文章编号: 1007-5321(2014)06-0049-05 DOI: 10.13190/j.jbupt.2014.06.010

结合时间上下文挖掘学习兴趣的协同过滤推荐算法

鄂海红， 宋美娜， 李 川， 江周峰
(北京邮电大学 计算机学院，北京 100876)

摘要: 提出了一种基于时间上下文的协同过滤推荐(TCCF-LI)算法，实现了基于高校图书馆图书借阅记录数据上的学生学习兴趣挖掘. 在传统协同过滤算法上引入时间上下文信息，既考虑了大尺度用户群体爱好的趋同性，又兼顾了小尺度个体用户爱好的短时相关性，获得了更高的推荐性能. 在实际数据集上的实验结果表明，该算法在推荐精准度、召回率等方面比传统推荐算法有较好表现.

关 键 词: 推荐系统; 协同过滤; 时间上下文; 学习兴趣挖掘
中图分类号: TP311 **文献标志码:** A

A Collaborative Filtering Recommendation Algorithm with Time Context for Learning Interest Mining

E Hai-hong， SONG Mei-na， LI Chuan， JIANG Zhou-feng
(School of Computer Science, Beijing University of Posts and Telecommunications, Beijing 100876, China)

Abstract: A time context based collaborative filtering recommendation (TCCF-LI) algorithm this paper was proposed, and students' learning interest mining from university library borrow record was implemented. The time context information was imported into the traditional collaborative filtering recommendation algorithm, in which, both interest homoplasy of large scale user groups and short-term correlation of small scale user groups was considered. Good recommendation performance was gained. According to the experiments on real dataset, TCCF-LI algorithm presents higher precision and recall rate compared with traditional recommendation algorithm.

Key words: recommender system; collaborative filtering; time context; learning interest mining

基于协同过滤(CF, collaborative filtering)上下文感知推荐生成技术[1]是基于“集体智慧”的思想，将上下文信息融入到基于用户相似性、项目相似性和基于模型的 CF 中. 已有的基于内容上下文感知推荐生成技术的主要研究思路是，将上下文信息融入基于内容的推荐方法，着重考虑用户偏好、上下文与项目属性的匹配度. 基于社交关系数据来提高推荐系统的性能方面，近年来研究的较多. 在文献[2-5]中，研究者加入用户间的社交网络上下文信息; 文献[6-7]基于用户签到位置轨迹与社交活动数据进行融合推荐; 文献[8-9]将用户位置上下文信息与用户位置签到行为上下文相结合，实现基于位置上下文的 CF.

基于时间上下文的协同过滤推荐(TCCF-LI, time context based collaborative filtering recommendation)算法是融合借阅时间作为推荐系统的另一个

收稿日期: 2014-04-25
基金项目: 高等学校博士学科点专项科研基金项目(20110005120007); 北京高等学校青年英才计划项目(YETP0445); 北京市教育委员会共建项目专项资助项目
作者简介: 鄂海红(1982—)，女，讲师，E-mail: ehaihong@bupt.edu.cn.

上下文维度,进一步地研究了在不同的时间上下文环境下,推荐系统的精度以及在算法上的改进.

1 学习兴趣挖掘的推荐算法建模

1.1 实验数据集分析及评分模型构建

实验采用北京邮电大学图书馆提供的借阅记录作为数据集,具体情况如表1和表2所示.

表1 图书借阅记录原始数据集情况

数据项	原始数据集	测试数据集
借阅记录总数/条	1 773 268	529 367
其中,借书记录数/条	743 304	
还书记录数/条	743 247	
续借记录数/条	241 945	
馆藏图书种数总计/册	404 624	46 945
被借阅过的图书总计/册	101 543	46 945
学生读者数/名	46 854	16 922
有借阅行为的学生读者数/名	28 951	16 922

表2 图书借阅记录原始数据集数据分布情况

按借书次数统计的读者人数分布			按被借次数统计的图书册数分布		
借书次数	原始集	测试集	被借次数	原始集	测试集
1~10	8 778	7 529	1~10	67 324	32 043
11~50	10 419	6 036	11~50	27 682	14 034
51~100	5 220	2 140	51~100	5 134	760
101~200	4 352	1 015	101~500	1 572	108****
201~500		202**			
501~1 000	47	无	501~1 000	22***	无
>1 000	1*	无			

注: * 原始集中最高借阅次数的读者累计借阅1 381次; **测试集中最高借阅次数的读者累计借阅471次; ***原始集中图书被借次数最高的图书累计被借阅930次; ****测试集中图书被借次数最高的图书累计被借阅402次.

