文章编号:1001 - 9081(2007)09 - 2302 - 02

基于时间加权的协同过滤算法

王 岚[1,2],翟正军[2]
(1. 洛阳师范学院 计算机科学系,河南 洛阳 471022;
2. 西北工业大学 计算机学院, 西安 710072)
(lanlanluoyang@163.com)

摘 要:协同过滤是个性化推荐系统中采用最广泛的推荐技术,但已有的方法是将用户不同时间的兴趣等同考虑,时效性不足。针对此问题,提出了一种改进的协同过滤算法,使得越接近采集时间的点击兴趣,在推荐过程中具有更大的权值,从而提高了推荐的准确性。

关键词:协同过滤;个性化推荐;邻居用户;时间权值
中图分类号: TP311 **文献标志码**:A

Collaborative filtering algorithm based on time weight

WANG Lan[1,2], ZHAI Zheng-jun[2]
(1. *Department of Computer Science, Luoyang Normal University, Luoyang Henan* 471022, *China*;
2. *School of Computer Science, Northwestern Polytechnical University, Xi' an Shaanxi* 710072, *China*)

Abstract: Collaborative filtering is the most widely used recommendation technology in the personalized recommendation system. However, the user's interests in different time have been taken into equal consideration with the method being used, which leads to the lack of effectiveness in the given period of time. In view of this problem, this paper presented an improved collaborative filtering algorithm to make the click interests, approaching the gathering time, have bigger weight in the recommendation process, thereby to improve the accuracy of the recommendation.

Key words: collaborative filtering; individual recommendation; neighbor user; time weight

0 引言

现实生活中用户的兴趣是会经常改变的,用户希望网站推荐的是自己感兴趣的新颖的资源。也就是说用户访问浏览日志中的项目总在不断更新,减少或增加项目会导致日志数据发生变化,从而产生新的感兴趣的资源,导致兴趣度的变化,这样推荐信息也应发生新的变化。但是现有的算法并没有考虑到这个问题,这就使得某一时段的高兴趣度在整个类集中显得并不突出,发现的很可能是高兴趣度的过时信息,以致造成推荐结果并不是最新出现的感兴趣的项目。这显然不符合 Web 挖掘的目的,也就是说我们不能发现最新出现的新颖信息,由此也就产生了新项目的敏感性问题。

本文就是在传统协同过滤算法的基础上,提出了基于用户兴趣的新颖资源的协同过滤推荐算法。并利用实验数据集对本文算法与传统协同过滤算法的性能进行了比较。

1 相关工作

最近邻协同过滤推荐算法是目前较成功的推荐技术。其基本思想是基于评分相似的最近邻居的评分数据向目标用户产生推荐[1]。协同过滤推荐算法的实现过程可以分为 3 步[2],即输入数据表示、邻居形成和推荐生成。

第 1 步:输入数据表示

数据表示主要完成浏览数据的描述,通常表述为一个 $m \times n$ 的用户—项评价矩阵 $\boldsymbol{R}$,m 表明用户数,n 表明了项数,R_{ij} 是第 i 个用户对第 j 项的兴趣度,就本文来讲,R_{ij} 是用户对网页的兴趣度,表明用户是否浏览了该网页以及用户对该网页的喜好程度。

第 2 步:邻居的形成

邻居形成主要完成目标用户最近邻居(或最相似用户)的识别[3]。协同过滤需要分析用户之间的相似性,形成当前目标用户的邻居集,从而根据"邻居"的信息进行推荐。协同过滤在推荐系统中实现的核心就是为一个需要推荐服务的当前目标用户寻找其最相似的"最近邻居"集,即:对一个用户 Tu,要产生一个根据相似度大小排列的"邻居"集合 $Neighbor_{Tu} = \{N_1, N_2, \cdots, N_n\}$,$Tu \notin Neighbor_{Tu}$,从 N_1 到 N_n,表示用户 Tu 与邻居用户的相似性大小的值 $\mathrm{Sim}(Tu, N_t)$,并有 $\mathrm{Sim}(Tu, N_1) > \mathrm{Sim}(Tu, N_2) > \cdots > \mathrm{Sim}(Tu, N_n)$,从 N_1 到 N_n,表示用户 Tu 与邻居用户的相似性大小的值 $\mathrm{Sim}(Tu, N_i)$ 从大到小排列。

第 3 步:产生推荐

设目标用户 Tu 的最近邻居集合用 $Neighbor_{Tu}$ 表示,则目标用户 Tu 对项目 i 的预测评分可以通过邻居用户 n 对项目 i 的评分得到,计算方法为[4]:

$$P_{Tu,i} = \overline{R_{Tu}} + \frac{\sum_{n \in Neighbor_{Tu}} \mathrm{Sim}(Tu, n) \times (R_{n,i} - \overline{R_n})}{\sum_{n \in Neighbor_{Tu}} (|\ \mathrm{Sim}(Tu, n)\ |)} \quad (1)$$

