SPCF:一种基于内存的传播式协同过滤推荐算法

赵琴琴 鲁 凯 王 斌
(中国科学院计算技术研究所 北京 100190)

摘 要 基于内存的协同过滤是当前互联网推荐引擎中的核心技术.然而,目前该技术的发展面临着严重的用户评分稀疏性问题.该文通过采用传播的思想对数据稀疏性问题进行了有益的探索和研究,并提出了一种改进的基于内存的协同过滤推荐算法SPCF.该算法通过相似度传播,寻找到更多、更可靠的邻居,然后在此基础上,从用户和项目两方面信息考虑对用户进行推荐.在Movie Lens和Yahoo Music数据集上的实验结果表明,SPCF在MAE指标上比传统的基于内存的协同过滤推荐算法有明显的提高.

关键词 推荐系统;相似度传播;基于内存的协同过滤
中图法分类号 TP391 **DOI号** 10.3724/SP.J.1016.2013.00671

SPCF: A Memory Based Collaborative Filtering Algorithm via Propagation

ZHAO Qin-Qin LU Kai WANG Bin
(*Institute of Computing Technology, Chinese Academy of Sciences, Beijing* 100190)

Abstract Memory based Collaborative Filtering (CF) plays an important role in current Internet recommendation engines. However, this technology suffers from serious data sparsity of user-item rating matrix. This paper proposes a kind of improved model called SPCF, which is based on the state-of-the-art methods. The key property of the proposed model is that it finds more reliable users through similarity propagation and recommends items from both user and item information. Our experimental results on the two datasets — Movie Lens and Yahoo Music show that the proposed model achieves at least 3% lift in MAE relative to traditional collaborative filtering algorithms.

Keywords recommend system; similarity propagation; memory-based collaborative filtering

1 引 言

随着互联网的飞速发展,网络上的信息严重"过载",用户很难在大量的信息中找到自己真正需要的信息.推荐系统根据用户个人的习惯和偏好向用户推荐其有可能感兴趣的项目(item),是解决信息"过载"问题的主要工具.协同过滤是现行推荐系统中应用最广泛、也最成功的技术之一.它主要基于如下假设:具有相似兴趣的用户会喜欢相似的项目.这样,它可以根据用户对项目的评分信息(通常表示成用户-项目评分矩阵,参考图1)找到用户或项目之间的关联,然后根据这种关联向用户推荐项目.协同推荐主要包括基于内存和基于模型两大类技术.前者直接利用评分信息计算用户或项目之间的相似度,后者对数据进行训练得到数据模型然后进行推荐.前者由于可解释性强而受到广泛关注,也是本文的研究对象.

基于内存的协同过滤CF(Collaborative Filtering)方法主要利用用户或者项目的相似度来进行推荐,数据的稀疏性使得用户或者项目之间的重叠性降低从而严重影响相似度计算.实际应用中广泛存在用户-项目评分矩阵的稀疏性,这严重影响了基于内存CF方法的应用.

收稿日期:2012-06-30;最终修改稿收到日期:2012-10-23.**赵琴琴**,女,1985年生,硕士,主要研究方向为信息检索和协同过滤.E-mail:zqqfy100637@163.com.**鲁 凯**,男,1988年生,硕士,主要研究方向为信息检索.**王 斌**,男,1972年生,博士,副研究员,主要研究方向为信息检索和自然语言处理.

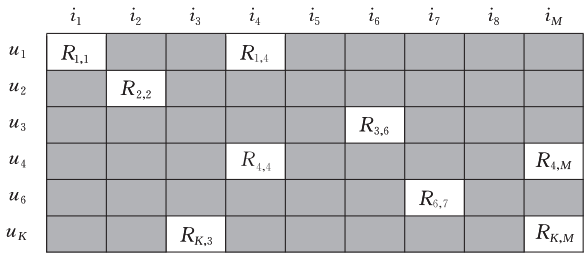

图 1 用户-项目评分矩阵

为了克服数据稀疏性问题,研究人员开展了一系列的研究.然而,现有方法中使用的相似度计算方法可靠性和推荐进度都有待进一步提高.针对上述问题,本文利用相似度传播的思路,提出了一种新的基于内存的协同推荐算法 SPCF(Similarity Propagation based Collaborative Filtering).在两个公共数据集 Movie Lens 和 Yahoo Music 上的实验结果表明,我们的方法优于传统的基于内存的协同过滤推荐算法.

