第 31 卷第 9 期 系统工程理论与实践 Vol.31, No.9
2011年 9 月 Systems Engineering — Theory & Practice Sept., 2011

文章编号: 1000-6788(2011)09-1718-08 中图分类号: F724.6 文献标志码: A

基于图论和信息最大化保留的在线推荐方法

李永立[1,2], 吴 冲[2], 王崑声[1]

(1. 中国航天科技集团公司 航天 710 所, 北京 100037; 2. 哈尔滨工业大学 管理学院, 哈尔滨 150001)

摘 要 随着电子商务的发展, 研究一套高效准确的推荐方法不仅便利了网上购物, 也有助于加速商品流通, 促进经济发展. 既有的方法主要从商品的相似性或顾客的相似性出发进行推荐, 没能将两者很好的结合, 不能充分利用既有的评价信息. 鉴于此, 提出了基于图论的推荐方法, 将人和物的相似性信息结合起来, 构成综合的评估图模型, 并转化为与之等价的评估矩阵. 在评估信息最大化保留的优化目标下, 以评估矩阵为基础建立推荐算法, 并与既有的推荐方法进行比较. 实验结果表明: 本文的方法具有计算时间短、准确度高的特点, 可以用于实时的在线推荐.

关键词 推荐方法; 图论模型; 电子商务; 数据挖掘; 统计学习

On-line recommendation method based on graph model and maximizing information retention

LI Yong-li[1,2], WU Chong[2], WANG Kun-sheng[1]

(1. Institute of No.710, China Aerospace Science and Technology Corporation, Beijing 100037, China;
2. School of Management, Harbin Institute of Technology, Harbin 150001, China)

Abstract The development of E-commence calls for an effective and accurate recommendation method which not only convinces customers, but also accelerates circulation of commodities and promotes economic development. The existed recommendation methods paid attention to either the similarity of goods or that of customers, thus could not trade off the two aspects of information and make full use of them. In view of the above, this paper proposed recommendation method on the basis of graph model, which synthesized the similarity of customers and goods. The method built a comprehensive assessment model able to be transformed into its equivalent evaluation matrix and established an algorithm based on the above evaluation matrix with the aim of maximizing the retention of information. What's more, this paper compared it with the benchmark methods. As a result, the numerical experiments show that the method has short-time calculations, high accuracy and is proper for real-time online recommendation.

Keywords recommendation method; graph model; E-commence; data mining; statistical learning

1 引言

随着计算机技术和互联网技术的发展, 网上购物走入人们的日常生活. 在网上购物的同时, 通常都会邀请购买者发表对所购商品的评论, 由此, 关于这些评论作用的研究成为信息系统管理领域一个重要的研究内容, 已有很多研究证实了这些评价会对购买者的购买行为产生影响[1-4]. 以评价信息能够影响购物行为为逻辑基础, 研究者开始关注对这些信息的利用, 以期对潜在的客户需求进行挖掘, 进行在线推荐方法的研究就是其中一个重要的方面. 既有的推荐方法大体上可以分为两类: 一类是基于物与物的相似性进行推荐, 即一个人购买了商品中的某一个, 从而为其推荐与他购买的商品相似度比较高的商品; 一类是基于人与人之间的推荐, 即当两个购买者有着比较高的相似程度, 当其中的一人购买了某种商品, 另一人就会被推荐这种商品.

收稿日期: 2011-01-09
资助项目: 国家自然科学基金 (60979016); 高等学校博士点专项基金 (20092302110060); 教育部新世纪优秀人才支持项目 (NCET-08-0171)
作者简介: 李永立 (1985–), 男, 汉, 辽宁沈阳人, 博士研究生, 研究方向: 管理科学与工程, E-mail: 0440004@fudan.edu.cn; 吴冲 (1971–), 男, 汉, 黑龙江哈尔滨人, 教授, 博士生导师, 研究方向: 管理科学与工程; 王崑声 (1960–), 男, 汉, 黑龙江齐齐哈尔人, 研究员, 博士生导师, 研究方向: 系统工程与信息化技术.

在这两类研究方法中, 都有着研究被跟进, 主要讨论的议题是不同的相似度定义 [5-7] 与推荐的算法实现机制 [8-10].

