服务于定向信息推荐的模糊聚类协同推荐算法*)

辛治运[1] 马兆丰[2] 顾 明[1]

(清华大学计算机系 北京 100084)[1] (北京邮电大学信息工程学院 北京 100876)[2]

摘 要 面对金融领域信息量和用户数量的不断增加,现有的金融信息推荐算法不能很好地满足金融用户的信息需求,推荐结果的及时性和准确性有待进一步提高。在分析现有协同推荐算法的基础上,本文提出了金融信息模糊聚类协同推荐算法,将模糊聚类和协同推荐算法相结合,以用户-项目评价矩阵为研究基础,对有相似信息需求兴趣的用户进行模糊聚类,用户组群的兴趣爱好代表并预测个人的兴趣爱好,能为用户提供和发现新的信息资源,很好地满足金融用户信息需求的多兴趣性和时效性。最后对提出的算法进行实验,实验结果表明了算法具有良好的推荐效果。

关键词 信息推荐,协同过滤,模糊聚类,用户聚类

Collaborative-filtering Recommendation Algorithm Based on Fuzzy Clustering for Domain Information Service

XIN Zhi-Yun[1] MA Zhao-Feng[2] GU Ming[1]

(Department of Computer Science and Technology, Tsinghua University, Beijing 100084)[1]

(School of Information Engineering, Beijing University of Post and Telecommunicatons, Beijing 100876)[2]

Abstract In face of the increase of Web information and the number of user in financial information domain, present financial information recommendation algorithm can't meet the users' information demand. After the analysis of present collaborative recommendation algorithm, a algorithm which combines the fuzzy clustering and collaborative-filtering method, based on the user- item rating-matrix, clusters user with similar interest in the way of fuzzy clustering is presented in this paper. It can stand for and forecast the personal information need with the information interest of the group which these users belonged to. The algorithm can find new information and interest for financial user. From the experiment result, the recommendation efficiency of algorithm is improved; it can meet the users' information need in the side of Multiple-interests and time.

Keywords Information recommendation, Collaborative-filtering, Fuzzy clustering, User clustering

协同推荐系统是利用用户的历史偏好信息实现个性化服务的系统,它已经成为电子商务和信息获取领域中的重要应用。然而,随着网络规模的扩大和信息量的巨增,用户很难在短时间内找到需要的信息,特别是在时效性要求比较高的信息领域,如金融信息领域,信息的重要性和时间有着重要的关系,且属于典型的多兴趣(Multiple-interests)、多项目(Multiple-items)的信息推荐。如何在短时间内能为用户推荐合理有效的信息资源仍然是一个十分重要的课题。

1 金融领域信息模型

金融信息涉及的内容广泛,领域繁多。相对于某个特定的金融用户,其需求的信息可能相对集中在某个领域,如证券用户可能更多地关注证券和股票信息,而对银行、期货、保险等领域关注较少。文[1]提出了用户信息模型的概念,将信息空间划分为不同层次的领域信息需求和原子信息需求。而原子信息需求更能较准确地描述金融用户的信息需求。

金融信息领域可以看作一个全局信息空间 F_{Global},全局信息空间是许多金融领域信息空间 F_{DI} 的并集,描述如下:

$$F_{Global}=F_{DI}\cup F_{D2}\cup F_{D3}\cup\cdots\cup F_{Dn}$$

每个特定的金融信息领域 FDI(Financial Domain Information)又是多个原子信息领域 PFD(Personal Financial Domain)的并集,可用下式表示:

$$FDI=FDI_1\cup FDI_2\cup FDI_2\cup\cdots\cup FDI_i$$

通常,一个金融用户可能对若干信息领域感兴趣,用层次化的金融信息模型更能合理、真实、准确地反映用户的实际需求,使系统能够准确、主动地检索信息并且进行推荐服务,提高了信息推荐的效率[2]。本文在文[1,2]工作基础上,分析金融用户原子信息需求,对兴趣相似的用户进行聚类,使用户的需求对应于模型中的一个或多个原子信息需求,更好地提高协同推荐系统的服务质量。图 1 给出了金融领域的信息模型。

