

基于局部主成分分析的协同过滤推荐模型

郁 雪, 李敏强

(天津大学管理与经济学部信息管理与信息系统系, 天津 300072)

摘 要: 根据传统协同过滤算法中用户数据的高维稀疏特点, 提出一种基于局部主成分分析协同过滤推荐模型, 采用基于语义分类和主成分分析的二阶段降维技术, 分别对各类主题页面进行局部降维处理, 以保留对某类主题真正感兴趣的用户群, 加速最近邻的搜索过程。通过对真实 Web 日志数据的测试, 证明该模型具有较高的预测精度。

关键词: 推荐系统; 协同过滤算法; 维数约简; 局部主成分分析

Collaborative Filtering Recommendation Model Based on Local Principle Component Analysis

YU Xue, LI Min-qiang

(Dept. of Information Management and Information System, School of Management, Tianjin University, Tianjin 300072)

【Abstract】 According to the high dimensionality and sparsity of rating matrix in traditional collaborative filtering recommendation system, a new collaborative filtering recommendation model based on Local Principle Component Analysis(LPCA) is proposed which combines taxonomy technique and local principle component analysis method to make dimension reduction for different subject genre respectively, and remains the real interested users in one specific subject of the Web pages which accelerates the neighbor searching process. Experiment on real log data indicates the new model can improve the predication quality.

【Key words】 recommendation system; collaborative filtering algorithm; dimensionality reduction; Local Principle Component Analysis(LPCA)

1 概述

传统的协同过滤算法的主要思路是利用与最近邻的相似度来加权预测当前用户对某一资源的评分, 其优点是对所推荐的资源类型没有特殊要求, 可以实现跨类别的推荐。但是随着用户和资源数量不断地膨胀增加, 传统算法将面临维数灾难^[1]所带来的各种问题, 例如, 最近邻相似性度量和稀疏性问题等, 使推荐质量难以保证。不少研究者提出使用维数约简技术, 如潜在语义标引(Latent Semantic Indexing, LSI)^[2]、奇异值分解(Singular Value Decomposition, SVD)^[3]、主成分分析(Principle Component Analysis, PCA)^[4]和基于项目聚类降维方法^[5], 用于减少项目空间的维数, 缓解高维评分矩阵的稀疏性, 但传统的降维手段会导致信息的损失, 同时降维的效果与数据集的特征紧密相关。

本文提出一种结合语义分类技术, 并使用局部主成分分析的协同过滤推荐方法, 对原始评分矩阵进行二次降维。当为用户预测对某一项目的评分时, 算法使用对该项目所在类别感兴趣的用户群进行预测, 采集电子政务网的真实日志数据进行算法验证, 实验结果证明新降维方法的有效性。

2 问题提出

传统协同过滤算法是比较用户在整个项目空间上的偏好作为判定最近邻的依据, 用对其他类别项目的偏好来预测当前项目的偏好, 且对任一项目的偏好预测, 其最近邻保持不变, 会造成误差, 原因是缺乏考虑项目之间的潜在关联。例如, 用户对唱片的兴趣可能是多方面的, 比如有的用户既喜欢摇滚又喜欢古典, 一部分用户只喜欢摇滚或只喜欢古典唱片。如表 1 所示, 预测用户 U6 对古典唱片 M2、C2 和 BE1

的偏好, 算法搜索全部空间上(假设只研究这 9 张唱片的评价记录)的兴趣相似度, 最终确定 U6 的最近邻为 U1、U3 和 U4(设最近邻的个数 $k=3$, 相似度使用 Pearson 相关系数), 很显然这些用户在摇滚音乐方面与 U6 有共同的偏好, 用这些最近邻来预测 U6 另外一个方面的偏好, 是不具说服力的。

表 1 用户对唱片的评分

用户	Rock(摇滚类)				Classical(古典类)				
	B2	B4	Q2	S3	M2	C2	C3	BE1	BE3
U1	4	3	5	2	1	2	0	2	0
U2	0	1	2	0	5	5	4	3	4
U3	2	3	1	2	0	4	3	5	4
U4	4	5	4	3	0	2	0	1	3
U5	1	0	0	0	0	2	1	0	0
U6	4	5	3	5	??	??	5	??	3

为了解决上述的问题, 针对电子政务门户网站信息网站的应用特点, 本算法提出首先根据网站结构的主题语义进行 Web 页面分类, 使原始兴趣矩阵的维数得到了缩减, 可以理解为第 1 阶段的降维, 与文献[6-7]的思路一致, 但是语义分类后发现每个主题类的页面数量依然很多, 仍然是超高维的主题兴趣矩阵, 于是针对每个兴趣主题类别进行第 2 次的降维, 采用局部主成分分析技术达到消除噪声数据, 减少冗余的目的。

