文章编号: 1000-5862(2012)01-0106-05

改进的聚类模式过滤推荐算法

高灵渲, 张　巍, 霍颖翔, 滕少华*

(广东工业大学计算机学院, 广东 广州　510006)

摘要: 通过对用户推荐项目模式进行建模, 预测新用户的推荐项目归属类别, 从而推测出目标用户对具体推荐项目的评分. 实验结果表明: 该方法可以提高推荐服务的效率, 实用性较高.

关键词: 过滤推荐; 聚类; *k*-means 算法; 欧氏距离; 用户-项目模式

中图分类号: TP 301.4　　文献标志码: A

0 引言

互联网已经成为人们日常生活中必不可少的重要环节, 数以亿计的网民接受互联网传递的信息. 如何在庞大的数据流中获取自己需要的信息, 已成为人们关注的焦点. 过滤推荐技术可帮助用户有效地解决网络上的"信息迷失"问题. 推荐系统是利用用户的历史偏好、习惯来向其推荐适合和需要的信息或商品. 目前, 推荐系统已成为电子商务应用系统领域中一个备受研究者关注的研究方向[1].

在众多的推荐系统中, 协同过滤推荐是当前较为广泛应用的技术之一[2], 其基本思想是由评分相似的最近邻居的具体评分结果向目标用户来产生推荐; 其主体思想是以分类用户评价的相似性作为理论基础, 通过计算目标用户最为接近的邻居, 并由此邻居类别对于目标项目的评分平均值作为目标用户对该项目的评分的结果, 同时也参考该评分的加权情况. 当空间和数据不断增大时, 寻找最近邻居的难度不断加大, 此时推荐系统的实时性及扩展性难度也相应增加[3-4].

针对上述问题, 本文提出了改进的聚类模式推荐算法, 它是运用用户历史数据建模. 由模型推测新用户的类别, 进而给用户适当的推荐. 由于推荐算法的主要时间都在聚类迭代的计算部分, 因此本文侧重于使用改进的聚类算法对模式建立过程进行简化, 减少了聚类的计算时间, 从而提高了推荐算法的效率.

1 相关工作

基于聚类技术的过滤推荐是以用户的兴趣的类似性作为分类标准, 由聚类对数据进行初始的分类处理. 数据聚类的结果作为其他用户对项目可能评价的预测基础, 并在相应的规则下产生需求的项目评价结果. B. Mobasher 等[5]提出数据源来自于由服务器隐形录入的日志等资料, 在这些数据集下进行聚类和关联规则等不同的技术, 并由此获取并建立用户的浏览特征模型, 最后由数据集合产生个性化的推荐. 戴亚娥等[6]提出了结合模糊聚类的过滤推荐, 由用户对项目评分的相似性对项目进行模糊聚类, 建立模糊聚类矩阵, 由此基础上搜索最近邻居, 从而缩小最近邻的查找范围并产生推荐结果. 董喜梅等[7]提出了基于聚类模式的推荐算法, 采用 *k*-means 聚类方式对用户及推荐项目进行聚类, 减少在协同过滤推荐过程中产生的时间冗余. 龚松杰[8]提出的基于用户模糊聚类的两阶段协同过滤推荐, 将推荐范围细分到在线和离线 2 个阶段. 周涛等[9]提出了基于用户情景的协同过滤推荐, 引入了用户情景等人物因素.

过滤推荐的实际应用过程中, 也在不断的出现

收稿日期: 2011-11-26

基金项目: 广东省自然科学基金(06021484, 9151009001000007), 广东省科技计划(2008A060201011)和韶关市科技计划(2010CXY/C05)资助项目.

作者简介: 滕少华(1962-), 男, 江西南昌人, 教授, 博士, 主要从事协同工作、网络安全和数据挖掘的研究.

