云南师范大学学报(自然科学版)
2017年7月37卷4期(Vol.37 No.4)

Journal of Yunnan Normal University
DOI:10.7699/j.ynnu.ns-2017-051

基于用户兴趣变化的隐语义协同过滤算法*

汪佩， 梁立， 甘健侯

(云南师范大学 信息学院,民族教育信息化教育部重点实验室,云南 昆明 650500)

摘 要： 协同过滤是推荐系统中广泛使用的算法.协同过滤模型没有考虑用户兴趣的动态变化,影响推荐质量.为提高推荐准确度,提出新的推荐算法——将基于动态时间窗口的协同过滤推荐与高斯概率隐语义模型结合,利用动态时间窗口捕捉用户的兴趣变化,并结合概率隐语义模型分析用户长期兴趣,进而为用户提供更准确的推荐.实验表明,该算法同其他协同过滤算法相比具有更高的准确度.

关键词： 协同过滤;兴趣变化;动态时间窗口;概率隐语义分析

中图分类号： TP301.6 **文献标志码：** A **文章编号：** 1007-9793(2017)04-0039-05

1 引 言

Web 2.0时代,信息量呈爆炸式增长,信息过载问题日益严重.推荐系统作为解决上述问题的工具变得越来越重要.

推荐系统采用的算法主要包括基于内容的过滤算法和协同过滤算法两类[1].基于内容的过滤算法是根据用户已有评分项目提取用户兴趣,同时对目标项目进行特征提取,计算两者之间的相似度完成推荐[2].协同过滤算法包括基于记忆算法和基于模型算法.基于记忆算法主要利用用户的历史行为,发现兴趣相似的用户群、项目集;根据之间的相似性,向用户推荐其他用户感兴趣的项目,或已评分项目中相似度大的项目;基于记忆算法存在稀疏性、冷启动等问题.基于模型的算法通过对用户评分数据和隐含的偏好建立用户的行为模型,根据模型进行预测.矩阵分解[3]推荐算法将 n 个用户、m 个项目的评分矩阵进行奇异值分解,得到 $n\times k$ 和 $k\times m$ 两个矩阵,分别表示用户和隐含主题,以及隐含主题和项目之间的关系,以此对项目进行评分预测.概率隐语义模型(PLSA)[4]方法也是通过提取隐含主题实现对用户偏好的建模,该算法具有较高的准确率.上述算法都只考虑用户历史数据之间的相关性,并没有考虑时间因素,然而在实际生活中,用户的兴趣是随时间动态变化的.

文献[5-6]借鉴心理学中艾宾浩斯曲线跟踪用户的兴趣变化,并利用数学分析工具,提出了基于指数函数的兴趣遗忘函数,解决了用户兴趣变化的难题,提高了推荐准确性.由于PLSA算法通过用户、项目和主题之间的相关性来计算,因此利用兴趣遗忘函数处理兴趣变化并不适合.文献[7-8]提出了基于评分时间窗口的改进算法,利用固定的时间窗口捕获用户短期兴趣,并结合PLSA算法,提升推荐准确度.然而,每个用户兴趣变化快慢不一,固定的时间窗口不能准确衡量每个用户的短期兴趣.因此,提出基于动态时间窗口的改进算法,根据每个用户的访问频率动态调整时间窗口,以准确表示用户兴趣的偏

* **收稿日期:**2017-06-15

基金项目:国家自然科学基金资助项目(61562093);云南省应用基础研究计划重点资助项目(2016FA024).

作者简介:汪 佩(1990-),男,安徽砀山人,硕士研究生,主要从事推荐系统方面研究.

通信作者:梁 立.E-mail:liangli@ynnu.edu.cn.

移,并结合高斯概率隐语义模型(PLSA),利用 PLSA 获取用户长期兴趣,对两者线性加权.实验表明,该算法在推荐准确度方面有明显提高.

