基于用户聚类的电子商务推荐系统

潘宇，林鸿飞，杨志豪

（大连理工大学计算机科学与工程系，大连 116024 ）

摘 要： 推荐系统是利用用户的一些历史个性偏好信息实现个性化服务的系统，它已经成为电子商务领域中的重要技术之一。协同过滤是推荐系统中采用最为广泛和成功的推荐技术，但随着电子商务系统用户数目和商品数目的增加，在整个用户空间上搜索目标用户的最近邻居的耗时也急剧增加，导致系统性能下降。本文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，有效地提高推荐系统的实时响应速度。

关键词： 电子商务；推荐系统；协同过滤；聚类

A Recommendation System in E-Commerce based on user Clustering

Pan Yu, Lin HongFei

(Department of Computer Science and Engineering, Dalian University of Technology , Dalian 116024)

Abstract: Recommendation system use customers' historical individual information to realize personal service and it has become an important technology in E-commerce. Collaborative filtering is the best recommendation system at present and it has been used widely. But with the gradual increase of users and commodities in E-Commerce, the time consuming will increase greatly to search the nearest neighbor of the target user in the whole user space. This paper introduces a collaborative filtering system based on user-item of a kind matrix .It can solve the problem of new-item in a certain extent. And the experiment suggests that this method can effectively improve the real-time performance of the recommendation system.

Keywords: E-Commerce; Recommendation System; Collaborative filtering; Clustering

1 引言

近年来，随着电子商务的快速发展，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的挑选出真正适合自己需要的商品[1]。

电子商务推荐系统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，实现信息服务的个性化，充分体现了以人为本。目前，几乎所有的大型电子商务系统，如 Amazon、CDNOW、eBay、当当网上书店等都不同程度的使用了各种形式的推荐系统[2]。

基金资助：国家自然科学基金 60373095

作者简介：潘宇（1982—），女，辽宁辽阳，硕士研究生在读，panyu1217@163.com.cn

林鸿飞（1962-），男，辽宁，教授，博士，hflin@dlut.edu.cn .

有效的帮助用户解决信息“超载”问题，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，而提供个性化的服务机制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，其优点是可以发现用户可能感兴趣的新信息，而且它能处理难以进行机器自动内容分析的信息，比如音乐，电影等，也能基于一些复杂的、难以表达的概念比如质量、品质、品味等进行推荐，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，随着电子商务系统规模越来越大，用户数量和商品的数量会急剧的增多，这就使得在整个用户空间上搜索目标用户的最近邻居所需时间也随着急剧增加，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，根据商品所属类别不同以及用户对各个商品的评分，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目标用户尽可能多的最近邻居，最后根据用户最近邻居对商品的评分预测目标用户的评分完成推荐。由于聚类可以定期离线进行，所以本文提出的方法能有效的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。

2 协同过滤推荐系统

2.1 研究背景

Typestry 是最早提出来的协同过滤推荐系统，是 Xerox PARC 研究中心提出的一个研究型推荐系统，用于过滤电子邮件、推荐电子新闻[3]。GroupLens 是基于用户评分的自动化合作推荐系统，用于推荐电影和新闻。Ringo 推荐系统和 Video 推荐系统通过电子邮件的方式推荐音乐和电影。MovieLens 是 Minnesota 大学开发的基于 Web 的研究型协同过滤推荐系统，用于推荐电影[4]。另外协同过滤推荐系统在商业上也取得了巨大的成就。

2.2 协同过滤算法

协同过滤是基于这样一个假设[5]：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；其次，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。

2.2.1 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。评价分为显式评价(Explicit Rating) 和隐式评价(Implicit Rating) 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，系统获得这些初始信息后，就将用户加入到用户数据库中，随着用户不断使用协作过滤系统，用户的信息不断积累和更新。隐式评价希望从用户行为中获取信息。目前一些研究者利用 agent 技术通过分析用户网上购物记录、阅读文章的时间和浏览行为等数据来获取用户信息。

2.2.2 相似度计算

度量用户 i 和 j 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).度量两个用户之间的相似性，一般有 3 种方法：余弦相似性、相关相似性、修正的余弦相似性。

余弦相似性：把用户评分看作是 n 维项目空间上的向量。如果用户对某个项目没有评分，则将此评分假设为 0。通过计算两个向量之间的夹角余弦来度量两个用户之间的相似性。见公式 2.1：

$$sim(i,j) = \frac{\sum_{k=1}^{n} R_{i,k} * R_{j,k}}{\sqrt{\sum_{k=1}^{n} R_{i,k}^{2} * \sum_{k=1}^{n} R_{j,k}^{2}}} \qquad (2.1)$$

$R_{i,k}, R_{j,k}$：用户 i，j 对项目 k 的评分。

相关相似性：通过 Peason 相关系数来度量两个用户的相似性。计算时，首先找到两个用户共同评分过的项目集 $I_{i,j}$，然后计算这两个向量的相关系数。见公式 2.2

$$sim(i,j)=\frac{\sum_{c\in I_{i,j}}(R_{i,c}-\overline{R_i})(R_{j,c}-\overline{R_j})}{\sqrt{\sum_{c\in I_{i,j}}(R_{i,c}-\overline{R_i})^2*\sum_{c\in I_{i,j}}(R_{j,c}-\overline{R_j})^2}} \qquad (2.2)$$

