DOI：10.13196/j.cims.2010.12.215.yanjw.011

第16卷第12期 计算机集成制造系统 Vol. 16 No. 12
2010年12月 Computer Integrated Manufacturing Systems Dec. 2010

文章编号: 1006－5911(2010) 12－2757－06

基于本体用户兴趣模型的个性化推荐算法

严隽薇, 黄 勋, 刘 敏, 朱延波, 倪亥彬
(同济大学 CIMS 研究中心, 上海 201804)

摘 要: 针对目前个性化服务中用户模型稳定性低、推荐结果不尽人意的现状, 在建立基于本体的用户兴趣模型基础上, 通过模型更新提高稳定性, 建立用户群实现用户模型管理。提出利用矩阵聚类降维分解技术的个性化推荐算法, 引入偏好方差的概念计算用户最近邻, 进而产生推荐, 避免了传统协同过滤算法的数据稀疏性缺陷, 提高了推荐质量。结合面向电影的个性化推荐系统, 验证了模型及算法的有效性。

关键词: 本体; 用户兴趣模型; 稳定性; 数据稀疏; 偏好方差
中图分类号: TP311 文献标志码: A

Personalized recommendation algorithm for user interest model based on ontology

YAN Jun-wei, HUANG Xun, LIU Min, ZHU Yan-bo, NI Hai-bin
(CIMS Research Center, Tongji University, Shanghai 201804, China)

Abstract: To deal with low stability of user model and unsatisfied recommendation result existing in personalized service nowadays, ontology-based User Interest Model (UIM) was set up. Model stability was improved by update and model management was realized by establishing user group. To avoid sparse data in traditional collaborative filtering algorithm, personalized recommendation algorithm utilizing matrix clustering dimensionality-reduction decomposition was proposed. Nearest neighbors were calculated according to preferences variance, and the recommendation could be obtained subsequently. Then, the recommendation quality was improved. Finally, effectiveness of the model and algorithm was proved through a personalized movie recommendation system.

Key words: ontology; user interest model; stability; data sparseness; preferences variance

0 引言

随着 Internet 技术的发展, 网络资源正在飞速增长, 如何使广大客户在信息的海洋里方便、迅速和准确地获取所需信息, 成为了人们普遍关注的问题[1]。个性化推荐服务把不同的服务策略提供给不同用户, 实现了“信息找人, 按需服务”的目标[2]。作为个性化服务的基础和核心, 用户模型的质量直接关系到个性化服务的质量。目前, 传统的用户模型逐渐过渡到基于本体的用户模型, 它是建立在领域本体上的基础的用户模型[3-5], 能够明确地描述领域涉及的概念、概念的含义, 以及概念之间的关系, 为简单的术语赋予明确的背景知识, 因而有利于知识的共享和重用, 也最适合于复杂和异构环境下的信息存储和检索, 因此可改善传统用户模型的语义信息不足的缺陷, 提高用户模型的质量。本文在相关领域本体用户兴趣模型的基础上, 针对传统协同过滤算法中数据稀疏性的缺陷, 提出利用矩阵聚类降维分解技术的个性化推荐算法, 并引入偏好方差的概念计算用户最近邻, 克服了传统协同过滤算法中

收稿日期: 2010-06-12; 修订日期: 2010-11-18。 Received 12 June 2010; accepted 18 Nov. 2010.
基金项目: 国家自然科学基金资助项目(61073090); 上海市科学技术委员会科研计划资助项目(09DZ1122302); 广东省教育部产学研结合资助项目(2009GJE00026, 2009B090300429); 上海市重点学科建设资助项目(B004). **Foundation items:** Project supported by the National Natural Science Foundation, China(No. 61073090), the Shanghai Science and Technology Commission Research Program, China (No. 09DZ1122302), the Guangdong Provincial Ministry of Education University-Industry Cooperation Foundation, China(No. 2009GJE00026, 2009B090300429), and the Shanghai Leading Academic Discipline Program, China(No. B004).