3 基于局部主成分分析的协同过滤算法

在推荐算法领域, 常用的维数约简方法是 SVD 和 PCA,

基金项目: 高等学校博士学科点专项科研基金资助项目(2002 0056047)

作者简介: 郁 雪(1977 -), 女, 讲师、博士, 主研方向: 信息系统, 电子商务, Web 智能; 李敏强, 教授、博士生导师

收稿日期: 2009-12-03 **E-mail:** yuki@tju.edu.cn

两者都是全局的降维技术(Global Dimension Reduction, GDR)技术,只有当数据集是全局相关时效果较好。由于在实际的数据集中并不一定总是全局相关,在这种情况下如果采用GDR技术降维,容易引起大量距离信息的丢失,降维效果不一定好,因此研究人员提出各自在局部相关的数据簇上分别进行降维^[8],即一种局部降维技术(Local Dimension Reduction, LDR),需要首先发现数据集的局部相关性,形成局部数据簇,并对不同的数据簇分别实施降维,得到不同的维集。

3.1 算法描述

由于传统协同过滤算法忽略了网页之间潜在的语义相关性,当网页在主题内容上差别较大时,预测质量会下降。本文以电子政务门户网站为应用背景,Web页面的语义分类体系总结如表2所示(截取其中9个主题)。

表2 按主题分类后类别的页面数量

类	所在目录	主题内容	页面数量
1	http://tj.gov.cn/zjtj	走进天津	64
2	http://tj.gov.cn/zwgk	政务公开	342
3	http://tj.gov.cn/bhxq	滨海新区	64
4	http://tj.gov.cn/tztj	投资天津	19
5	http://tj.gov.cn/hdxq	互动社区	35
6	http://tj.gov.cn/smsb	市民生活	52
7	http://tj.gov.cn/jmjj	津门聚焦	177
8	http://tj.gov.cn/wsbs	网上办事	84
9	http://tj.gov.cn/tblm	特别栏目	237

新算法在2个方面做了改进:

(1)根据主题分类对原始兴趣矩阵进行降维

主题分类树常被用来组织Web资源,其中,每个Web页面属于一个主题类别,本文依据电子政务网的一级栏目划分,为页面进行主题分类,认为属于同一个栏目下的页面具有较高的主题相关性。于是原始兴趣矩阵 $R_{m \times n}$ 就分解成9类主题的兴趣矩阵 $R_{m \times n_{ci}}^i$,其中, $i=1,2,\dots,9$; $n_{ci} \ll n$ 。

$$R_{m \times n_{ci}}^i = \begin{bmatrix} r_{11} & \cdots & r_{1j} & \cdots & r_{1n_{ci}} \\ \vdots & \ddots & \vdots & & \vdots \\ r_{i1} & \cdots & r_{ij} & \cdots & r_{in_{ci}} \\ \vdots & & \vdots & \ddots & \vdots \\ r_{m1} & \cdots & r_{mj} & \cdots & r_{mn_{ci}} \end{bmatrix}_{m \times n_{ci}}$$

在每个基于主题的兴趣矩阵中如果用户的访问页面数量小于 θ ($\theta=5$),可认为该用户对此类主题页面不感兴趣,其值可以忽略不计,从矩阵中删除,以确保矩阵中的用户对该类型的主题具有一定的偏好,能够对当前用户提供有效的推荐作用。于是每个主题所对应的兴趣矩阵为 $R_{m_{ci} \times n_{ci}}^i$, $m_{ci} < m$,完成了第1步的降维。

$$R_{m_{ci} \times n_{ci}}^i = \begin{bmatrix} r_{11} & \cdots & r_{1j} & \cdots & r_{1n_{ci}} \\ \vdots & \ddots & \vdots & & \vdots \\ r_{i1} & \cdots & r_{ij} & \cdots & r_{in_{ci}} \\ \vdots & & \vdots & \ddots & \vdots \\ r_{m_{ci}1} & \cdots & r_{m_{ci}j} & \cdots & r_{m_{ci}n_{ci}} \end{bmatrix}_{m_{ci} \times n_{ci}}$$

(2)完成局部主成分分析

在上述页面通过主题进行分类后,每个主题类别内的页面具有很强的内容相关性,虽然经过了语义分类的降维,但是子类的兴趣矩阵仍然是超高维的,于是对每个主题子类的兴趣矩阵单独做主成分降维,降维后 $n_{d1} \sim n_{d9}$ 分别对应每个主题子类的主成分个数。图1是算法的整体流程,其中浅色箭头部分表示模型的离线计算阶段,深色箭头表示模型的在

