第29卷 第7期
2007年7月

武 汉 理 工 大 学 学 报
JOURNAL OF WUHAN UNIVERSITY OF TECHNOLOGY

Vol. 29 No. 7
Jul. 2007

融合用户和项目相关信息的协同过滤算法研究

王惠敏,聂规划
(武汉理工大学经济学院,武汉 430070)

摘 要: 针对User-based协同过滤和Item-based协同过滤算法的不足,提出了一种新的推荐算法。该算法融合用户-项目评分数据集所包含的用户相关和项目相关的信息来推荐商品,并且利用模糊聚类技术分别将相似的项目和相似的用户聚类,改善传统推荐算法的数据稀疏性和可扩展性问题。实验结果表明,将用户相关和项目相关的信息融合能够提供更好的推荐。

关键词: 协同过滤; 模糊聚类; 推荐系统; 信息融合
中图分类号: TP 391 文献标志码: A 文章编号: 1671-4431(2007)07-0160-04

Research on Collaborative Filtering Algorithm Based on Fusing User and Item's Correlative Information

WANG Hui-min, *NIE Gui-hua*
(School of Economy, Wuhan University of Technology, Wuhan 430070, China)

Abstract: Aiming at the disadvantages of user-based collaborative filtering and item-based collaborative filtering algorithms, the paper proposed a novel recommendation algorithm that generated item's recommendation by fusing user and item's correlative information inhering in the user-item rating dataset. The algorithm also involved the fuzzy clustering of similar items and similar users to improve the data sparsity and scalability of traditional collaborative filtering algorithms. Experiments showed a better recommendation could be provided by fusing user and item's correlative information.

Key words: collaborative filtering; fuzzy clustering; recommendation system; information fusion

电子商务推荐系统是基于可得到的信息资源向用户推荐适合其需要的信息或商品的系统[1]。架构推荐系统的推荐方法主要有2种:基于内容的推荐和协同过滤推荐[2]。协同过滤作为目前最成功的推荐算法被广泛地应用,其目标是根据具有相似偏好的用户的观点向目标用户推荐新的商品[3]。Memory-based协同过滤算法利用整个用户-项目评分数据集来产生推荐,系统利用统计技术搜寻一组用户,称为邻居,他们与目标用户有一致的历史偏好[4]。Memory-based协同过滤算法主要有User-based协同过滤推荐算法和Item-based协同过滤推荐算法。User-based协同过滤根据评分相似的最近邻居的评分数据向目标用户产生推荐。通常在电子商务网站中,用户购买或评分的商品相对于总商品数量仅占有限的百分比,这导致用户-项目评分数据集稀疏。在这种数据量大而评分数据又极端稀疏的情况下,一方面难以成功地定位邻居用户集,影响推荐精度;另一方面在整个用户空间上计算相似用户群的过程不可避免地成为了算法的瓶颈[5]。针对User-based协同过滤推荐算法面临的问题,研究者们提出了Item-based协同过滤推荐算法[4]。Item-based协同过滤推荐依赖于项目的相似度来决定推荐。算法的不足之处是只能推荐那些和用户当前购买的商品相类似的商品,不能挖掘用户的潜在兴趣,作出"跨类型"的推荐。

收稿日期:2007-03-08.
基金项目:国家自然科学基金(70572079).
作者简介:王惠敏(1971-),女,讲师,博士. E-mail: huiminwangbj@126.com

?1994-2015 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

User-based 协同过滤和 Item-based 协同过滤只利用了用户-项目评分数据集中的部分信息来推荐商品。User-based 协同过滤算法在预测过程中没有考虑目标用户已经购买商品的相似商品也可能是该用户感兴趣的; 而 Item-based 协同过滤算法没有考虑与目标用户拥有共同兴趣爱好的相似用户的爱好对目标用户的影响。

提出的协同过滤算法结合了 User-based 协同过滤推荐算法和 Item-based 协同过滤推荐算法的基本思想, 综合考虑用户的相似性和项目的相似性对目标用户推荐产生的影响; 并将该算法与模糊聚类算法相结合, 利用模糊聚类技术分别将相似的项目和相似的用户聚类。

