第 30 卷第 10 期 系统工程理论与实践 Vol.30, No.10
2010年 10 月 Systems Engineering — Theory & Practice Oct., 2010

文章编号: 1000-6788(2010)10-1850-05 中图分类号: TP311.13 文献标志码: A

基于 PCA-SOM 的混合协同过滤模型

郁 雪, 李敏强

(天津大学 管理与经济学部, 天津 300072)

摘 要 针对推荐系统中协同过滤技术面临的数据稀疏性和推荐实时性难以保证的问题, 提出一种基于主成分分析 (Principle component analysis) 和 SOM (Self-organizing map) 聚类的混合协同过滤模型. 首先对原始评分数据进行全局降维, 并在转换后的主成分空间上进行用户聚类, 缩小了目标用户的最近邻搜索空间, 减少了在线计算时间复杂度, 最后对真实的电子政务门户网站 Log 日志数据进行了几种常用的推荐算法的比较, 实验结果证明新的推荐模型具有较好的预测精度.

关键词 推荐系统; 协同过滤算法; 主成分分析; 自组织映射; 聚类技术

Effective hybrid collaborative filtering model based on PCA-SOM

YU Xue, LI Min-qiang

(College of Management & Economics, Tianjin University, Tianjin 300072, China)

Abstract To alleviate data sparsity and scalability issues of collaborative filtering technique in recommendation systems, a new hybrid collaborative filtering model based on Principle Component Analysis and Self-Organizing Map cluster method was proposed. In our approach, dimension reduction technique was first performed on whole data space. The clusters were generated from relatively low dimension vector space transformed by the first step, and then used for neighborhood selection in stead of searching in the whole user space, which can reduce the computation complexity in online recommendation. The experiments were based on web log data from E-government portal web site, and the results indicate that the proposed algorithm can provide better prediction accuracy compared with some exiting collaborative filtering algorithms.

Keywords recommendation system; collaborative filtering; principle component analysis; self-organizing map; clustering technique

1 引言

协同过滤推荐技术是推荐系统中最广泛使用和最成功技术之一, 近些年来在理论研究和实践中都取得了快速的发展, 但是随着用户数量和系统规模的不断扩大, 使用传统的基于用户的协同过滤算法 (User-based CF) 需要搜索整个用户空间[1], 推荐系统的实时性难以保证. 另外在大型的 Web 应用中, 推荐系统将面临更严重的数据稀疏性、超高维和冷启动等方面的挑战[2]. 为了解决上述的问题, 一些研究者提出 Model-based 协同过滤算法[3−6], 利用历史评分矩阵离线训练好模型, 当目标用户到达时, 通过与模型的匹配情况来进行预测. 基于模型的协同过滤算法使系统的可扩展性有了很大的提高, 但在预测精度上不如传统的 k 最近邻技术.

本文以电子政务门户网站的页面推荐系统为背景, 提出一种基于降维技术与 SOM 自组织神经网络聚类的混合协同过滤页面推荐算法, 新算法可以有效的缓解数据稀疏性的问题, 提高预测精度, 并且能够改善系

收稿日期: 2009-06-19
资助项目: 高等学校博士学科点专项科研基金 (20020056047)
作者简介: 郁雪 (1977–), 女, 天津人, 讲师, 博士, 主要研究方向: 信息系统、Web 智能; 李敏强 (1965–), 男, 河北人, 教授, 博士生导师, 主要研究方向: 系统工程与信息系统, 人工智能.

统的可扩展性问题. 最后通过对真实的电子政务门户网站匿名访问日志进行实验测试, 结果证明了新算法的有效性.

2 基于 PCA-SOM 的混合协同过滤模型

2.1 算法描述

本文提出的新算法主要有两个方面的改进, 首先利用主成分分析[7] 的维数约简技术对高维数据进行降维处理, 使原始数据转换到主成分空间上, 缓解了数据稀疏性的特点, 并且为后面的聚类减少了计算复杂度; 随后在低维的向量空间上进行 SOM 聚类, 得到用户的偏好模式 $\{C_1, C_2, \cdots, C_j\}$, 由于上述算法的时间复杂度较高, 因此可以作为系统的离线模型, 改善系统的伸缩性. 当用户到达时, 在其变换的低维主成分空间上可以通过离线模型的聚类结果搜索相似邻居, 预测用户对未知页面的兴趣度. 算法的具体流程如图 1 所示, 其实现的具体步骤如下:

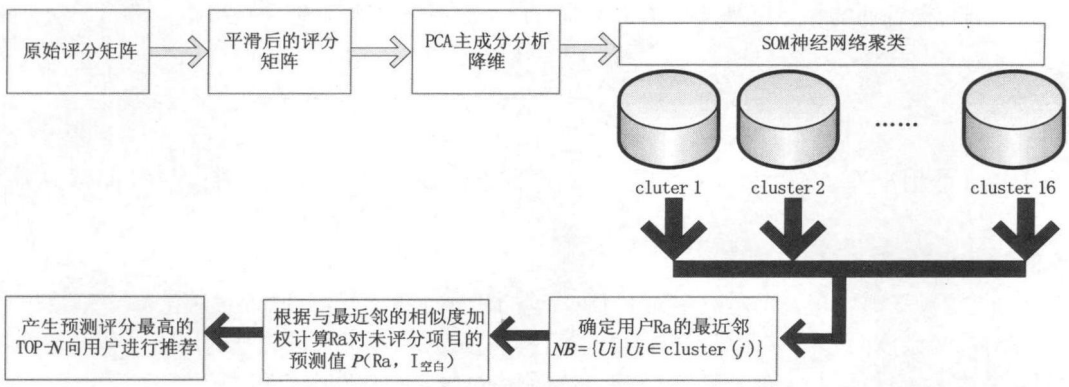

图 1　算法流程图

1) 收集日志数据, 经过数据过滤、用户识别、会话识别、路径补充和事务识别等预处理过程, 删除噪声数据. 考虑到外在环境的复杂性, 影响用户停留时间的因素比较多, 因此本文在分析用户对页面的兴趣度方面采用了考虑页面的访问频率, 通过对 Web 日志的预处理得到用户的原始评分矩阵:

$$PM_{m\times n} = \begin{bmatrix} p_{1,1} & \cdots & p_{1,j} & \cdots & p_{1,n} \\ p_{2,1} & \cdots & p_{2,j} & \cdots & p_{2,n} \\ \vdots & & \vdots & & \vdots \\ p_{m,1} & \cdots & p_{m,j} & \cdots & p_{m,n} \end{bmatrix} \tag{1}$$

用户的评分可以用 $m \times n$ 的矩阵来表示, 其中的元素 $p_{u,p}$ 表示用户 u 对页面 p 的兴趣值, 代替传统协同过滤算法中的用户评分值, 计算公式为:

$$p(u,p) = freq(u,p)|_{si} \tag{2}$$

其中 $freq(u,p)|_{si}$ 是指用户 u 在会话 si 期间对页面 p 的访问次数和.

2) 对原始评价矩阵中未评分的项目进行平滑填充, 得到一个无缺失值的评分矩阵, 缺失评分项目的预测方法采用基于项目的协同过滤算法[8], 公式为:

$$\hat{\mathrm{Pref}}(U,p) = \sum_{q\in NBS} Sim(p,q) \times P_{U,q} \Big/ \sum_{q\in NBS} (|Sim(p,q)|) \tag{3}$$

其中 $Sim(p,q)$ 为两个项目之间的评分相似度; NBS 是目标项目 p 的最近邻项目集合; $P_{U,q}$ 是用户 U 在邻居项目上的各自评分;

3) 应用主成分分析 (PCA) 全局降维技术对平滑后的的评分矩阵进行空间变换, 提取主成分因子, 使降维后的主成分能够代表大部分的评价信息, 并且消除原页面之间的相互影响, 新的特征空间维数为 $d(d << n, n$为原始项目数);

4) 对降维后的主成分向量 (d 维) 进行SOM 神经网络聚类[9], 生成评分习惯相似的用户模式 $C_1, C_2, \cdots, C_j$, SOM 聚类过程如下:

设: 输入空间的输入向量记为: $\boldsymbol{x}=[x_1,x_2,\cdots,x_d]^{\mathrm{T}}$, d 为输入向量的维数.

输出神经元 j 的突触权值向量记为: $w_j=[w_{j1},w_{j2},\cdots,w_{jd}]^{\mathrm{T}}$, $j=1,2,\cdots,l$, 其中 l 为输出神经元的总数. 获胜神经元 $i(\boldsymbol{x})$ 的领域函数为 $h_{j,i(\boldsymbol{x})}(t)$; 学习率参数 $\eta(t)$, 初始值为 η_0, 并随着时间 t 递减.