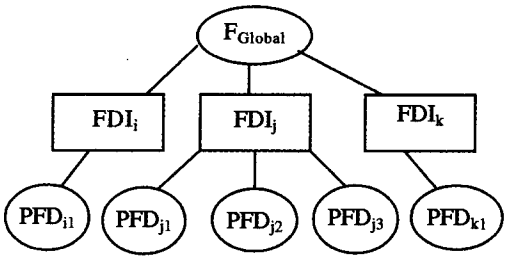

图 1 金融领域信息模型图

2 协同过滤的推荐算法概述

协同过滤技术是根据用户的相似性来推荐信息,国内外

*)本课题得到国家"九七三"重点基础研究项目(2004CB719406)。辛治运 博士研究生,研究方向:信息处理,计算智能。

已经有了很多基于协同过滤的推荐系统，Tapestry[3]是最早提出的协同过滤推荐系统，目标用户需要明确指出与自己行为比较类似的其他用户。GroupLens[4]是基于用户评分的自动化协同过滤推荐系统，用于推荐电影和新闻，Ringo 推荐系统[5]和 Video 推荐系统[6]通过电子邮件的方式推荐音乐和电影。

目前，协同推荐出现了很多改进算法，如基于粗糙集的协同过滤算法[7]，曾艳等提出基于用户-项目评价的协同过滤算法[8]、O'Connor. M 等[9]提出对项目进行聚类，然后在对应的聚类中搜索目标用户的最近邻居、基于降维的协同推荐；林鸿飞等将内容推荐和协同推荐两种方式结合进行合作推荐[10]、孙汝杰等提出的基于时间序列的协同推荐[11]等。上述算法在不同程度上解决了原始算法的一些问题，但协同推荐中的最初评价问题[12]、数据稀疏问题[13]、实时性推荐问题在一定程度上仍然存在。

由于用户可能具有多个兴趣点，不能笼统说某个用户就属于某个类别，可能属于这个类，也能说属于另外一个类。对用户的描述只能是通过用户对某些领域的感兴趣度或从属度来刻画显得更为合理。本文提出的算法将模糊聚类与协同推荐算法相结合，对相似兴趣的用户进行模糊聚类用户组群的兴趣爱好代表并预测个人的兴趣爱好。

3 金融信息模糊聚类协同推荐

3.1 金融信息模糊聚类协同推荐模型

模糊聚类是利用模糊等价关系将给定的对象分为一些等价类，这种聚类方法不需要事先确定聚类的数目，而是通过一定的阈值来确定对象的相似类别。基于用户的协同过滤是个性化推荐中应用最为广泛的方法，它是基于邻居用户的兴趣爱好预测目标用户的兴趣偏好。算法先使用模糊聚类寻找与目标用户有相同喜好的邻居，然后根据目标用户的邻居的偏好产生向目标用户的推荐。这种方法利用项目之间在相似群组的相似性来初步预测用户未知项目的喜好，在此基础之上再完成基于用户的协同过滤推荐算法，基于模糊聚类的协同推荐推荐模型如图 2 所示。