其中 $I_{i,j}$ 为用户 i 和 j 共同评分过的项目集，$R_{i,c}$ 为用户 i 对项目 c 的评分，$\overline{R_i}$ 为用户 i 对资源的平均评分。

修正的余弦相似性：在余弦相似性中没有考虑不同用户的评分尺度问题。修正的余弦相似性通过减去项目的平均评分来弥补这种不足，见公式 2.3：

$$sim(i,j)=\frac{\sum_{c\in I_{i,j}}(R_{i,c}-\overline{R_c})(R_{j,c}-\overline{R_c})}{\sqrt{\sum_{c\in I_{i,j}}(R_{i,c}-\overline{R_c})^2*\sum_{c\in I_{i,j}}(R_{j,c}-\overline{R_c})^2}} \qquad (2.3)$$

$\overline{R_c}$ 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 $N=\{U_1, U_2, \cdots U_t\}$，$0\leqslant t\leqslant m$，$u$ 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。

2.2.3 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公式 2.4 计算得到：

$$P_{u,i}=\overline{R_u}+\frac{\sum_{m=1}^{n}(R_{m,i}-\overline{R_m})*sim(u,m)}{\sum_{m=1}^{n}sim(u,m)} \qquad (2.4)$$

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。

3 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其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。

3.1 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，比如少女购买的商品主要是服饰，已婚的有孩子的女士，她购买的商品主要是玩具等儿童用品，而购买昂贵商品的用户主要是高收入者。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：

$$PC_{u,j} = \frac{\sum_{\in I_u} PI_{u,i} * \mu_j(x_i)}{\sum_{\in I_u} \mu_j(x_i)} \qquad j=1,2,3...$$

$PC_{u,j}$ 代表用户 u 对类别 j 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。

协同过滤存在于冷开始[6]问题，如果一个商品还没有人购买或没有人评价它，则这个商品肯定就得不到推荐，而新商品信息往往又是用户比较关注的。通过分析用户对于不同类别商品的关注程度，能够分析用户的购买行为和心理。而用户对于他以往主要购买类别的商品最新消息往往也给予很高的关注。比如，只购买少女服饰的用户，对于一些类似“06 新款服饰”、“最新上市少女装”等商品信息是决不会错过的。这样在计算得到了用户对不同类别商品的偏好值后，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。

3.2 聚类得到目标用户所在簇

K-Means 聚类算法的计算量很小，可以有效的处理多变量、大样本数据而不占用太多的内存空间和计算时间，对于电子商务网站用户和商品数目都很庞大的情况比较适合，同时在分析时可以人为的指定起始中心位置，或者将曾做过的聚类分析结果作为起始位置引入分析，提高聚类的有效性。

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，将预测评分值较高的商品信息推荐给目标用户。

4 实验及结果分析

实验采用的数据集是 MovieLens(http://www.grouplens.org)。MovieLens 中的数据是 Minnesota 大学进行 GroupLens Research 项目时收集的，每星期都有上百的用户访问该系统，进行电影评价和获得关于电影的推荐。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（代号、电影名和所属类别），users.dat 中包含 943 位用户的详细信息（用户 ID、性别、年龄和从事的职业），ratings.dat 中包含 943 位用户对 1682 部电影的 100，000 条评分记录（用户 ID、电影代号、评分值和时间戳），评分值为从 1 到 5 的整数，数值越高，表明用户对该电影的喜爱程度越高。本文将 MovieLens 数据集的 80%作为训练集而另 20%作为测试集。

根据[7]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。因此，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：1-8952/ (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，使用修正的余弦相似性计算目标用户的最近邻居，并且最近邻居数选择为 10。为了对比说明本文所提出方法的效果，首先在整个用户空间上搜索目标用户的最近邻居，然后使用 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下表所示：

表 1 最近邻居个数

Tab.1 The number of the nearest neighbor

用户 ID / 邻居个数 / 聚类个数	82	111	445	681	904

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%的用户空间上可以搜索得到目标用户 74%的最近邻居；当聚类数目为 4 时，在 28.95%的用户空间上可以搜索得到目标用户 64%的最近邻居；当聚类数目为 5 的时候在 25.45%的用户空间上可以搜索得到 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。

5 结论

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，可以有效的解决协同过滤过滤推荐算法面临的可扩展性差的问题，更好的满足用户的实时性要求。

参考文献：

[1] 程岩，肖小云，吴洁倩. 基于聚类分析的电子商务推荐系统[J].计算机工程与应用，2005，24：1

[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] D. Goldberg, D. Nichols, D. Terry. Using collaborative filtering to weave an information tapestry[J].Communications of the ACM,1992, 35(12): 61-70.

[4] B. Dahlen, J. Konstan, J. Herlocker. Jump-starting MovieLens: User benefits of starting a collaborative filtering system with "Dead Data"[J].University of MinnesotaTR 98-017, 1998.

[5] J. Breese, D. Heckerman, C. Kadie. 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.

[6] 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）：4

[7] 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