

一种新颖的协同推荐算法研究

李 慧¹, 胡 云^{1,2}, 李存华¹

(1 淮海工学院 计算机工程学院, 江苏 连云港 222002; 2 南京大学 信息工程学院, 江苏 南京 110004)

摘 要: 为了克服协同推荐系统中的用户评分数据稀疏性和推荐实时性差的问题, 提出了一种高效的基于粗集的个性化推荐算法. 该算法首先利用维数简化技术对评分矩阵进行优化, 然后采用分类近似质量计算用户间的相似性形成最近邻居, 从而降低数据稀疏性和提高最近邻寻找准确性. 实验结果表明, 该算法有效地解决用户评分数据极端稀疏情况下传统相似性度量方法存在的问题, 显著地提高推荐系统的推荐质量.

关键词: 推荐系统; 最近邻; 用户相似性; 维数简化

中图分类号: TP39

文献标识码: A

文章编号: 1000-7180(2012)03-0069-04

A New Collaborative Recommendation Algorithm Research

LI Hui¹, HU Yun^{1,2}, LI Cun-hua¹

(1 Department of Computer Science, Huaihai Institute of Technology, Lianyungang 222002, China;

2 Department of Information Engineering, Nanjing University, Nanjing 110004, China)

Abstract: In order to overcome the problem of traditional collaborative filtering algorithm faces severe challenge of sparse user ratings and real-time recommendation, a high efficient personalization recommendation algorithm based on rough set is proposed. The algorithm refine the user ratings data using dimensionality reduction, then uses the quality of approximation of classification to find the target users' neighbors. Thus the sparsity can be decreased and the accuracy of searching nearest neighbor can be improved. The experimental results show that this method can efficiently improve the extreme sparsity of user rating data, and provide better recommendation results than traditional collaborative filtering algorithms.

Key words: recommendation system; nearest neighbor; user similarity; dimensionality reduction

1 引言

随着互联网的膨胀和电子商务的出现, 信息迷茫困扰着每个用户, 为了帮助用户获取真正所需的信息, 必然会出现一种推荐系统. 最近邻协同过滤推荐是当前最成功的推荐技术^[1], 目前主要有两类协同过滤推荐算法: 基于用户的协同过滤推荐算法^[2]和基于项目的协同过滤推荐算法^[3]. 基于用户的协同过滤推荐算法根据目标用户的最近邻居(最相似的若干用户)对某个项目的评分逼近目标用户对该项目的评分^[2]. 基于项目的协同过滤推荐算法认为, 用户对不同项目的评分存在相似性, 当需要估计用

户对某个项目的评分时, 可以用用户对该项目的若干相似项目的评分进行估计^[3]等. 在此基础上, 研究人员提出了一些改进方法, 例如基于项目的协同过滤及其改进算法^[4]、基于矩阵降维的协同过滤^[5]、基于神经网络的协同过滤^[6]等.

然而, 随着电子商务站点用户和商品项数量的不断增加, 协同过滤面临严峻的用户评分数据稀疏性和推荐实时性挑战, 导致推荐质量迅速下降. 针对这种情况, 本文提出了一种新颖的基于粗集的协同推荐算法: 该算法首先利用维数简化技术对评分矩阵进行优化, 在此基础上, 采用分类近似质量计算目标用户的最近邻居, 使得最近邻搜索更加准确, 使推

荐系统的实时性和准确性得到改善. 实验结果表明本方法在用户评分项目较少的情况下, 推荐效果优于传统的协同过滤方法, 显著提高了推荐质量.

2 问题描述与相关定义

2.1 问题描述

协同过滤推荐基于这样一个假设: 如果用户对一些项目的评分相似, 则他们对其他项目的评分也较为相似. 基于该思想可以将给定的用户项目矩阵映射为一个不完备信息系统, 下面先给出不完备信息系统的概念.

四元组 $S = \langle U, A, V, f \rangle$ 称为信息系统, 其中 U 为非空对象集合; $A = C \cup D$ 为非空的属性集合, C 表示条件属性, D 表示决策属性; V 为属性值的集合; $f: U \times A \rightarrow V$ 为信息函数, 它指定 U 中每一个对象 x 的属性值. 对象 x_i 在属性 a 上的取值用 $a(x_i)$ 表示. 若信息系统 S 中至少存在一个对象其属性值缺省, 则称 S 为不完备信息系统. 对于对象 x_i 在属性 a 的缺省值, 表示为 $a(x_i) = *$.