其中:$\mathrm{Sim}(Tu, n)$ 表示目标用户 Tu 与最近邻居用户 n 的相似性,$R_{n,i}$ 表示用户 n 对资源 i 的兴趣度(n 是"最近邻居"集中的

收稿日期:2007 - 03 - 16;**修回日期**:2007 - 06 - 06。
基金项目:河南省高校杰出科研人才创新基金资助项目(2006KYCX004);河南省青年骨干教师基金资助项目(134)。
作者简介:王岚(1967 -),女,河南洛阳人,副教授,硕士研究生,主要研究方向:数据挖掘、电子商务; 翟正军(1965 -),男,河南洛阳人,教授,主要研究方向:多媒体、智能信息处理、数据挖掘。

用户)。$\overline{R_{Tu}}$ 和 $\overline{R_n}$ 分别表示用户 Tu 和用户 n 对资源的平均兴趣度。

通过上述方法预测当前目标用户对所有未浏览资源的兴趣度,然后选择预测兴趣度最高的若干项推荐给目标用户[5]。

2　时间加权协同过滤算法

传统的协同过滤方法不能将兴趣的变化表现出来。例如,用户的兴趣是经常变化的:一位家长可能对各大学的介绍、招生、录取信息感兴趣,但学生考上大学后,这方面兴趣可能逐渐减弱;一个女人可能对育儿知识感兴趣,而在一年前却对韩剧更感兴趣。为了提高对新项目预测的敏感性,取得实时的预期效果,我们讨论用目标用户目前的点击率而不是用过去的点击率来反映他们将来的喜好。为此,本论文提出了时间加权协同过滤算法以得到兴趣变化的新颖信息。

由于旧的兴趣度是不太可靠的,不够准确的,因此我们需要寻求算法的改进,以便能处理协同过滤中变动的数据,也就是说此算法必须减少那些重现用户过去喜好的数据的影响,从而得到更精确的预测效果,实现更准确的实时推荐。

2.1　算法的提出

在时间加权协同过滤算法中,充分考虑到"时间效应"的影响,即越早发生的点击兴趣,其重要性越小,为了降低它对推荐的影响,本文引入时间加权函数 $f(t)$(t 为时间变量)到兴趣预测中,并将目标用户 Tu 对项目 i 的加权预测评分改进为 $P'_{Tu,i}$:

$$P'_{Tu,i} = \overline{R_{Tu}} + \frac{\sum_{n\in Neighbor_{Tu}} \mathrm{Sim}(Tu,n) \times (R_{n,i} - \overline{R_n}) \times f(t_{ni})}{\sum_{n\in Neighbor_{Tu}} |\mathrm{Sim}(Tu,n)| \times f(t_{ni})} \quad (2)$$

$\mathrm{Sim}(Tu,n)$ 表示目标用户 Tu 与最近邻居用户 n 的相似性,$R_{n,i}$ 表示用户 n 对资源 i 的兴趣度(n 是"最近邻居"集中的用户)。$\overline{R_{Tu}}$ 和 $\overline{R_n}$ 分别表示目标用户 Tu 和用户 n 对资源的平均兴趣度。t_{ni} 表示用户 n 对项目 i 产生兴趣的时间,我们假设,时间函数 $f(t)$ 是单调递减函数,它随时间 t 一直在降低,而且时间权值保持在(0,1)范围内,也就是说,虽然所有数据都有利于推荐项目,而最新数据贡献更大,旧数据反映用户以前的喜好,它在推荐的预测上占较小的权值。本文为时间选择指数来取得目标,指数时间功能被广泛用于实际中,它更有希望得到渐进的过去行为变化的趋势。时间函数:

$$f(t_{ni}) = e^{-t_{ni}} \quad (3)$$

从式(3)能观察到时间函数范围是(0,1),它随时间而降低,数据越新,时间函数值越大,指数函数正满足了我们的需求。

2.2　协同过滤改进算法的实现

在协同过滤推荐算法中,用户间的相似度的计算和推荐的产生是在一个处理过程之中,为了提高推荐的速度,本文把用户之间相似度的计算放在离线处理部分,减少了在线推荐的计算量。

下面介绍协同过滤推荐部分的实现,推荐部分的实现过程如图 1 所示。

首先获取目标用户 IP,然后判断该用户是第一次访问,还是已经访问过该网站。如果用户是第一次访问,选择访问频率高的前 k 项作为推荐内容。如果用户以前访问过该网站,根据与聚类中心的相似度找出该用户所属类以及类内所有用户,再计算出目标用户和类内其他用户之间的相似系数,最后通过式(2)预测出目标用户对未访问项的评分,把评分比较高的前 k 项作为推荐结果推荐给用户。

