第 36 卷 第 23 期 计 算 机 工 程 2010 年 12 月
Vol. 36 No. 23 Computer Engineering December 2010
·软件技术与数据库· 文章编号:1000—3428(2010)23—0082—03 文献标识码:A 中图分类号:TP309.2

协同过滤系统隐私保护和推荐准确性研究

徐 南[1]， 王新生[2]
(1. 秦皇岛职业技术学院信息工程系，河北 秦皇岛 066100；2. 燕山大学信息科学与工程学院，河北 秦皇岛 066004)

摘 要：针对协同过滤推荐系统在预测过程中容易泄漏用户概貌数据的问题，在不影响推荐准确性的前提下，提出一种用户数据混淆策略，使响应用户的评分数据在计算用户相似度之前被假数据代替，用户尽量少泄露(或不泄露)个人评分信息，进而实现用户隐私的保护。通过实验分析数据混淆策略对协同过滤推荐准确性的影响，证明该策略的有效性。

关键词：协同过滤；隐私保护；推荐系统；准确性

Research on Privacy-preserving and Recommendation Accuracy of Collaborative Filtering System

XU Nan[1]， WANG Xin-sheng[2]
(1. Department of Information Engineering, Qinhuangdao Institute of Technology, Qinhuangdao 066100, China;
2. College of Information Science and Engineering, Yanshan University, Qinhuangdao 066004, China)

【Abstract】On the basis of the users' profile-exposing in the prediction generation process, a data obfuscation policy without affecting the accuracy of recommender system is proposed, in which the ratings of the responding user are substituted by false data before computing the similarity degree of users. The process does not(or a little) expose the ratings of user and preserves users' privacy data. Experimental results demonstrate the impact of obfuscation policies on the accuracy of the generated predictions, and show the improvement is effective.

【Key words】collaborative filtering; privacy-preserving; recommender system; accuracy

1 概述

协同过滤推荐是个性化服务中最成功、使用最广泛的推荐技术之一[1]。协同过滤推荐系统分为基于用户的推荐系统和基于项目的推荐系统 2 种，基于用户的协同过滤推荐系统基于这样一种假设：过去表现出相似兴趣的用户将对相似的产品或服务感兴趣，用户的兴趣表现为用户对项目集的评分向量，协同过滤系统通过收集大量的用户兴趣信息，通过计算这些用户与目标用户的相似度产生目标用户的最近邻居，然后通过计算最近邻居对目标项目评分的加权平均值产生推荐。而基于项目的推荐系统与前者不同之处在于计算项目之间的相似性，认为用户对相似的项目感兴趣。但是无论哪种推荐系统，用户需要经常主动提供个人相关信息，这些数据信息不断上传，并通过网络中的信息系统保存和处理信息。服务机构需要利用 Cookies 等网络信息采集技术进行数据挖掘，并通过机器挖掘等比较隐蔽的方式监视用户的信息搜索与浏览过程，从而收集尽可能多且准确的用户信息来为推荐做准备，推荐精度取决于收集到的用户信息的广度和精度，所以，推荐系统质量和用户的隐私之间存在着不可避免的矛盾。因此，用户的隐私问题成为个性化推荐系统发展的瓶颈。为了解决这个问题，本文提出了一种数据混淆的协同过滤推荐方法。

2 相关知识

目前协同过滤中隐私保护技术基本可以分为基于密码学的方法和数据变换两大类。文献[2]研究了 P2P 环境下的协同过滤推荐隐私保护问题。文中采用了 SVD 技术和极大似然技术产生推荐，并设计了一个基于安全多方计算的通信协议。所有的用户拥有对自己的数据的完全控制权。同一个社区中的所有用户可以通过加密协议计算出他们数据的聚集而不用暴露个人的隐私数据。社区内和社区外的用户最终都可以通过计算得到推荐。文献[3]扩展了文献[2]的工作，提出了一种基于双向聚类的隐私保护方法。其主要改进有用基于交叉最小化的双向聚类来替代 SVD 技术，简化了同态加密技术的复杂性，可实现增量计算，以及对原有加密系统的简化和改进，进一步提高了算法实现效率。

文献[4]提出一种集中式协同过滤保护方法，通过对用户数据添加一些不确定的因素，保护用户的隐私，在把用户评分转发到服务器之前，每个用户使用随机数据修改技术对其进行扰乱。因此，服务器不能找出真实的评分，只有这些被修改了的数据。后来文献[5]提出了运用随机扰乱的数据变换技术进行基于用户相关性的协同过滤的隐私保护。不同于文献[2]中的系统，文献[4-5]中的系统是集中式的，不是每个用户都参与计算。用户将经过随机扰乱的数据发送到服务器进行运算。

