文章编号:1001 - 9081(2010)04 - 1079 - 04

协同过滤系统的矩阵稀疏性问题的研究

曾小波[1],魏祖宽[1],金在弘[2]
(1. 电子科技大学 计算机科学与工程学院,成都 610054; 2. 永同大学校 计算机工学科,韩国 永同郡 370701)
(zxb19840101@gmail.com)

摘 要:应用奇异值算法得到一个无缺失的矩阵,引进了一种增强的、基于参数的 Pearson 相关系统算法来提高相关性算法的准确性。提出一个基于奇异值分解和增强 Pearson 系数的"HybridSVD"算法,用 MovieLens 数据集来评价该算法,并和其他经典的传统算法做了比较。实验结果证明,"HybridSVD"算法比其他传统算法能更好地处理协同过滤中的稀疏性问题。

关键词:协同过滤;矩阵稀疏性;奇异值分解;增强的 Pearson 相关系数
中图分类号: TP391 **文献标志码**:A

Research of matrix sparsity for collaborative filtering

ZENG Xiao-bo[1], WEI Zu-kuan[1], KIM Jae-hong[2]
(1. *School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu Sichuan* 610054, *China*;
2. *Department of Computer and Information Engineering, Youngdong University, Yongdong-Gun Chungbuk* 370701, *Korea*)

Abstract: This paper applied singular value decomposition to predict the missing data. An enhanced Pearson correlation coefficient algorithm based on parameter was introduced to increase the accuracy when computing the similarity of user and items. Finally, a new algorithm called "HybridSVD" was explored, which was based on singular value decomposition and our novel similarity model. In the experiment section, the authors evaluated this new algorithm using the dataset MoiveLens and the results suggest that the new algorithm can better handle this matrix sparsity problem.

Key words: collaboration filtering; matrix sparesity; ingular Value Decomposition (SVD); Enhanced Pearson Correlation Coefficient (EPCC)

0 引言

推荐系统已被广泛应用在通过现场互动来供电影、产品提供个性化建议。其中在推荐系统中应用最成功的是协同过滤方法。例如亚马逊的推荐系统给顾客提供了关于书的建议,而这些建议是其他顾客通过推荐系统来告诉亚马逊的。总的来说协同过滤包括了基于内存和基于模型的两种不同方法。

1)基于内存的协同过滤。该算法先用统计的方法得到具有相似兴趣的邻居用户,再用基于邻居的方法来计算,所以该方法也称为基于邻居的协同过滤,其中最常使用的方法是 K 近邻(K-Nearest Neighbor,KNN)方法。

基于内存的协同过滤包括数据规范化、邻居选择和决定插值的权重三个主要步骤。目前基于用户和基于项目的近邻选择方法已被广泛研究。基于用户的协同过滤首先计算活动用户和其他用户之间的相似度,并从相似用户来得到活动用户的评分矩阵。基于项目的协同过滤基于这样的假设:能够引起用户兴趣的项,必定与之前评分高的项相似。基于项目的协同过滤和基于用户的协同过滤的唯一区别是基于项目的协同过滤是找出每个项目的相似项目,而基于用户的协同过滤则是找出活动用户的相似用户。基于内存的协同过滤相比基于模型的协同过滤拥有更少的参数的优点,但是稀疏性问题仍然存在。如图 1、图 2 所示(? 表示的是基于用户的协同过滤计算的预测评分)。

	i_1	i_2	i_3	i_4	i_5	i_6	i_7	i_8	i_9
u_1		3	4		1		2	5	4
u_2	4		3	2		2	1		
u_3			?				3		2
u_4	4			2		3			
u_5									
u_6		5	4		2		3		
u_7	4		3	2		3	4	5	

图 1 基于用户的协同过滤

	i_1	i_2	i_3	i_4	i_5	i_6	i_7	i_8	i_9
u_1		3	4		1		2	5	4
u_2	4		3	2		2	1		
u_3			?				3		2
u_4	4			2		?	3		
u_5									
u_6		5	4		2		3		
u_7	4		3	2		3	4	5	