将给定的用户项目矩阵映射为一个不完备信息系统 $S = \langle U, A, V, f \rangle$, $U = \{x_1, x_2, \dots, x_m\}$ 为用户组成的论域 (不包括目标用户 x_0); $A = \{a_1, a_2, \dots, a_{n-1}\} \cup \{a_n\}$, $\{a_n\}$ 为决策属性, 表示目标用户 x_0 未评分的项目, $\{a_1, a_2, \dots, a_{n-1}\}$ 为条件属性, 表示 x_0 已评分的项目集合; $V = \bigcup_{a_i \in A} V_{a_i}$ 是评分值的集合, V_{a_i} 表示项目 a_i 评分值的范围; 信息函数 $f: U \times A \rightarrow V$ 指定每一个用户对项目的评分, $a_k(x_i)$ 表示用户 x_i 对项目 a_k 的评分. 若目标用户 x_0 对项目 a_i 未评分, 则令 $a_i(x_0) = 0$. 然后运用粗集理论中分近似质量, 计算用户项目分类间的近似质量, 度量用户间的相似性, 形成最邻居集.

2.2 相关定义

粗集理论是一种处理不精确、不一致、不完备信息问题的方法. 这里先简要的介绍一下与本文相关的粗集概念.

定义 1 近似空间: 二元序对 $K = \langle U, R \rangle$ 称为近似空间. 这里 U 有限的非空对象集合, 称为论域; R 是 U 上的一个等价关系, 将 U 划分为若干个等价类. 对于 $\forall x \in U, [x]_R$ 表示包含 x 的 R 的等价类.

定义 2 近似集: 给定知识表达系统 $S = \langle U, A, V, f \rangle$, U 是一个有限的非空对象集合, 称为论域, 对于每个子集 $X \subseteq U$ 和不分明关系 R , X 的上近

似集和下近似集可以定义如下:

$$R_-(X) = \{x \in U: [x]_R \subseteq X\} \quad (1)$$

$$R^-(X) = \{x \in U: [x]_R \cap X \neq \emptyset\} \quad (2)$$

$R^-(X)$ 是根据知识 R, U 中一定能归入 X 的元素的集合, $R_-(X)$ 是根据知识 R, U 中可能归入 X 的元素的集合, $[x]_R$ 表示包含 x 的 R 的等价类.

定义 3 分类近似质量: 设集合簇 $F = \{X_1, X_2, \dots, X_n\}$, $(U = \bigcup_{i=1}^n X_i)$ 为论域上的知识, U 是一个有限的非空对象集合, 称为论域, R 为一个属性子集, F 在 R 下近似集簇 $R_-(F) = \{R(X_1), R(X_2), \dots, R(X_n)\}$, 则定义 R 在分类 F 的分类近似质量 $r_R(F)$ 为:

$$r_R(F) = \frac{\sum_{i=1}^n |R_-(X_i)|}{|U|} \quad (3)$$

3 基于粗集的协同推荐算法

为了解决用户评分数据的极端稀疏性, 有必要对这种矩阵表示方式做优化, 维数简化是其中较好的一种方法, 本文应用单值分解技术对用户评分矩阵进行维数简化. 而且为了有效地解决用户评分数据极端稀疏情况下传统相似性度量方法存在的不足, 本文提出了一种新颖的算法来产生目标对象的最近邻居, 使得最近邻搜索更加准确.

下面详细介绍基于粗集的协同推荐算法. 算法主要分为三步: 维数简化、寻找最近邻居和产生推荐.

3.1 维数简化

单值分解 (singular value decomposition, SVD) 是一种矩阵分解技术^[7], 它可将一个 $m \times n$ 的矩阵 R 分解为 3 个矩阵.

$$R = T_0 S_0 D'_0, S_0 = \text{diag}(\sigma_1, \dots, \sigma_r) \quad (4)$$

其中, $\sigma_1 \geq \dots \geq \sigma_r \geq 0$, T_0 和 D_0 分别是 $m \times r$ 和 $n \times r$ 的正交矩阵, r 是矩阵 R 的秩 ($r \leq \min(m, n)$). S_0 是一个 $r \times r$ 的对角矩阵, 所有的 σ_r 大于 0 并按照大小顺序排列, 称为单值 (singular value). 通常对于矩阵 $R = T_0 S_0 D'_0$, T_0, S_0, D_0 必须是满秩的. 但单值分解有一个优点, 它允许存在一个简化的近似矩阵. 对于 S_0 , 保留 k 个最大的单值, 将其余的用 0 来替代, 这样, 就可以将 S_0 简化为仅有 k 个单值的矩阵 ($k < r$). 因为引入了 0, 可以将 S_0 中的值为 0 的行和列删除, 得到一个新的对角矩阵 S , 如果矩阵 T_0, D_0 据此简化得到矩阵 T, D , 那么有重构的矩阵