的一些弊端,提高了算法效率。

1 本体用户兴趣模型

1.1 相关定义

定义 1 用户兴趣模型(User Interest Model, UIM)是进行个性化兴趣服务的关键部分,是提供个性化信息服务的依据,可以将得到的用户兴趣喜好用结构化的形式保存为用户个体的兴趣模型[6]。

定义 2 领域本体(domain ontology)在一个特定的领域内可以重用,它们提供该领域特定的概念定义和概念之间的关系,提供该领域中发生的活动以及该领域的主要理论和基本原理等。本文使用领域本体的一个子集,即一个小型的领域本体来构建用户的初始个性化用户兴趣本体。使用文本挖掘技术来构建领域本体,具体做法参考文献[5]实现领域本体的构建。

定义 3 个性化用户兴趣本体 Personal IO(personal interest ontology)是基于用户研究领域的领域本体构建的初始个性化用户兴趣本体,是领域本体的一个子集。个性化兴趣本体 Personal IO 是领域本体在不同用户需求描述的基础上通过本体投影获取的。

定义 4 本体投影(ontology projection)是领域本体根据不同用户的需求描述生成不同的 Personal IO,此过程称为本体投影,此时不同的用户需求相当于不同的投影面。本文采用文献[7]中的三层树形结构来表示,能够精确表示用户的兴趣。

定义 5 浅层参考本体(simple reference Ontology)。Guarino[8]曾经提出以详细程度对本体进行分类,详细程度高的本体称为参考本体。此处的浅层参考本体是在用户模型学习更新过程中,根据用户的浏览或检索信息日志文件所建立起来的本体,主要用于用户模型的学习更新。

定义 6 本体归并是指将不同的本体合并为一个本体的过程,只是该合并过程更侧重于将一个小型本体并入到已有的相对大型的本体中,而不是通常的同级合并。

1.2 用户兴趣模型框架及表现形式

基于本体的用户模型构建框架如图 1 所示,由个性化用户兴趣本体即用户模型的获取、更新和用户群的构建三部分组成。其中,个性化用户兴趣本体的获取包括获得用户的个人信息、构建领域本体等;用户模型的更新是根据用户浏览或检索信息的行为构建参考本体,并把它归并到个性化用户兴趣本体中,实现用户模型的学习更新;用户群是每个个性化用户兴趣本体通过相似度计算得到的。