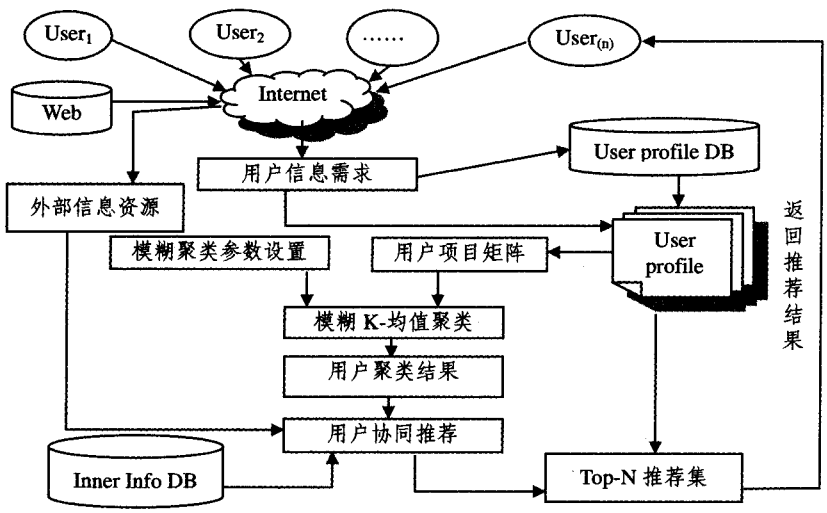

图 2 基于模糊聚类的协同推荐模型

(1)金融用户模糊聚类

构造用户对类别的评价偏好，一是对用户-项目偏好矩阵降维，因为簇的数目是远远小于项目的数目，低维矩阵的计算必定比高维矩阵的计算时间要少，从而提高推荐方法的可扩展性；二是在实际应用中，一般来说用户对同类项目的喜好程度基本上是相一致的，那么由用户对类别的偏好代替用户对单个项目的偏好并不会影响到用户偏好的表达；三是用户-项目评价偏好是稀疏的而用户-模糊簇的评价偏好是密集的，这可以解决评价数据稀疏性造成的相似群度量不准确的问题。

聚类的主要思想是通过用户描述文件，即用户对项目的感兴趣情况进行用户聚类，用户的兴趣用矢量空间模型表示，首先选择一个有代表的用户作为该类用户的聚类中心，然后计算通过计算目标用户和聚类中心的相似度，不断选择聚类中心，直到满足设定的阈值，产生用户的聚类，计算相似度可以采用下面的公式：

$$sim(u,v)=\frac{\sum_{j=1}^{c}(PC_{u,j}-\overline{PC_u})(PC_{v,j}-\overline{PC_v})}{\sqrt{\sum_{j=1}^{c}(PC_{u,j}-\overline{PC_u})^2}\sqrt{\sum_{j=1}^{c}(PC_{v,j}-\overline{PC_v})^2}}$$

其中，$PC_{u,j}$代表用户 u 对模糊簇 j 的偏好值，$\overline{PC_u}$代表用户 u 在所有模糊簇的平均评价偏好值。用户 v 的表示与用户 u 一样。

(2)金融信息 Top-N 推荐集合的产生

用户聚类产生以后，选择待推荐用户所在聚类中与该用户相似的 k 个用户，并由这 k 个近邻对目标项目的评分值来完成目标用户 u 对目标项目 t 的预测。预测公式如(1)式所示。

$$p_{u,t}=\overline{PI_u}+\frac{\sum_{v=1}^{k}sim(u,v)*(PI'_{(v,t)}-\overline{PI_v})}{\sum_{v=1}^{k}|sim(u,v)|} \quad (1)$$

其中，$\overline{PI_u}$代表用户 u 的平均评分值，如果用户 v 对目标项目已评分，$PI'_{(v,t)}$就为 $PI_{(v,t)}$即用户 v 对项目 t 的实际评分值；否则利用项目之间属性特征的相似性和用户 v 已经评分过的项目，对目标项目 t 进行初步预测。产生最后的 Top-N 推荐集合，并将这些推荐信息以合适的方式返回给用户。

3.2 金融信息模糊聚类协同推荐算法

算法 服务于定向信息推荐的模糊聚类的协同推荐算法

输入 待进行信息推荐的用户 $User(i)$和一个相关信息需求用户集 $U=\{U|User_1,User_2,User_3,\cdots,User_i,\cdots\}$；