$R_k = TSD', R_k \approx R$. 单值分解能够生成初始矩阵 R 的所有秩等于 k 的矩阵中与矩阵 R 最近似的一个.

本文将单值分解应用到推荐系统中,首先将矩阵 R 中评估值为 0 的稀疏项用相关列的平均值代替,即项的平均评估值.接着将矩阵每行规范化为相同长度,用 $r_{ij} - r_i'$ 代替原来的 r_{ij} (r_i' 是相关列的项的平均评估值).进行规范化的目的是因为选择不同数量项的用户对相似度计算结果的影响不同,容易造成偏差,规范化为相同长度后,选择项数目较多的用户对相似度计算结果的影响降低了.经过这样的处理,得到矩阵 R' 并将其做为之后近邻计算与产生推荐的评分矩阵.

应用单值分解技术对用户评分矩阵进行维数简化后,大大降低了它的数据稀疏性.

3.2 寻找最近邻居

经过评分矩阵优化后,不完备信息系统转化为完备信息系统 $S' = \langle U', A', V, f' \rangle$. 为了计算用户间的相似性,寻找目标用户 x_0 的最近邻居,具体步骤如下:

第一步:将 $\{x_0\}$ 加入用户集中,即 $U' = U' + \{x_0\}$;

第二步:并将 x_0 未评分的项目 a_n 从项目集合 A 中删除;

第三步:将信息系统 S' 对应的用户项目矩阵转置生成新的完备信息系统 $S = \langle U, A, V, f \rangle$, 其中 $U = A' - \{a_n\} = \{a_1, a_2, \dots, a_{n-1}\}$, $A = U' \cup \{x_0\} = \{x_0\} \cup \{x_1, x_2, \dots, x_m\}$, 其中 $\{x_0\}$ 为决策属性集合, $\{x_1, x_2, \dots, x_m\}$ 为条件属性集合. 为了统一,用户 x_i 对项目的评分仍记作 $a_k(x_i)$.

第四步:对于新生成的完备信息系统 S , x_0, x_i 根据项目评分对项目集 U 进行分类,令 x_0 在 U 的分类为 F , $F = \{Y_1, Y_2, \dots, Y_n\}$, x_i 在 U 的分类为 R_i , $R_i = \{X_1, X_2, \dots, X_m\}$, 则根据 R_i, F 的近似质量

做为 $r_{R_i}^{(F)} = \frac{\sum_{j=1}^n |R_i - (X_j)|}{|U|}$, 并且把用户 x_0, x_i 的

相似性 $\text{sim}(x_0, x_i)$ 定义为: $\text{sim}(x_0, x_i) = r_{R_i}^{(F)}$.

第五步:生成最近邻居集依次计算出目标用户 x_0 与用户空间中其他用户的相似性,然后取相似性从大到小排列的前 K 个用户作为 x_0 的最近邻居集 $NBS = \{x_1, x_2, \dots, x_K\}$.

3.3 产生推荐

通过 3.2 节中提出的产生近邻的方法得到目标用户的最近邻居,下一步需要产生相应的推荐. 设用

户 u 的最近邻居集合用 NBS_u 表示,则用户 u 对项目 i 的预测评分 $P_{u,i}$ 可以通过用户 u 对最近邻居集合 NBS_u 中项目的评分得到,计算方法如下:

$$P_{u,i} = \bar{R}_u + \frac{\sum_{n \in NBS_u} \text{sim}(u, n) \times (R_{n,i} - \bar{R}_n)}{\sum_{n \in NBS_u} (|\text{sim}(u, n)|)} \quad (5)$$

$\text{sim}(u, n)$ 表示用户 u 与用户 n 之间的相似性, $R_{n,i}$ 表示用户 n 对项目 i 的评分, $\bar{R}_n$ 分别表示用户 u 和用户 n 对项目的平均评分. 然后按 P 值从大到小取前 N 个项目组成 $top-N$ 推荐集 $rec = \{i_1, i_2, \dots, i_N\}$ 并推荐给目标用户 u , 从而完成整个推荐过程.