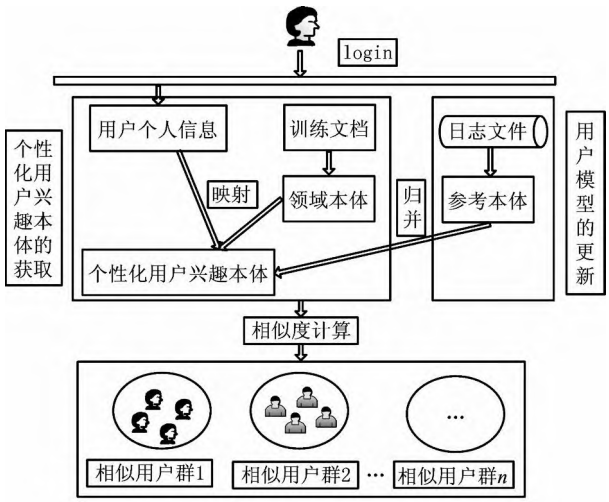

图1 用户兴趣模型框架

使用 Prot g 4.0 构建的关于电影的本体用户兴趣模型如图 2 所示,其结构为树型的表现形式。

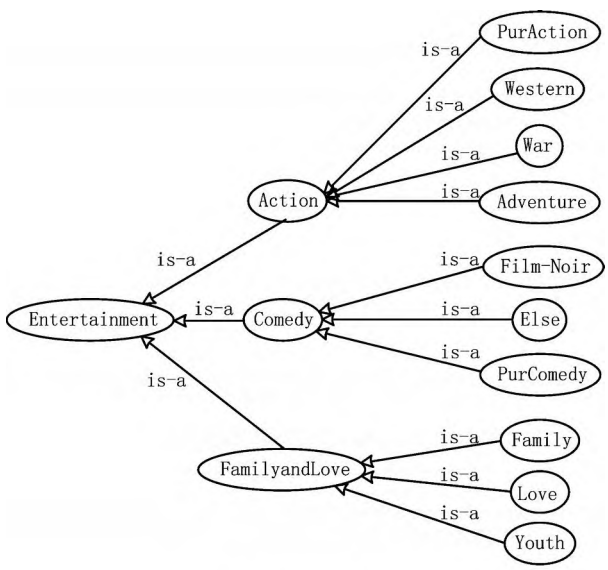

图2 本体用户兴趣模型表现形式

2 基于本体 UIM 的个性化推荐算法

2.1 UIM 的算法表示及矩阵聚类降维分解

目前,协同过滤技术已经被成功运用到各种推荐系统中,但随着资源种类的不断膨胀与用户的日益增加,用来评判的数据矩阵越来越稀疏,严重影响了推荐质量[9],即对于两个用户之间没有都评过分的项目,这两个用户之间的相似度就无法求取,这样就产生了稀疏性,在形成目标用户的最近邻集合时,

往往会造成信息的丢失, 从而降低推荐效果。

因此, 在传统协同过滤算法的基础上, 提出利用矩阵聚类降维分解技术的个性化推荐算法, 并引入偏好方差的概念计算用户最近邻。其流程图如图 3 所示。

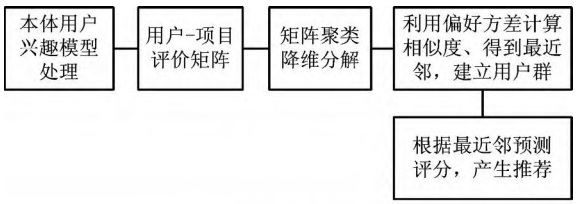

图3　优化算法的流程图

首先对兴趣树形式的本体模型进行抽取, 将其转化为矩阵的形式, 其实质就是对一棵一般树的基于引用的实现。“用户 1”节点的三个子节点“Action”, “Comedy”和“FamilyAndLove”是兄弟。最左的子节点被称为“用户 1”的大子节点。要实现该树, 需要使用基于引用的二叉树使用的节点结构。即每个节点有两个引用, 左引用指向节点的第一个子节点, 右引用指向节点的下一个兄弟。可使用图 4 所示的数据结构来实现图 2 中的树。

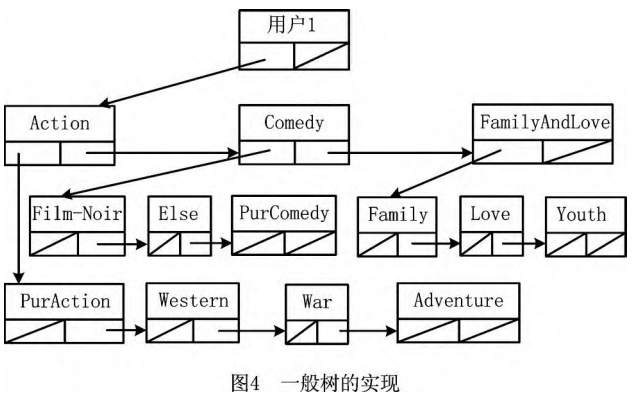

图4　一般树的实现

采用前序或中序遍历, 将图 4 的一般树实现中第一个用户的遍历结果放于一个 $m \times n$ 数组的第一行, 第二个用户遍历结果放于第二行, 以此类推, 构成了一个 $m \times n$ 维, 诸如 $\begin{bmatrix} R_{11} & R_{12} & \cdots & R_{1n} \\ R_{21} & R_{22} & \cdots & R_{2n} \\ \vdots & \vdots & & \vdots \\ R_{m1} & R_{m2} & \cdots & R_{mn} \end{bmatrix}$ 形式的用户 —项目评价矩阵。矩阵中 m 行代表 m 个用户, n 列代表 n 个项目, 第 i 行第 j 列元素 R_{ij} 代表第 i 个用户对项目 j 的评分。