2 基于动态时间窗口的隐语义模型

2.1 概率隐语义模型(PLSA)

自从 Netflix Prize 比赛举办以来,LFM(Latent factor model)隐语义模型逐渐成为推荐系统中耳熟能详的名词[9],核心思想是通过隐含的特征联系用户和项目,广泛用于文本分类.Hofmann 首先使用隐语义模型进行个性化推荐.在 PLSA 推荐模型中定义隐含特征 $Z=\{z_1,z_2,\cdots,z_k\}$,该模型认为用户对项目的兴趣度是由隐含特征决定的.隐含特征可以理解为项目的潜在分类,或者是不同兴趣用户的群组划分.在有 n 个用户、m 个项目、k 个隐含特征的推荐系统中,定义用户 u 对项目 w 评分为 v 的概率为:

$$P(v\mid u,w)=\sum_{z\in Z}p(z\mid u)\,p(v\mid z,w) \tag{1}$$

概率 $P(z\mid u)$ 、$P(v\mid z,w)$ 分别表示用户 u 与隐含特征 z 、隐含特征 z 对项目 w 评分为 v 的概率.假设 $P(v\mid z,w)$ 服从标准正态分布,因此通过正态函数进行初始赋值.利用 EM 算法求解相关参数,Hofmann[4]给出了详细的推导过程,简单描述如下.

(ⅰ)定义损失函数

$$\begin{aligned}R(\theta)&=-\frac{1}{N}\sum\nolimits_{\langle u,z,w\rangle}\log P(v\mid u,w)\\&=-\frac{1}{N}\sum_{\langle u,z,w,v\rangle}[\log p(z\mid u)]+\log p(v\mid z,w)\end{aligned}$$

(ⅱ)E-Step

$$p(z\mid u,v,w)=\frac{p(z\mid u)\,p(v\mid z,w)}{\sum_{z'\in Z}p(z'\mid u)\,p(v\mid z',w)}$$

(ⅲ)M-Step

$$p(z\mid u)=\frac{\sum_{<u',v,w>;u'=u}P(z\mid u,v,w)}{\sum_{z'\in Z}\sum_{<u',w>;u'=u}P(z'\mid u,v,w)}$$

$$\mu_{w,z}=\frac{\sum_{<u,v,w'>;w'=w}v\cdot P(z\mid u,v,w)}{\sum_{<u,v,w'>;w'=w}P(z\mid u,v,w)}$$

$$\sigma_{w,z}=\frac{\sum_{<u,v,w'>;w'=w}(v-\mu_{w,z})^2\cdot p(z\mid u,v,w)}{\sum_{<u,v,w'>;w'=w}p(z\mid u,v,w)}$$

利用随机函数初始化 $P(z\mid u)$ 和 $P(v\mid z,w)$,循环执行 E-step 和 M-step,直到损失函数增加或达到迭代次数为止.用户 u 对项目 w 的预测评分即是评分概率 $P(v\mid z,w)$ 的期望:

$$\hat{r}_{u,w}=E[P(v\mid z,w)]=\int v\cdot P(v\mid z,w)\,\mathrm{d}v=\sum\nolimits_z p(z\mid u)\,\mu_{z,w}$$

2.2 动态时间窗口

由于 PLSA 算法是对用户历史行为进行分析,提取隐含特征的方式没有考虑到时间因素;而现实中近期的行为更能体现用户的兴趣变化和需求,现有解决兴趣变化的方法是利用时间衰减函数处理,增加近期行为的权重,从而提高推荐质量.但是,该方法不适用于 PLSA 算法[8],因为 PLSA 预测是根据用

户所属各个社区的平均分计算.因此，提出时间窗口，根据用户兴趣变化的频率给每个用户动态的设置时间窗口 Windows，时间窗口的大小反映出用户受近期行为影响度的不同.

表 1　用户－项目－时间矩阵

Table 1　User-item-time matrix

用户	w_1	w_2	…	w_m
u_1	23	4	…	76
u_2	70	120	…	0
…	…	…	…	…
u_n	0	20	…	0

表 1 中每行代表一位用户，每列代表用户评分项目，矩阵中的元素代表用户评分时间，如果为 0 表示没有进行评分.将用户已经评分项目按时间顺序降序排列，根据用户访问频率设置时间窗口大小.定义用户时间窗口函数为：