目前各种方法还停留在实验阶段，难以得到广泛应用。而且每种方法都有应用的局限性，因此，本文提出一种基于数

作者简介：徐 南(1976—)，女，讲师、硕士，主研方向：推荐系统，数字水印；王新生，教授
收稿日期：2010-02-27 **E-mail**：xunan356@yahoo.com.cn

据混淆的协同过滤推荐，在不妨碍推荐系统的准确性的前提下，更好地保护用户的隐私。

3 基于数据混淆的协同过滤推荐

3.1 用户概貌的存储形式

在分布式 P2P 系统中，用户以完全分散的方式保存他们的个人概貌信息。用户的评分矩阵存储在中心协同过滤系统中，被一个虚拟的矩阵所代替，其中，矩阵的行（用户的评分向量）被用户以分布式的方式存储。图 1 对比了中心式存储和分布式存储 2 种方式。

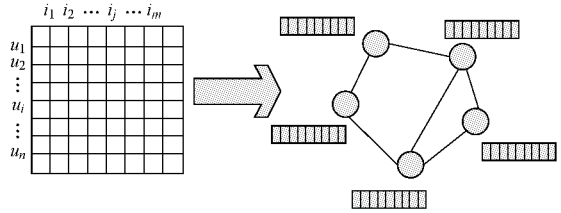

图 1 中心式和分布式存储方式的比较

在这种设置中，用户是他们个人信息的拥有者，在预测生成期间他们可以直接彼此通信，并且互不干涉地决定具体评分和要泄露给其他用户的信息。这个预测生成的过程分 3 个阶段：

(1)过程由目标用户发起，他泄露自己的概貌信息并且广播一个对某一项目的预测请求给其他用户。在这个阶段应该确定 2 个参数：

1)应该泄露概貌的哪一部分。为了更好地保护目标用户的隐私，泄露的评分数目应该尽可能少。然而，减少评分的数目就妨碍了相似性的计算，也就影响了生成预测的准确性。

2)这个请求应该发送给哪些用户。理论上，这个请求应该发送给所有的用户，因为网络中任何一个用户都有可能是最近的邻居之一。实际上，这导致了比较严重的通信负荷，因此，需要限制一个接收请求的用户集。

(2)当接收到请求时，每个用户自动决定是否该响应。如果用户决定响应该请求，他将根据接收到的目标用户概貌信息，采用余弦相似性计算公式计算与目标用户的相似度。当相似度计算出来之后，将该值和对请求项目的评分都发送给目标用户。在这个过程中，响应用户的两部分信息泄露了：一是请求项目的评分被直接泄露了；二是计算的相似度，利用该值可以推断响应用户的部分概貌。

(3)根据收集到的响应，选择出最相似的 K 个用户，目标用户为预测的产生建立一个相似用户邻居集。最后，按照他们的相似度，通过聚集邻居用户对该项目的评分计算出一个加权平均值，目标用户为请求项目生成一个预测。

为了总结预测生成过程，需要强调协同过滤保护用户隐私的方式，然而仍然允许他们支持其他用户发起的预测生成。

3.2 数据混淆策略

根据分布式系统过滤预测的过程，用户概貌可能在 2 种情况下被泄露。第 1 种就是作为部分预测请求的目标用户概貌被广播给其他用户，这种情况下的泄露是不可避免的，因为为了得到响应用户的可靠相似性计算，目标用户必须泄露概貌信息的重要部分。第 2 种情况是当其他用户决定参与并响应其他用户发起的预测生成过程时，概貌信息的泄露把请求项目的评分和两者的相似度发送给目标用户。尽管这种情况下，响应用户泄露的仅仅是一部分信息，但是这仍然会构成一种隐私破坏，因为攻击者使用多样的预测请求通过系统恶意攻击可以使大部分概貌泄露。

为了减轻这种隐私破坏，本文提出混淆用户概貌中的数据，也就是说存储在概貌中的评分的一个子集被假数据替代。工作的重点主要集中在修改响应用户的概貌数据，因为修改目标用户的概貌会大大地降低相似性计算的准确性。因此，响应用户的评分在计算相似度响应请求之前就被假数据代替了。尽管修改用户的概貌不能阻止攻击者收集响应用户的概貌和重现他们的概貌，但是这些收集的评分不能反映真正概貌的内容。