因为原始数据集图书(item)和读者(user)的数量级较高,不适合在算法测试和验证阶段进行全集数据的运算. 因此在数据预处理阶段,将借阅记录为0和1的用户过滤出去,并将借阅次数高于500的用户和图书过滤去除. 因为这其中包含了部分系统非有效数据(图书借阅系统自身测试数据等噪声数据).

构建推荐系统必须建立用户模型(user model)保存用户的偏好. 这里用 $U=\{u_1,2,\cdots,u_n\}$ 代表学生借阅用户集,用 $D=\{d_1,2,\cdots,d_m\}$ 代表图书集. 初始读者用户对图书的评分用借阅记录表示: 一次借阅表示用户对这本书的喜欢(或者评分)为1,续借或非连续地借阅2次则评分为2,并逐次累加.

通过借阅记录构建用户对图书的评分矩阵 $\boldsymbol{R}$. $\boldsymbol{R}$ 代表评分项 $r_{i,j}$ 的 $n\times m$ 评分矩阵,这里 $i\in\{1,2,\cdots,n\}$, $j\in\{1,2,\cdots,m\}$. 如果用户 i 没有借阅过图书 j,视为对该图书没有评分,则对应的矩阵项 $r_{i,j}$ 为空.

为了减小较高借阅次数对相似性计算的误差,本实验中将用户的借阅次数通过式(1)转化为(0,5)的评分.

$$r_{\text{rate}}=a-a\mathrm{e}^{-\frac{c_{\text{count}}}{b}} \tag{1}$$

其中: r_{rate} 为最后的得分, c_{count} 为书籍的借阅次数, a 和 b 分别为调整系数. 为了将得分分布在(0,5)之间, a 赋值为5. 通过调整 b 的值来适应数据的分布规律,实验中 b 赋值为4.5. 通过式(1)可以将分布(1,402)的评分映射到(0,5)之间.

2 TCCF-LI 算法设计

引入时间上下文信息在基于借阅记录的学生兴趣挖掘上有重要的意义. 学生的学习兴趣与其所在学年和学期课程内容,以及领域技术更新的情况具有较高的时间效应. 读者用户在相隔很短的时间内感兴趣的书籍具有更高相似度,也可以建模为读者在相隔很短的时间内喜欢的图书具有更高相似度. 因此,在预测阶段将读者近期借阅行为相比用户很久之前的行为加重了权值.

Item-based CF 算法,设 U 为所有图书 i 和 j 评分的读者集,则图书 i 和 j 之间的相似度 $\text{sim}(i,j)$ 为

$$\text{sim}(i,j)=\frac{\sum_{u\in N(i)\cap N(j)}1}{\sqrt{|N(i)||N(j)|}} \tag{2}$$

其中: $N(i)$ 为借阅图书 i 的读者集合, $N(j)$ 为借阅图书 j 的读者集合. 因此,式(2)的分子部分表示同时借阅图书 i 和图书 j 的读者用户数; 分母部分中 $|N(i)|$ 表示借阅图书 i 的用户数, $|N(j)|$ 表示借阅图书 j 的用户数. 而在给读者 u 做推荐时,读者 u 对图书 i 的兴趣 $p(u,i)$ 为

$$p(u,i)=\sum_{j\in N(u)\cap S(i,k)}\text{sim}(j,i)\,r_{uj} \tag{3}$$

其中: $N(u)$ 为读者 u 借阅图书的集合, $S(i,K)$ 为和图书 i 最相似的 K 个图书集合, r_{uj} 为读者 u 对图书 j 的评分.