图 2 基于项目的协同过滤

2)基于模型的协同过滤。主要是通过统计或机器学习的方法,用用户的历史偏好记录来构造用户偏好模型,进而来产生推荐。目前比较流行的方法包括聚类模型、SVD 等。

收稿日期:2009 - 08 - 26;**修回日期**:2009 - 11 - 13。

作者简介:曾小波(1983 -),男,四川渠县人,硕士研究生,主要研究方向:空间数据库技术; 魏祖宽(1968 -),男,四川成都人,副教授,博士,主要研究方向:空间数据库、3G 应用技术、数据库技术; 金在弘(1960 -),男,韩国永同郡人,副教授,博士,主要研究方向:地理信息系统、数据库技术。

本展示了一个基于奇异值分解和增强 Pearson 相关系数的特征递增算法(HybridSVD)。在这种方法中,一种方法的输出作为另一种方法的输入。HybridSVD 算法首先使用奇异值分解方法得到预测的评分矩阵,然后使用基于近邻的方法来得到活动用户的邻居,最后使用增强的 Pearson 相关系数方法来得到活动用户的最终预测值。

1 注记和相关工作

1.1 注记

首先给出本文中所有使用的注记。用一个 m 用户和 n 项目的评分矩阵来标示所有用户对所用项目的评分。

$\boldsymbol{R} = \{r_{u,i}\}(1 \leqslant u \leqslant m, 1 \leqslant i \leqslant n)$,如图3所示,其中 $R_{u,i}$ 表示用户 u 对项目 i 的评分。如果用户 u 没有对项目 i 评分,那么其值为0。如:

$$R_{u,i} = \begin{cases} R_{u,i}, & rating \\ 0, & not - rating \end{cases}$$

	i_1	i_2	i_3	i_4	i_5	i_6	i_7	i_8	i_9
u_1		3	4		1		2	5	4
u_2	4		3	2		2	1		
u_3							3		2
u_4	4			2			3		
u_5									
u_6		5	4		2		3		
u_7	4		3	2		3	4	5	

图3 用户 — 项目评分矩阵

1.2 稀疏性问题

在实际中,用户和项目的数量都非常大。在这种情况下,评分矩阵就会极度的稀疏,这个问题就是通常说的稀疏性问题,对协同过滤的算法有着消极的影响。由于这个问题的存在,两个用户之间的相似度非常有可能为0。这种情况称邻居传递损失。例如,如果用户 u 和用户 v 具有很高的相关性,并且用户 v 和用户 w 也具有很高的相关性,那么用户 u 和用户 w 不一定具有很高的相关性,因为他们可能具有很少的共同的评分,甚至可能由于具有很少的评分而导致具有负的相关性。

1.3 奇异值分解

奇异值分解[4,8]是线性代数中最重要的矩阵分解,它已经被广泛引用到信息检索和信号处理领域。

定义1 给定一个 $m \times n$ 的矩阵 $\boldsymbol{R}$,其中 $\text{rank}(\boldsymbol{R}) = r$,则 $\boldsymbol{R}$ 的奇异值分解 $SVD(\boldsymbol{R})$,定义为:

$$\boldsymbol{R} = \boldsymbol{U} \cdot \boldsymbol{S} \cdot \boldsymbol{V}^{\mathrm{T}}$$

其中:$\boldsymbol{U}$ 和 $\boldsymbol{V}$ 是两个大小分别为 $m \times r$ 和 $n \times r$ 的正交矩阵;并且 $\boldsymbol{S} = \text{diag}(\sigma_1, \cdots, \sigma_r)$,$\sigma_1 \geqslant \sigma_2 \geqslant \cdots \geqslant \sigma_r \geqslant 0$;$r$ 是矩阵 $\boldsymbol{R}$ 的秩;σ_r 是 $\boldsymbol{R}$ 的奇异值。

定义2 设 $\boldsymbol{R}$ 的奇异值分解为 $\boldsymbol{R}_k$。$k = \text{rank}(\boldsymbol{R}) \leqslant r = \min(m,n)$ 并且定义:

$$\boldsymbol{R}_k = \sum_{i=1}^{k} \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^{\mathrm{T}}$$

$$\min_{\text{rank}(\boldsymbol{R})=k} \| \boldsymbol{R} - \boldsymbol{B} \|_F^2 = \| \boldsymbol{R} - \boldsymbol{R}_k \|_F^2 = \sigma_{k+1}^2 + \cdots + \sigma_r^2$$