$$win_i = Windows \times \frac{|w_i|}{\frac{1}{n}\sum_{i=1}^{n}|w_i|} \tag{2}$$

$Windows$ 为设定的滑动窗口数目，$\{w_i\}$ 表示用户已评分项目集合，$|w_i|$ 为已评分项目集合的数目，公式(2) 根据当前用户的已评分项目数与全局平均的不同来实现窗口的动态调整.结合时间窗口预测用户 u 对项目 w 的评分为：

$$\hat{r}_{u,w} = \alpha \cdot r_1(u,w) + (1-\alpha) \cdot r_2(u,w) \tag{3}$$

式中 $r_1(u,y)$ 为 PLSA 算法的预测评分，而 $r_2(u,y)$ 表示用户窗口中的项目预测出的短期兴趣评分.将其线性加权求得最终预测评分，参数 α 动态调整两者所占评分的权重值，以应对不同数据集.短期兴趣评分公式如下：

$$r_2(u,w) = \bar{r}_u + \frac{\sum_{x \in windows(u)} sim(x,w) \cdot (r_{u,x} - \bar{r}_x)}{\sum_{x \in windows(u)} sim(x,w)} \tag{4}$$

$\bar{r}_u$ 为用户 u 的平均评分，$\bar{r}_x$ 为项目 x 的平均评分，$windows(u)$ 为用户 u 动态时间窗口内的评分项目，$sim(x,w)$ 表示项目之间的相似性大小.公式(4)反映了短期兴趣对预测项目的评分值.

2.3　改进算法描述

输入：用户－项目－评分矩阵，用户－项目－时间矩阵

输出：用户的推荐列表

Step1：由期望极大化(EM)算法，训练 PLSA 模型，求出 $p(z \mid u)$，$\mu_{z,w}$.

Step2：计算各个项目之间的相似度，并根据公式(2)求出用户的动态时间窗口函数.

Step3：利用 PLSA 模型求出长期兴趣对项目的预测评分值，根据公式(4)预测短期兴趣对项目的预测评分值.

Step4：生成推荐列表，由公式(3)求出最终的预测评分，形成基于 Top-N 的推荐列表.

参数 α 以及 $Windows$ 窗口数通过推荐模型交叉验证离线求得.

3 实验结果及分析

实验数据来自 MovieLens 数据集.该实验数据集包含了 943 名用户对 1682 部电影共 10 万条记录,每条记录包括评分值和评分时间.根据二八法则,本次实验将数据的 80%作为训练数据集,剩下 20%作为测试数据集.

3.1 评测指标

评分预测的准确度是评价推荐系统的重要指标,一般通过平均绝对差(MAE)(即用户真实评分)与预测评分之间偏差的绝对值来衡量.MAE 值越小,说明推荐算法效果越好.令 T 为测试数据用户集合,r_{ui} 为用户 u 对项目 i 的实际评分,$\hat{r}_{ui}$ 为用户 u 对项目 i 的预测评分.MAE 计算公式如下:

$$MAE=\sqrt{\frac{\sum_{u,i\in T}(r_{ui}-\hat{r}_{ui})^2}{|T|}}$$

3.2 实验结果

针对 MovieLens 的数据集,通过多次实验表明隐含特征个数取 14 时能够得到较高的准确率,在接下来的实验中隐含特征个数统一设置为 14.因此在确定评分窗口大小为 30 情况下,对 α 参数进行考察,实验结果如图 1 所示.

根据图 1 可知,当 α 取值为 0.7 时算法效果最好;α 取值越大,算法越接近 PLSA 算法;当 α 取值为零时,预测由窗口中的数据决定,算法退化为基于项目的协同过滤.考虑时间窗口对算法性能的影响,令 $\alpha=0.7$(此时算法最优),时间窗口数大小取(5—60),改变窗口大小得到实验结果如图 2 所示.