引入时间衰减项 $f(|t_{ui}-t_{uj}|)$,其中 t_{ui} 为读者 u 借阅图书 i 发生的时间. f 函数的含义是,读者对图

书 i 和图书 j 借阅行为发生的时间越远,则 $f(|t_{ui}-t_{uj}|)$ 越小. 实验中引入衰减函数为

$$f(|t_{ui}-t_{uj}|)=\sum_{(u,i,t\in G,t<T)} N\mathrm{e}^{-\mu(t_{ui}-t_{uj})} \tag{4}$$

其中: t_{ui} 及 t_{uj} 取值为借阅记录日期; N 和 μ 为时间衰减参数,它们的取值在不同系统中不同; G 为实验仿真中的测试集,限定了 i、j、t 的取值空间. N 一般取值为 1. 如果一个系统用户兴趣变化很快,就应该取比较大的 μ,反之需要取比较小的 μ,在实验 3 中对 μ 取值进行了测试和讨论.

在得到时间信息(用户对物品产生行为的时间)后,可以通过式(5)改进相似度的计算.

$$\mathrm{sim}(i,j)=\frac{\sum_{u\in N(i)\cap N(j)} f(|t_{ui}-t_{uj}|)}{\sqrt{|N(i)||N(j)|}} \tag{5}$$

3　实验

3.1　仿真场景建立

下面将通过仿真来验证采用不同相似性计算方法的 User-based CF,以及采用余弦相似性的 User-based CF 和 TCCF-LI 算法的性能. 这里将实验数据集的数据划分为训练集(training set)和测试集(test set). 训练集和测试集的划分是 9∶1,按照借阅时间排序取最早借阅记录的 90% 作为训练集,余下的为测试集.

3.2　算法评价指标

推荐预测是为学生读者推荐感兴趣的图书,也就是经典的 Top-N 问题,评价这类问题可以通过精准度 $P_{\mathrm{precision}}$ 和召回率 R_{recall} 来衡量推荐的准确度,即

$$P_{\mathrm{precision}}=\frac{\sum_u |B_u\cap A_u|}{\sum_u |B_u|} \tag{6}$$

$$R_{\mathrm{recall}}=\frac{\sum_u |B_u\cap A_u|}{\sum_u |A_u|} \tag{7}$$

其中: B_u 为给读者 u 的推荐列表集合, A_u 为在测试集中读者 u 感兴趣的图书集合.

为了同时考查精准度和召回率,实验中还使用了 F 指标, F 指标定义为

$$F=\frac{2P_{\mathrm{precision}}R_{\mathrm{recall}}}{P_{\mathrm{precision}}+R_{\mathrm{recall}}} \tag{8}$$

通过 F 指标可以选择合适的 Top-N 的推荐长度,以兼顾精准度和召回率的表现.

3.3　仿真结果与分析

笔者将所提出的 TCCF-LI 算法与已有的 User-based CF 进行相关的对比.

实验 1　本实验应用场景中,CF 算法中不同相似性计算方法的性能表现.

实验中将续借的记录作为用户对书借阅的权值,来实现带权余弦的计算,并在不同的推荐列表长度的情况下,对 3 种算法进行比较,图 1、图 2 反映了实验结果. 其中,图 1 显示了 3 种算法在推荐列表 K 长度依次为 1、3、7、10 时的精准度、召回率对比,得出以下结论.

