文章编号: 1009－2269(2009)03－0011－03

用户聚类的电子商务推荐系统研究*

解亚萍[1], 牛广文[2]

(1. 兰州资源环境职业技术学院 计算机中心, 甘肃 兰州 730020;
2. 兰州工业高等专科学校 电气工程系, 甘肃 兰州 730050)

摘要: 随着电子商务规模越来越大, 协同过滤推荐算法的可扩展性差的问题也越来越受到人们的重视, 提出了一种基于用户项目类偏好值矩阵聚类的合作推荐方法, 解决了"冷开始"问题, 并且由于只在目标用户所属类别中搜索其最近邻居, 减少了搜索空间, 有效地提高了推荐系统的实时响应速度.

关 键 词: 电子商务; 协调过滤; 算法; 推荐系统
中图分类号: TP 273 文献标识码: A

0 引言

近年来, 电子商务系统中的信息"超载"现象越来越严重, 面对海量的商品信息, 消费者很难快速准确地挑选出真正适合自己需要的商品. 电子商务推荐系统是在了解和学习客户的需求与喜好的基础上为用户提供商品信息和建议, 向用户推荐其可能感兴趣的商品, 帮助用户完成购买过程, 提供个性化服务已经成为进一步提高网络内容服务质量急需解决的重要课题之一, 也是未来网络内容服务的一个发展方向.

协同过滤是通过比较用户之间的相似性, 把和目标用户具有相似兴趣的其他用户(也称邻居)的意见提供给目标用户完成推荐, 是目前最成功的推荐技术.

但协同过滤需要在整个用户空间上搜索目标用户的最近邻居, 随着用户数量和商品数量的急剧增多, 在整个用户空间上搜索最近邻居所需时间也急剧增加, 越来越难以满足用户的实时响应的要求. 本文提出了一种基于用户聚类的合作推荐实现方法, 首先计算得到用户对商品类别的评分, 再基于用户对商品类别的评分来对用户进行聚类, 只在用户所属类别中搜索用户的最近邻居, 从而能够在尽量少的用户空间上搜索尽可能多的最近邻居. 由于聚类可以定期离线进行, 所以本文提出的方法能有效地提高推荐系统的实时响应速度, 解决推荐系统随着用户数目增多而导致系统性能下降的问题.

1 协同过滤推荐系统

协同过滤是基于这样一个假设: 如果用户对一些项的评分比较相似, 则他们对其他项的评分也比较相似. 协同过滤的实现一般分两步: 首先, 获得用户信息, 即获得用户对某些信息条目的评价; 然后, 分析用户之间的相似度并预测特定用户对某一信息的喜好.

1) 用户信息的获取.

用户信息的获取主要通过用户对给定信息的评价. 评价分为显式和隐式两种. 显式评价需要用户有意识地表达自己对某一信息的认同程度, 一般用整数值来表示喜欢的不同程度. 隐式评价则希望从用户行为中获取信息.

* 收稿日期: 2008-12-05
作者简介: 解亚萍(1965-), 女, 陕西周至人, 副教授.

2) 相似度计算.

度量用户和之间的相似性方法如下, 首先得到用户 i 和 j 评分过的所有项, 然后通过不同的相似性度量方法计算它们之间的相似性, 记为 sim (i, j). 本文采用修正的余弦相似性计算方法:

$$\mathrm{sim}(i,j)=\frac{\sum_{c\in I_{ij}}(R_{i,c}-\overline{R})(R_{j,c}-\overline{R}_c)}{\sqrt{\sum_{c\in I_{ij}}(R_{i,c}(R_{i,c}-\overline{R}_c)^2\times\sum_{c\in I_{ij}}(R_{j,c}-\overline{R}_c)^2}}. \tag{1}$$

$R_{i,c}$ 为用户 i 对项目 c 的评分, $\overline{R}_c$ 为项目 c 的平均评分. 计算完用户之间的相似度后, 对一个用户 u, 产生一个按照相似度大小排列的"邻居" 集合, $N\leqslant=\{U_1,U_2,\cdots,U_t\}$, $0\leqslant t\leqslant m$, u 不属于 N, 从 U_1 到 U_t, sim$(u,U_i)(1\leqslant i\leqslant t)$ 从大到小排列.

3) 产生推荐.

产生推荐主要解决从最近邻居信息中获得目标用户对未评分项目兴趣程度的预测. 用户兴趣度的预测可以通过如下计算得到:

$$P_{u,i}=\overline{R}_u+\frac{\sum_{m=1}^{n}(R_{m,i}-\overline{R}_m)\times\mathrm{sim}(u,m)}{\sum_{m=1}^{n}\mathrm{sim}(u,m)}. \tag{2}$$

其中 $\overline{R}_u$ 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): 用户 u 和 m 的相似度. 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反馈给目标用户.

2 基于用户聚类的电子商务推荐系统

本文提出的基于用户聚类的电子商务推荐系统, 是在协同过滤系统基础上, 通过对用户进行聚类, 将对商品偏好比较相似的用户加入同一类中, 然后仅在目标用户所属类别中查找最近邻居并进行推荐, 从而减小了搜索空间, 提高了系统的实时响应速度.