对于大多数用户尤其是研究型用户来说, 他们的研究兴趣往往集中在某一个或几个领域中, 对于研究领域中的信息资源具有较多的评价, 对研究领域以外的信息资源评价很少甚至几乎没有。就是说在某段时间, 用户的兴趣是相对稳定的。用户的评分数据集中在其感兴趣的领域。因此, 如果把对同一类资源感兴趣的用户聚到一类, 则在这一类中, 矩阵的稀疏性会大大地降低。

聚类算法中的数据对象为 n 项资源, 每项资源表示为一个由 m 个变量组成的向量, 每个用户作为资源的一个属性变量, 变量的值为用户对资源的打分。利用 K–均值聚类算法把 n 项资源分成 k 类, 每项资源属于且只属于一个簇, 每一个簇至少包含一项资源。按这样的思想对之前得到的用户 —项目评价矩阵进行分块。

K–均值算法的聚类准则是: $J = \sum_{j=1}^{k} \sum_{i=1}^{n} \| x_i^{(j)} - c_j \|^2$, $x_i^{(j)} \in S_j$, c_j 是聚类 S_j 的聚类中心。

该算法的主要思想是:

(1)为每个聚类确定一个初始的聚类中心, 这样 k 个聚类存在 k 个聚类中心。

(2)将样本集中的每一个样本按照最小距离原则 $D_i = \min\{ \| x - c_j \| \}$, $x \in DataSet$, $i = 1, 2, \cdots, k$ 分配到 k 个聚类中的某一个。

(3)使用每个聚类中所有样本的均值作为新的聚类中心。

(4)如果聚类中心有变化, 则重复第(2)步和第(3)步, 直到聚类中心不再变化为止。

(5)最后得到的 k 个聚类中心就是聚类的结果。

以下是尝试从一个数据集中找到两个聚类的 K–均值聚类算法的伪代码实现:

```
import random
def kcluster(rows, distance=pearson, k=2):
  # 确定每个点的最小和最大值
  ranges=[(min([row[i]for row in rows]),max([row[i]for
row in rows]))
  for i in range(len(row[0]))]
  # 随机创建 k 个中心点
  clusters=[[random.random()*(ranges[i][1-ranges[i]
[0]]+ranges[i][0]
  for i in range(len(rows[0]))]for j in range(k)]
  lastmatches=none
  for t in range(100):
    print 'iteration %d' % t
    bestmatches=[[]for i in range(k)]
  # 在每一行中寻找距离最近的中心点
  for j in range(len(rows)):
    row=rows[j]
```

```
        bestmatch= 0
        for i in range(k):
            d= distance(clusters[i], row)
            if d< distance(clusters[bestmatch], row): bestmatch= i
        bestmatches[bestmatch]. append(j)
    #  如果结果与上一次相同,则整个过程结束
    if bestmatches= = lastmatches: break
    lastmatches= bestmatches
    #  把中心点移到其所有成员的平均位置处
    for i in range(k):
        avgs= [0. 0]* len(row[0])
        if len(bestmatches[i]) > 0:
    for rowid in bestmatches[i]:
                for m in range(len(rows[rowid])):
                    avgs[m]+ = rows[rowid][m]
            for j in range(len(avgs)):
                avgs[j]/= len(bestmatches[i])
            clusters[i]= avgs
    return bestmatches
```