但是, 由于推荐方法是利用既有的样本进行统计学习的问题, 针对以上两类方法的研究, 一个问题在于这些研究是否充分利用了既有的评价信息, 如果信息利用的不充分, 会影响推荐方法的可信性和准确度. 比如, 基于人的相似性的推荐就没能用到物的相似性的信息, 以上两种推荐方法都会出现顾此失彼的问题. 有没有相对既有的推荐方法更加全面利用既有信息的方法和机制? 在这个观点指导下, 一部分研究者提出了综合利用以上两方面信息进行推荐的方法 [11-14], 其中, 基于协同过滤方法 (collaborative filtering, CF) 的研究 [11-12] 综合考虑了物和人的相似性信息, 给出了一个综合两方面信息的相似性度量用以指导推荐; 而 Zhang[13] 和 Liu[14] 则提出了一种网络模型, 以信息传递的物理模型为指导, 通过网络分析的方法指导推荐. 然而, 从建模的合理性和充分利用信息的角度考察这些模型可以发现: 对于 CF 方法中将两种信息进行集结的相似性公式, 其合理性还有待讨论; 通过网络方法建立模型的思想对本文的图模型建立有重要启示, 不过已有的文献 [13-14] 在处理信息时, 没有从信息最大化保留的角度进行推荐算法的设计, 而充分利用已有信息对于一个统计学习的问题则是重要的. 基于此, 本文以 Zhang[13] 和 Liu[14] 的网络建模思想为基础, 提出了基于图论的推荐模型, 从信息有效利用和最大化保留的角度进行推荐算法的创建, 并与既有的基准方法进行比较.

2 既有推荐方法的回顾

从方法继承和实验对比的角度考虑, 首先简要回顾推荐方法的基准模型, 也即基于物的相似性或基于人的相似性的推荐方法. 这类基准模型以定量化的评分数据为建模对象, 利用既有的评估信息训练模型, 并对新的样本进行推荐.

2.1 基于物的相似性的推荐模型

$$sim(\alpha,\beta)=\frac{\sum\limits_{i\in U_\alpha\cap U_\beta}(v_{i,\alpha}-\bar{v}_\alpha)\cdot(v_{i,\beta}-\bar{v}_\beta)}{\sqrt{\sum\limits_{i\in U_\alpha\cap U_\beta}(v_{i,\alpha}-\bar{v}_\alpha)^2}\sqrt{\sum\limits_{i\in U_\alpha\cap U_\beta}(v_{i,\beta}-\bar{v}_\beta)^2}} \tag{1}$$

这里, U_α 和 U_β 分别表示评过物品 α 和 β 的个体的集合, $v_{i,\alpha}$ 和 $v_{i,\beta}$ 分别表示第 i 个人对第 α 和第 β 件物品的打分, $\bar{v}_\alpha$ 和 $\bar{v}_\beta$ 分别表示第 α 和第 β 件物品的得分均值, 由此定义了两个物品的相似度 $sim(\alpha,\beta)$.

定义 Ω_i 为第 i 个人打过分的全部物品的集合, 选取其中与待评估物品 γ , 根据 (1) 式计算出的相似度, 令相似度最高的 k 个物品组成的集合 Θ, 则有 $\Theta\in\Omega_i$, 从而预测第 i 个人对第 γ 件物品的打分 $p_{i,\gamma}$ 为:

$$p_{i,\gamma}=\bar{v}_\gamma+\frac{\sum\limits_{\alpha\in\Theta}sim(\gamma,\alpha)(v_{i,\alpha}-\bar{v}_\alpha)}{\sum\limits_{\alpha\in\Theta}sim(\gamma,\alpha)} \tag{2}$$

其中, $\bar{v}_\gamma$ 表示第 γ 件物品已有得分的均值.

2.2 基于人的相似性的推荐模型

$$sim(i,j)=\frac{\sum\limits_{\alpha\in U_i\cap U_j}(v_{i,\alpha}-\bar{v}_i)\cdot(v_{j,\alpha}-\bar{v}_j)}{\sqrt{\sum\limits_{\alpha\in U_i\cap U_j}(v_{i,\alpha}-\bar{v}_i)^2}\sqrt{\sum\limits_{\alpha\in U_i\cap U_j}(v_{j,\alpha}-\bar{v}_j)^2}} \tag{3}$$

这里, U_i 和 U_j 分别表示第 i 个人和第 j 个人评过的全部物品的集合, $\bar{v}_i$ 和 $\bar{v}_j$ 分别表示第 i 和第 j 个个体评估全部物品的评分均值, 由此定义两个人的相似性 $sim(i,j)$. 注意到这个相似性是基于人们对物品评分的相近程度定义的.