输出 待推荐用户得到的 Top-N 信息推荐集合，$I=\{item_i|i\in[1,N]\}$；

Step1：取得待聚类用户的 User profile，对描述用户兴趣的数据进行预处理和数据变换；

Step2：设置最大类别数目 Cmax、实际类别数目 Cmin、初始化用户聚类中心 $V^0_{c\max}$、设定 $\varepsilon(\varepsilon>0)$、$1=0$，$k=0$，$C(l)=$ Cmax，最近邻居个数 N；

Step3：For，$\forall i,\forall j,u_{ji}\in[0,1]$计算 $\mu^{(k)}_{CD}$ 根据

$$\mu_{ji}=1/\sum_{i=1}^{k}\left(\frac{X_j-C_i}{X_j-C_l}\right)^{\frac{2}{m-1}};$$

Step4：For，$\forall j$，计算聚类中心 $V^{(k+1)}_{CD}$ 根据

$$V_i=\sum_{j=1}^{n}(\mu_{ji})^m X_j/\sum_{j=1}^{n}(\mu_{ji})^m;$$

Step5：如果 $\| V^{(k+1)}_{CD}-V^{(k)}_{CD}\|>\varepsilon$，置 $k=k+1$，返回 Step2，如果满足设定条件，结束用户聚类算法。

Step6：找到待推荐的用户 $User_i$ 所属聚类，再根据用户-项目矩阵计算用户相似度找到该聚类中和目标用户兴趣最相似的 N 个用户，相似度计算采用公式：

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

Step7：在用户群组的兴趣中找出待推荐用户没有发现但可能感兴趣的信息资源集合 $I=\{item_1,item_2,item_3,\cdots,item_i,\cdots\}$；

Step8：根据用户组对这些项目的感兴趣程度，推算目标用户可能感兴趣的 Top-N 推荐集合，采用公式(2)计算

$$prediction-N=\bar{u}+\frac{\sum_{i=1}^{n}(corr_i)\cdot(rating_i-\bar{i})}{\sum_{i=1}^{n}(corr_i)} \qquad (2)$$

Step9：将产生的 Top-N 信息推荐集合返回给目标用户，完成推荐；

Step10：结束。

4 算法评价和实验分析

4.1 实验环境和检验指标

为了验证算法的有效性，实验环境操作系统为 Windows NT2000，Inter(R) Pentium(R) CPU 2.4GHz，内存：1.0GB。程序执行环境采用 Visual C++ 6.0，实验数据来自中国金融网(http://www.cifiew.com/2004-6-v/index.jsp)金融论坛用户对各种金融数据的评价数据，选择和兴趣点相似的 5 类用户包括外汇牌价、国债利率、股票价格、股票指数、期货价格。文中采用平均绝对偏差 *MAE* 作为推荐质量的度量标准[11]。平均绝对偏差 *MAE* 是计算所预测的用户评分与实际的用户评分之间的偏差来度量预测的准确性 *MAE* 越小，推荐质量越高。平均绝对偏差 *MAE* 定义为下式，其中，通过预测得到的用户评分集合表示为 $\{p1,p2,\cdots,p_n\}$，实际的用户评分集合为 $\{q1,q2,\cdots,q_n\}$。

$$MEA=\frac{\sum_{i=1}^{n}|p_i-q_i|}{N}$$

4.2 实验结果和算法分析

将本算法和基于用户(User-based)的协同推荐做了比较实验，以从平均绝对偏差 *MAE* 作为推荐质量的度量标准得到如图 3 所示结果。

可以看出，随着用户邻居数目的不断增加，基于模糊聚类的推荐总是比基于用户的协同推荐具有较小的 *MEA* 值，表明提出的方法在信息推荐的精度上的确有所改进。将相似用户集合，以用户群组的兴趣和需求为参考来进行目标用户的推荐，也能在一定程度上克服目标用户对信息资源评价的数据稀疏问题。在模糊聚类的基础上得到某个用户相对与各个用户群组的隶属度，从而能得到比较精确的推荐和比较高的查全率。为了选择一个最佳的聚类阈值，分别在三个不同的阈值上进行了实验，三个阈值分别是：0.7、0.8 和 0.95。