以此得出,由矩阵 $\boldsymbol{R}$ 的 k 个最大奇异值组成的矩阵 $\boldsymbol{R}_k$ 是矩阵 $\boldsymbol{R}$ 的秩为 k 的相似矩阵中最近似的一个。实际上,对于 Frobenius 范式,奇异值分解提供了原始矩阵 $\boldsymbol{R}$ 的最佳低阶近似矩阵。由此,可以将矩阵 $\boldsymbol{S}$ 的维数进行简化得到仅有 k 个奇异值的新的对角矩阵,其中 $k < r$。按照同样的方法,如果同时将矩阵 $\boldsymbol{U}$ 和 $\boldsymbol{V}$ 简化,那么就可得到更加简化的公式 $\boldsymbol{R}_k = \boldsymbol{U}_k \boldsymbol{S}_k \boldsymbol{V}_k' \circ \boldsymbol{R}_k \approx \boldsymbol{R}$。Sarwar 等人首先将奇异值分解应用到协同过滤系统中,并用相应列的平均评分替换矩阵中评分为0的项目,然后用规范化技术得到一个规范矩阵 $\boldsymbol{R}_{\text{norm}}$。

2 相似度计算

在协同过滤系统中,相似度计算是通过计算项目之间或用户之间的相关性来选择最近似的项目或用户时非常关键的部分。计算相关性主要分余弦相关性、Pearson 相关相关性和修正余弦相关性三类,在本节中,简要介绍这三种相关性的概念。

2.1 余弦相关性

在这类方法中,两个项目被认为两个 m 维的用户空间向量,它们之间的相关性是用这两个向量的余弦来度量的。基于余弦的相关性可以描述为:

$$sim(i,j) = \cos(\vec{i}, \vec{j}) = \frac{\vec{i} \cdot \vec{j}}{\| \vec{i} \|_2 * \| \vec{j} \|_2}$$

2.2 Pearson 相关相关性

在这类方法中,两个项目或用户之间的相关性是通过计算两个项目或用户的 Pearson 相关系数来确定的。可以描述为:

$$sim(i,j) = \frac{\sum_{u \in U} (R_{u,i} - \overline{R_i})(R_{u,j} - R_j)}{\sqrt{\sum_{u \in U} (R_{u,i} - \overline{R_i})^2} \sqrt{\sum_{u \in U} (R_{u,j} - R_j)^2}}$$

其中:U 表示了用户集合;$R_{u,i}$ 表示了用户 u 对项目 i 的评分;$\overline{R_i}$ 表示项目 i 的平均评分。

2.3 修正的余弦相关性

用余弦相关性来计算两个项目或用户的相关性有个很大的缺点是没有考虑的评分尺度问题。修正的余弦相关性可以通过减去用户对项目的平均评分来改善上述缺陷。修正的余弦相关性可以描述为:

$$\text{sim}(i,j) = \frac{\sum_{u \in U} (R_{u,i} - \overline{R_u})(R_{u,j} - R_u)}{\sqrt{\sum_{u \in U} (R_{u,i} - \overline{R_u})^2} \sqrt{\sum_{u \in U} (R_{u,j} - R_u)^2}}$$

其中 $\overline{R_u}$ 表示用户 u 的平均评分。

3 HybridSVD 协同过滤算法

通常情况下,因为用户只对大量项目的一部分评分而造成大部分评分都是缺失的。在这种情况下,如果用户只对很少一部分评分,Pearson 相关系数会过度地估计它们之间的相关性。因此本文提出基于奇异值和增强 Pearson 相关系数的特征递增算法(HybridSVD)来克服稀疏性问题。首先,用奇异值方法来预测缺失的评分来得到一个无缺失的评分矩阵,然后用增强的 Pearson 相关相关性方法来计算相关性来形成邻居来获得缺失评分的预测值。

3.1 相关性计算

在协同过滤系统中应用最多的模型是基于邻居的模型。它最基本的两种形式是基于用户和基于项目的协同过滤。其中应用最广泛的是基于项目的协同过滤。它具有更好的扩展性和正确性,因此本算法集中在基于项目的方法,而关键步骤是相关性计算。通常采用 Pearson 相关系数,表示为 $sim(i,$

j),但仍然有过于估计相关性的缺点,于是给出一个基于参数的更可靠的相关性公式,可以描述为:

$$sim'(i,j) = \frac{|N_i \cap N_j|}{|N_i \cap N_j| + \alpha} \cdot sim(i,j)$$

其中:$|N_i \cap N_j|$ 是在项目 i 和项目 j 上同时评分的用户数量;α 是调整相关性计算的参数。在 MovieLens 数据集中这个典型值为 80。