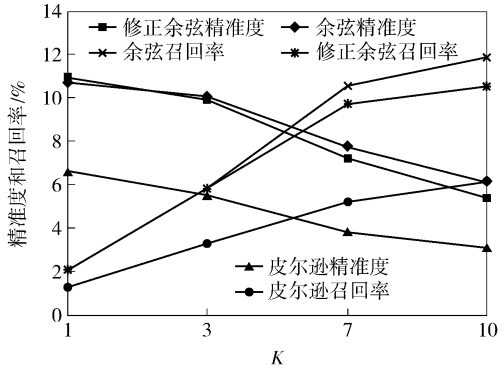

图 1　推荐长度 K 取 1 ~ 10 下的精准度和召回率

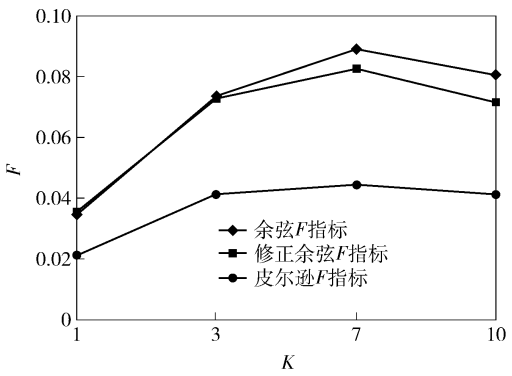

图 2　推荐长度 K 取 1 ~ 10 下的 F 指标

1) 从整体上看,推荐的精准度随着推荐列表的增加而降低,召回率正好相反. 其中,余弦算法的整体效果要好于修正余弦,而这 2 种算法的推荐效果要明显优于皮尔逊算法. 余弦算法的推荐精准度最高达到了 10.72%,最低也有 6.1%.

2) 召回率中,表现最好的是余弦算法为 11.87%,最差的是皮尔逊算法.

3) 一般推荐列表长度越长,召回率越高,精准度很有可能会随之下降. 而图 2 综合考虑了召回率与精准度. 对于表现最好的余弦算法,精准度和召回率相交在 5 ~ 6 之间. 实验仿真验证了北邮读者用户通常平均每次借阅书籍 5 ~ 6 册.

实验2 对比TCCF-LI算法与实验1中选出的性能最好的CF的性能表现.

TCCF-LI算法中,时间上下文是以用户借阅图书行为发生时间为参数的. 因为在数据集中,同一个用户对同一本书产生了多次借阅,取值为用户最后一次借阅该图书的时间. 实验中计算时间 $t_{ui}-t_{uj}$ 及 t_0-t_{uj} 时,采用用户借阅时间的时间戳来计算. 差值再按照2个时间的时间戳的毫秒差折算为天数来计算,从而获得的 $t_{ui}-t_{uj}$ 及 t_0-t_{uj} 的天数差值的取值空间是(0,548)(借阅记录数据集的时间范围). 时间衰减参数 μ 的取值在实验3中测试,这里取其性能表现最优的值. 如图3和图4所示,可以得到以下结论.

1) TCCF-LI与传统单一模型的CF算法相比,在精准度和召回率方面都具有较好的表现,在精准度上最好可以是12.8%,召回率最好可以是15.5%.

2) 图4中TCCF-LI在推荐长度 K 取7~10时,F 指标都有较好的表现,且明显优于单一模型CF算法.

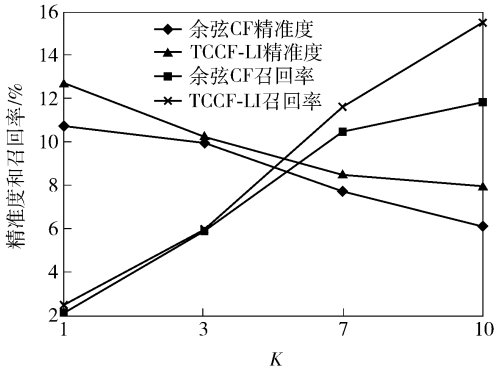

图3 TCCF-LI与余弦CF精准度和召回率比较

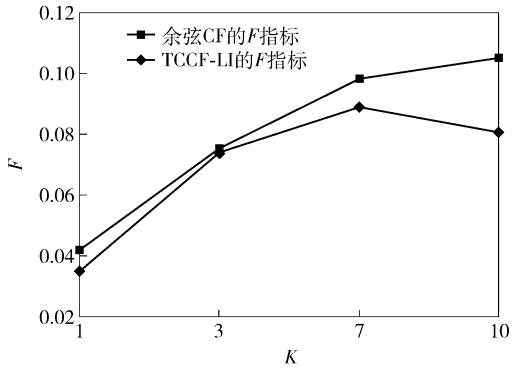

图4 TCCF-LI与余弦CF的 F 指标比较

实验3 TCCF-LI算法中时间衰减因子 μ 对预测结果的影响.

实验中通过调整时间衰减因子 μ 以及推荐长度 K 来对比结果. 由表3和表4,可以得到以下结论.

1) 在 $K=5$ 且衰减因子 μ 初始逐步调低时,精准度和召回率都在逐步提升;在 μ 取值为0.005达到波峰时,$P_{\text{precision}}$ 为10.28%,R_{recall} 为6.01%. 因此,在本实验应用场景中 μ 取值为0.005.