①初始化权值向量 $w_j(0)$, 赋予其随机值;

②取输入向量$\boldsymbol{x}$, 使用最小 Euclid 距离准则找出与当前输入向量$\boldsymbol{x}$最匹配的神经元为获胜神经元 $i(\boldsymbol{x})$:

$$i(\boldsymbol{x})=\arg\min_j\|\boldsymbol{x}(t)-w_j\|,\quad j=1,2,\cdots,l \tag{4}$$

③更新获胜神经元的邻域内所有神经元的权值向量:

$$w_{j(new)}=w_{j(old)}+\eta(t)h_{j,i(\boldsymbol{x})}(\boldsymbol{x}(t)-w_{j(old)}) \tag{5}$$

④重复步骤 2) 直到最大训练步长, 或训练误差达到预设值.

5) 当预测目标用户 U_a 对空白项目的评分值时, 首先计算 U_a 的历史评分向量在主成分空间上的坐标, 输入到 SOM 训练模型判断 U_a 所属的类 C_i; 在 C_i 中搜索最近邻, 其中最近邻个数 $k=25$;

6) 最后根据与最近邻的相似度来加权预测当前用户 U_a 对未知项目的评分 $P_{Ua,item}$, 计算公式[10] 为:

$$P_{Ua,item}=\overline{R_x}+\frac{\sum\limits_{y\in NBS}sim(U_a,y)\times(R_{y,i}-\overline{R_y})}{\sum\limits_{y\in NBS}(|sim(U_a,y)|)} \tag{6}$$

其中, $sim(U_a,y)$ 是用户 U_a 与用户 y 的相似性, 本文采用 Pearson 相关系数计算; NBS 是用户 U_a 的最近邻集合.

2.2 算法的时间复杂度

传统协同过滤算法需要在整个用户空间上 (m 个基本用户) 搜索最近邻, 通过对比 n 个项目进行相似度的比较, 因此时间复杂度为 $O(mn)$. 基于用户聚类的协同过滤算法则需要计算当前用户与各聚类中心的相似度, 则时间复杂度为 $O(kn)$, k 为类别数目. 然后在当前用户所属的类别中找到 l 个最近邻, 由于 k 和 l 都远小于 n, 所以最后产生推荐的时间复杂度近似为 $O(n)$. 本文提出的推荐算法由于事先对 n 进行了降维处理, 因此在线的时间复杂度可以近似为 $O(d)$, 其中 $d<<n$.

3 实验及讨论

3.1 数据集

本文采用了电子政务门户网 (http://www.tj.gov.cn/) 匿名日志数据集, 经过预处理后, 考虑到 Web 日志数据噪声大的特点, 我们又采用了如下的处理过程:

1) 删除管理员的访问记录. 管理员访问站点会产生大量的日志记录, 由于这些记录的主要目的是维护和测试 web 系统, 因此不在我们研究的目标内.

2) 删除点击次数少于 15 次, 或浏览页面小于 10 的用户. 这些用户的访问记录可以忽略不计, 可以认为对政务网的提供的信息没有特别的兴趣, 因此作为干扰数据可以过滤掉.

3) 删除总的点击次数小于 10 的页面. 这些页面可以认为没有提供值得用户关注的重要信息, 因此可以不作为推荐的对象.

最终形成 1486 个用户对 1430 个页面的浏览信息. 在此数据集中随机抽取出 1250 条记录, 分成训练集合 1000 条和测试集 50 条 (分成 5 组当前用户), 分别进行五次交叉验证. 为了测试不同信息量下的预测质量, 随机抽取每组可见的用户兴趣, 用户的可见兴趣页面数依次从 5–20 个页面, 分别命名为 Given5-Given20. 该数据集的稀疏等级 ψ_{gl} 为:

$$\psi_{gl}=1-50561/1250\times1430=0.9717,$$

可见该数据集比标准的 MovienLens 数据集 (ψ_{ml} =0.9369) 更为稀疏, 属于严重稀疏级别.

3.2 评价指标

MAE (Mean absolute error, 平均绝对误差) 经常用来协同过滤算法的预测精度[5], 本文采用 MAE 测量新算法对未知页面兴趣度预测的准确性.

$$MAE=\frac{\sum\limits_{u\in T}\left|Pref(u,p)-\tilde{P}ref(u,p)\right|}{|T|} \tag{7}$$

其中 $Pref(u,p)$ 为用户的真实兴趣值, $\tilde{P}ref(u,p)$ 是通过算法预测的页面的兴趣值, T 是测试集, $|T|$ 是测试集的元素个数, MAE 值越小说明越接近于真实兴趣值, 预测就越准确.