令 Ω_α 为评过第 α 件物品全部的评分者的集合, 而对于第 i 个人, 其没有评过这件物品, 先要依据人与人的相关性来推测第 i 个人对第 α 件物品的评分, 由此定义集合中与 i 相关度最高的 k 个人组成的集合 Θ, 由此, 第 i 个人对第 α 件物品的评分的预测值 $p_{i,\alpha}$ 为:

$$p_{i,\alpha}=\bar{v}_i+\frac{\sum\limits_{j\in\Theta}sim(i,j)(v_{j,\alpha}-\bar{v}_j)}{\sum\limits_{j\in\Theta}sim(i,j)} \tag{4}$$

2.3 对基准方法的评述

从以上的符号描述可见, 基准方法对既有评价信息的利用有一定的不足: 根据物的相似性的方法会丢掉人之间相似性的信息, 根据人的相似性的方法会丢掉物之间相似性的信息. 在一个定量化的评价信息集中, 这两方面的信息都是存在的, 如果只考虑部分的信息, 对于一个良好的推荐方法来说是一种缺陷. 究其根源, 可以发现以上两种方法的思维是一种线性思维, 在一维空间中考察相似, 如果能够建立一个二维的模型, 包含两方面的信息, 则可以实现对信息更有效的利用, 以期提高推荐能力. 由此, 本文提出以下图模型.

3 评估图模型的建立

将在线的评分和推荐系统抽象为如下的模型: 含有 M 个个体 (将其编成 1 到 M 号) 和 N 种物品 (将其编成 1 到 N 号), 评价采用 α 分制 (α 为整数), 通常网站上采用的是 5 分制. 根据个体 m 对这些物品的评分情况, 将每个个体的评分向量 $v^m_{\alpha N\times 1}$ 定义如下:

$$\left.\begin{array}{l}\text{当第 } m \text{ 个个体没有评价第 } k \text{ 件物品时:}\\ \qquad v^m_{l\times 1}=0, l\in\{\alpha k-\alpha+1,\alpha k-\alpha+2,\cdots,\alpha k\}\\ \text{当第 } m \text{ 个个体对第 } k \text{ 件物品作出 } z \text{ 的评分时:}\\ \begin{cases} v^m_{l\times 1}=1, l=\alpha k-z+1\\ v^m_{l\times 1}=0, l\in\{\alpha k-\alpha+1,\alpha k-\alpha+2,\cdots,\alpha k\}\backslash\{\alpha k-z+1\}\end{cases}\\ \text{其中}, m\in\{1,2,\cdots,M\}, k\in\{1,2,\cdots,N\}, z\in\{1,2,\cdots,\alpha\}\end{array}\right\} \tag{5}$$

为了便于进一步分析, 将以上的评分向量转化为与之含有同样信息的第 m 个个体的评分矩阵 $A^m_{\alpha N\times\alpha N}$ 如下:

$$A^m_{\alpha N\times\alpha N}=\frac{v^m_{\alpha N\times 1}(v^m_{\alpha N\times 1})^{\mathrm{T}}-\mathrm{diag}(v^m_{\alpha N\times 1})}{n^m-1} \tag{6}$$

这里 n^m 表示第 m 个个体总共评价的物品数, $\mathrm{diag}(v^m_{\alpha N\times 1})$ 表示对角线为 $v^m_{\alpha N\times 1}$ 的对角矩阵, 同时规定当分子为零矩阵时, 除式为零矩阵. 综合全部 M 个个体的评价矩阵, 则有:

$$A_{\alpha N\times\alpha N}=\sum_{m=1}^{M}A^m_{\alpha N\times\alpha N}=\sum_{m=1}^{M}\frac{v^m_{\alpha N\times 1}(v^m_{\alpha N\times 1})^{\mathrm{T}}-\mathrm{diag}(v^m_{\alpha N\times 1})}{n^m-1} \tag{7}$$

矩阵 $A_{\alpha N\times\alpha N}$ 就是本文对于评分系统信息处理的工作: 将评估信息转化为矩阵的表述形式. 其基于图的理论看待这样一个评价问题, 这里将物品视为点 (这些点有着不同的分数属性), 将评估视为边将这些点联系起来, 构成了一个图的形式. 这里以一个 2 分制的评价系统为例, 令个体 m_1 评估了物品 1、物品 2 和物品 3 三种物品, 评分如图 1 所示, 个体 m_2 评估了物品 1 和物品 2 两种物品, 评分如图 2 所示.