本文提出了 3 种用户混淆用户概貌评分的策略：

(1)默认混淆。用一个固定值 x 代替用户概貌中的真实评分。

(2)随机混淆。用数据集中选出来的一系列评分的随机值代替用户概貌的真实评分。

(3)集中混淆。用集中评分的选择值替代用户概貌的真实评分。

4 实验与评价

4.1 实验环境及实验数据

实验数据采用 MovieLens 站点（http://movielens.umn.edu/）提供的数据集。其中包括 943 个用户对 1 682 部电影所做出的 100 000 个评分，评分范围为 1～5，每个用户至少对 20 部电影做出评分。

在实验中，上面提到的 3 个混淆策略通过 5 个具体策略实例化。

(1)积极。以数据集中最积极的评分代替实际评分，即评分 5。

(2)消极。以数据集中最消极的评分代替实际评分，即评分 1。

(3)中等。以数据集中中等的评分，即最大和最小可能评分之间的平均，代替实际评分，即评分 3。

(4)随机。以数据评分范围中的一个随机值代替实际评分，即从评分 1 到 5 的一个随机值。

(5)分布。以反映数据集中评分整体分布（即平均值和方差）的一个值代替实际评分。

其中，积极的、消极的和中等的策略是默认策略的实例；随机策略是随机策略的实例；分布策略是集中混淆的策略。

4.2 度量标准

平均绝对误差 MAE 是推荐系统中最常用的一种推荐质量度量方法。MAE 通过计算用户评分的真实值和预测值之间的偏差来度量预测的准确性。显然，MAE 越小，推荐质量越高。本文采用下式计算 MAE 值：

$$MAE=\frac{1}{\|A\|}\sum_{a\in A}\frac{\sum_{j\in T}|v_{a,j}-p_{a,j}|}{\|T\|}$$

其中，$v_{a,j}$ 是用户 a 对项目 j 的评分；$p_{a,j}$ 是系统对项目 j 的预测分数；A 和 T 分别为测试用户和训练用户集合，$\|A\|$ 和 $\|T\|$ 分别为集合的大小。

4.3 实验结果及分析

为了验证本文提出的混淆策略对生成预测的准确性的影响，做了下面的实验：选出 10 000 个评分作为测试集，剩下的作为训练集，用第 3 节描述的分布式协同过滤技术预测它们的值，然后计算出 MAE 值。逐渐增加数据混淆率（也就是增加用户概貌信息中被混淆的数据），从 0 增加到 90%。图 2

是 *MAE* 值随混淆率变化的情况。

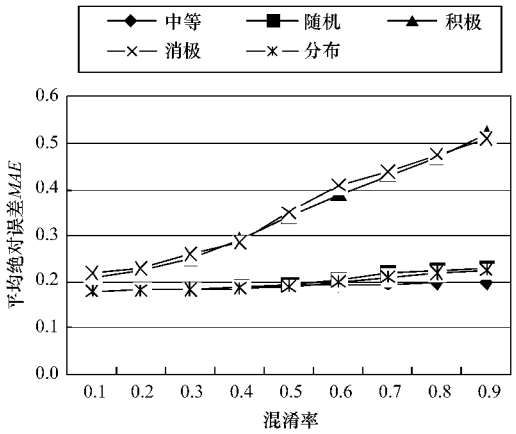

图 2　*MAE* 随混淆率的变化情况

从图 2 中可以看出，随机、中等和分布策略的影响是相似的：*MAE* 随混淆率线性增加。尽管如此，当混淆率比较小时，系统预测仍然相当准确，这是因为这 3 种策略并没有明显地修改用户概貌信息（修改后的数据和真实数据是相似的），所以对 *MAE* 的影响也比较小，但是当混淆率比较大时，这 3 种策略的 *MAE* 比较接近非个性化预测的 *MAE* 值。相反，积极和消极策略修改用户概貌数据程度比较大，最终生成的预测也相对不太准确，并且随着混淆率的增大，*MAE* 急剧增加。

为测试数据混淆对各种类型评分的预测是否会有影响，通过以下实验进行验证：按照评分 1、2、3、4、5，把数据集中的可用评分分为 5 组，每组中选出 1 000 个评分作为测试集，然后根据协同过滤预测对这些评分进行预测。混淆率从 0 到 90%逐渐增长，每组的评分分别计算 *MAE*，其变化情况如图 3所示。为了使结果更清晰，本文主要分析了混淆率为 0、0.3、0.6、0.9 的情况。