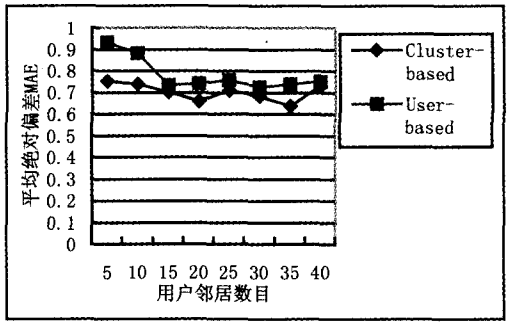

图 3 user-base 和 cluster-based 推荐的 MAE 比较

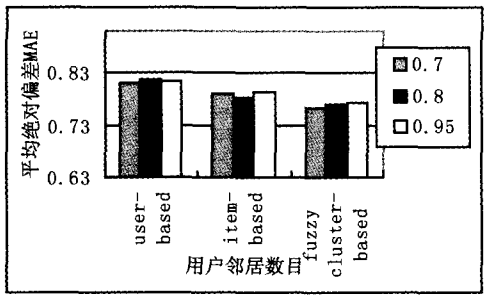

图 4 不同阈值对应的平均绝对偏差 MAE 比较

通过对设定的不同聚类阈值对应的平均绝对偏差 *MAE* 进行比较(图 4)，发现基于模糊聚类的协同推荐比基于用户(User-based) 和基于项目(Item-based)的平均绝对偏差 *MAE* 都相对较低，充分证明了对相似用户进行模糊聚类的推荐算法具有较好的推荐效果。

协同推荐能为用户发现新的信息，为了进一步讨论算法的有效性和对协同推荐的改进效果，以个性化推荐的一个评价标准：查全率(Recall ratio)进行了进一步的实验，图 5 是随用户类别个数增加，Cluster-based 和 User-based 两种协同推荐方法的查全率的数据。

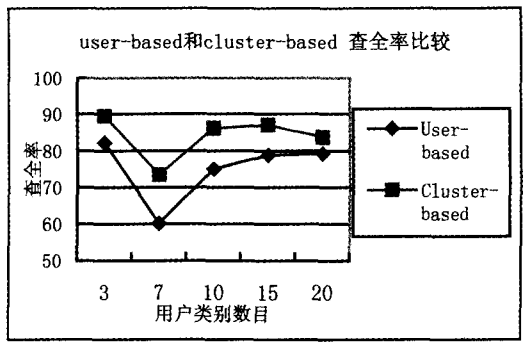

图 5 user-based 和 cluster-based 查全率比较

总结 本文在分析现有各种协同推荐算法基础上，提出

(下转第 166 页)

4 实验与分析

为了评价本文提出的混沌差分进化算法(CDE)对组合优化问题的求解性能，分别将 CDE 中的差分进化模式使用本文第 2 部分的(6)(记为 DE_1)、(7)(记为 DE_2)，得到的混沌优化算法分别为分别为 CDE_1、CDE_2。测试的数据是长度为 359 的 PSTVd 碱基序列。

5′-CGGAACUAAA[10]CUCGUGGUUC[20]CUGUGGUUCA[30]
CACCU GACCU[40]CCU GAGCAGA[50]AAAGAAAAAA[60]
GAAGGCGGCU[70]CGGAGGAGCG[80]CUUCAGGGAU[90]
CCCCGGGGAA[100]ACCUGGAGCG[110]AACUGGCAAA[120]
AAAGGACGGU[130]GGGGAGUGCC[140]CAGCGGCCGA[150]
CAGGAGUAAU[160]UCCCGCCGAA[170]ACAGGGUUUU[180]
CACCCUUCCU[190]UUCUUCGGGU[200]GUCCUUCCUC[210]
GCGCCCGCAG[220]GACCACCCCU[230]CGCCCCCUUU[240]
GCGCUGUCGC[250]UUCGGCUACU[260]ACCCGGUGGA[270]
AACAACUGAA[280]GCUCCCGAGA[290]ACCGCUUUUU[300]
CUCUAUCUUA[310]CUUGCUUCGG[320]GGCGAGGGUG[330]
UUUAGCCCUU[340]GGAACCGCAG[350]UUGGUUCCU[359]—3′