本文第 2 节介绍相关工作;第 3 节介绍 SPCF 的具体实现;第 4 节给出实验和结论;第 5 节对全文进行总结和展望.

2 相关工作

为了解决基于内存的 CF 方法中用户-项目评分信息的数据稀疏问题,国内外研究者进行了一系列研究,提出了多种解决方法.这些方法大都采用各种技术对矩阵进行填充.最简单的填充办法是将用户对未评分项目的评分设为一个固定的缺省值,或者设为其他用户对该项目的平均评分[1],然而用户对未评分项目的评分不可能完全相同,这种简单的办法并不能从根本上解决稀疏性问题.更多研究采用预测填充技术.文献[2-4]采用 BP 神经网络来进行评分预测.这种方法对噪声数据有较强的承受能力[3],可以有效降低用户-项矩阵的稀疏性.然而,BP 算法的效率会导致近邻查找时间的延长[4].文献[5-8]基于朴素贝叶斯方法估计某个项目所属的类别,然后利用此类别中其它项目的评分情况来预测未评分项目的评分.文献[9-10]利用基于项目的属性联系以及项目所处的地位、相互关系和项目的元数据等内容计算项目之间的内容相似度,而不依赖用户对项目的评分.这类方法的应用面较窄.文献[11]于 2005 年提出了一种基于聚类的方法,即通过计算用户之间的相似度把用户分成 k 类,然后选择离目标用户最近的一个类中的所有用户作为邻居,计算目标用户与其邻居的相似度.该方法中的聚类数目不易确定.另外一些工作[12-14]采用奇异值分解解决矩阵的稀疏性问题,将原始空间转换到另一空间来避免稀疏性.但是一方面矩阵分解的开销太大,另一方面降维会导致用户-项矩阵中的信息丢失.

上述基于矩阵填充的方法没有考虑前次填充结果所带来的连锁反应,因此实际上填充结果并不稳定.在基于模型的 CF 方法中也有一些基于传播思路的方法[15-16],但是其目标是学习到模型参数,和本文的方法是截然不同的.

相对于传统的基于内存的 CF 方法,本文方法同时考虑了目标用户的邻居对目标项目的评分和目标用户对目标项目的邻居的评分,从而进一步提高相似度计算的准确性.

3 本文工作

本文的主要思想是利用相似度传播技术,寻找更多的邻居信息,然后在此基础上结合用户和项目两方面的信息向目标用户推荐,我们将该算法命名为 SPCF.下面对其进行详细的介绍.

3.1 算法思路

我们通过一个例子阐明本文算法思路.图 1 给出了一个稀疏的用户-项目评分矩阵,其共有 K 个用户和 M 个项目,其中灰色部分表示用户对项目未评分.假设我们要对 u_1 进行推荐,按照传统的基于用户协同过滤推荐算法,首先计算跟 u_1 兴趣偏好相同的邻居,可以找到 u_4.然后根据 u_4 的评分预测 u_1 对 i_M 的评分.很显然,该预测值只是基于一个用户的评分,预测的精度值得怀疑.而我们的思想是:在每位用户都已经找到了自己的邻居之后,可以利用相似度传播寻找每位用户更多的邻居.比如 u_1 的邻居为 u_4, u_4 的邻居为 u_1 和 u_K.此时我们认为 u_1 和 u_K 应该在兴趣偏好上也相似.之后我们可以同时利用 u_4 和 u_K 的评分对 u_1 进行推荐,这样在很大程度上可以提高预测的准确度.

算法的工作流程如图 2 所示.

