} \quad (6)$$

式中, $\#(A_i \cap A_j)$ 为项目 i 和 j 共有的属性数量, $\#(A_i \cup A_j)$ 为项目 i 和 j 的全部属性数量。

则两项目的相似性计算式为:

$$\text{sim}(i, j) = (1 - \lambda) \text{sim}_r(i, j) + \lambda \text{sim}_a(i, j) \quad (7)$$

式中, λ 为比例系数。

根据项目之间的相似性大小,即可寻找与目标项目最相似的 K 个项目,并计算出最终的项目预测等级 $R^*(u, d)$:

$$R^*(u, d) = \frac{R(u, d) + \sum_{d_k \in d_{K,d}} \text{sim}(d, d_k) R(u, d_k)}{1 + \sum_{d_k \in d_{K,d}} \text{sim}(d, d_k)} \quad (8)$$

式中, $R(u, d)$ 为根据改进 URP 模型对用户 u 推荐项目 d 的预测等级; $d_{K,d}$ 为项目 d 的 K 近邻集合, d_k 为其中的元素。

3.3 推荐算法流程

根据推荐系统的工作原理, 结合 3.1 节和 3.2 节所述方法, 给出本文推荐算法的流程图, 如图 3 所示, 具体步骤说明如下:

- 1) 对于一个推荐项目, 根据 3.2 节所述方法, 利用其评分数据和自身属性找出其 K 个近邻项目。
- 2) 根据 3.1 节所述的改进 URP 模型, 利用 Gibbs 抽样方法预测出目标用户对于推荐项目及其 K 个近邻项目的预测评分等级。
- 3) 根据 3.2 节中式(8)对第 2)步得到的评分数据进行处理, 且将得到的结果作为用户对项目的最终预测评分等级。
- 4) 根据前 3 个步骤可以得出目标用户对所有项目的评分数据; 并将这些预测评分按照从大到小的顺序进行排序, 最后将前 N 个项目作为最终的推荐结果推荐给用户。

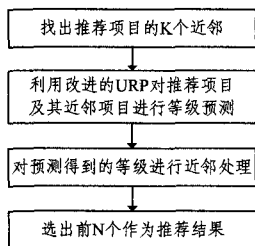


图3 本文推荐算法流程

4 实验

4.1 实验数据集

为了评测本文算法的推荐准确性, 使用 Grouplens 提供的 MovieLens 数据集进行实验。该数据集包括随机抽取的 100000 个评分数据, 评分范围为 1~5, 包括 943 个用户和 1682 部电影, 用户-商品的评分矩阵密度为 6.3%。将实验数据按 4:1 的比例分为训练集和测试集。

4.2 度量准则

对于推荐质量的度量, 实验采用统计精度度量方法中的平均绝对误差 MAE(mean absolute error)。平均绝对误差通过计算预测的用户评分与实际用户评分之间的偏差来度量预测的准确性, MAE 值越小, 推荐精度越高。该实验将推荐度作为预测等级进行 MAE 值的统计, MAE 计算式如式(9)所示:

$$MAE = (\sum_{i=1}^N |p_i - q_i|) / N \quad (9)$$

式中, p_i 为预测的用户的第 i 项评分, q_i 为实际用户的第 i 项评分, 其中 N 为测试物品的数量。

4.3 实验结果

首先确定本文方法的比例系数 λ 和项目的近邻数目 K 。 λ 以 0.05 为步差, K 以 5 为步差, 测得不同 λ 和 K 情况下的 MAE 值, 如表 1 所列。实验过程为: 1) 从测试集中抽取一个数据组并用本文方法进行等级预测; 2) 重复步骤 1) 共 M 次; 3) 根据式(9)统计 MAE。

表1 本文方法 MAE 随 λ 和 K 的变化情况

$\lambda \backslash K$	5	10	15	20	25	30	35	40
0.05	1.55	1.49	1.38	1.22	1.20	1.18	1.17	1.17
0.10	1.52	1.47	1.36	1.15	1.10	1.08	1.07	1.09
0.15	1.44	1.37	1.28	1.09	0.93	0.82	0.80	0.80
0.20	1.31	1.22	1.11	0.99	0.83	0.72	0.72	0.73
0.25	1.36	1.25	1.17	1.06	0.92	0.84	0.82	0.83
0.30	1.41	1.35	1.23	1.07	0.91	0.84	0.86	0.84

通过表 1 选出最优的 λ 和 K 值, 即使 MAE 最小的 λ 和 K 值。本文 λ 和 K 的值分别选择 0.20 和 30。当 λ 和 K 值选定以后, 将本文方法与 URP 模型^[5] 及 GPLSA^[4] 两种方法进行对比。