2 2 UIM 用户群

用户群是指具有相似兴趣爱好的用户的集合,建立用户群有助于用户模型的管理和用户兴趣的发现,是在构建好的个性化兴趣本体的基础上通过相似度的计算建立的。用户群的建立流程图如图 5 所示。

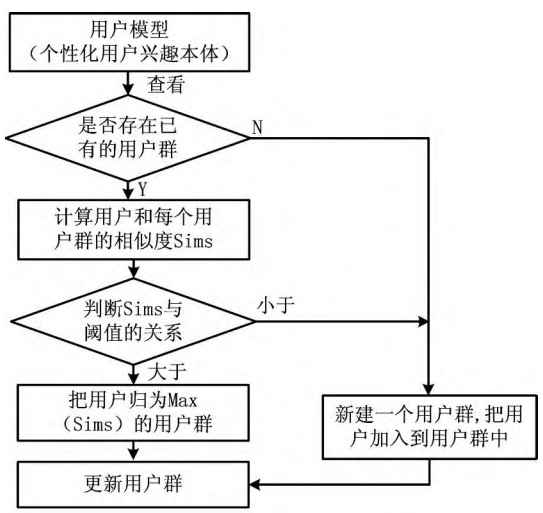

图5 用户群建立流程图

偏好方差是用来显示用户在偏好领域内相互之间偏好差异的一个概念。具体做法为:

(1) 在已经分好的若干小矩阵中,找到用户 A 和 B 各自存在的矩阵 $\boldsymbol{I}_1, \boldsymbol{I}_2, \cdots, \boldsymbol{I}_m$,和它们共同存在的矩阵 $\boldsymbol{I}_{m+1}, \cdots, \boldsymbol{I}_n$。

(2) 分别计算 A 和 B 在这些矩阵中的相似度 $S_1, S_2, \cdots, S_n$,用余弦相似度来计算两个用户之间的相似度,计算公式为:

$$sim(A,B) = \frac{\sum_{k=1}^{m} R_{B,k} \times R_{A,k}}{\sqrt{\sum_{k=1}^{m}(R_{A,k})^2 \times \sum_{k=1}^{m}(R_{B,k})^2}}。$$

式中 $R_{A,k}$, $R_{B,k}$ 分别为用户 A, B 对项目 k 的评分。之后计算这些相似度的方差 $D_{A,B}$。

(3) 当方差 $D_{A,B}$ 大于某一阈值时,则认为 A 和 B 不是最近邻居,记 $sim(A,B)= 0$;否则,取 $E(X)$ 为 A 和 B 的相似度,$E(X)$ 为离散型随机变量的数学期望。

在 A, B 单独存在的矩阵中,A 和 B 的相似度为 0,即 $S_1= S_2= \cdots= S_m= 0$,之后采用 Top-N 方法将相似度最高的 N 个邻居降序排列出来。采用这种做法,可以在一定程度上避免共同项很少时,最近邻计算不准确的问题,这是因为两个用户是最近邻居,当且仅当这两个用户在其每个偏好领域内都相似。

2 3 产生推荐

通过偏好方差找到最近邻后,下一步需要产生相应的推荐。用户兴趣度的预测可以通过公式计算得到,

$$P_{u,i} = \overline{R_u} + \frac{\sum_{m=1}^{n}(R_{m,i} - \overline{R_m}) \times sim(u,m)}{\sum_{m=1}^{n} sim(u,m)}。$$

式中:$\overline{R_u}$ 为用户 u 对资源的平均评分,$R_{m,i}$ 为用户 m 对项目 i 的评分,$\overline{R_m}$ 为用户 m 对资源的平均评分,$sim(u,m)$ 为用户 u 和 m 的相似度,$P_{u,i}$ 为用户 u 对未评分项的预测评分。

通过上述方法预测用户对所有未评分项的评分,然后选择预测评分最高的前若干个项作为推荐结果提交给用户。

3 UIM 稳定性分析

用户模型稳定性指在用户的个性化兴趣树中的非叶子节点上的兴趣度很少或根本没有变化。稳定性分析有助于提高用户模型的质量,主要包括用户模型和用户群的稳定性分析。分析过程如下:

为了对用户模型的稳定性进行分析,本文把用户的个性化兴趣树中非叶子节点转化为一个向量 ***Interest***,它的值代表用户长期兴趣的兴趣度。随着用户模型的不断更新,***Interest*** 不断变化,记更新后

的用户兴趣度向量为 $\boldsymbol{Interest}'$，当 $d=\lim_{n\to\infty}\sqrt{\prod_{i=1}^{length(Interest)}(\boldsymbol{Interest}'(i)-\boldsymbol{Interest}(i)^2)}$ 很小时，即随着时间的推移，用户的长期兴趣变化很少或根本不改变时，用户模型趋于稳定，它能更精确地代表用户的兴趣。

而用户群稳定性分析是指用户群中的用户平均兴趣度变化很小或根本没有变化。

在用户群语义矩阵 $\boldsymbol{G}$ 中，$\boldsymbol{G}$ 的行数代表用户群中用户的数量 M，列数代表用户对兴趣树上节点的概念数目 N。矩阵中 G_{ij} 代表用户 i 对兴趣 j 的兴趣度。

在 $\boldsymbol{G}$ 中，把从第一行到第 M 行的平均值放在第 $M+1$ 行中，这样就形成矩阵 $\boldsymbol{G}'$，$G'_{M+1j}(j=1,\ \cdots,\ N)$ 代表用户群中的用户对兴趣 j 的平均兴趣度。取树中的非叶子节点上的兴趣度作为长期兴趣的兴趣度，假设在 $G_{M+1j}(j=1,\ \cdots,\ N)$ 中有 T 个非叶子节点。把 $G_{M+1j}(j=1,\ \cdots,\ T)$ 赋给一个向量 $\boldsymbol{Avt}$(用户平均兴趣度)。随着用户模型的不断更新，得到更新后的 $\boldsymbol{Avt}'$，然后计算 $\boldsymbol{Avt}$ 和 $\boldsymbol{Avt}'$ 之间的距离，$d=\sqrt{\prod_{i=1}^{T}(\boldsymbol{Avt}_i-\boldsymbol{Avt}'_i)^2}$。当 d 很小时，说明随着时间的推移，用户群中的用户平均兴趣度变化很小，用户群趋于稳定，这时用户群中的用户相似性更高，有助于用户兴趣的发现。

4　验证案例

上文已经构建了关于电影的本体用户兴趣模型。系统采用的数据集取自于 MovieLens。这是一个基于 Web 的研究性推荐系统。MovieLens 数据集包含 movies. dat, users. dat 和 ratings. dat 三个文件。movies. dat 中包含了 3 952 部电影的详细描述信息，users. dat 中包含 6 040 位用户的详细信息，ratings. dat 中包含 6 040 位用户对 3 952 部电影的约 1 000 000 条评分记录。

根据 MovieLens 数据集中三个. dat 文件中的数据在 mysql 数据库中建立了三张数据表，分别为 users 表，movies 表和 ratings 表。表关系图如图 6 所示。

在开发的面向电影的个性化推荐系统中，用户可以对上千部电影进行浏览、搜索和评分，系统会根据用户兴趣推荐电影。面向电影的个性化推荐系统分为用户登陆注册模块、个人信息维护模块、电影浏览搜索评分模块和电影推荐模块四个模块。系统总体框架如图 7 所示，推荐算法时序图如图 8 所示。