1　相关工作

协同过滤是预测某客户对某个他没有评价过的项目的喜爱程度, 预测的依据是过去用户群体对一系列项目的历史评分。用户-项目评分数据可用一个 $m\times n$ 阶矩阵 $\boldsymbol{R}(m, n)$ 表示, m 行代表 m 个用户, n 列代表 n 个项目, 第 i 行第 j 列的元素 R_{ij} 代表用户 i 对项目 j 的评分数值。用户-项目评分数据也可用行向量 $\boldsymbol{R}=[\boldsymbol{u}_1, \boldsymbol{u}_2, \cdots, \boldsymbol{u}_m]^{\mathrm{T}}$, $\boldsymbol{u}_k=[R_{k1}, R_{k2}, \cdots, R_{kn}]^{\mathrm{T}}$, $k=1, 2, \cdots, m$ 或列向量 $\boldsymbol{R}=[\boldsymbol{t}_1, \boldsymbol{t}_2, \cdots, \boldsymbol{t}_n]$, $\boldsymbol{t}_l=[R_{1l}, R_{2l}, \cdots, R_{ml}]$, $l=1, 2, \cdots, n$。$\boldsymbol{u}_k$ 代表某一用户对所有项目的评分, $\boldsymbol{t}_l$ 代表所有用户对某一项目的评分。

User-based 协同过滤根据评分相似的最近邻居的评分数据向目标用户产生推荐。该算法的核心部分是为一个需要推荐服务的目标用户寻找最相似的最近邻居集。最近邻居查找的效果和效率很大程度上决定了 User-based 协同过滤推荐算法的效果和效率。余弦相似性和相关相似性是常用的 2 种相似性度量方法。采用相关相似度量方法, 其相关相似度 $\mathrm{sim}(x, y)$ 表示为

$$\mathrm{sim}(x, y)=\frac{\sum_{s\in S_{xy}}(R_{xs}-R_x)(R_{ys}-R_y)}{\sqrt{\sum_{s\in S_{xy}}(R_{xs}-R_x)^2\sum_{s\in S_{xy}}(R_{ys}-R_y)^2}}$$

式中, R_{xs} 和 R_{ys} 分别是用户 x 和用户 y 对项目 s 的评分; R_x 和 R_y 分别是用户 x 和用户 y 的平均评分; S_{xy} 是用户 x 和用户 y 评分项目的交集。

基于相似性度量找出最相似的用户之后, 需要根据目标用户的评分, 预测其对未评分项目的评分。用户 x 对未评分项目 p 的预测评分可表示为

$$P_{xp}=R_x+\frac{\sum_{i=1}^{l_{\mathrm{u}}}(R_{ip}-R_i)\times\mathrm{sim}(x, i)}{\sum_{i=1}^{l_{\mathrm{u}}}\mathrm{sim}(x, i)}$$

式中, l_{u} 为相似的用户数。

Item-based 协同过滤采用了与 User-based 协同过滤方法同样的思想, 只是根据历史评分数据计算项目的相似度代替用户的相似度。完成项目的相似度计算之后, 选择最相似的项目集, 按如下公式预测用户 u 对未评分项目 s 的评分 P_{us}。

$$P_{us}=\frac{\sum_{d=1}^{l_{\mathrm{t}}}R_{ud}\times\mathrm{sim}(s, d)}{\sum_{d=1}^{l_{\mathrm{t}}}\mathrm{sim}(s, d)}$$

式中, l_{t} 为相似的项目数; 项目相似度 $\mathrm{sim}(s, d)$ 可通过相关相似性度量方法计算获得。

2　融合用户和项目信息的协同过滤算法

融合用户和项目信息的协同过滤算法的基本思想是考虑充分利用用户-项目评分数据集所包含的用户相关和项目相关的信息来推荐商品。由于基于项目的协同过滤算法考虑了项目相关的信息, 因此针对由于评分数据的极端稀疏导致难以定位相似的用户集合, 首先对用户-项目评分数据集按项目进行模糊聚类, 在聚类的簇中依照基于项目的协同过滤算法预测未评分数据, 并将其值填充在用户评分数据集中。填充后的