4 实验结果与分析

4.1 数据集与度量标准

MovieLens(<http://movielens.lima.edu>) 是一个基于 Web 的研究型推荐系统,其中包含 943 位用户对 1682 部电影的 100000 条评分数据,每位用户至少对 20 部电影进行了评分,所有电影分属于 19 种电影类别. 为了分析实验数据集的样本量成倍增加时对算法性能的影响,本文从这个数据集中随机抽取 100, 200, 300 名用户的评分数据组成 3 个数据集,分别记为 DB100, DB200, DB300, 表 1 给出了这 3 个数据集各自的用户数量、电影数量和评分数量等信息. 实验采取 All but one 协议将实验数据集每个用户的评分数据随机隐藏 1 个组成测试集,然后基于其他评分数据(即训练集)来预测这些被隐藏评分.

表 1 实验数据集

统计项目	数据集		
	DB100	DB200	DB300
用户数量	100	200	300
电影数量	1292	1436	1486
评分数量	10143	19255	29826

本文采用平均绝对偏差 MAE 作为度量标准. 平均绝对偏差 MAE 通过计算预测的用户评分与实际的用户评分之间的偏差度量预测的准确性, MAE 越小, 推荐质量越高. 设预测的用户评分集合表示为 $\{p_1, p_2, \dots, p_N\}$, 对应的实际用户评分集合为 $\{q_1, q_2, \dots, q_N\}$, 则平均绝对偏差 MAE 定义为:

$$\text{MAE} = \frac{\sum_{i \in N} |P_i - q_i|}{N} \quad (6)$$

4.2 实验结果与分析

本组实验基于稀疏等级最小的 DB100 数据集进行. 实验以传统的协同过滤推荐算法作为对照, 在传统的协同过滤推荐算法中, 分别以余弦相似性和相关相似性作为相似性度量标准, 计算其 MAE, 邻居个数从 4 增加到 20, 间隔为 4, 然后与本文的 rough-based CF 作比较, 实验结果如图 1 所示.

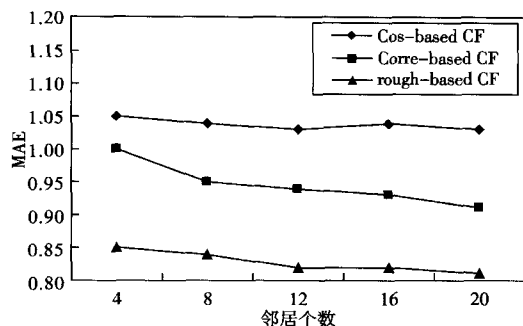


图1 推荐算法推荐精度比较

实验结果表明, 采用相关相似性这种相似性度量方法的推荐质量最低. 如果采用余弦相似性, 则用户对所有未评分项目的评分均相同, 这显然是不精确的. 在本文提出的基于粗集的协同过滤推荐算法中, 首先通过维数简化技术对用户评分矩阵进行补齐, 使得用户之间拥有更多的共同评分项目, 从而降低用户项目矩阵的稀疏性. 然后利用近似分类质量进行最近邻居的计算, 使得推荐系统中邻域的计算更加准确. 实验结果证明, 本文提出的协同过滤推荐算法推荐质量最高.

由表 1 可知, DB100, DB200, DB300 评分数量分别为 10143, 19255, 29826, 其比例约为 1 : 2 : 3. 本组实验在这 3 个数据集上运行本文算法并计算相应的 MAE. 实验结果如图 2 所示.

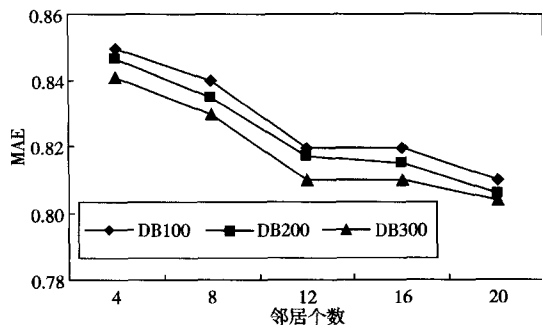


图2 数据集大小对算法的影响

从图 2 可以看出, 本文提出的基于粗集的协同推荐算法 rough-based CF 在 DB300 上的 MAE 最小, 而在 DB100 上的 MAE 最大, 在最近邻数量增

加到 20 时, 算法在三个数据集上的 MAE 最为接近. 由此可知, 当实验数据集的样本量成倍增加时 rough-based CF 的推荐质量随样本量增加而得到小幅度提高. 因此当实验数据集变大, 而用户及评分数量随之增多时, 本文的 rough-based CF 算法更容易找到相似性高的邻居集, 从而预测用户评分项更准确, 使用户能够得到更加满意的推荐.