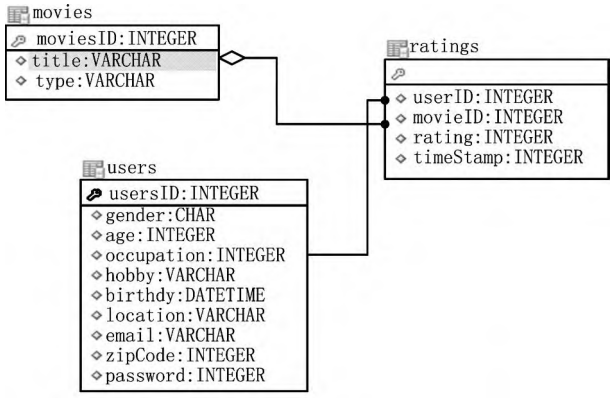

图6　表关系图

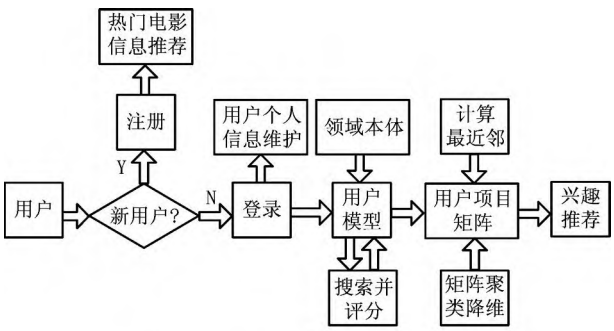

图7　面向电影的个性化推荐系统总体架构

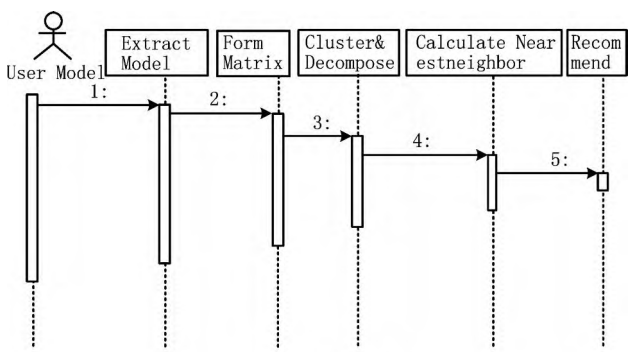

图8　电影推荐算法时序图

根据本体用户兴趣模型中的信息，运用优化的个性化推荐算法，按照用户的兴趣度和所属的用户群，对用户进行电影推荐，推荐结果如图 9 所示。

图9　电影推荐结果

5 对比评价

平均绝对误差(Mean Absolute Error, MAE)通过系统对目标项的预测值与用户对目标项的实际评分值进行比较, 获得对系统预测准确性的评价。为凸显较大的差值, 对 MAE 公式稍作修改, 修改后的 MAE 公式为: $\frac{\sum_{i=1}^{N} |p_i - q_i|^2}{N}$, 所得值越小说明推荐算法的预测精度越高。其中$(p_i, q_i)$为实际值—预测值对, N 为推荐电影数目。

对于传统用户兴趣模型和本体用户兴趣模型, 分别运用传统的协同过滤算法和本文所提出的优化的个性化推荐算法。

推荐电影数目选择的多少影响到系统的推荐精度。实验中, 选择了 25 部、50 部、100 部、500 部和 1 000部电影, 并将优化的协同过滤算法与传统的协同过滤算法进行了比较, 如图 10 所示。