新的技术[10-11], 柯丽等[12]提出了基于频率共现熵的跨语言网页自动分类模型, 提出了更好性能的算法对数据集的分类过程进行自动化, 实现批量的自动分类过程, 而不断的增强分类与推荐的自动化与持久程度也是推荐系统不断发展的目标[13-15]. 范波[16]等提出提升单一评分相似度到多个评分相似性来提高评分正确性的方法, 其基础思想是增大用户相似度的范围及准确程度, 目的也是避免少量数据的低关联性给推荐带来的瓶颈. 利用项目分类对推荐的项目进行优化[17]也是十分可取的方面, 通过建立推荐项目的删除规则可以较好的避免出现不合时宜的推荐结果的情况. 在用户多兴趣的情况下, 对兴趣进行分类对于实际的情况有更好应用意义, 项目的分类过程中考虑用户多兴趣的影响, 并对兴趣的区分度加以标记则[18-19], 从而实现类别的集中以及评价预测的针对性.分类的单一性的向多维性的正在逐步体现, 其现实需求的根源则是由于电子商务系统中的数据日渐增加, 多维化情况日趋明显.

基于聚类模式的过滤推荐算法的核心部分是对用户及推荐项预分类, 可以由用户对推荐项的对应类的评分, 同时由此用户对应的类别判定出用户模式的概率, 以及此用户对于推荐项目的评分所属类别, 来预测该用户对某项目可能的评分. 考虑对用户进行分类, 得到各类中心的取值, 先确定此用户所属类型, 与该类中心比较其相似性, 这样简化了对实时数据的处理, 缩短了等待时间. 若将用户划分为不同的类型, 则需要应用聚类算法. 本文同时采用计算较高的改进的 k-means 算法对实际的数据进行相应的减少复杂性的处理.

2　改进的聚类推荐算法

2.1　改进的聚类方法

对用户项目的评分结果模式进行聚类, 首先在聚类阶段使用基于欧式空间的改进的 k-means 算法, 其中使用 n 表示的是模式的数量, k 则表示类别数量, 在初始时从样本空间中随机的取出 k 个不重合样本作为初始的聚类中心, 聚类算法[10]的具体描述如下所述:

(1)算法的功能: 用于划分的 k 均值算法, 每个簇的中心用簇中对象的均值表示.

(2)输入: k: 簇的数目; D: 包含 n 个对象的数据集.

(3)输出: k 个簇的集合.

(4)算法具体过程如下:

(i)从 D 中任意选择 k 个对象作为初始簇中心;

(ii)repeat

①根据簇中对象的均值, 将每个对象指派到最相似的簇;

②更新簇的均值, 即计算每个簇中对象的均值;

(iii)until 不再发生变化.

聚类过程中对数据使用的是欧几里得距离的计算方法, 在本文中, 主要是对 3 维空间即用户、项目和评分数据进行聚类, 设3 维空间中的用户-项目-评分数据点为 $u(x,y,z)$, x 为用户编号, y 为项目编号, z 为评分, 计算距离公式为

$$d(u_1,u_2)=\sqrt{(x_2-x_1)^2+(y_2-y_1)^2+(z_2-z_1)^2}\ . \quad (1)$$

实际计算时, 由于计算结果是用于比较大小, 因而可以对公式(1)进行改进, 省去开方步骤, 减少数据的计算量, 即转换为如下计算公式为

$$D(u_1,u_2)=(x_2-x_1)^2+(y_2-y_1)^2+(z_2-z_1)^2\ . \quad (2)$$

(2)式也可在高维的情况下进行改写与应用, 假若此时空间的点为 4 维的的情况, 可以将距离的公式后加入两点间位于第 4 维的距离差平方, 同理在开方时可以减少计算步骤, 增加计算效率.

对数据点进行聚类的分类过程中, 可采用部分比较的方式进行划分, 具体原理如下: 设一个待分类的点为 $m(x_m,y_m,z_m)$, 选取聚类的中心点时, 附近有 2 个中心点 $C_1(x_1,y_1,z_1)$ 及 $C_2(x_2,y_2,z_2)$, 则该点到 2 个中心 C_1 和 C_2 的距离值分别为 $D(C_1,m)$ 和 $D(C_2,m)$, 参考公式(2). 在计算过程中, 仅先计算 $D(C_1,m)$ 以及 $\Delta x(C_2,m)=(x_m-x_2)^2$ 的值, 若出现 $\Delta x(C_2,m)>D(C_1,m)$ 时, 该点则可放弃划入中心为 C_2 的聚类.