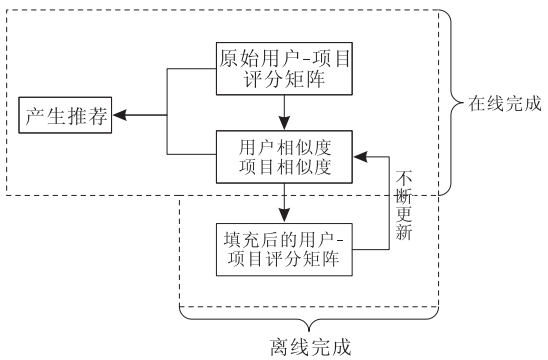

图 2 SPCF 的算法流程

3.2 算　法

本节从两个步骤来探讨基于内存的相似度传播的协同过滤推荐算法，即邻居的生成和推荐的生成.

3.2.1　邻居的生成

为了充分利用原始的用户-项目评分数据(毕竟原始用户的评分数据才能真正代表用户的兴趣偏好)，在计算用户和项目之间的相似度时，我们采用了之前提出的 SimTrans 算法[17].

(1)对于用户来说，设用户 u 和用户 v 在 n 维空间上的共同评分组成的向量分别表示为 $\boldsymbol{U},\boldsymbol{V}$，并且用户 u 和用户 v 共同评分过的项目集合用 $I_{uv}=I_u\cap I_v$ 表示. 用户 u 和用户 v 的相似性 $SimTrans(u,v)$ 为

$$SimTrans(u,v)^{(h+1)}=\begin{cases}\frac{1}{2}[sim(u,v)+sim(u,v)], & h=0\\ \frac{C}{2}[sim(u,v^{(h)})+sim(u^{(h)},v)], & h>0\end{cases}\tag{1}$$

其中，$C(0<C<1)$ 是一个置信系数或阻尼系数，表示用户相似度随着迭代次数增加的传播衰减率；h 表示迭代次数.

(2)对于项目来讲，设项目 i 和项目 j 在 m 维的用户空间上的共同评分组成的向量分别表示为 $\boldsymbol{I},\boldsymbol{J}$，并且项目 i 和项目 j 共同被评过分的用户集合用 $U_{ij}=U_i\cap U_j$ 表示. 项目 i 和项目 j 的相似性 $SimTrans(i,j)$ 为

$$SimTrans(i,j)^{(h+1)}=\begin{cases}\frac{1}{2}[sim(i,j)+sim(i,j)], & h=0\\ \frac{C}{2}[sim(i,j^{(h)})+sim(i^{(h)},j)], & h>0\end{cases}\tag{2}$$

其中 C 的含义同式(1)相同，表示项目相似度随着迭代次数增加的传播衰减率.

式(1)和式(2)中使用的 sim 相似度函数可以用传统基于内存的协同过滤推荐算法的相似度度量函数替代.

3.2.2　推荐的生成

跟传统的基于内存的协同过滤推荐算法不同，基于内存的相似度传播的协同过滤推荐算法从用户和项目两方面考虑预测目标用户对目标项目的评分. 根据目标用户对目标项目的预测评分，推荐给目标用户预测评分最高的一个项目或者 N 个项目作为结果. 假设目标用户 a 的最近用户邻居集合用 A 表示，目标项目 t 的最近项目邻居集合用 T 表示，则目标用户 a 对目标项目 t 的预测评分 $P_{a,t}$ 可以通过目标用户 a 由最近用户邻居集合 A 中邻居对项目的评分和目标用户对目标项目 t 的最近项目邻居集合 T 的评分得到，方法如下：

$$P_{a,t}(h)=\frac{1}{2}\left(\bar{R}_a+\frac{\sum\limits_{a'\in A}SimTrans(a,a')^{(h)}\times(R_{a',t}-\bar{R}_{a'})}{\dfrac{1}{\sum\limits_{a'\in A}SimTrans(a,a')^{(h)}}}+\bar{R}_t+\frac{\sum\limits_{t'\in T}SimTrans(t,t')^{(h)}\times(R_{a,t'}-\bar{R}_{t'})}{\dfrac{1}{\sum\limits_{t'\in T}SimTrans(t,t')^{(h)}}}\right)\tag{3}$$

其中 $P_{a,t}$ 为目标用户 a 对项目 t 的预测评分值. $SimTrans(a,a')$ 为目标用户 a 和用户 a' 之间的兴趣偏好相似度，$SimTrans(t,t')$ 为目标项目 t 和项目 t' 被用户喜欢的程度大小.