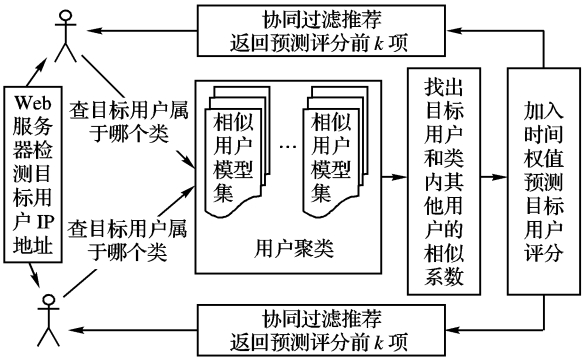

图 1　改进的协同过滤推荐引擎实现过程

利用改进的基于用户的时间加权协同过滤(*TWCF*)推荐方法预测目标用户对新项目的评价分,据此进行推荐:

算法 $TWCF(Tu,k,Neighbor_{Tu},i,\xi)$

输入:目标用户 Tu,推荐的项目数 k,邻居用户数据集 $Neighbor_{Tu}$,项目 i,评分阈值 ξ。

输出:目标用户的 k 个推荐项目。

方法:

对于任意用户 Tu 和项目 i,预测用户 Tu 对第 i 个项目的评价分,记预测值为:

1) 计算目标用户 Tu 和任一邻居用户 $n(n \in Neighbor_{Tu})$ 的相似性 $\mathrm{sim}(Tu,n)$。

2) 求邻居用户在项目 i 上的评价分,由式(2)预测目标用户 Tu 对项目 i 的评价值 $P'_{Tu,i}$。

3) 根据评分 $P'_{Tu,i}$ 的大小顺序排列,采用下列两种方法之一确定用户 Tu 最感兴趣的项目。

(1) 取前 k 个最大的 $P'_{Tu,i}$ 所对应的项目为用户 Tu 最感兴趣的项目;

(2) 对于预先设定评分阈值 ξ,$P'_{Tu,i} > \xi$ 所对应的项目为用户 Tu 最感兴趣的项目。

3　实验结果及其分析

为了验证算法的效果,本文用 Delphi 编程实现了上述算法,数据采用河南某大学校园文化网日志数据。在将网站日志数据经过数据清洗、用户识别、页面识别这几个步骤预处理之后,选择 4 582 条兴趣评分数据作为实验数据集,实验数据集中共包含 422 个用户和 312 个网页。本文将检验本方法对预测推荐准确性的影响。

在本文的研究中,我们希望最后的算法能够准确的预测未评价的项目的评价分,从而为用户做出比较准确地推荐,算法的精确度是评价推荐算法的一个主要指标。像大多数文献的做法一样,我们以预测值和实际值的平均绝对误差(*MAE*)作为衡量算法精确度的一个标准[6]。

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

(下转第 2326 页)

表示的图变换为空间域表示的图。

IDCT 变换可用下面的公式表示为：

$$S_{yx} = \frac{1}{4}\sum_{u=0}^{7}\sum_{v=0}^{7} C_u C_v S_{vu} \cos\frac{(2x+1)u\pi}{16} \cos\frac{(2y+1)v\pi}{16}$$

其中：S_{yx} 表示 Y,Cb,Cr 二维数组中位于 (x,y) 坐标处的像素颜色值。当 $u = v = 0$ 时，$C_u = C_v = 0.17071$；当 u,v 为其他情况时，$C_u = C_v = 1$。直接按照上面的公式计算，将会有很大的运算量，势必影响 JPEG 的解码速度，为此可以将二维的 IDCT 分为 2 个一维的 IDCT，转换如下：

$$S_{yx} = \frac{1}{2}\sum_{u=0}^{7} C_u G_{ux} \cos\frac{(2x+1)u\pi}{16}$$

$$G_{ux} = \frac{1}{2}\sum_{v=0}^{7} C_v S_{vu} \cos\frac{(2y+1)v\pi}{16}$$

这样对 8×8 的二维数组进行 IDCT 的计算可以转化为先对该数组的行分别进行 8 次一维 IDCT，再对列分别进行 8 次一维 IDCT，这样就简化了计算复杂度。

4.5 颜色空间变换

JPEG 算法本身与颜色空间无关，因此“YUV 到 RGB 变换”不包含在 JPEG 算法中。但由于作为输出的位图数据一般要求 RGB 表示，所以将颜色空间变换也表示在算法框图中。用于测试此解码器的图片水平下采样和垂直下采样均为 1∶1，所以解码出的 YUV 数据的比例为 4∶4∶4，可直接采用以下公式转换为 RGB 三原色显示。

$$R = Y + 1.402 \times (Cr - 128)$$

$$G = Y - 0.34414 \times (Cb - 128) - 0.71414 \times (Cr - 128)$$

$$B = Y + 1.772 \times (Cb - 128)$$

5 结语

本文描述的解码方法在 X86 体系结构的嵌入式微处理器 Vortex86 下通过测试（CPU 200 MHz，内存 16 MB）。编码调试环境为 Windows CE 和 Embedded VC++ 4.0。采用上述解码方法，在 Windows CE 系统下成功解码显示一幅 1 100 万像素 JPEG 图片需要内存约 112 kB，仅占完全解码所需内存（29 MB）的 0.37%，极大地减少了资源消耗，以这样的资源消耗可以在目前绝大多数嵌入式产品比如手机、PDA 中解码显示大尺寸 JPEG 图片。