判定之后, 若有 3 个中心点出现, 即 $C_3(x_3,y_3,z_3)$, 可将 $D(C_1,m)$ 与 $\Delta x(C_3,m)=(x_m-x_3)^2$ 进行比较, 如果此时出现 $D(C_1,m)>\Delta x(C_3,m)$, 可再计算出 $\Delta y(C_3,m)=(y_m-y_3)^2$ 的值; 比较 $D(C_1,m)$ 与 $\Delta x(C_3,m)+\Delta y(C_3,m)$, 如果此时出现 $D(C_1,m)<\Delta x(C_3,m)+$

$\Delta y(C_3,m)$，则肯定该点不能划入中心为 C_3 的聚类.

通过上述 2 个对聚类计算方法的改进，可以减少计算量，并较快产生聚类，在能保障聚类精确性的基础上，加快聚类的时间效率，并且此方法也可用于多维的计算，空间维度越高，其体现的效率改进越加明显.

2.2 计算用户-项目属性类别的概率

现阶段度量用户相似性的方法主要由 2 种：余弦相似性和相关相似性[11]，本文选取使用余弦相似性作为计算相似性的方法，假设有用户 u_1 和用户 u_2 在 m 维的项目评分分别可以表示为向量 u_1 和 u_2，则用户之间的相似性 $sim(u_1,u_2)$ 的计算公式为

$$sim(u_1,u_2)=\cos(u_1,u_2)=\frac{u_1\cdot u_2}{\|u_1\|\cdot\|u_2\|}. \tag{3}$$

通过上述方法可求得用户与各用户类中心间的距离类及项目类中心间的距离，再由此距离计算用户匹配分类模式的概率，也需计算项目匹配模式的概率[7]为

$$P(u_iZ_{uj})=\frac{D(u_i,C_{U_j})}{\sum_{a=1}^{n}D(u_i,C_{U_a})}, \tag{4}$$

$$P(I_iZ_{Ij})=\frac{D(I_i,C_{I_j})}{\sum_{b=1}^{n}D(I_i,C_{I_b})}, \tag{5}$$

其中 $P(u_iZ_{uj})$ 则可计算出用户 u_i 分属用户模式 Z_{uj} 的概率，$P(I_iZ_{Ij})$ 则可计算为项目 I_i 分属项目模式 Z_{Ij} 的概率. $\sum_{a=1}^{n}D(u_i,C_{U_a})$ 为用户向量及用户类别 a 从 1 到 n 的距离之和，$\sum_{b=1}^{n}D(I_i,C_{I_b})$ 为项目向量及项目类别 b 从 1 到 n 的距离之和.

2.3 预测目标用户分属的类别

完成预测评分之前，首先需要计算用户类 U 对于属于项目类 I 中的项目估分，计算方法[7]为

$$R(Z_u,Z_m)=\frac{\sum_{x\in U}P(u|Z_u)R(u,m)}{\sum_{x\in U}P(m|Z_m)}, \tag{6}$$

其中 U 为用户全集表示，$R(u,m)$ 为某用户 u 对某项目 m 的评分，$P(m|Z_m)$ 则为某用户 u 属于聚类的用户模式 Z_u 的概率. 此结果则可以去计算每个用户类别 C 对每个项目类别 M 的预测评分

$$R(Z_u,Z_m)=\frac{\sum_{x\in U}P(m|Z_m)R(Z_u,m)}{\sum_{x\in U}P(m|Z_m)}, \tag{7}$$

$P(m|Z_m)$ 为项目 m 属于类别 Z_m 的概率，$R(Z_u,m)$ 为用户对项目 m 的评分.

2.4 产生推荐值

经过上述步骤，可计算用户 u 对项目 m 的预测评分公式为

$$R(u,m)=P(m|Z_m)R(Z_u,Z_m)P(u|Z_u), \tag{8}$$

其中 $P(u|Z_u)$ 为用户 u 属于用户模式 Z_u 的概率，$P(m|Z_m)$ 为项目 m 属于项目模式 Z_m 的概率，$R(Z_u,Z_m)$ 为用户模式 Z_u 对电影模式 Z_m 的评分.