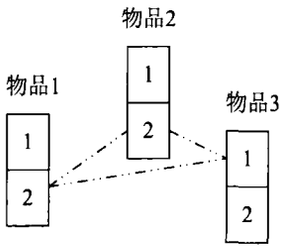

图 1 第 m_1 个个体对三种商品的评分图

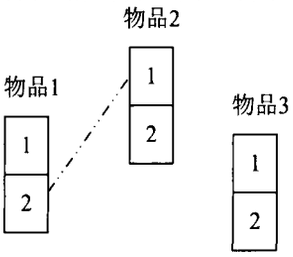

图 2 第 m_2 个个体对三种商品的评分图

则根据公式 (5), 第 m_1 个个体的评分向量为 $v^{m_1}_{6\times 1}=(1\ 0\vdots 1\ 0\vdots 0\ 1)^{\mathrm{T}}$, 第 m_2 个个体的评分向量为 $v^{m_2}_{6\times 1}=(1\ 0\vdots 0\ 1\vdots 0\ 0)^{\mathrm{T}}$, 由此根据公式 (6) 即可算得:

$$A_{6\times 6}^{m_1}=\begin{array}{c}\begin{array}{cccccc}\text{商品 1}&&\text{商品 2}&&\text{商品 3}&\\ \text{2 分}&\text{1 分}&\text{2 分}&\text{1 分}&\text{2 分}&\text{1 分}\end{array}\\ \left[\begin{array}{cc:cc:cc}0&0&1/2&0&0&1/2\\0&0&0&0&0&0\\\hdashline 1/2&0&0&0&0&1/2\\0&0&0&0&0&0\\\hdashline 0&0&0&0&0&0\\1/2&0&1/2&0&0&0\end{array}\right]\end{array}\begin{array}{l}\text{2 分}\\\text{1 分}\\\text{2 分}\\\text{1 分}\\\text{2 分}\\\text{1 分}\end{array}\begin{array}{l}\\\text{商品 1}\\\\\text{商品2}\\\\\text{商品3}\\\end{array}$$

注意到这个矩阵分块的方式对应着物品的种类, 在每个种类中对应着相应的分数, 如第一行第三列的 1/2 意味着对第 1 种物品打了 2 分, 对第 2 种物品打了 2 分, 其对应着图 1 中物品 1 和物品 2 之间相应的连线, 又由于共评价了 3 个物品, 所以除以 2. 同理可得到 $A_{6\times 6}^{m_2}$ 及 $A_{6\times 6}$.

由这个例子可以发现: $A_{\alpha N\times\alpha N}^{m}$ 的定义式实际上是对第 m 个个体评分图的矩阵表达, 这个矩阵与图示表达了相同的信息, 可以读出该个体评估的物品的评分及其评估的物品的数量 (考查各量分母的值加 1 即可), 而评分图就是该个体在系统中评分的刻画, 这个过程没有损失与评价有关的信息: 即通过图或矩阵, 可以完全还原已有的评价信息. 由此可知, 本文的信息处理方式相对于既有的推荐方法对已有信息的利用来说, 保留了更多的原始信息.

将全部个体的 $A_{\alpha N\times\alpha N}^{m}$ 求和就得到了整个系统的评分矩阵 $A_{\alpha N\times\alpha N}$, 这个矩阵刻画了系统中商品之间的联系, 而这些联系的产生, 即图的边是由每个个体通过评价产生的, 该矩阵包含了评价个体和被评价物品两方面的信息. 以上矩阵有如下的一些性质:

1) $A_{\alpha N\times\alpha N}^{m}$ 是对称矩阵, 行和为 1 或 0, 为 1 时表示第 m 个个体对相应行所对应的物品打了某一分数, 为 0 时表示没有对物品作出该行所对应分数的评价, 由于矩阵的对称性质, 列和也是一样;

2) $A_{\alpha N\times\alpha N}$ 的行和是所有 $A_{\alpha N\times\alpha N}^{m}$ 行和的和, 这一点可以从式 (7) 中得出, 这个行和表示在全部的个体中, 对于某一物品某个分数的支持人数.