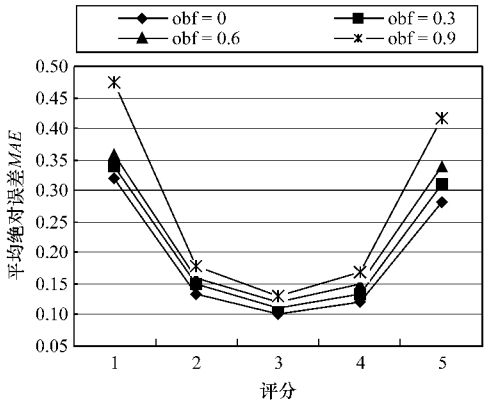

图 3　数据混淆对评分预测的影响情况

从图 3 中可以看出，数据混淆策略对各种类型评分的预测的影响是不同的。对于中度评分，影响比较小，随着混淆率的变化，*MAE* 基本保持不变。相反，对于极端评分，影响非常大，*MAE* 随着混淆率的增大而增大。因此，当对用户概貌数据进行混淆时，极端评分预测的准确性严重破坏了。

从以上 2 个实验可以得出这样的结论：采用随机、中等和分布策略对预测准确性的影响并不是太大，并且混淆率越低，对结果的影响越小。因此，为了尽可能减小对推荐的准确度的影响，更好地保护用户隐私，采取随机、中等和分布策略对用户评分数据进行混淆，令混淆率为 0.3，与文献[2]中 SPA 算法和文献[5]中的 PPCF-SVD 算法进行了对比，其比较结果如图 4 所示。

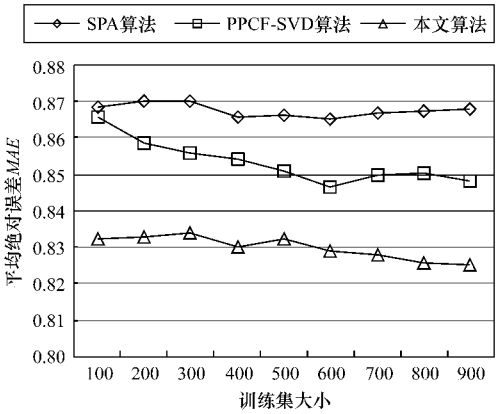

图 4　3 种算法的推荐准确性比较

从图 4 中可以看出，本文算法误差最小，且随着训练集用户的增加，误差呈下降趋势。混淆策略对用户评分数据的修改幅度并不大，没有降低推荐的准确性，而且保护了用户的隐私。因此，采用数据混淆策略在某种程度上可以有效保护用户隐私。

通过这 3 个实验可以看出，对非极端评分的数据混淆对生成预测的准确性影响不大，但是它却改进了用户信息的隐私保护问题，用户个人信息泄露得少多了。因此，采用数据混淆策略在某种程度上可以有效保护用户隐私。

5　结束语

协同过滤技术是推荐系统未来的研究重点。本文提出了一种基于数据混淆的协同推荐技术，更好地实现了隐私保护和协同过滤推荐准确性之间的权衡。下一步主要研究如何充分利用极端评分对推荐准确性的有利影响，从而进一步提高推荐系统的效率和准确率。

参考文献

[1] 郭艳红，邓贵仕，雒春雨. 基于信任因子的协同过滤推荐算法[J]. 计算机工程，2008，34(20)：1-3.

[2] Canny J. Collaborative Filtering with Privacy[C]//Proc. of IEEE Symposium on Security and Privacy. [S. l.]：IEEE Computer Society，2002：45-57.

[3] Ahmad W，Khokhar A. Phoenix：Privacy *P* Reserving Biclustering on Horizontally Partitioned Data Amid Malicious Adversaries[C]//Proc. of ACM SIGKDD International Workshop of Privacy，Security and Trust in KDD. San Jose，USA：[s. n.]，2007.

[4] Polat H，Du Wenliang. Privacy Preserving Collaborative Filtering Using Randomized Perturbation Techniques[C]//Proc. of the 3rd IEEE International Conference on Data Mining. Melbourne，Florida：[s. n.]，2003：625-628.

[5] Polat H，Du Wenliang. SVD-based Collaborative Filtering with Privacy[C]//Proc. of the 20th ACM Symposium on Applied Computing，Track on E-commerce Technologies. Santa Fe，New Mexico，USA：[s. n.]，2005：791-795.

编辑：张正兴