3 实验及结果分析

3.1 数据集

采用 www. grouplens. org 站点提供的数据，该数据是由网站收集的用户对于观看电影的评分，数据由 4 项构成：用户编号、电影编号、评分、时间戳，其中评分标准为{1, 2, 3, 4, 5}. 时间戳为数据产生时间，未将其加入具体聚类过程. 具体的数据信息如表 1 所示.

表 1 实验数据格式

数据集	用户数/个	电影数/集	评分记录数/条	数据大小/M
数据集 1	934	1 682	100 000	2.72
数据集 2	6 040	3 900	1 000 209	23.4
数据集 3	71 567	10 681	10 000 054	252

3.2 实验结果

为了检验算法的实际运行效果，使用改进前的算法和改进后的算法分别对数据进行实验，并对实验结果进行比较，如图 1 所示，3 维空间下聚类时间图.

图 3 中数据 1 对应改进前的点为(0.1, 2 495)，改进后的对应点为(0.1, 2 028)；数据 2 对应改进前的点为(1, 44 625)，改进后的点为(1, 34 382)；数据 3 对应改进前的点为(10,2 296 842)，改进后的点为(10, 1 531 228).

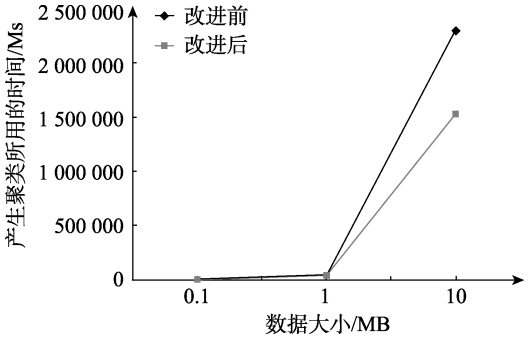

图 1　3 维空间下聚类推荐所用时间比较

3.3　算法性能分析

改进的聚类算法比未改进算法提高了 1.2~1.5 倍的速率, 而聚类是推荐算法中耗时最多的处理步骤. 同时 *k*-means 算法是迭代类型的算法, 需要多次计算, 由此可得, 该算法在大数据量情况下, 能大大提高推荐效率, 并且不会影响推荐的精度.

在具体处理过程中出现了当数据量过大时, CPU 不能满载计算的情况, 只能达到 80%左右的计算负载, 其原因是数据量较大, 运算单元需要等待而不能满载荷工作. 因此在大数据量的情况下, 其聚类所用的计算时间高于线性比例所需计算时间.

4　结束语

随着个性化推荐系统的发展和数据的多种类及大规模增长, 给推荐系统带来很多的挑战与问题, 算法的时间及空间复杂度的指数级增长为计算带来很大的瓶颈, 同时甚至可能会造成计算出的数据不能正确的进行推荐, 甚至可能会产生数据的溢出等消极的情况, 因此提高推荐算法的效率是推荐系统不容忽视的发展方向. 本文在欧式距离上及中心点判定阶段的改进算法, 可以实现聚类及建模阶段的效率提升, 并在实验的基础上体现工作的改进效果. 当数据在高维情况下也可实现这种技术, 同时也是发展的方向, 并且当高维情况出现, 体现的效果更加明显. 理论及实现结果都表明, 本文提出的改进算法对推荐效率的提高有所帮助.

5　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(9): 1621-1628.

[2] Herlocker J, Konstan J, Terveen I, et al. Evaluating Collaborative Filtering Recommender systems [J]. ACM Trans on Information Systems, 2004, 22(1): 5-53.

[3] 李涛, 王建东, 叶飞跃, 等. 一种基于用户聚类的协同过滤推荐算法 [J]. 系统工程与电子技术, 2007, 29(7): 1178-1182.

[4] Sarwar B, Karypis G, Konstan J. Item-based collaborative filtering recommendation algo rithms [EB/OL].[2011-10-18].http://www.citeulike.org/user/mboehmer/article/3752299.