4 推荐算法的建立

4.1 以信息最大化保留为目标的推荐方法

对于矩阵 $A_{\alpha N\times\alpha N}$, 为使列向量 $X_{\alpha N\times 1}$ 能反映出矩阵 $A_{\alpha N\times\alpha N}$ 最多的信息, 意味着下式达到最大值, 即:

$$\max_{X_{\alpha N\times 1}}\frac{(A_{\alpha N\times\alpha N}X_{\alpha N\times 1},X_{\alpha N\times 1})}{(X_{\alpha N\times 1},X_{\alpha N\times 1})}\tag{8}$$

对于目标函数的解释: 最大化问题相当于矩阵 $A_{\alpha N\times\alpha N}$ 在向量 $X_{\alpha N\times 1}$ 方向上的投影值, 该值越大, 反映向量 $X_{\alpha N\times 1}$ 包含了 $A_{\alpha N\times\alpha N}$ 更多的信息. 注意到: 上式求解的直接结果导致了求矩阵 $A_{\alpha N\times\alpha N}$ 的最大特征值及与之对应的特征向量 [15]. 当 $A_{\alpha N\times\alpha N}$ 是非负矩阵时, 根据 Perron-Frobinus 定理 [16], 若限定 $X_{\alpha N\times 1}$ 的列和为 1, 以上最优化问题存在唯一的非负解.

4.2 推荐算法的实现机制

为了得到对第 m 个个体的推荐, 需要对原始的 $A_{\alpha N\times\alpha N}$ 做些加工, 删去与第 m 个个体评估不一致的信息, 由此得到 $\tilde{A}_{\alpha N\times\alpha N}^{m}$, 其定义如下:

$$\left.\begin{array}{l}\text{令 } v_i^m \text{ 为第 } m \text{ 个个体的评分向量 } v_{\alpha N\times 1}^m \text{ 的第 } i \text{ 个元素,}\\ \text{当 } i\in\{i|v_i^m=0\} \text{ 时, } \tilde{A}_{i,:}^m=A_{i,:}, \text{即 } \tilde{A}_{\alpha N\times\alpha N}^m \text{ 中的第 } i \text{ 行与 } A_{\alpha N\times\alpha N} \text{ 的第 } i \text{ 行一样;}\\ \text{当 } i\in\{i|v_i^m\neq 0\} \text{ 时, } \tilde{A}_{i,:}^m=0, \text{即 } \tilde{A}_{\alpha N\times\alpha N}^m \text{ 中的第 } i \text{ 行元素全部取零.}\end{array}\right\}\tag{9}$$

以上的定义删去了 $A_{\alpha N\times\alpha N}$ 与第 m 个个体不一致的评分, 保留了该个体没有评过的物品的评分信息及与该个体一致的评分信息.

对于 (8) 式, 将其中的 $A_{\alpha N\times\alpha N}$ 换做 $\tilde{A}_{\alpha N\times\alpha N}^{m}$, 求出相应的解的 $X_{\alpha N\times 1}^{m}$, 定义 x_i 为 $X_{\alpha N\times 1}^{m}$ 的第 i 个

元素, 将对某个物品各个档次评分的向量元素进行归一化处理:

$$\hat{x}_i = \frac{x_i}{\sum\limits_{j=1}^{\alpha} x_{(\lceil \frac{i}{\alpha} \rceil - 1)\times\alpha + j}} \tag{10}$$

这里 $\lceil \frac{i}{\alpha} \rceil$ 表示对 $\frac{i}{\alpha}$ 取上整.

则第 m 个人对第 n 个物品的预测评分 g_n^m 计算如下:

$$g_n^m = \sum_{i=1}^{\alpha} \hat{x}_{\alpha(n-1)+i} \times (\alpha + 1 - i) \tag{11}$$

由此根据预测评分的大小排序, 可以对第 m 个人进行商品的推荐.