$\bar{R}_a$ 为原始用户-项目评分矩阵中目标用户 a 对以前评过分的项目的平均评分，$\bar{R}_t$ 为原始用户-项目评分矩阵中所有用户对目标项目 t 的平均评分.

4 实验和结论

4.1 实验数据

实验数据采用的是美国明尼苏达州立大学的 Group Lens 研究小组提供的 Movie Lens 数据集(http://www.grouplens.org). 它包含了 943 位用户对 1682 部电影的 100000 个评分. 该数据集的原始用户-项目评分矩阵的稀疏度为 93.7%. 然而为了更贴近现实，同时也为了更能突出我们提出的算法的优势，我们从该数据集中首先随机抽取了 500 个用户，其中 300 个用户作为训练数据，记作 Training300，其余 200 个用户作为测试数据. 在 Training300 的数据基础上，我们又随机抽取了 200 个用户的数据作为另一组训练数据，记作 Training200. 我们按照用户的打分个数(5 个，10 个和 20 个)分成 3 个不同的测试数据，记作 Given5、Given10 和 Given20.

另外，在 Yahoo Music① 的数据上也抽取了小部分数据来验证我们提出的算法的有效性. 同样地，我们仍然采用随机抽取的方法在训练数据上抽取了 500 个用户以及这 500 个用户在对应的测试数据上的测试数据. 但是有一点不同的是：由于 Yahoo Music 数据集的庞大，我们只选用了前 2000 个音乐作为项目集合.

① http://webscope.sandbox.yahoo.com/

4.2 评价指标

实验评估标准的选择是实验的重要组成部分.合理的评估标准能够很好地评价算法性能,发现算法的有待改进之处.本实验从推荐的准确性方面考虑,选取了协同过滤中常用的平均绝对误差 *MAE* (*Mean Absolute Error*)作为实验结果的评估标准.

平均绝对误差 *MAE* 是评价推荐算法质量的标准之一,它通过计算预测评分与真实评价数据上的差别来衡量推荐结果的准确性.*MAE* 的值越小,推荐准确性越高.假设预测的用户评分集合表示为 $\{p_1,p_2,\cdots,p_N\}$,对应的实际用户评分集合为 $\{q_1,q_2,\cdots,q_N\}$,则具体的 *MAE* 计算公式为

$$MAE=\frac{\sum_{i=1}^{N}|p_i-q_i|}{N} \tag{4}$$

4.3 实验方案

我们引入了传统的基于内存的推荐算法进行对比.由于推荐算法涉及到相似度计算和预测评分两个关键部分.我们考察了 5 种相似度度量方法和 2 种预测评分函数的 10 种组合方法.其中 5 种相似度度量方法包括基于用户的皮尔逊相关系数 UPCC (User-based Pearson Correlation Coefficient)、基于用户的余弦夹角相关性 UVS(User-based Vector Similarity)、基于项目的皮尔逊相关系数 IPCC (Item-based Pearson Correlation Coefficient)、基于项目的余弦夹角相关性 IVS(Item based Vector Similarity)和基于项目的修正余弦夹角相关性 IAVS (Item-based Adjusted Vector Similarity).2 种预测评分函数分别为简单的加权平均 SWA(Simple Weighted Average)和基于相似用户评分的加权求和 WSOR(Weighted Sum of Others' Ratings).最后取上述组合中最好的 3 个结果和我们的方法对比.

整个实验分为两个部分:

实验 1. 考察基于内存的相似度传播的协同过滤推荐算法的收敛性.本论文采用的收敛判断标准是平均偏差的比率,即偏差比,此时的平均偏差为迭代后的预测结果与迭代前的预测结果之间的平均偏差.偏差比等于 1 时的迭代次数就为算法收敛的迭代次数.

实验 2. 对比 SPCF 算法和传统的基于内存的协同过滤推荐算法.