线预测流程。

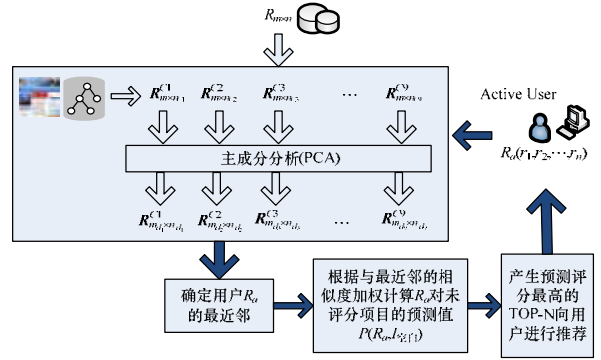


图1 新算法的预测流程

当用户 R_u 到达时,在线部分预测对项目 i 的兴趣值 $Pred(r_{u,i})$,算法的流程如下:

(1)由推荐引擎分析需要预测项目 i 的主题类别,找到其对应的主题兴趣矩阵。

(2)计算 R_u 的历史评分向量在该主题兴趣矩阵中局部主成分空间内的坐标 $R_{1 \times n_{d1}}^{ci}$ 。

(3)在新的变换空间($R_{m_{ci} \times n_{d1}}^{ci}$)上寻找最近邻,保证对该类主题有兴趣的用户的推荐比较有效,其中相似度采用Pearson相关系数计算。

(4)根据与最近邻的相似度,采用经典的最近邻预测算法^[9]:

$$p_{Ua, item} = \bar{R}_x + \frac{\sum_{y \in NBS(ci)} \text{sim}(U_a, y) \times (R_{y,i} - \bar{R}_y)}{\sum_{y \in NBS(ci)} (|\text{sim}(U_a, y)|)}$$

其中, $\text{sim}(U_a, y)$ 是用户 U_a 与用户 y 的相似性; NBS 是用户 U_a 在主题类别 ci 上的最近邻集合。

(5)选取预测值最高的Top-N推荐给用户。

3.2 算法说明

算法说明如下:

(1)由于PCA的时间复杂度较高,为 $O(d^2N) + O(d^3)$,因此模型运行的时间较长,但可以离线进行。

(2)由于经过二阶段的降维处理,搜索最近邻的范围与计算向量的维度都大大减少,与传统算法不同的是每次预测的项目不同,需要重新计算其最近邻,因此算法具有较高的预测精度。

(3)该算法对数据的稠密度要求较高。

4 实验结果与分析

4.1 数据集

算法的实验数据集使用了电子政务门户网(http://www.tj.gov.cn/)匿名日志数据,对数据进行基本预处理和清洗后,考虑到Web日志数据噪声大的特点,又采用如下处理过程:

(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,因此不在研究范围内。

(2)由于增加了对稠密性的要求,实验仅保留访问页面数量大于25的用户作为训练集。

(3)删除点击次数小于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。

经过预处理后,最终形成1276个用户对1074个页面的浏览信息。在此数据集中随机抽取1250条记录,分成训练集合1000条和测试集50条,进行5次交叉验证,并随机

抽取若干个每组可见的用户兴趣值，分别命名为 Given5-Gvien20^[9]。

4.2 评价指标

采用平均绝对误差(MAE)来衡量用户对未知页面兴趣度预测的准确性：

$$MAE = \frac{\sum_{u \in T} |Pref(u, p) - \tilde{Pref}(u, p)|}{|T|}$$

其中， $Pref(u, p)$ 为用户的真实兴趣值； $\tilde{Pref}(u, p)$ 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。

4.3 实验结果及讨论

在验证算法的有效性时，本文做了 3 组测试，分别对每类主题相关页面进行局部降维，累积贡献率取 90%~70%，表 3 显示了在 3 种累积贡献率的参数下，各主题类别对应的保留主成分个数情况。图 2 为各类别的原始维度(页面数量)与局部降维后(主成分累积贡献率 90%)维度的对比情况。

表 3 不同累积贡献率下保留的主成分个数比较

类	主题目录	页面数量	保留主成分个数		
			降维后主成分个数	降维后主成分个数	降维后主成分个数
			累积贡献率为 90%	累积贡献率为 80%	累积贡献率为 70%
1	http://tj.gov.cn/zjtj	64	19	11	7
2	http://tj.gov.cn/zwgk	342	78	35	20
3	http://tj.gov.cn/bhxq	64	19	13	8
4	http://tj.gov.cn/tztj	19	5	4	2
5	http://tj.gov.cn/hdxq	35	10	7	5
6	http://tj.gov.cn/smsb	52	8	5	4
7	http://tj.gov.cn/jmjj	177	37	19	12
8	http://tj.gov.cn/wsbs	84	6	3	2
9	http://tj.gov.cn/tblm	237	12	5	3