4.3 推荐方法的一个算例

人为给定一个 2 分制的含有 3 种物品和 34 个个体参与的评分样本, 经整理得到如下的系统评分矩阵:

$$A = \left[\begin{array}{cccccc} 0 & 0 & 0 & 1 & 23/2 & 3/2 \\ 0 & 0 & 1 & 0 & 3/2 & 23/2 \\ \hline 0 & 1 & 0 & 0 & 7/2 & 1/2 \\ 1 & 0 & 0 & 0 & 7/2 & 1/2 \\ \hline 23/2 & 3/2 & 7/2 & 7/2 & 0 & 0 \\ 3/2 & 23/2 & 1/2 & 1/2 & 0 & 0 \end{array}\right]$$

如果已经得到某个人的评估信息, 比如该人喜欢第三种物品, 给了该物品 2 分的评价. 那么根据本文的推荐方法, 根据 4.2 中的说明, 由于上面矩阵 A 的第 6 行表示对商品 3 评价为 1 分, 与该人给出的评价信息不符, 将 A 的第 6 行的元素全部归零, 得到:

$$\tilde{A} = \left[\begin{array}{cccccc} 0 & 0 & 0 & 1 & 23/2 & 3/2 \\ 0 & 0 & 1 & 0 & 3/2 & 23/2 \\ \hline 0 & 1 & 0 & 0 & 7/2 & 1/2 \\ 1 & 0 & 0 & 0 & 7/2 & 1/2 \\ \hline 23/2 & 3/2 & 7/2 & 7/2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{array}\right]$$

根据将 $\tilde{A}$ 代入 (8) 式及 X 行和为 1 的约束可以求得:

$$X = (0.3422, 0.0515; 0.1051, 0.1276; 0.3735, 0)$$

对每个物品归一化后的结果为: $\hat{X} = (0.8692, 0.1308; 0.4517, 0.5483; 1, 0)$

由此基于 (11) 式可以求得该人对于另外两种物品的打分: 第一种物品为 1.869, 第二种物品为 1.452. 同时可以发现, 该方法保持了该人对第三种物品的打分没变: 还为 2. 通过这个结果可以推断, 该人对于第一种物品也倾向于喜欢, 而对第二种物品的评价不高.

4.4 推荐算法的程序实现

为提高计算的速度和在线评估的需要, 本文采用以下方法计算 (8) 中的最大化问题, 这个算法就是求矩阵最大特征值和特征向量的方法, 对于稀松矩阵来说, 简单高效, 对对称与非对称的矩阵问题都是适用的, 是整个推荐算法最核心的计算步骤.

5 实验分析及方法对比

5.1 实验数据的来源

选取一个在线电影客户的打分数据集用以比较这几种推荐方法, 在该集合中包含评论者对电影的打分数据, 该打分的体制为 1 分到 5 分, 1 分最差, 5 分最好. 数据集的网址为 www.gruoplens.org, 该网址上提供了两个不同大小的打分数据集, 这里选取有 943 个人参与, 对 1682 部电影打分的数据, 总的评分数据量约为 100000. 根据第 3 部分评估图模型的建立原理, 将其读成相应的等价矩阵的形式.

推荐算法的伪代码

输入: A;　A 是一个的非负矩阵
输出: λ, V;　λ 是矩阵的最大特征值, V 是相应的特征向量
CalProcess:
$V \leftarrow [1.0, 1.0, \cdots, 1.0]^{\mathrm{T}}$
$TR \leftarrow sum(A)$;　' 对 A 的每行求和存入 TR 中
WHILE(TRUE)
{
　$R \leftarrow TR$;
　$\min \leftarrow \{\min(R) \backslash 0\}$;　求 R 中不为 0 的最小分量
　$\max \leftarrow \max(R)$;　　求 R 中的最大分量
　$V \leftarrow V \cdot (R/\max)$;
　$delta \leftarrow (\max - \min)/\max$;
　If $\{\delta < 0.0000001\}$, then BREAK;
　FOR $i \leftarrow 0$ to $n-1$
　　If $(R[i] = 0)$
　　　CONTINUE;
　FOR $j \leftarrow 0$ to $n-1$
　　$A[i][j] \leftarrow A[i][j]/R[i]$;
　　$A[i][j] \leftarrow A[i][j] \times R[j]$
　$TR \leftarrow sum(A)$;
}
$\lambda = (\max + \min)/2$;
返回 λ, V;

5.2 推荐的实验结果

为了验证推荐算法的有效性和准确性, 这里对原始的数据集进行一个随机的划分, 将其分成 80% 的样本集和 20% 的测试集. 将推荐算法作用在样本集上, 得到相关的评估矩阵, 而后计算测试集中出现的人及其对相应电影的评分, 将计算出的评分与测试集中真实的评分进行对比, 选择对比的指标 MAE 定义如下:

$$MAE = \frac{\sum_{n=1}^{N} |v_{i,\alpha}^{n} - round(\hat{v}_{i,\alpha}^{n})|}{N} \tag{12}$$

N 是测试集中所有的评分数据, $v_{i,\alpha}^{n}$ 表示其中的第 n 条数据, 反映第 i 个人对第 α 部电影的真实评分, $\hat{v}_{i,\alpha}^{n}$ 表示根据推荐算法得到的第 i 个人对第 α 部电影评分的预测值, $round()$ 表示按四舍五入的方式取整.