其真实的二级结构匹配碱基总数为 246，茎总数为 25 个，具体如下：

s_1(3,357,7)，s_2(14,348,8)，s_3(25,337,4)，
s_4(30,331,6)，s_5(39,322,4)，s_6(44,317,6)，
s_7(52,309,4)，s_8(60,300,9)，s_9(69,289,5)，
s_{10}(80,282,7)，s_{11}(90,270,3)，s_{12}(93,266,5)，
s_{13}(103,255,8)，s_{14}(114,246,4)，s_{15}(121,240,5)，
s_{16}(128,234,7)，s_{17}(136,226,3)，s_{18}(140,221,2)，
s_{19}(143,218,4)，s_{20}(148,213,2)，s_{21}(152,209,5)，
s_{22}(159,202,2)，s_{23}(162,199,4)，s_{24}(168,193,4)，
s_{25}(173,186,5)。。

表 1 算法预测结果的比较

算法	碱基配对正确率(平均)	茎区正确率(平均)
DE_1	78.3%	70.1%
CDE_1	86.3%	81.8%
DE_2	80.91%	72.4%
CDE_2	89.93%	84.3%

表 1 是 4 种算法 DE_1、DE_2、CDE_1、CDE_2 独立运行 20 次，预测获得的 RNA 二级结构碱基配对正确率(平均)和茎区正确比率(平均)的统计结果

实验结果表明本文的混沌差分进化算法的精确度比差分进化算法要高，其中碱基配对预测正确率要高 10%左右、茎区预测正确率要高 16%左右，验证了算法的有效性。

结论 为了借助非线性混沌本质来有效改善差分进化算法，提高算法的性能，本文将混沌优化搜索技术融入到差分进化算法，提出了混沌差分进化算法，该算法不仅保持了差分进化算法简单的优点，而且充分利用了混沌的随机性、遍历性和规律性等特点，有效克服了算法的早熟的缺陷，提高了全局最优解的计算效率，是一种高效的差分进化算法，具有较大的使用价值。另外据笔者所知，本文是混沌差分进化算法在 RNA 二级结构预测中的首次应用，是一种很好的应用尝试。对混沌差分进化算法进行理论上的分析以及在其它方面的应用是我们未来研究的方向。

参 考 文 献

1 Cai L, Malmberg R L, Wu Y. Stochastic modeling of RNA pseudoknotted structures: a grammatical approach. Bioinformatics, 2003, 19: 66～73
2 TAN Guang-Ming, FENG Sheng-Zhong, SUN Ning-Hui. An optimized and efficiently parallelized dynamic programming for RNA secondary structure prediction. J Software, 2006, 17(7): 1501～1509
3 Hofacker I L, Schuster P. Combinatorics of RNA Secondary Structure. Discr Appl Math, 1998, 88: 207～237
4 Nebel M E. Identifying Good Predictions of RNA Secondary Structure. In: Proceedings of the Pacific Symposium on Biocomputing 2004, 2004. 423～434
5 Kirkpatrick S, Gelatt C D, Vecchi M P. Optimization by simulated annealing. Science, 1983, 220: 671～680
6 DE Homepage. Http: //www. icsi. Berkeley. Edu/storn/code. htm
7 Shi Yan-jun. Teng Hong-fei, Li Zi-qiang. Cooperative Co-evolutionary Differential Evolution for Function Optimization. Lecture Notes in Computer Science, 2005, 1075～1083
8 Ali M M, Fatti L P. A differential free point generation scheme in the differential evolution algorithm. Journal of Global Optimization, 2006, 35 (4): 551～572
9 Zhang Huanguang, Wang Zhiliang, Huang Wei. Control theory of chaos system[M]. Shen Yang: Publishing House of Northeast University, 2003. 1～45
10 王翼飞，等. 生物信息学—智能计算算法及其应用. 北京：化学工业出版社，2006. 172～210