3.2　HybridSVD 算法

首先用奇异值分解方法得到 $\boldsymbol{R}_{\text{norm}}$,然后使用相关模型来计算 $\boldsymbol{R}_{\text{norm}}$ 得到预测矩阵 $\boldsymbol{R}_{\text{pred}}$,该算法的主要结构可以描述为:

输入:评分矩阵 $\boldsymbol{R}$。

输出:预测矩阵 $\boldsymbol{R}_{\text{pred}}$。

变量:$\alpha = 80$。

1) 分解矩阵 $\boldsymbol{R}$ 并且规范化得到 $\boldsymbol{R}_{\text{norm}}$。

2) 使用奇异值分解方法分解 $\boldsymbol{R}_{\text{norm}}$ 得到矩阵 $\boldsymbol{U},\boldsymbol{S},\boldsymbol{V}$。

3) 将矩阵 $\boldsymbol{S}$ 进行维数简化到 k 得到矩阵 $\boldsymbol{S}_k$($k < r$, $\text{rank}(\boldsymbol{R}_{\text{norm}}) = r$)。

4) 同样将矩阵 $\boldsymbol{U},\boldsymbol{V}$ 进行简化得到矩阵 $\boldsymbol{U}_k$, $\boldsymbol{V}_k$。

5) 计算 $\boldsymbol{S}_k$ 平方根得到矩阵 $\boldsymbol{S}_k^{1/2}$。

6) 使用下面的公式计算矩 $\boldsymbol{U}_k\boldsymbol{S}_k^{1/2}$ 的第 u 行和矩阵 $\boldsymbol{S}_k^{1/2}\boldsymbol{V}_k^{\mathrm{T}}$ 的第 i 行的点积,得到用户 u 在项目 i 的预测评分:

$$R_{u,i} = \overline{R_u} + \boldsymbol{U}_k \sqrt{\boldsymbol{S}_k}'(u) \cdot \sqrt{\boldsymbol{S}_k}\boldsymbol{V}_k'(i)$$

7) 使用新的相关模型来计算项目 i 和其他项目之间的项目性。

$$sim'(i,j) = \frac{|N_i \cap N_j|}{|N_i \cap N_j| + \alpha} \cdot sim(i,j)$$

8) 用户 u 对项目 i 的评分 $P_{u,i}$ 可以描述为:

$$P_{u,i} = \bar{i} + \frac{\sum_{j \in S(i)} sim'(i,j) \cdot (R_{u,j} - \bar{j})}{\sum_{j \in S(i)} sim'(i,j)}$$

HybridSVD 算法利用了用户和项目之间的潜在关系来对维空间进行了简化。它用奇异值分解来增加了数据的密度,使用增强的 Pearson 相关系数来进行相关性计算。这样不仅克服了协同过滤中的稀疏性问题,而且保存了 KNN 方法的优点。

4　实验

本文使用了三个实验来评价新模型的性能。在第一个实验中,使用数据集来测试在奇异值分解的情况下维数 k 的敏感性得到最佳维数 k。在第二个实验中测试了协同过滤系统的系统怎样随参数 α 变化。最后将本文模型和传统经典的算法做了比较。

4.1　数据集

本文的实验采用来自 MoiveLens 的数据集,MoiveLens 的数据集包括了超过 35 000 用户对超过 3 000 部电影的评分。此数据集的密度为:

$$\frac{10\,100}{943 \times 168} = 6.3\%$$

从数据集中随机选择了 1 000 条记录。记录被定义为〈user,item,rating〉三元组。将所选的记录分成训练集和测试集,其中 800 条记录为训练集,剩下的 200 条记录为测试记录。

4.2　评估指标

使用平均绝对误差(Mean Absolute Error,MAE)指标来评价 HybridSVD 算法和其他算法的性能。平均绝对误差定义为:

$$MAE = \frac{\sum_{(u,i) \in TestSet} |R_{u,i} - \hat{R}_{u,i}|}{|TestSet|}$$

4.3　维数 k 的影响

在奇异值分解方法中,维数 k 的确定是很关键的。如果 k 太小,则不能得到评分矩阵的重要结构;如果 k 太大,则失去了维数简化的意义。因此在采用 HybridSVD 方法之前,必须得到维数 k 的值。