3.3 实验结果与讨论

按照新算法的步骤, 首先平滑训练模型 m, 对未有兴趣值的页面通过基于项目的协同过滤算法进行初步预测, 得到填充后的训练模型 m'; 对 m' 进行主成分分析后, 其累计贡献率如下图 2 所示, 图中的实线部分表示平滑后的模型 m' 的各主成分的累积贡献率. 分别选取累积贡献率 80%和 90%的前 58 个和前 147 个主成分作为转换后向量的维度, 并在低维空间内进行聚类, 表 1 描述了本次实验中所用到的对比算法的命名.

表 1　各算法命名描述

算法命名	算法说明
UPCC	基于用户的协同过滤算法 (User_based CF)
KCLUST-CF	基于用户聚类的协同过滤算法 (K-means Clustering)
PCA90-UPCC	主成分降维后基于用户的协同过滤算法 (PCA+UPCC)
PCA-KM90	主成分降维后基于用户聚类的协同过滤算法 (PCA+K-means Clustering)
PCA-SOM90	主成分降维后基于用户聚类的协同过滤算法 (PCA+SOM Clustering)

为了证明本文提出的二阶段预测算法的准确性, 我们设计了 2 组实验, 第一组实验分别用不同聚类算法 (K-means 聚类与 SOM 聚类) 对比预测精度, 并同时比较了经典的协同过滤算法 UPCC (基于用户的协同过滤)、基于用户聚类的协同过滤算法. 第二组实验比较了取 90%与 80%不同的主成分累积贡献率, 对比其在预测质量上的影响. 图 3 记录了在 Given10 实验条件下采用经典 K-means 快速聚类方法, 聚类数目对 MAE 的影响, 初步判断最优的聚类数目 k 之应该在 8–12 之间, 以后的实验我们不妨设 $k=9$. 实验中 K-means 聚类使用 Cosine 距离测度, SOM 自组织神经网络聚类算法的输出为 3×3 的二维神经元节点, 训练步长为 1000, 学习率为指数衰减函数.

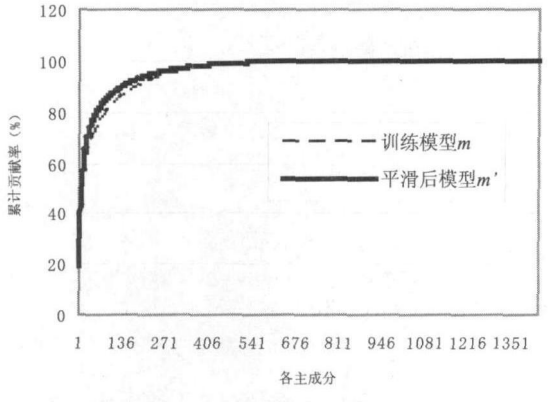

图 2　m 与 m' 的主成分累积贡献率的曲线对比

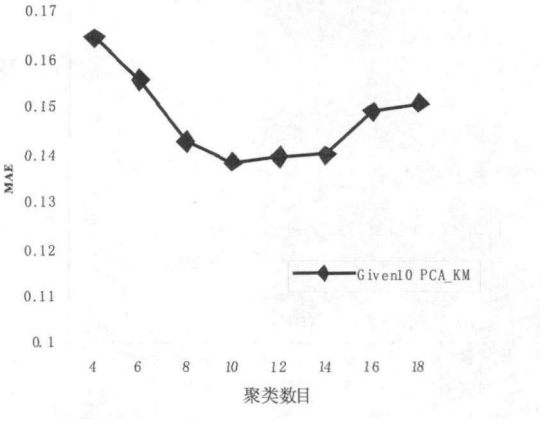

图 3　不同的聚类数目对推荐精度的影响

从图 4 的实验结果可见: 随着测试集中事先给出的历史兴趣信息量越来越多, 算法的总体预测质量是越来越准确的, 页面兴趣度预测采用 PCA 降维与 SOM 神经网络聚类算法的结合效果最好, 可见对于极度稀疏矩阵, SOM 算法更适合在变换后的低维空间进行聚类. 另外, 与基于用户聚类的协同过滤算法 (KCLUST-CF) 相比, 传统的 UPCC 最近邻算法仍然具有较好的预测准确性.

图 5 反映了取不同的主成分累积贡献率对预测质量的影响, 我们取最常用的 80%与 90%累积贡献率进行比较, 在四种不同的实验条件下, 总的趋势为后者比前者的预测准确率稍高, 原因是降维后所包含的原矩阵的信息量较多, 使用户相似性计算更加准确, 提高最后的预测精度.