5 结束语

针对用户评分数据稀疏问题, 本文提出的结合维数简化和基于粗集的个性化推荐方法, 不仅减少了计算耗费, 同时使最近邻寻找更加准确. 实验结果表明, 本文提出的基于粗集的协同过滤推荐算法有效地解决数据稀疏问题, 能有效提高推荐质量, 具有一定的有效性和可行性.

参考文献:

- [1] You Wen, Ye Shui-sheng. A survey of collaborative filtering algorithm applied in E-commerce recommender system[J]. Computer Technology and Development, 2006, 16(9): 70-72.
- [2] Chuan Y, Jie-ping X. Recommendation algorithm combining the user-based classified regression and the item-based filtering[C]//Processing of the International Conference on Electronic Commerce. Canada: New Brun Swick, 2006: 574-578.
- [3] Kim B M, Li Q, Park C S, et al. A new approach for combining content-based and collaborative filters[J]. Journal of Intelligent Information System, 2006, 27(1): 79-91.
- [4] 项武. 基于评分区间相似性的协同过滤推荐算法[J]. 微电子学与计算机, 2010, 27(7): 125-132.
- [5] Zhou Junfeng, Tang Xian, Guo Jingfeng. An optimized collaborative filtering recommendation algorithm [J]. Journal of Computer Research and Development, 2004, 41(10): 1842-1847.
- [6] Zhang Feng, Chang Huiyou. Employing BP neural networks to alleviate the sparsity issue in collaborative filtering recommendation algorithms [J]. Journal of Computer Research and Development, 2006, 43(4): 667-672.
- [7] Zhang S, Wang W H, Ford J, et al. Using singular value decomposition approximation for collaborative filtering[C]//Proceedings of 7th IEEE International Conference on E-commerce Technology. Germany, 2005: 257-264.

(下转第 77 页)

- [4] 朱义毅, 缪淮扣, 唐新香. 一种基于 Object-Z 规格说明的集成测试用例生成方法[J]. 上海大学学报: 自然科学版, 2009, 15(1): 60-65.
- [5] Ashraf A, Nadeem A. Automating the generation of test cases from Object-Z specifications[C]// 30th Annual International Computer Software and Applications Conference (COMPSAC '06). Chicago, Compsac, 2006: 101-104.
- [6] Vilkomir S A, Swain W T, Poore J H. Combinatorial test case selection with markovian usage models[C]// Fifth International Conference on Information Technology: New Generations (itng 2008). Las Vegas, IEEE Computer Society, 2008: 3-8.
- [7] Prowell S J. Computations for markov chain usage models[R]. Tennessee: The University of Tennessee UT-CS-03-505, 2003.
- [8] 于敏, 何正友, 钱清泉. 基于 Markov 过程的硬/软件综合系统可靠性分析[J]. 电子学报, 2010, 38(2): 473-479.
- [9] 缪淮扣, 李刚, 朱关铭. 软件工程语言——Z[M]. 上海: 上海科学技术文献出版社, 2000.
- [10] Sapna P G, Mohanty H. Automated test scenario selection based on levenshtein distance[C]// The 6th International Conference on Distributed Computing and Internet Technologies (ICDCIT 2010). Bhubaneswar, Springer Berlin/Heidelberg, 2010: 255-266.

作者简介:



仲晓敏 女, (1977—), 硕士, 讲师. 研究方向为软件工程、人工智能.

(上接第 72 页)

作者简介:



李 慧 女, (1979—), 硕士研究生, 讲师. 研究方向为数据挖掘、智能信息处理.

胡 云 女, (1978—), 硕士研究生, 讲师. 研究方向为人工智能.

李存华 男, (1963—), 博士研究生, 研究员, 教授. 研究方向为数据挖掘、人工智能.