从图 4 可以看出,k 的值对推荐结果的性能有重要影响,当 k 的值为 15,该数据集上的 MAE 最小,所以对于这个数据集,维数 $k = 15$ 是最佳的。

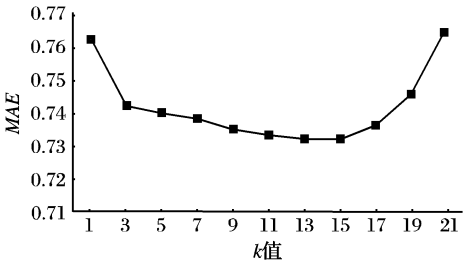

图 4　维数为 k 的预测性能

4.4　参数 α 的影响

在本文模型中,使用调整参数来使相关性计算更加合理,设 k 为 15,如图 5 所示。

从图5可以看出,当 $\alpha = 80$,$k = 15$,协同过滤系统具有最好的性能。

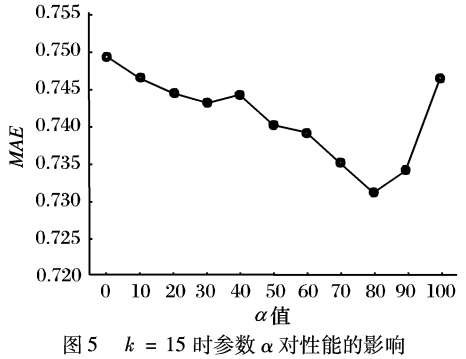

图 5　$k = 15$ 时参数 α 对性能的影响

4.5　性能比较

为了展示本文方法作用在推荐系统时性能的提高,将本方法和传统的协同过滤算法:pure SVD 、item-based 协同过滤方法[5] 和 user-based 协同过滤方法[6] 对了对比,结果如图 6 所示。

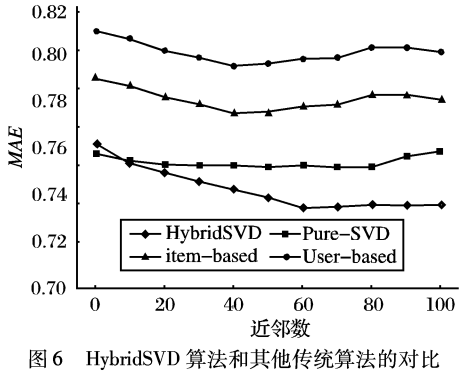

图 6　HybridSVD 算法和其他传统算法的对比

从图 6 可以看出,当邻居数量在[14,18]区间的时候,HybridSVD 的 MAE 比其他的更小,因此 HybridSVD 算法比其

他传统算法拥有更好的性能。

5 结语

本文提出的新的协同过滤算法性能好,预测准确性高,是对传统的经典协同过滤算法是很好的补充,能够满足专业应用领域的要求。

参考文献:

[1] ADOMAVICIUS G, TUZHILIN A. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions[J]. IEEE Transactions on Knowledge and Data Engineering, 2005, 17(6): 734 - 749.

[2] BENNET J, LANNING S. The netflixprize [EB/OL]. [2009 - 06 - 20]. http://www.netflixprize.com/assets/NetflixPrizeKDD_to_appear.pdf.

[3] BELL R, KOREN Y. Improved neighborhood-based collaborative filtering[EB/OL]. [2009 - 06 - 20]. http://public.research.att.com/~volinsky/netflix/cfworkshop.pdf.

[5] SARWA B, KARYPIS G, KONSTAN J, *et al.* Item-based collaborative filtering recommendation algorithms[C]// Proceedings of the 10th International Conference on World Wide Web. New York: ACM, 2001: 285 - 295.

[6] WANG JUN, de VRIES A P, REIDERS M J T. Unifying user-based and item-based collaborative filtering approaches by similarity fusion [C]// Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM, 2006: 501 - 508.

[7] MA HAO, KING I, LYU M R. Effective missing data prediction for collaborative filtering[C]// Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM, 2007: 39 - 46.