4.4 实验结果及分析

(1)基于内存的相似度传播的协同过滤推荐算法的收敛性问题

从图 3 可以看出,本文提出的方法在迭代一定次数之后,均表现出一定的稳定性.需要指出的是,该算法的收敛性还有待严格的理论证明.但是通过实验基本能给出算法收敛性的经验结论.同时,根据我们提出来的判断算法的收敛标准以及这 3 幅图的结果,我们把基于内存的相似度传播的协同过滤推荐算法的迭代次数定为 10.

(2)不同方法推荐质量精度的对比

首先,确定算法的邻居个数.该值采用的是传统基于内存的协同过滤推荐算法判定出的值.图 4 分别显示了测试数据邻居个数和 *MAE* 的曲线.

从图 4 中我们得出邻居个数为 60 的情况下推荐效果最佳.

其次,表 1(a)、(b)分别显示了 3 种传统基于内存的协同过滤推荐算法和 SPCF 算法在不同实验数据上 *MAE* 的对比.

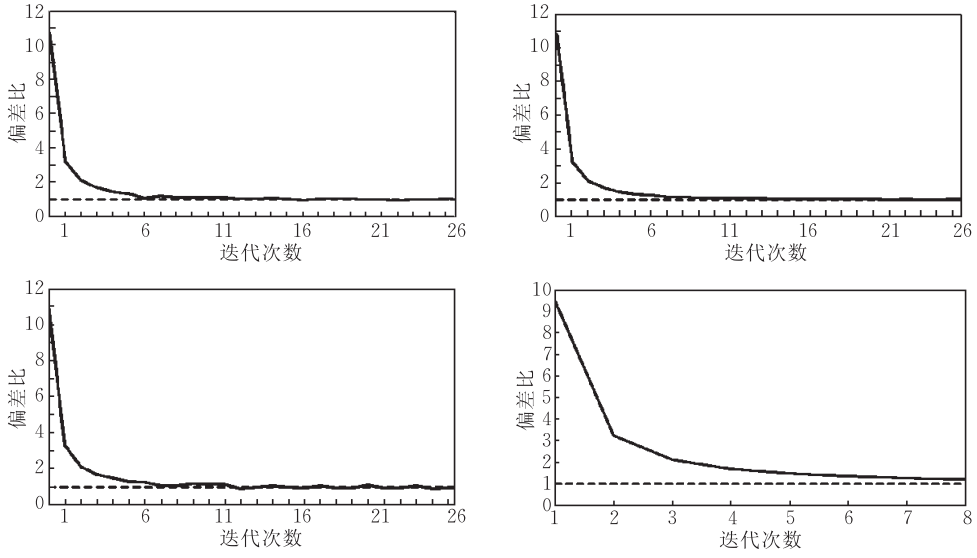

图 3 Training300 的迭代次数与偏差比对比

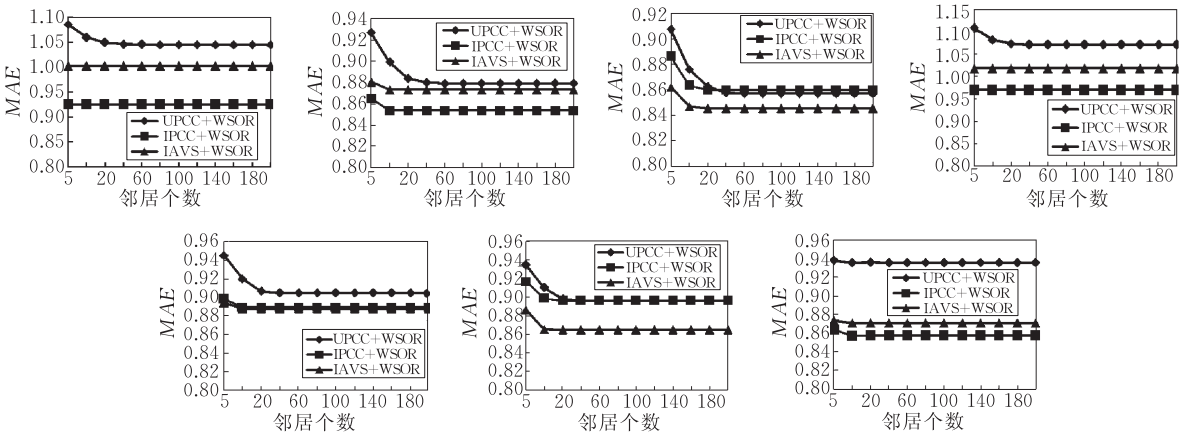

图 4　邻居个数与 *MAE* 的取值

表　1

(a) 各种算法在 *MAE* 指标上的性能(Yahoo Music)

训练数据	方法	Given5
Yahoo music	SPCF	0.843(+50%)
	UPCC+WSOR	1.888
	IPCC+WSOR	1.693*
	IAVS+WSOR	1.726

(b) 各种算法在 *MAE* 指标上的性能(Movie Lens)

训练数据	方法	Given5	Given10	Given20
Training 300	SPCF	0.853 (+7.6%)	0.814 (+4.5%)	0.811 (+4%)
	UPCC+WSOR	1.045	0.879	0.857
	IPCC+WSOR	0.924*	0.853*	0.859
	IAVS+WSOR	1.002	0.873	0.845*
Training 200	SPCF	0.870 (+10.2%)	0.832 (+6.2%)	0.836 (+3.2%)
	UPCC+WSOR	1.070	0.947	0.895
	IPCC+WSOR	0.969*	0.889	0.895
	IAVS+WSOR	1.018	0.887*	0.864*

表 1(b)中"*"号代表 3 种传统的基于内存的协同过滤推荐算法中 *MAE* 值最小的那个方法，即推荐精确度最高的方法. 括号中的百分数表示的是我们提出的基于相似度传播的方法比带"*"号的方法在推荐精确度方面的提高比率.