(1) 对于新用户问题, 本文比较 3 种方法的 MAE 随用户对项目评分数量的变化, 如图 4 所示。实验过程为: 1) 从数据集中随机抽取一个用户 u ; 对于用户 u , 从其评分项目集中随机抽取 N 项作为评分项目集合, 其余项目作为未评分项目集合; 2) 对用户 u 预测其未评分项目的等级; 3) 重复步骤 1) 和步骤 2) 共 M 次; 4) 根据前 3 步骤统计不同 N 情况下的 MAE。

(2) 对于新项目问题, 本文比较 3 种方法的 MAE 随对项目评分的用户数量的变化, 如图 5 所示。实验过程为: 1) 从数据集中随机抽取一个项目 i ; 对于项目 i , 从对其评分的用户集合中随机抽取 N 位作为评分用户集合, 其余用户作为未评分用户集合; 2) 预测将项目 i 推荐给其未评分用户的等级; 3) 重复步骤 1) 和步骤 2) 共 M 次; 4) 根据前 3 步骤统计不同 N 情况下的 MAE。

(3) 对于准确性问题, 本文统计 3 种方法的平均 MAE 值, 如表 2 所列。实验过程与确定 λ 和 K 的过程相同。

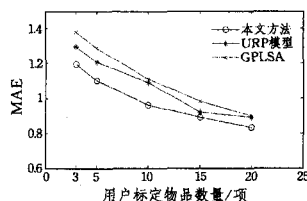


图4 MAE 随用户对项目评分数量的变化

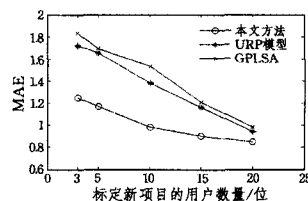


图5 MAE 随对项目评分的用户数量的变化

表2 3 种方法的平均 MAE

方法	改进 URP	URP	pls
平均 MAE	0.72	0.76	0.78

综上实验对比可知, 本文方法能够有效地解决新用户和新项目问题, 并且显著提高推荐结果的准确性。因此本文方法具有良好的物品推荐品质。

结束语 针对推荐技术当中存在的冷开始和准确性等问题, 提出一种基于改进 URP 模型和 K 近邻的推荐方法。本方法通过改进的 URP 模型作为预测等级的基础, 并通过 K 近邻法对预测等级进行处理, 可以有效地解决新用户、新项目 and 准确性等问题。实验结果表明, 该方法具有良好的物品推荐品质。

参考文献

- [1] 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 and Applications, 2010, 9(1):73-83

(下转第 299 页)

$\Delta t=0.2$ 的情况下 SNR 值随迭代次数 N (横坐标) 变化的曲线图, 迭代次数 $N>10$ 之后 SNR 值趋于稳定; 图 5(b) 则为本文模型分别在 $(a,b) \in [(0.4, 0.8), (4, 0.8), (4, 4)]$ 3 种情况下 SNR 值随迭代次数 N (横坐标) 变化的曲线图 ($a=4, b=0.8$ 为最佳参数), 从图中可以看出 $N>6$ 就可以使 SNR 值趋于稳定的一个值, 迭代次数更少, 可以一定程度上减少算法执行时间。

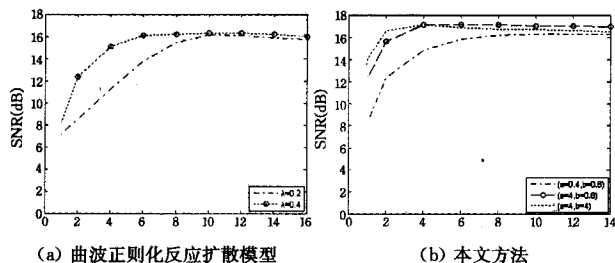


图 5 不同算法在不同调和参数下的时间复杂度的度量

目标函数中 a 和 b 分别为两个反应项的调整参数, 用于调整扩散项和两个正则项之间的平衡。从图 5(b) 可以看出, 当 b 值不变时, 增加 a 值会一定程度上减少迭代次数, 使其尽快地进入平稳状态, 但随着 a, b 值的过度增加反而导致噪声去除不彻底使 SNR 值下降, 因此调整参数应取偏小的值。