2.1 计算用户项目类偏好值

用户对于商品的兴趣在一定的时间内是相对固定的, 电子商务系统中的数据库记录了每个客户的交易数据, 每个交易数据中记载了客户购买的商品, 而每个商品又有其类别属性, 这样就可以利用这些数据以及用户对于商品的评价信息计算得到用户对不同商品类别的偏好值, 具体做法如下:

$$PC_{u,j}=\frac{\sum_{i\in I_u}PI_{u,i}\times\mu_j(x_i)}{\sum_{i\in I_u}\mu_j(x_i)}\quad j=1,2,3,\cdots \tag{3}$$

式中: $PC_{u,j}$ 代表用户 u 对类别的偏好值; $PI_{u,i}$ 代表用户 u 对商品 i 的评分值; I_u 代表用户 u 已评估的商品集合; $\mu_j(x_i)$ 代表商品 i 对类别 j 的隶属度.

协同过滤存在冷开始问题. 通过分析用户对于不同类别商品的关注程度, 只将用户偏好值较高的商品类别的新商品信息推荐给用户, 而对于用户不太关注的类别的新商品信息则不推荐, 从而解决了冷开始问题.

2.2 聚类得到目标用户所在簇

对于随着用户空间增大而导致系统性能下降的问题本文采用聚类方法解决. 根据计算得到的用户项目类的偏好值矩阵, 利用 K-Means 聚类算法将用户划分到不同的簇中, 在目标用户所在的簇中搜索目标用户的若干个最近邻居, 再根据其最近邻居对商品的评价信息预测目标用户对未购买的商品的评分值, 最后将预测评分值较高的商品信息推荐给目标用户.

3 实验及结果分析

实验采用的数据集是 MovieLens(http://www.grouplens.org). MovieLens 数据集包含 movies.dat、ratings.dat 和 users.dat. movies.dat 中包含了 1682 部电影的详细描述信息, users.dat 中包含 943 位用户的详细信息, rating.dat 中包含 943 位用户对 1682 部电影的 100,000 条评分记录, 评分值为从 1 到 5 的整数.

根据定义稀疏等级的概念为用户评分数据矩阵中未评分条目所占的百分比. MovieLens 数据集的稀疏等级为: $1-100000/(1682*943)=0.936953$. 首先对 943 位用户分别计算对 18 类电影的偏好

值, 得到 8952 条记录, 因此用户项目类偏好值矩阵的稀疏等级为: *l*-8952/(18 *943) = 0.472605, 降低了数据集的稀疏性.

本文中随机选取 ID 为 82、111、445、681、904 的用户作为目标用户, 并且最近邻居数选择为 10. 使用 K-Means 聚类算法对 943 个用户进行聚类, 聚类数目分别选择为 2, 3, 4, 5, 然后对每一个目标用户只在其所在的类别中搜索其最近邻居, 查找到的最近邻居数目如表 1 所示.

表 1　最近邻居个数

聚类个数	用户 ID				
	82	111	445	681	904
	邻居个数				
2	9	9	9	9	9
3	7	8	8	7	7
4	7	6	6	6	7
5	7	6	6	6	7

当聚类数目为 2 时, 在 67.23 %的用户空间上可以搜索得到目标用户 88 %的最近邻居; 当聚类数目为 3 时, 在 36.48 %的用户空间上可搜索得到 74 %的最近邻居; 当聚类数目为 4 时, 在 28.95 %的用户空间上可搜索得到 64 %的最近邻居; 当聚类数目为 5 的时候在 25.45 %的用户空间上可搜索得到 64 %的最近邻居; 平均计算在 39.52 %的用户空间可以搜索得到目标用户 72.5 %的最近邻居, 因此对用户进行聚类后可以在较小的用户空间上搜索出目标用户的大部分最近邻居. 如果想要提供推荐精度, 搜索出更多的最近邻居, 可以计算目标用户与聚类中心的距离, 选择与目标用户距离小的若干簇, 适当增大搜索的用户空间以得到更多的最近邻居.

4　结语

随着电子商务规模越来越大, 协同过滤推荐算法的可扩展性差的问题也越来越受到人们的重视, 本文提出了一种基于用户聚类的电子商务推荐系统, 可以有效地解决协同过滤推荐算法面临的可扩展性差的问题, 更好地满足用户的实时性要求.

参考文献:

[1] Breese J, Heckerman D, Kadie C. Empirical analysis of predictive algorithms for collaborative filtering. In: Proceedings of the 14th Conference on Uncertainty in Artificial Intelligence, San Francisco, CA, July 1998: 43-52.

[2] 潘红艳. 个性化信息服务的研究与实现[D]. 大连: 大连理工大学, 2005.

[3] 余　力, 刘　鲁. 电子商务个性化推荐研究[J]. 计算机集成制造系统, 2004, 10(10): 1306-1312.

[4] 高凤荣, 杜小勇, 王　珊. 一种基于稀疏矩阵划分的个性化推荐算法[J]. 微电子与计算机, 2004, 21(2): 58-62.

[5] 黎星星, 黄小琴, 朱庆生. 计算机工程与科学[J]. 2004, 5(2): 164-166.

Research on User Clustering of Recommendation System in E-commerce

XIE Ya-ping[1], NIU Guang-wen[2]

(1. The Computer Center of Lanzhou Resources and Environment Vocational School, Lanzhou 730020, China;
2. The Electric Engineering Department of Lanzhou Polytechnic College, Lanzhou 730050, China)

Abstract: Nowadays, with the scale of E-commerce is getting larger and larger, more importance has been attached to the problem of poor expansibility appearing in collaborative filtering recommendation algorithm. This paper describes a method of cooperative recommendation based on user-item preference values and matrix clustering, which has solved "cold start" problem to some extent. This method only searching for its nearest neighbor in the classification of object user, so the search space is reduced and the real-time performance of the recommendation system can be effectively improved.

Key words: E-commerce; collaborative filtering; algorithm; recommendation system