定义推荐的准确率 D 如下, 其中, $\aleph(|v_{i,\alpha}^{n} - round(\hat{v}_{i,\alpha}^{n})| = 0)$ 表示评估准确的数目.

$$D = \frac{\aleph(|v_{i,\alpha}^{n} - round(\hat{v}_{i,\alpha}^{n})| = 0)}{N} \tag{13}$$

根据以上算法[1] , 得到四种评估方式的 MAE 值及准确率 D 列表如表 1.

表 1　几种推荐方法的准确度对比

推荐方法名称	MAE	D
本文方法	0.78	37.1%
基于物的相似性	0.88	31.6%
基于人的相似性	0.86	32.3%
随机推荐	1.38	21.7%

同时, 将错评的分数 $|v_{i,\alpha}^{n} - round(\hat{v}_{i,\alpha}^{n})|$ 做饼图如图 3.

1. 这里的计算是基于 5 次随机划分分别计算各次的 MAE 和 D 而后取平均的结果. 其中在应用基于物或人的相似的算法时, 有一个对相似个体数量的选择问题, 这里的选取标准是基于该方法有最低的 MAE 选取相应的在 (2) 和 (4) 式中的 k 值以确定相应的集合 Θ , 这样做保证了基于物或人的相似性的方法能达到最优的效果, 由此进行比较才更有说服力.

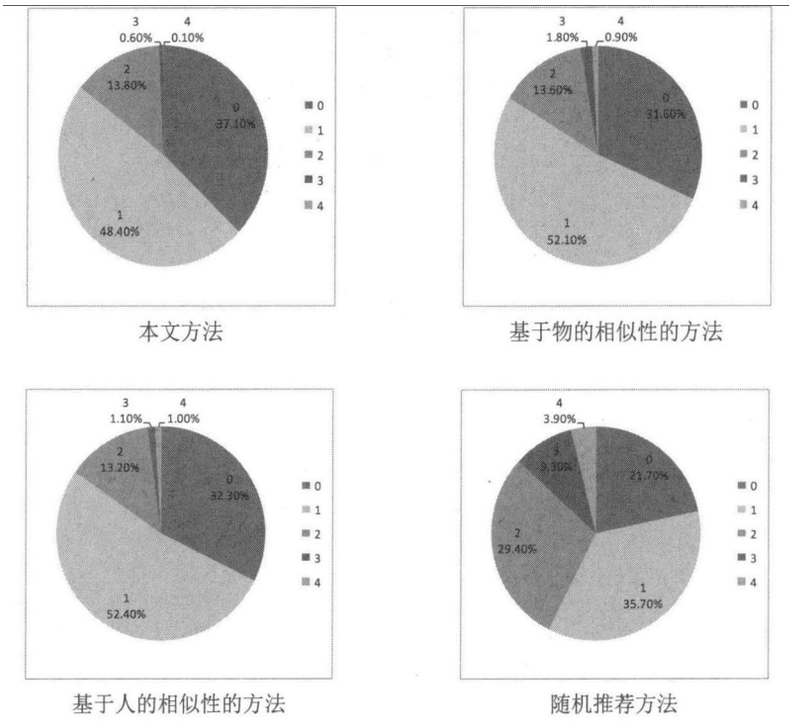

图 3 各种方法错评分数的饼图

读表及图可以发现: 本文的推荐方法正确率占比最高, 同时错评在 2 以上的占比最小, 具有最小的平均错误率 MAE, 而无论哪种方法都显著优于随机指定一个评分的推荐方法, 说明以上方法都是有效的.