2) 在 μ 取值为0.005时,可以发现在 K 取15时,F 指标最好,为10.63%. 此时,推荐系统的精准度可以达到20.8%,召回率为7.14%,实验性能有较大提升.

表3 不同时间衰减因子 μ 取值下的精准度和召回率

$K=5$	$P_{\text{precision}}$	R_{recall}
$\mu=0.5$	0.063 7	0.061 9
$\mu=0.05$	0.075 6	0.073 6
$\mu=0.001$	0.091 0	0.088 5
$\mu=0.005$	0.102 8	0.060 1
$\mu=0.000\,5$	0.090 2	0.087 7

表4 μ 取值0.005时不同 K 的精准度和召回率

$\mu=0.005$	R_{recall}	$P_{\text{precision}}$	F 指标
$K=1$	0.127 6	0.024 9	0.041 7
$K=3$	0.102 8	0.060 1	0.075 9
$K=5$	0.091 6	0.089 1	0.090 3
$K=7$	0.085 5	0.116	0.098 4
$K=9$	0.081 1	0.142	0.103 2
$K=10$	0.079 5	0.155	0.105 1
$K=15$	0.071 4	0.208	0.106 3
$K=20$	0.064 7	0.252	0.103 0

4 结束语

笔者针对高校学生个性化培养平台建设的目标,研究并提出了一种TCCF-LI算法. 该算法将时间上下文信息引入协同过滤方法中以获得更好的推荐精确度,由此提出了基于一种新的时间衰减函数的时间上下文CF算法. 通过详细实验测试,①验证了CF算法在所提应用场景实践中适合的相似性计算方法为余弦相似性;②验证了所提出的TCCF-LI相比传统CF算法具有更好的性能表现;③通过实验给出了所提出的TCCF-LI算法中时间衰减因子的合理取值.

参考文献：

[1] Chen Annie. Context-aware collaborative filtering system: predicting the user's preference in the ubiquitous computing environment [J]. Lecture Notes in Computer Science, 2005, 3479: 244-253.

[2] Liu Fengkun, Lee H J. Use of social network information to enhance collaborative filtering performance [J]. Expert Systems with Applications, 2010, 37(7): 4772-4778.

[3] Yang Shuanghong, Long Bo, Smola A, et al. Like like alike: joint friendship and interest propagation in social networks [C] // Proceedings of the 20th International Conference on World Wide Web. Bangalore India: ACM, 2011: 537-546.

[4] Hasan S, Zhan Xianyuan, Ukkusuri S V. Understanding urban human activity and mobility patterns using large-scale location-based data from online social media [C] // Proceedings of the 2nd ACM SIGKDD International Workshop on Urban Computing. Chicago USA: ACM, 2013: 1-8.

[5] 胡祥，王文东，龚向阳，等. 基于流形排序的社会化推荐方法 [J]. 北京邮电大学学报，2014，37(3)：18-22.
Hu Xiang, Wang Wendong, Gong Xiangyang, et al. Social recommendation based on manifold ranking [J]. Journal of Beijing University of Posts and Telecommunications, 2014, 37(3): 18-22.

[6] 孙甲申，王小捷. 一种用于社会化标签推荐的主题模型 [J]. 北京邮电大学学报，2014，37(3)：38-42.
Sun Jiashen, Wang Xiaojie. A topic model for social tag recommendation [J]. Journal of Beijing University of Posts and Telecommunications, 2014, 37(3): 38-42.

[7] Ying Josh Jia-Ching, Lee W C, Ye Mao. User association analysis of locales on location based social networks [C] // Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Location Based Social Networks. Chicago USA: ACM, 2011: 69-76.

[8] Yuan Quan, Cong Gao, Ma Zongyang. Time-aware point-of-interest recommendation [C] // Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval. Dublin Ireland: ACM, 2013: 363-372.

[9] Zheng Ning, Jin Xiaoming, Li Lianghao. Cross-region collaborative filtering for new point-of-interest recommendation [C] // Proceedings of the 22nd International Conference on World Wide Web Companion. Seoul Korea: [s. n.], 2013: 45-46.