[8] SARWAR B M, KARYPIS G, KONSTAN J A, *et al.* Application of dimensionality reduction in recommender systems: A case study. [EB/OL]. [2009 - 06 - 20]. http://robotics.stanford.edu/~ronnyk/WEBKDD2000/papers/sarwar.pdf.

[9] LINDEN G, SMITH B, YORK J. Amazon.com recommendations: item-to-item collaborative filtering[J]. IEEE Internet Computing, 2003, 7(1): 76 - 80.

[10] HOFMANN T, PUZICHA J. Latent class models for collaborative filtering[C]// Proceedings of the 16th International Joint Conference on Artificial Intelligence. San Francisco: Morgan Kaufmann, 1999: 688 - 693.

[11] KOREN Y. Factorization meets the neighborhood: A multi-faceted collaborative filtering model[C]// Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM, 2008: 426 - 434.

[12] RESNICK P, IACOVOU N, SUCHAK M, *et al.* Grouplens: An open architecture for collaborative filtering of net news[C]// Proceedings of the 1994 International Conference on Computer Supported Cooperative Work. New York: ACM, 1994: 175 - 186.

[13] PARK S T, PENNOCK D M. Applying collaborative filtering techniques to movie search for better ranking and browsing[C]// Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM, 2007: 550 - 559.

(上接第1078页)

3)当用户相异度在0.025 3以上时,MAE反弹。原因是当用户相异度逐渐增大时,其聚类中的用户数目也会随之增多,当达到最优后,各类中用户数目的增多反而降低了项目相异度的真实性。

4 结语

本文提出了一种结合用户情境和基于项目的协同推荐方法。本方法从用户自身作为出发点,按照对用户的情境因素进行聚类后,在同类中的用户进行资源得分预测并推荐,取得了优于传统协作推荐算法的性能,为现有的推荐技术提供了一种新的思路。

同时,由于人的情景因素是在不断发生变化的,而本文研究的是相对静止的情景因素,所以在以后的研究中应该着重考虑构建一个用户情景因素的模型,可以随着各种外部条件的改变而动态的修改用户自身的情景因素值,以达到对用户正确分类的目的。

参考文献:

[1] 李聪,梁昌勇,马丽.基于领域最近邻的协同过滤推荐算法[J].计算机研究与发展,2008,45(9):1532 - 1538.

[2] 张光卫,康建初,李鹤松,等.面向场景的协同过滤推荐算法[J].系统仿真学报,2006,18(z2):595 - 601.

[3] 王磊.协同推荐技术及其在科技文献个性化推荐系统中的应用研究[D].南京:南京理工大学,2007.

[4] 邓爱林,左子叶,朱扬勇.基于项目聚类的协同过滤推荐算法[J].小型微型计算机系统,2004,25(9):1665 - 1670.

[5] 赵亮,胡乃静.个性化推荐算法设计[J].计算机研究与发展,2002,39(8):986 - 991.

[6] SARWAR B, KARYPIS G, KONSTAN J, *et al.* Incremental singular value decomposition algorithms for highly scalable recommender systems [EB/OL]. [2009 - 07 - 01]. http://www.grouplens.org/papers/pdf/sarwar_SVD.pdf.

[7] 邓爱林,朱扬勇,施伯乐.基于项目评分预测的协同过滤推荐算法[J].软件学报,2003,14(9):1621 - 1628.

[8] SARWAR B, KARYPIS G, KONSTAN J, *et al.* Item-based collaborative filtering recommendation algorithms [C]// Proceedings of the 10th International Conference on World Wide Web. New York: ACM, 2001: 285 - 295.

[9] ADOMAVICIUS G, TUZHILIN A. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions [J]. IEEE Transactions on Knowledge and Data Engineering, 2005, 17(6): 734 - 749.

[10] LEMIRE D, MACLACHLAN A. Slope one predictors for online rating-based collaborative filtering [C]// Proceedings of the SIAM Data Mining Conference. Newport Beach: Society for Industrial Mathematics, 2005: 21 - 25.

[11] 姚忠,吴跃,常娜.集成项目类别与语境信息的协同过滤推荐算法[J].计算机集成制造系统,2008,14(7):1449 - 1456.

[12] 赵明清,蒋昌俊,陶树平.基于等价相异度矩阵的聚类[J].计算机科学,2004,31(7):183 - 184.