在 Yahoo Music 数据集上的提高结果非常可观，达到了 50%. 而在 Movie Lens 数据集的提高也高于传统方法. 这两者提高幅度的差距可能和这两个语料的具体数据不同有关. 表 1(b)的数据同时也表明，当已知打分信息减少时，SPCF 相对于传统方法的提高幅度逐渐提高，这也表明本文方法在稀疏情况下能够体现出更强的优势.

5　总结和展望

本文介绍了一种改进的基于内存的协同过滤推荐算法 SPCF，该算法通过不断更新原始用户-项目评分矩阵为目标用户和目标项目找到更多的最近邻居集合，对目标用户给出推荐结果. 实验结果验证了 SPCF 方法的有效性.

未来的工作包括：

(1) 对本文方法的收敛性进行理论上的证明.

(2) 本文方法主要关注精度，但是算法的时空开销较大，下一步要研究降低开销的方法.

(3) 将该方法和基于内容的 CF 方法相结合.

(4) 分析噪音数据在相似度传播过程中的影响程度.

(5) 考虑与在 Movie Lens 数据集上表现较好的方法的对比，比如 Model-based CF 和矩阵分解的一些方法.

致　谢　在撰写论文期间，王斌老师谆谆教导，实验室同学给予了鼓励，特别是李亚楠师兄丰富的想法启发并影响我完成了这篇论文. 向所有支持、关心和帮助过我的人表示最诚挚的谢意！

参　考　文　献

[1] Deng Ai-Lin, Zhu Yang-Yong, Shi Bai-Le. A collaborative filtering recommendation algorithm based on item rating prediction. Journal of Software, 2003, 14(9): 1621-1628

[2] Zhang Feng, Chang Hui-You. A collaborative filtering algorithm embedded bp network to ameliorate aparsity issue// Proceedings of the International Conference on Machine Learning and Cybernetics. Guangzhou, China, 2005: 1839-1844

[3] Chen Gang, Liu Fa-Sheng. Method for data mining based on BP neural network. Computer and Modernization, 2006, 10(6): 20-22

[4] Jia Li-Hui, Zhang Xiu-Ru. Analysis and improvements of BP algorithm. Computer Technology and Development, 2006, 16(10): 101-103, 107