(上接第 130 页)

一种基于模糊聚类的协同信息推荐算法，通过实验结果和数据分析，基于模糊聚类的协同推荐方法较一般的协同推荐(User -based 和 Item-based)其查全率有了很大的提高，提高了推荐的质量和精度，而又有相对较小的平均绝对偏差 MAE，实验中和传统的基于用户的和基于项目的协同推荐进行了比较，充分证实了实行模糊聚类推荐的有效性。用目标用户相对于聚类后的用户组群的兴趣隶属度来描述用户兴趣更能真实地反映金融用户的需求，实验中的聚类算法采用了模糊 c 均值聚类法进行用户聚类，对传统的协同推荐方法做了一步改进。

然而，随着 Internet 信息的不断增长，如何更为有效地组织信息资源，加上金融领域信息的时效性特点，虽然以用户组群的兴趣爱好为参考，能解决一点数据稀疏的问题，如何客观描述用户在某个特定领域的最小最全需求，进一步解决用户-项目矩阵的数据稀疏问题，提高系统的推荐效率仍然是协同推荐中一个必须解决的问题。

参 考 文 献

1 Ma Zhaofeng, Feng Boqin. Support Vector Machines Learning for Adaptive and Active Information Retrieval. Advanced Web Technologies and Applications (APWEB'04), Lecture Notes Computer Science, 2004, 3007: 89～99
2 马兆丰，冯博琴. 基于支撑向量机的自适应信息推荐算法. 小型微型计算机系统，2004，25(3)：384～387
3 Goldberg D, Nichols D, Oki B, et al. Using collaborative filtering to weave an information tapestry[J]. Communications of the ACM, 1992, 35 (12): 61～70
4 Konstan J, Miller B, Maltz D, et al. GroupLens: applying collaborative filtering to usenet news. Communications of the ACM, 1997, 40(3): 77～87
5 Shardanand U, Maes P. Social information filtering: algorithms for automating "Word of Mouth"[C]. In: Proceedings of ACM CH I'95 Conference on Human Factors in Computing System s, 1995. 210～217
6 Hill W, Stead L, Rosenstein M, et al. Recommending and evaluating choices in a virtual community of Use[C]. In: Proceedings of CH I'95, 1995. 194～201
7 张巍. 刘鲁，葛健. 一种基于粗糙集的协同过滤算法 小型微型计算机系统，2005，26(11)
8 曾艳，麦永浩. 基于内容预测和项目评分的协同过滤推荐. 计算机应用，2004，24(1)
9 O'ConnerM, Herlocker J. Clustering items for collaborative filtering[C]. In: Proceedings of the ACM SIGIR Workshop on Recommender System s. Berkeley, CA, 1999
10 林鸿飞，杨志豪，赵晶. 基于内容和合作模式的信息推荐机制. 中文信息学报，2005，19(11)：1003～0077
11 孙汝杰，张宇光. 基于时间序列的个性化信息协同过滤技术研究. 情报杂志，2006(8)
12 Mobasher B, Jin X, Zhou Y. Semantically enhanced collaborative filtering on the web[C]. In: Proceedings of the European Web Mining Forum, 2004
13 Hill F, Stead L, Rosenstein M, et al. Recommending and Evaluating Choices in a Virtual Community of Use[C]. In: Proceedings of ACM CHI'95Conference on human factors in computing systems, 1995. 210～217