同时, 由统计模型的计算时间可以发现: 对于本文实证的含有 100000 个评分数据的样本, 在 2G 内存的个人计算机上, 本文方法计算的平均用时为 8.8 秒, 而基于物或人的相似性推荐方法的平均用时分别为 6.5 秒和 4.9 秒, 随机方法不到 1 秒. 综合以上比较得知, 本文的方法准确度高, 同时计算时间短, 与两种相似性方法在一个水平上, 可用于实时的在线推荐.

6 结论及进一步的工作

在图论的原理下, 将评价人给出的评价信息转化为图的表示形式, 这样的表示形式保留了原有的评论信息, 为推荐奠定了基础. 根据图与矩阵的对应表达, 使得用数学的方法处理信息较为方便, 本文应用了矩阵投影原理, 使得矩阵的信息得到最大化地保留. 实验表明: 本文设计的算法是实用高效的, 可以在较短的时间内, 相比于既有的基于物和基于人的相似性的推荐算法, 较为准确的给出推荐信息, 准确率达 37%, 平均误差 0.78, 比上面两种方法在准确率方面提高了 5% 左右, 在平均误差方面提高了 0.10 左右.

本文方法只讨论了数据评估的形式, 进一步的工作可以从下面两个方面展开: 其一, 扩展可分析的评论方式, 比如包括文字性评价等; 其二, 进一步挖掘客户的信息, 客户之间的联系在本文中是通过对商品评分建立起来的, 其还可以有更多的联系方式, 比如性别、职业、爱好等, 可以考虑利用本文提出的图模型及信息最大化保留的推荐思想解决这类问题.

参考文献

[1] Sulin B, Paul A P. Evidence of the effect of trust building technology in electronic markets: Price premiums and buyer behavior[J]. MIS Quarterly, 2002, 26(3): 243–268.

[2] Paul A P, David G. Building effective online marketplaces with institution-based trust[J]. Information Systems Research, 2004, 15(1): 37–59.

[3] Paul A P, Liang H, Xue Y. Understanding and mitigating uncertainty in online exchange relationships: A principal-agent perspective[J]. MIS Quarterly, 2007, 31(1): 105–136.

[4] Park D H, Kim S. The effects of consumer knowledge on message processing of electronic word-of-mouth via online consumer reviews[J]. Electronic Commerce Research and Applications, 2008, 7(4): 399–410.

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

[6] Jonathan L H, Joseph A K, Loren G T, et al. Evaluating collaborative filtering recommender systems[C]//ACM Transactions on Information Systems, 2004, 22(1): 5–53.
[7] Pan X, Deng G S, Liu J G. Information filtering via improved similarity definition[J]. Chinese Physical Letters, 2010, 27(6): 068903.
[8] Wang J, Stephen R, Arjen P D V, et al. Probabilistic relevance ranking for collaborative filtering[J]. Information Retrieval, 2007, 11(6): 477–497.
[9] Azene Z, Anthony F N. Representation, similarity measures and aggregation methods using fuzzy sets for content-based recommender systems[J]. Fuzzy Sets and Systems, 2009, 160(1): 76–94.
[10] Asela G, Guy S. A survey of accuracy evaluation metrics of recommendation tasks[J]. The Journal of Machine Learning Research, 2009, 10(12): 2935–2962.
[11] Panagiotis S, Alexandros N, Apostolos N P, et al. Nearest-balusters' collaborative filtering based on constant and coherent values[J]. Information Retrieval, 2008, 11(1): 51–75.
[12] Liu R R, Jia C X, Zhou T, et al. Personal recommendation via modified collaborative filtering[J]. Physica A: Statistical Mechanics and Its Applications, 2009, 388(4): 462–468.
[13] Zhang Z K, Zhou T, Zhang Y C. Personalized recommendation via integrated diffusion on user-item-tag tripartite graphs[J]. Physica A: Statistical Mechanics and Its Applications, 2010, 389(1): 179–186.
[14] Liu J G, Zhou T, Wang B H, et al. Degree correlation of bipartite network on personalized recommendation[J]. International Journal of Modern Physics C, 2010, 21(1): 137–147.
[15] 袁玉波. 数据挖掘与最优化技术及其应用 [M]. 北京: 科学出版社, 2007.
Yuan Y B. Data Mining & Optimization Technology and Its Applications[M]. Beijing: Science Press, 2007.
[16] David G L. Introduction to Dynamic Systems — Theory, Models and Applications[M]. 袁天鑫, 黄午阳, 译. 上海: 上海科学技术文献出版社, 1985.