?1994-2015 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

评分数据集作为基于用户的协同过滤算法的输入数据，计算其相似的用户集，然后完成未评分数据的最后预测。在计算相似用户集时，采用模糊聚类法将用户-项目评分数据集按用户进行划分，在聚类的簇中计算用户相似度。

2.1 未评分项目的评分的初始预测

聚类分析的目的是将评分数据相似的项目聚集到一个簇中，在簇中采用基于项目的协同过滤方法预测用户未评分项目的评分。其优点是减少评分数据的稀疏性，并且可降低计算空间的维度，使计算速度加快。算法如下：

1）将模糊 c-均值算法应用于用户-项目评分数据矩阵 $\boldsymbol{R}$，将项目划分为 c 类，$\boldsymbol{R}_1$，$\boldsymbol{R}_2$，…，$\boldsymbol{R}_c$，且 $\boldsymbol{R}_1 \cup \boldsymbol{R}_2 \cup \cdots \cup \boldsymbol{R}_c = \boldsymbol{R}$，$\boldsymbol{R}_i \cap \boldsymbol{R}_j = \phi$。

2）查找目标用户 u_a 的未评分数据 $R_{u_a t_a}$ 所属的类，若 $R_{u_a t_a} \in \boldsymbol{R}_i$，那么 $\boldsymbol{R}_i$ 中的项目作为用户未评分项目的邻居。

3）在类 $\boldsymbol{R}_i$ 中采用相关相似性函数作为度量函数，用户未评分项目 t_a 与其邻居项目 t_b 的相似度可表示为

$$\mathrm{sim}(t_a, t_b) = \frac{\sum_{u_s \in U} (R_{u_s t_a} - \bar{R}_{t_a})(R_{u_s t_b} - \bar{R}_{t_b})}{\sqrt{\sum_{u_s \in U} (R_{u_s t_a} - \bar{R}_{t_a})^2 \sum_{u_s \in U} (R_{u_s t_b} - \bar{R}_{t_b})^2}}$$

式中，U 为所有已评分项目 t_a 和 t_b 的用户的交集；$\bar{R}_{t_a}$ 为项目 t_a 的平均评分；$\bar{R}_{t_b}$ 为项目 t_b 的平均评分。

4）预测用户 u_a 对未评分项目 t_a 的评分 P，并将其预测值填充在用户-项目评分数据矩阵中。P 可表示为

$$P = \frac{\sum_{t_b \in \boldsymbol{R}_i} \mathrm{sim}(t_a, t_b) \cdot R_{u_a t_b}}{\sum_{t_b \in \boldsymbol{R}_i} \mathrm{sim}(t_a, t_b)}$$

2.2 未评分项目的评分预测

填充后的用户-项目评分矩阵作为输入数据，运用模糊 c-均值聚类方法将具有相似兴趣爱好的用户分配到相同的簇中，根据簇中其他用户对某商品的评价预测目标用户对该商品的评价。具体算法如下：

1）将模糊 c-均值算法应用于填充后的用户-项目评分矩阵 $\boldsymbol{A}$，将用户划分为 g 类，$\boldsymbol{A}_1$，$\boldsymbol{A}_2$，…，$\boldsymbol{A}_g$，且 $\boldsymbol{A}_1 \cup \boldsymbol{A}_2 \cup \cdots \cup \boldsymbol{A}_g = \boldsymbol{A}$，$\boldsymbol{A}_i \cap \boldsymbol{A}_j = \phi$。

2）决定目标用户 u_a 的邻居，如果 $u_a \in \boldsymbol{A}_i$，那么 $\boldsymbol{A}_i$ 中的用户作为用户 u_a 的邻居。

3）在类 $\boldsymbol{A}_i$ 中计算目标用户 u_a 与其邻居用户的相似度。在类 $\boldsymbol{A}_i$ 中查找目标用户与其邻居用户 u_b 在初始评分数据矩阵中评分项目的并集 S，该并集中未评分的项目用 2.1 节中的预测值填充。然后在此并集中运用相关相似性度量方法计算两用户的相似度。

$$\mathrm{sim}(u_a, u_b) = \frac{\sum_{t_s \in S} (R_{u_a t_s} - \bar{R}_{u_a})(R_{u_b t_s} - \bar{R}_{u_b})}{\sqrt{\sum_{t_s \in S} (R_{u_a t_s} - \bar{R}_{u_a})^2 \sum_{t_s \in S} (R_{u_b t_s} - \bar{R}_{u_b})^2}}$$