[5] Jung Kyung-Yong, Hwang Hee-Joung, Kang Un-Gu. Constructing full matrix through naive Bayesian for collaborative filtering//Proceedings of the Computational Intelligence. Kunming, China, 2006: 1210-1215

[6] Robles V, Larranaga P, Menasalvas E, Perez M S, Herves V. Improvement of Naïve Bayes collaborative filtering using interval estimation//Proceedings of the IEEE/WIC International Conference on Web Intelligence. Halifax, Canada, 2003: 168-174

[7] Ko Su-Jeong. Prediction of consumer preference through Bayesian classification and generating profile//Proceedings of the Conceptual Modeling for Novel Application Domain. IL, USA, 2003: 29-39

[8] Jung Kyung-Yong. User preference through Bayesian categorization for recommendation//Proceedings of the PRICAI 2006: Trends in Artificial Intelligence. Guilin, China, 2006: 112-119

[9] Chedrawy Zeina, Abidi Syed Sibte Raza. An adaptive personalized recommendation strategy featuring context sensitive content adaptation//Proceedings of the Adaptive Hypermedia and Adaptive Web-Based Systems. Dublin, Ireland, 2006: 61-70

[10] Kim B M, Li Q, Park C S et al. A new approach for combining content-based and collaborative filtering. Journal of Intelligent Information Systems, 2006, 27(1): 79-91

[11] Xue Gui-Rong, Lin Chenxi, Yang Qiang, Xi Wensi, Zeng Hua-Jun, Yu Yong, Chen Zheng. Scalable collaborative filtering using cluster-based smoothing//Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Salvador, Brazil, 2005: 114-121

[12] Zhang Sheng, Wang Weihong, Ford J, Makedon F, Pearlman J. Using singular value decomposition approximation for collaborative filtering//Proceedings of the 7th IEEE International Conference on E-commerce Technology. Munich, Germany, 2005: 257-264

[13] Vozalis M G, Margaritis K G. Applying SVD on item-based filtering//Proceedings of the 5th International Conference on Intelligent System Design and Applications. Wroclaw, Poland, 2005: 464-469

[14] Aggarwal Charu C. On the effects of dimensionality reduction on high dimensional similarity search//Proceedings of the 12th ACM SIGMOD-SIGACT-SIGART Symposium on Principles of database systems. Washington, USA, 2001: 256-266

[15] Pucci Augusto, Gori Marco, Maggini Marco. A random-walk based scoring algorithm applied to recommender engines//Proceedings of the Advances in Web Mining and Web Usage Analysis. San Jose, USA, 2007: 127-146

[16] Zhang Jiyong, Pu Pearl. A recursive prediction algorithm for collaborative filtering recommender systems//Proceedings of the RecSys. Minnesota, USA, 2007: 57-64

[17] Li Yanan, Wang Bin, Xu Sheng, Li Peng, Li Jintao. QueryTrans: Finding similar queries based on query trace graph//Proceedings of the IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technologies. Milano, Italy, 2009: 260-263

ZHAO Qin-Qin, born in 1985, M.S.. Her research interests include information retrieval(IR) and collaborative filtering(CF).

LU Kai, born in 1988, M. S.. His research interest is information retrieval(IR).

WANG Bin, born in 1972, Ph. D., associate professor. His research interests include information retrieval (IR) and natural language processing (NLP).

Background

This paper studies the problem which belongs to the field of information retrieval, and it is related to recommendation system. This field is very hot in the worldwide, with more commonly matrix filling and transfer technology.

With gradual popularization of the Internet and rapid development of the E-Commerce, Web appears serious information overload and it is hard for users to find their real needed products within a mass of product information. E-Commerce recommender systems are based on personal habits and preferences of a user to recommend him product which he might be interested in, and they are the main tools to solve the problem of information overload.

In previous work, predecessors mainly focued on the recommendation algorithm based on association rules, content, and collaborative filtering algorithm. Besides, different recommendation algorithms on different occasions may get different results.

This paper valuably explores memory-based collaborative filtering algorithms, particularly in data sparsity. In the future, we will do further research with big data and statistic methods.