结束语 复杂纹理图像去噪中纹理和边缘信息的保持是一个难点。为了更好地平衡去噪与纹理、边缘等信息的保持, 本文分析纹理、边缘与噪声的自身特点, 将多尺度几何分析方法和各向异性扩散模型相结合, 构建了一种采用双正则项各向异性扩散的反应扩散方程, 并完成了目标函数的离散化及其数值解的收敛性证明。目标函数定义为以波原子、曲波变换后邻域内梯度值为参数的扩散控制函数, 使扩散在图像信息丰富的纹理和边缘区域减弱, 并通过反应项对扩散进行调节, 以得到更好的平滑效果。实验结果证明, 本文方法较以往的反应扩散模型不仅可以提高图像的信噪比, 而且可以更好地保留图像边缘和纹理等细节信息。

参考文献

[1] Chan T F, Osher S, Shen J. The digital TV filter and nonlinear

denoising [J]. IEEE Transactions on Image Processing, 2001, 10 (2): 231-241

[2] Vese L, Osher S. Modeling textures with total variation minimization and oscillating patterns in image processing [J]. J. Sci. Comput, 2003, 19(1-3): 553-572

[3] Weickert J. Anisotropic diffusion in image processing [M]. Stuttgart: B. G. Teubner-Verlag, 1998

[4] 王志明, 张丽. 自适应的快速非局部图像去噪算法 [J]. 中国图象图形学报, 2009, 14(4): 1-7

[5] Buades A, Morel J M. A non-local algorithm for image denoising [C]//Proceeding of IEEE Computer Society Conference on Pattern Recognition. San Diego, CA, USA, 2005: 60-65

[6] 刘芳, 刘文学, 焦李成. 基于复小波邻域隐马尔科夫模型的图像去噪 [J]. 电子学报, 2005, 33(7): 1284-1287

[7] Donoho D L. Adaptive wavelet threshold for image denoising [J]. Electronics Letters, 2005, 41(10): 586-587

[8] 刘镇强, 李涛, 杜慧敏, 等. 基于 context 模型的 contourlet 域图像去噪 [J]. 计算机科学, 2012, 39(3): 243-245

[9] Plonka G. A digital diffusion-reaction type filter for nonlinear denoising [J]. Results Math, 2009, 53(3/4): 371-381

[10] de O Casaca W C, Boaventura M. A Regularized Nonlinear Diffusion Approach for Texture Image Denoising [C]//XXII Brazilian Symposium on Computer Graphics and Image Processing. 2009: 1-5

[11] Plonka G, Ma J. Nonlinear regularized reaction-diffusion filters for denoising of images with textures [J]. IEEE Transactions on Image Processing, 2008, 17(2): 1283-1294

[12] Ma Jian-wei, Plonka G. Combined Curvelet Shrinkage and Nonlinear Anisotropic Diffusion [J]. IEEE Transactions on Image Processing, 2007, 16(9): 2198-2206

[13] 隆刚, 肖磊, 陈学俭. Curvelet 变换在图像处理中的应用综述 [J]. 计算机研究与发展, 2005, 42(8): 1331-1337

[14] Demanet L, Ying L. Wave atoms and sparsity of oscillatory patterns [J]. Appl. Comput. Harmon. Anal, 2007, 23(3): 368-387

[15] 宋宜美, 宋国乡. 结合波原子和 Cycle Spinning 的图像去噪 [J]. 西安电子科技大学学报: 自然科学版, 2010, 37(2): 1-2

(上接第 278 页)

[2] Lee J-S, Olafsson S. Two-way cooperative prediction for collaborative filtering recommendations [J]. Expert Systems with Applications, 2009, 36(3): 553-561

[3] Wang Ke-bin, Tan Ying. A New Collaborative Filtering Recommendation Approach Based on Naive Bayesian Method [C]//Lecture Notes in Computer Science, 2011, 6729: 218-227

[4] Hofman T. Collaborative filtering via Gaussian probabilistic latent semantic analysis [C]//Proceedings of international conference on SIGIR'03. Toronto, Canada, July 2003

[5] Marlin B. Modeling user rating profiles for collaborative filtering [C]//Advances in Neural Information Processing Systems. 2004, 16: 627-634

[6] Savia E, Puolamäki K, Kaski S. Latent grouping models for user

preference prediction [J]. Mach Learn, 2009, 74(1): 75-109

[7] Bei D M, Laferty J D. Correlated topic models [M]. Advances in Neural Information Processing Systems 18, Cambridge, MA: MIT Press, 2006

[8] Mimno D, Wallach H M, McCallum A. Gibbs Sampling for Logistic Normal Topic Models with Graph-Based Priors [C]//NIPS Workshop on Analyzing Graphs. Whistler, BC, 2008

[9] Griffiths T L, Steyvers M. Finding scientific topics [J]. PNAS, 2004, 101(1): 5228-5235

[10] Zhang Fu-guo. Research on Recommendation List Diversity of Recommender Systems [C]//International Conference on Management of e-Commerce and e-Government. Nanchang, China, 2008: 72-76