4）在类 $\boldsymbol{A}_i$ 中预测未评分项目的评分。

$$P_{u_a t_a} = \bar{R}_{u_a} + \frac{\sum_{u_b \in \boldsymbol{A}_i} (R_{u_b t_a} - \bar{R}_{t_a}) \times \mathrm{sim}(u_a, u_b)}{\sum_{u_b \in \boldsymbol{A}_i} \mathrm{sim}(u_a, u_b)}$$

3 实验及结果分析

实验采用的数据集来自 Minnesota 大学 GroupLens Research 项目组收集的 MovieLens 数据集。MovieLens 站点（http://MovieLens.umn.edu/）是一个基于 Web 的研究型推荐系统，用于接收用户对电影的评分并提供相应的电影推荐列表。从 MovieLens 数据集中随机截取 200 个用户对 800 部电影的评分数据，然后

?1994-2015 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

将评分数据按 0.8 的比率划分为训练集和测试集, 从测试用户中的评分中随机选取 5 个评分作为可见的评分。

采用平均绝对偏差 MAE(Mean Absolute Error)作为评价推荐系统推荐质量的度量标准, 验证算法的有效性。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来度量预测的准确性, MAE 越小, 推荐质量越高。

对于测试集中的项目评分和用户, MAE 表示为

$$\mathrm{MAE}=\frac{\sum_{x,s}|P_{xs}-R_{xs}|}{L}$$

式中, P_{xs}表示用户 x 对项目 s 的预测评分; R_{xs}表示用户 x 对项目 s 的实际评分; L 为测试集合的基数。

实验以传统的基于用户的协同过滤推荐算法和基于用户聚类的协同过滤算法作为参照, 分别计算其平均绝对偏差 MAE, 然后与联合用户和项目信息的协同过滤推荐算法作比较, 实验结果为: 基于用户的协同过滤算法的 MAE 为 0.782 4; 基于用户聚类的协同过滤算法的 MAE 为 0.820 5; 基于融合用户和项目信息的协同过滤推荐算法的 MAE 为 0.761 2。实验结果表明联合用户和项目信息的协同过滤推荐算法具有的 MAE 较小, 完成的推荐质量要较高一些。

4　结　语

在提出的电子商务推荐算法中, 未评分项目的预测过程融合了用户-项目评分数据集中项目相关和用户相关的信息, 降低了评分数据的稀疏性, 预测结果更为准确。同时将模糊聚类应用于查找项目邻居和用户邻居的过程, 降低了维度, 减少了计算量, 加快了预测速度。

参考文献

[1] Weng L T, Xu Y, Li Y F. An Improvement to Collaborative Filtering for Recommender Systems[A]. Proceedings of the 2005 International Conference on Computational Intelligence for Modelling, Control and Automation, and International Conference on Intelligent Agents, Web Technological and Internet Commerce[C]. Washington: IEEE Computer Society, 2005: 792-795.

[2] Breese J, Heckerman D, Kadie C. Empirical Analysis of Predictive Algorithms for Collaborative Filtering[A]. Proceedings of the 14th Conference on Uncertainty in Artificial Intelligence[C]. Madison: Morgan Kaufmann, 1998: 43-52.

[3] Herlocker J L, Konstan J A, Terveen L G, et al. Evaluating Collaborative Filtering Recommender Systems[J]. ACM Trans Inf Syst, 2004(1): 5-53.

[4] Sarwar B, Karypis G, Konstan J. Item-based Collaborative Filtering Recommendation Algorithms[A]. Proceedings of the 10th International World Wide Web Conference[C]. New York: ACM, 2001: 285-295.

[5] 张海燕, 丁　峰, 姜丽红. 基于模糊聚类的协同过滤推荐方法[J]. 计算机仿真, 2005(8): 144-147.

?1994-2015 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net