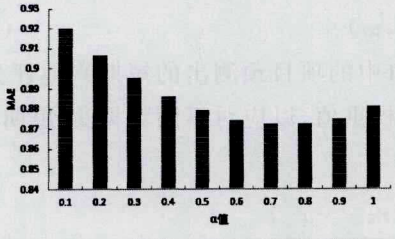

图 1 参数 α 对 MAE 的影响

Fig.1 The effect of parameter α in terms of MAE

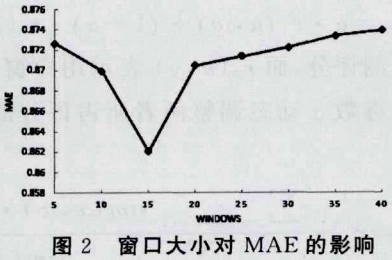

图 2 窗口大小对 MAE 的影响

Fig.2 The effect of windows in terms of MAE

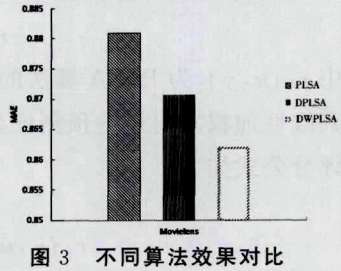

图 3 不同算法效果对比

Fig.3 Performance comparison with different algorithms

当窗口数过小时,能够反映短期兴趣的信息较少,计算误差较大;随着窗口数的增加,该模型能够正确反映用户的兴趣变化,效果较理想;而窗口过大时,由于长期兴趣与短期兴趣不好区分,因此模型质量下降.选择一个合适的窗口对模型至关重要,通过多次试验,在 MovieLens 数据集上,窗口数取值为 15 时,模型的效果最佳.

为了检测本文模型的推荐质量,取参数 α 为 0.7、窗口数为 15 同 PLSA 算法以及文献[8]作比较,实验结果如图 3 所示.由图 3 可知,本文提出的算法无论是对于单一的 PLSA 算法,还是文献[8]中基于时间窗口的算法,都在推荐准确度上得到明显提高.

4 结 论

针对传统的协同过滤算法推荐准确率低,不能反映用户兴趣变化问题.本文提出了基于用户的动态时间窗口机制,构建了用户的短期兴趣偏好,并融合隐语义模型,明显提高了推荐的准确度.对于近期时

间窗口的分析是以项目为基础的，然而实际中，相邻项目之间存在极大的相似性，进一步的研究将会考虑项目的时间划分尺度，实现对用户兴趣更加精准的建模.

参　考　文　献：

[1] 刘建国，周涛，汪秉宏.个性化推荐系统的研究进展[J].自然科学进展，2009，19(1)：1-15.
[2] 曹毅.基于内容和协同过滤的混合模式推荐技术研究[D].长沙：中南大学，2007.
[3] KOREN Y，BELL R，VOLINSKY C.Matrix factorization techniques for recommender systems[J].IEEE Computer Society，2009，42(8)：30-37.
[4] HOFMANN T，PUZIEHA J.Latent class models for collaborative filtering[C].Proceedings of the Sixteen International Joint Conference on Artifical Intelligence(IJCAI99)，Stockholm，Sweden，1999.
[5] 于洪，李转运.基于遗忘曲线的协同过滤推荐算法[J].南京大学学报：自然科学版，2010，46(5)：520-527.
[6] 李克潮，梁正友.适应用户兴趣变化的指数遗忘协同过滤算法[J].计算机工程与应用，2011，47(13)：154-156.
[7] 孙克雷，陈安东.基于用户兴趣的个性化推荐算法研究[J].安徽建筑大学学报，2017，25(1)：65-69.
[8] 吴成超，王卫平.考虑用户兴趣变化的概率隐语意协同推荐算法[J].计算机系统应用，2014，23(5)：162-166.
[9] 项亮.推荐系统实践[M].北京：人民邮电出版社，2012.

PLSA Collaborative Filtering Algorithm Based on User Interest Change

WANG Pei，LIANG Li，GAN Jian-hou

(College of Information，Key Laboratory of Education Informalization for Nationalities of Ministry of Education，Yunnan Normal University，Kunming 650500，China)

Abstract：Collaborative Filtering(CF) is widely used algorithm in recommender system. Collaborative Filtering model does not consider the dynamic change of user's interests，influence recommendation quality. In order to improve recommendation precision，a new recommender algorithm was proposed，which combines the Collaborative Filtering based on dynamic time windows that capture change of user's interest by dynamic time windows，and the Gaussian probabilistic latent semantic analysis(PLSA) that capture user's long-term interest together. The experimental results on Movielens dataset show that the new algorithm compares favorably with other collaborative filtering algorithm.

Keywords：Collaborative filtering；Interest change；Dynamic time window；PLSA