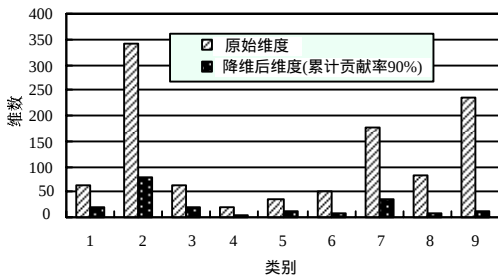


图 2 原始各主题类别进行局部降维后的维度对比

实验在 Given5-Given20 4 种条件下分别进行局部降维处理，用 MAE 值来评价预测的精度，预测结果如图 3 所示。

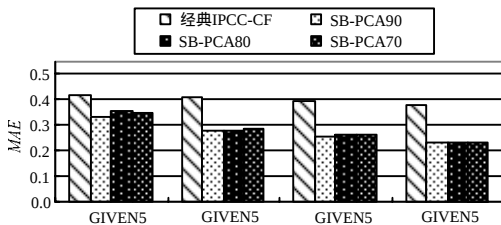


图 3 新算法预测精度的比较

与传统的协同过滤算法(IPCC-CF)相比，本文算法的总体

预测精度都有明显提高，MAE 的总体趋势是下降的，特别是在 Given20 条件下，由于目标用户的历史兴趣信息提供较多，每类主题在主成分累积贡献率 90%的前提下，新算法的 MAE 值下降了 0.144 1；而在 Given5 条件下，MAE 则下降 0.082 6，并且随着历史信息的增多(从 Given5 到 Given20)，算法预测精度越来越好(MAE 值变小)。另外由图中可以看出在同一实验前提下(图中的每一组条形)，各主题类别的主成分累积贡献率采用 3 种不同的取值，预测精度并没有明显变化。

以上实验中取项目的最近邻数量 $k1=15$ ，相似用户的最近邻数量 $k2=15$ 。每个主题类别中保留其兴趣页面数量 $n>10$ 的用户。

5 结束语

本文提出一种基于主题分类的局部降维算法，针对不同的主题页面分别进行降维，对未访问页面兴趣的预测在其所在主题类别的主成分空间上进行计算。实验结果证明了新的降维方法在改进协同过滤算法的预测质量上有明显提高，但是离线模型和在线运算耗费会有所增加，算法对测试数据集的稠密性要求较高，不同的稀疏度对算法实现的效果会有不同的影响。今后的工作是针对上述算法的广泛适用性进行更深入的验证。

参考文献

- [1] 杨风召. 高维数据挖掘技术研究[M]. 南京: 东南大学出版社, 2007.
- [2] Sebastiani F. Machine Learning in Automated Text Categorization[J]. ACM Computing Survey, 2002, 34(1): 1-47.
- [3] Sarwar B M, Karypis G, Konstan J, et al. Application of Dimensionality Reduction in Recommender System: A Case Study[C]//Proc. of ACM Workshop on Web Mining for E-commerce Challenges and Opportunities. Boston, USA: [s. n.], 2000.
- [4] Goldberg K, Roeder T, Gupta D, et al. Eigentaste: A Constant Time Collaborative Filtering Algorithm[J]. Information Retrieval, 2001, 4(2): 133-151.
- [5] 邓爱琳, 左子叶, 朱扬勇. 基于项目聚类的协同过滤推荐算法[J]. 小型微型计算机系统, 2004, 25(9): 1665-1670.
- [6] Gao Fengrong, Xing Chunxiao, Zhao Yong. An Effective Algorithm for Dimensional Reduction in Collaborative Filtering[C]//Proc. of the 10th International Conference on Asian Digital Libraries. Berlin, Germany: [s. n.], 2007.
- [7] 余力, 刘鲁, 李雪峰. 用户多兴趣下的个性化推荐算法研究[J]. 计算机集成制造系统, 2004, 10(12): 1610-1615.
- [8] Chakrabarti K, Mehrotra S. Local Dimensionality Reduction: A New Approach to Indexing High Dimensional Space[C]//Proc. of IEEE VLDB'00. [S. l.]: IEEE Press, 2000.
- [9] Breese J S, Heckerman D, Kadie C. Empirical Analysis of Predictive Algorithms for Collaborative Filtering[C]//Proc. of the 14th Int'l Conference on Uncertainty in Artificial Intelligence. Madison, Wisconsin, USA: [s. n.], 1998.

编辑 陈文

(上接第 36 页)

- [4] 买京京, 龚红艳, 宋纯贺. 集群系统中的动态反馈负载均衡策略[J]. 计算机工程, 2008, 34(16): 114-115.
- [5] Mitzenmacher M. How Useful Is Old Information[J]. IEEE

- Transactions on Parallel and Distributed System, 2000, 11(1): 6-20.
- [6] Dahlin M. Interpreting Stale Load Information[J]. IEEE Transactions on Parallel and Distributed System, 2000, 11(10): 1033-1047.

编辑 陈文