[5] Mobasher B, Dai Honghua, Luo Tao, et al. Discovery and evaluation of aggregate usage profiles for web personalization [EB/OL]. [2011-10-18].http://citeseerx.ist.psu.edu/viewdoc/summary? doi=10.1.1.15.9737.

[6] 戴亚娥. 个性化服务中基于模糊聚类的协同过滤推荐 [J]. 计算机工程与科学, 2009, 31(4): 110-116

[7] 董喜梅, 白振轩. 基于聚类模式的推荐算法的研究 [J]. 系统仿真技术, 2011, 7(1): 44-47.

[8] 龚松杰. 基于用户模糊聚类的两阶段协同过滤推荐 [J]. 计算机工程与科学, 2009, 31(4): 153-155.

[9] 周涛, 李华. 基于用户情景的协同过滤推荐 [J]. 计算机应用, 2010(4): 1076-1078.

[10] Han Jianwei, Micheleline K. 数据挖掘: 概念和技术 [M]. 北京: 北京机械工业出版社, 2007: 263.

[11] 王辉, 高利军. 个性化服务中基于用户聚类的协同过滤推荐 [J]. 计算机应用, 2007, 27(5): 1225-1227.

[12] 柯丽, 王明文, 何世柱, 等. 基于频率共现熵的跨语言网页自动分类研究 [J]. 江西师范大学学报: 自然科学版, 2011, 35(3): 240-245.

[13] 李军, 许丽佳, 游志宇. 一直带压缩因子的自适应权重粒子群算法 [J]. 西南大学学报: 自然科学版, 2011, 33(7): 118-122.

[14] 张洪梅, 周东岱, 钟绍春, 等. 一种基于鲁棒性分析的 Web 应用 PIM 建模方法 [J]. 东北师范大学学报: 自然科学版, 2010, 42(2): 50-56.

[15] 栗晓聪, 滕少华. 频繁项集挖掘的 Apriori 改进算法研究 [J].江西师范大学学报: 自然科学版, 2011, 35(5): 498-502.

[16] 范波, 程久军. 用户间多相似度协同过滤推荐算法 [J]. 计算机科学, 2012, 39(1): 23-26.

[17] 何光辉, 鲍丽山, 王蔚韬, 等. 协同过滤推荐项目优化处理的初步研究 [J]. 计算机科学, 2004, 31(10): 76-78.

[18] 杨芳, 潘一飞, 李杰, 等. 一种改进的协同过滤推荐算法 [J]. 河北工业大学学报, 2010, 39(3): 82-87.

[19] 陈孟建. 电子商务系统中协同过滤推荐算法研究 [J]. 商场现代化, 2008(14): 137-139.

Improved Clustering Filtering Recommendation Algorithm

GAO Ling-xuan, ZHANG Wei, HUO Ying-xiang, TENG Shao-hua*

(Department of Computer Science, Guangdong University of Technology, Guangzhou Guangdong 510006, China)

Abstract: The paper recommended by the project mode of user modeling, forecasting new user ownership of the recommended project categories, which speculated that the target users of the specific projects recommended rate. The experimental results indicate that this method can improve the efficiency of items in recommendation systems and with values of practical use.

Key words: filtering recommendation; clustering; k-means algorithm; Euclidean distance; user-item model

(责任编辑: 冉小晓)

(上接第 101 页)

A Self-Embedding Image Watermarking Algorithm Based on Contourlet Transform

ZHANG Gui-cang, YANG Jun-yan, QIN Na, LI Zhi

(College of Mathematics and Information Science, Northwest Normal University, Lanzhou Gansu 730070, China)

Abstract: First of all, briefly analyzed contourlet transform, and then proposed a self-embedding image watermarking algorithm based on contourlet. In the algorithm, the watermarking was extracted from image feature generated from the image singular value decomposition, and was embedded into contourlet domain by quantitative methods. The experimental results demonstrate that the proposed algorithm is invisible, and can distinguish non-malicious operations from malicious operations.

Key words: contourlet transform; self-embedding; image watermarking;

(责任编辑: 冉小晓)