4　结束语

本文提出了一种结合主成分分析降维与 SOM 神经网络聚类的混合协同过滤模型, 主要思路是首先利用 PCA 进行维数约简, 把原始高维向量变换到相对低维的主成分空间上, 抽取主要特征, 并在主成分空间上进

行聚类寻找当前用户最近邻的混合协同过滤算法. 新算法尝试从原始评分 (兴趣) 矩阵的列维 (项目角度) 和行维 (用户角度) 同时进行约简, 缩小了最近邻的搜索范围, 使特征信息更加集中, 在对真实 Web 日志数据的测试中获得了较好的推荐效果.

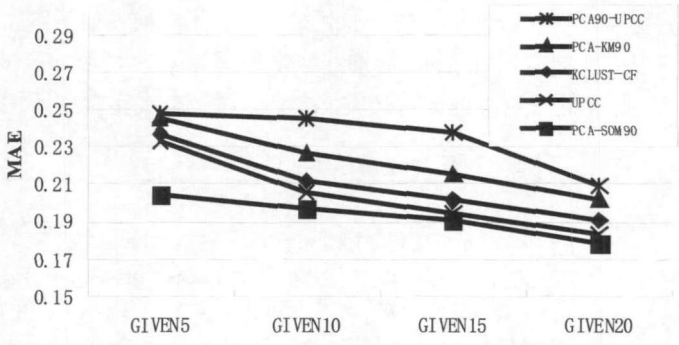

图 4 五种不同算法的预测精度比较

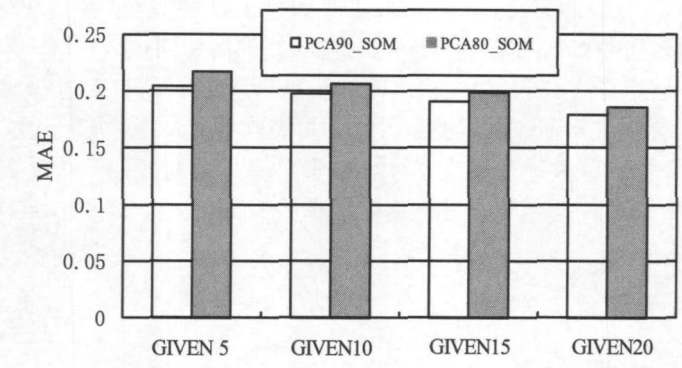

图 5 新算法中取不同主成分累积贡献率对预测的影响

参考文献

[1] Breese J S, Hecherman D, Kadie C. Empirical analysis of predictive algorithms for collaborative filtering[C]//Proceedings of 14th Conf Uncertainty in Artificial Intelligence, San Francisco: Morgan Kaufmann, 1998: 43–52.

[2] Sarwar B, Karypis G, Konstan J, et al. Item-based collaborative filtering recommendation algorithm[C]//Proceedings of the 10th International World Wide Web Conference, Hong Kong, 2001: 285–295.

[3] Kohrs A, Merialdo B. Clustering for collaborative filtering applications[C]// Proceedings of CIMCA'99, Vienna: IOS press, 1999: 199–204.

[4] Sarwar B, Karypis G, Konstan J, et al. Application of dimensionality reduction in recommender system — A case study[C]//ACM WebKDD 2000 Web Mining for E-Commerce Workshop, 2000: 82–90.

[5] Goldberg K, Roeder T, Gupta D, et al. Eigentaste: A constant time collaborative filtering algorithm[J]. Information Retrieval Journal, 2001, 4(2): 133–151.

[6] Xue G, Lin C, Yang Q, et al. Scalable collaborative filtering using cluster-based smoothing[C]//Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Brizil: ACM Press, 2005: 114–121.

[7] Johnson R A, Wichers D W. 实用多元统计分析 [M]. 陆璇, 葛余博, 赵衡秀, 等译. 北京: 清华大学出版社, 2005.
Johnson R A, Wichers D W. Applied Multivariate Statistical Analysis[M]. Beijing: Tsinghua University Press, 2005.

[8] Sarwar B, Karypis G, Konstan J, et al. Item-based collaborative filtering recommendation algorithms[C]//Proceedings of the 10th International Conference on World Wide Web. New York: ACM Press, 2001: 285–295.

[9] Haykin S. 神经网络原理 [M]. 叶世伟, 史忠植, 译. 北京: 机械工业出版社, 2004.
Haykin S. Neural Networks: A Comprehensive Foundation[M]. Beijing: China Machine Press, 2004.

[10] Sarwar B, Karypis G, Konstan J, et al. Analysis of recommendation algorithms for E-commerce[C]//Proceedings of 2nd ACM Conf on Electronic Commerce. New York: ACM Press, 2001: 158–167.