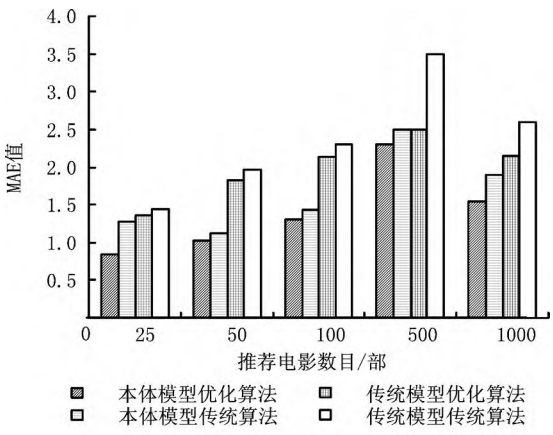

图10 推荐算法精度比较

通过实验对比数据可以发现, 基于本体用户兴趣模型及优化的个性化推荐算法在四种模型算法中预测精度最高, 体现了其优越的推荐性能。

6 结束语

本文将本体应用于用户兴趣模型的构建, 并在此基础上对传统的协同过滤算法进行改进, 引入了模型更新、模型稳定性分析、矩阵聚类降维分解及偏好方差等概念, 提高了个性化服务中的推荐精度, 改善了用户满意度。在下一步工作中, 可寻找更优的推荐算法以消除更多的不足, 同时, 可考虑将 UIM 抽取转化为除评价矩阵以外的矢量形式, 多元组形式等, 使模型处理更为全面。

参考文献:

[1] GAUCH S, CHAFFEE J, PRETSCHNER A. Ontology-based personalized search and browsing[J]. Web Intelligence and Agent System, 2003, 1(3/4): 219-234.

[2] CAMILLE G P D, HE Kai, DENG Xiaoheng. Personalized recommendation system Web mining research[J]. Enterprise Technology Development, 2010, 29(3): 1(in Chinese). [古丽拜天·卡米尔, 贺 恺, 邓晓衡. 个性化推荐系统中 Web 使用挖掘技术的研究[J]. 企业技术开发, 2010, 29(3): 1.]

[3] XU Zhenning. Ontology-based semantic information representation and processing method of Web data[D]. Changsha: National University of Defense Technology, 2002(in Chinese). [徐振宁. 基于本体的 Web 数据语义信息的表示与处理方法[D]. 长沙: 国防科技大学, 2002.]

[4] LI Yong. Ontology-based personalized user modeling and application in intelligent search[D]. Changsha: National University of Defense Technology, 2002(in Chinese). [李 勇. 智能检索中基于本体的个性化用户建模技术及应用[D]. 长沙: 国防科技大学, 2002.]

[5] GUAN Qingzhen, ZHOU Zhurong. Ontology-based user model research[D]. Chongqing: Southwest University, 2007(in Chinese). [关庆珍, 周竹荣. 基于 Ontology 的用户模型研究[D]. 重庆: 西南大学, 2007.]

[6] CHEN Lihua. Personalized information service technology analysis based on user model interest[J]. Software Guide, 2010, 9(1): 13(in Chinese). [陈丽花. 基于用户兴趣模型的个性化信息服务技术分析[J]. 软件导刊, 2010, 9(1): 13.]

[7] GAUCH T S. Improving ontology-based user profile[EB/OL]. (2004-09-09)[2010-03-09]. http://citeseer.vak.edu/publications/RIAO 2004.pdf.

[8] PATHEODOROU C, VASSILIOU A, SIMON B. Discovery of anthologies for learning resources using word-based clustering[C]//Proceedings of the World Conference on Education Multimedia, Hypermedia and Telecommunications. Norfdk, Va., USA: AACE, 2002: 1523-1528.

[9] YU Xue, LI Minqiang. An effective compound collaborative filtering algorithm for ease of data sparseness[J]. Computer Applications, 2009, 29(6): 1590(in Chinese). [郁 雪, 李敏强. 一种有效缓解数据稀疏性的混合协同过滤算法[J]. 计算机应用, 2009, 29(6): 1590.]

作者简介:

严隽薇(1946-), 女, 上海人, 教授, 博士生导师, 研究方向: 系统集成、智能生产系统、知识管理等, E-mail: jwyan@tongji.edu.cn;
黄 勋(1985-), 男, 硕士研究生, 研究方向: 智能生产系统;
刘 敏(1970-), 男, 副教授, 硕士生导师, 研究方向: 智能制造系统、分布式计算;
朱延波(1985-), 男, 硕士, 研究方向: 智能生产系统;
倪亥彬(1983-), 男, 硕士, 研究方向: 智能生产系统。