本文提出的解码方法显示一幅 1 100 万像素的 JPEG 图片约耗时 16.4 s，花费时间较多，因此该方法还需做进一步的优化，提高效率，使解码方法更有实用价值，最终应用到 PDA、智能手机、移动智能终端等嵌入式产品中。

参考文献：

[1] 求是科技. Visual C++ 音视频解码技术及实践[M]. 北京：人民邮电出版社，2006.

[2] 求是科技. Visual C++ 数字图像处理典型算法及实现[M]. 北京：人民邮电出版社，2006.

[3] 傅曦，陈黎，石卫华. Windows CE 嵌入式开发入门：基于 Xscale 架构[M]. 北京：人民邮电出版社，2006.

[4] 李于剑. Visual C++ 实践与提高图形图像编程篇[M]. 北京：中国铁道出版社，2001.

[5] ISO/IEC 10918-1, Information technology - digital compression and coding of continuous tone still images requirements and guidelines[S], 1994.

（上接第 2303 页）

其中：N 代表用户在该数据集中的点击数量，p_i 是用户对第 i 个项目的预期点击，q_i 是用户对第 i 项目的实际点击。根据实验需要，我们将数据集划分为训练集和测试集，我们选择训练集占数据集的 80%，测试集占 20%。*MAE* 通过计算预测的用户所浏览的网页与用户实际浏览的网页之间的偏差来度量预测的准确性。*MAE* 值越小，推荐质量越高。

预测用户对资源项的兴趣评分时，参与计算的最近邻邻居的多少影响着算法的 *MAE*。实验中，我们采取目标用户的最近邻居个数从 5 增加到 40，间隔为 5，查看不同的最近邻居集大小对预测准确度的影响，实验结果如图 2 所示。

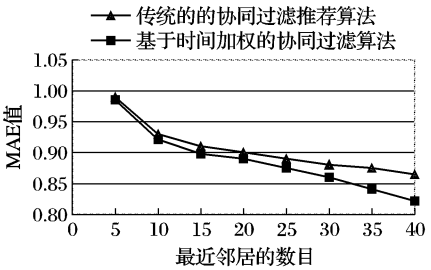

图 2 推荐算法的平均绝对误差 *MAE* 比较

由实验结果看出，大多数情况下，加时间权值的协同过滤（*TWCF*）的准确性比基于用户聚类的 *CF* 的准确性稍高，原因是它能更准确地反映出用户近期的兴趣变化。从实验结果中也可看出邻居数目越多的情况下，能反映用户兴趣变化的用户越多，可看出本算法的准确性越高。利用聚类实现最近邻的查询，理论上来说增加时间开销，但聚类的过程可以离线进行。一旦形成聚类后，可以提高推荐效率。本算法通过增加时间权值，能随用户兴趣的转移发现最近的兴趣所在进行推荐，与传统协同过滤算法相比推荐更具准确性。

4 结语

本文首先详细介绍了传统的基于用户协同过滤算法的实现过程，进而针对传统方法没有考虑用户兴趣变化的问题，提出了在传统的协同过滤算法中引入时间权值，目的是通过时间权值提高预测将来用户兴趣的精确度。用户新近点击的项目在预测上要比旧数据作用大。实验结果表明，新算法提高了用户协同过滤的准确性。

参考文献：

[1] 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.

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

[3] 周军锋，汤显，郭景峰. 一种优化的协同过滤推荐算法[J]. 计算机研究与发展，2004，41(10)：1842－1847.

[4] BREESE J, HECHERMAN D, KADIE C. Empirical analysis of predictive algorithms for collaborative filtering[C]// Proceedings of the 14th Conference on Uncertainty in Artificial Intelligence(UAI-98). San Francisco: Morgan Kaufmann Publishers, 1998：43－52.

[5] 黄光球，靳峰，彭绪友. 基于兴趣度的协同过滤商品推荐系统模型[J]. 微电子学与计算机，2005，22(3)：5－8.

[6] SARWAR B, KARYPIS G, KONSTAN J, *et al.* Item-based collaborative filtering recommendation algorithms[C]// Proceedings of the 10th International World Wide Web Conference. Hong Kong: [s.n.], 2001：285－295.