Truser:一种基于可信用户的服务推荐方法

何 鹏[1),3)] 吴 浩[1)] 曾 诚[1),3)] 马于涛[2)]

[1)](湖北大学计算机与信息工程学院 武汉 430062)

[2)](武汉大学计算机学院 武汉 430072)

[3)](湖北省教育信息化工程技术研究中心 武汉 430062)

摘 要 在服务推荐过程中,为排除不可信用户信息带来的干扰,确保推荐结果的精准性,该文从用户聚类的角度,通过两阶段的 ISODATA 聚类,将离群用户视为不可信用户进行过滤,再基于得到的可信用户提出一种改进的服务推荐方法.最后,在两个公开数据集 Last.FM 和 Delicious 上进行了实证分析.结果表明,该文所提方法在两个数据集上的推荐精度相较于已有基准方法分别提高 16.1%和 4.5%,且发现当第一阶段聚类的预期聚类中心为 6 时,推荐效果最好;同时,在推荐过程中为目标用户返回 Top-5 个可信用户,且向其推荐这 5 个用户中至少有 70%的人关注过的服务最为适宜.因此,围绕可信用户的数据进行推荐,能有效地提高服务推荐的质量.

关键词 ISODATA 聚类;协同过滤;服务推荐;服务计算

中图法分类号 TP311 **DOI 号** 10.11897/SP.J.1016.2019.00851

Truser: An Approach to Service Recommendation Based on Trusted Users

HE Peng[1),3)] WU Hao[1)] ZENG Cheng[1),3)] MA Yu-Tao[2)]

[1)](*School of Computer and Information Engineering, Hubei University, Wuhan* 430062)

[2)](*School of Computer Science, Wuhan University, Wuhan* 430072)

[3)](*Engineering Technology Research Center for Education Informatization, Hubei Province, Wuhan* 430062)

Abstract As a very important topic in the field of service computing, service recommendation has been paid much attention by researchers. To improve the service recommendation quality and ensure user's experience of services, various methods have been proposed successively. However, most existing methods mainly focus on how to improve the accuracy of recommendation models by introducing richer information or advanced modeling techniques, and there is a general lack of discussion on users' trustworthiness from the data quality point of view. In practice, untrusted users are common for a variety of reasons, such as fake comments. Data generated by untrusted users is often misleading and unhelpful for recommendations. Therefore, it is very necessary to eliminate noise from untrusted users before service recommendation; otherwise, the quality of recommendations will always be affected regardless of how the model is optimized. To achieve this, we identify untrusted users based on their abnormal labeling behavior compared to the public and attempt to filter out these outliers before service recommendation, from the perspective of a two-stage ISODATA (Iterative Self-Organizing Data Analysis Technique Algorithm) clustering, and then propose a novel approach to service recommendation based on the resulting trusted users, named as Truser. Compared with the conventional *K*-means clustering algorithm, ISODATA is

收稿日期:2018-06-04;在线出版日期:2018-09-26.本课题得到国家重点研发计划(2017YFB1400602)、国家自然科学基金(61572371)、湖北省技术创新重大专项(2018ACA13)、湖北省教育厅青年人才项目计划(Q20171008)资助.**何 鹏**,男,1988 年生,博士,副教授,中国计算机学会(CCF)会员,主要研究方向为软件工程、复杂网络.E-mail: penghe@hubu.edu.cn.**吴 浩**,男,1997 年生,学士,主要研究方向为软件工程.**曾 诚**,男,1975 年生,博士,教授,中国计算机学会(CCF)会员,主要研究领域为服务计算.**马于涛**(通信作者),男,1980 年生,博士,副教授,中国计算机学会(CCF)高级会员,主要研究方向为软件工程、服务计算.E-mail: ytma@whu.edu.cn.

more flexible and increases the operation of "merging" and "splitting" to adjust the clustering number. In other words, the clustering center K value of ISODATA can be adjusted dynamically according to the actual situation. In this paper, we first perform ISODATA clustering for the concerned users of each service to label its candidate untrusted users. Then we get the number of times each user is marked as candidate untrusted, and the number of times any two users are clustered into the same group. According to the information obtained above, we further utilize ISODATA to cluster users, so as to filter out untrusted users eventually. Finally, we return Top-k similar users for the target user in the set of trusted users, and implement service recommendation based on the selection of similar trusted users. To verify the feasibility of our approach, an empirical study is conducted on two public datasets: Last.FM and Delicious. We compare the proposed method with other state-of-the-art recommendation methods with the general performance metric: Root Mean Square Error (*RMSE*). Experimental results show that, compared with the existing methods, the accuracies of recommendation of our approach achieved on the two datasets are improved by 16.1% and 4.5%, respectively. By learning the parameters of the proposed algorithm, we find that the recommendation result is the best when the expected clusters k is set to six during the first-stage clustering. Meanwhile, the experimental results also suggest that it is most suitable for the target user to return the Top-5 most similar trusted users, and to recommend services followed by at least 70% of those similar users. In general, the results roughly coincide with the practice. Therefore, it is reasonable to conclude that the quality of service recommendations can be improved by filtering out untrusted users.

Keywords ISODATA clustering; collaborative filtering; service recommendation; service computing

1 引 言

互联网的快速发展促使 Web 服务的数量与日俱增,可供用户选择的服务也越来越多.但这种便利背后带来的负面影响,是用户不得不花费更多的时间来寻找满足自己需求的服务.推荐系统可有效地帮助用户解决以上信息过载问题,快速找到适合自己的服务,甚至还可以根据用户的特点,提供个性化推荐.典型的服务推荐应用有商品推荐[1]、电影推荐[2-3]、图书推荐[4]、音乐推荐[5],等等.

协同过滤(collaborative filtering)是服务推荐中最为常用的方法之一,主要是利用用户历史信息来预测目标用户可能感兴趣的服务,包括基于事项(item-based)的协同过滤和基于用户(user-based)的协同过滤两类.前者是先计算事项之间的相似度,再向目标用户推荐与其过去感兴趣的事项最为相似的 Top-N 个事项;基于用户的协同过滤则先计算用户之间的相似度,然后把与目标用户最为相似的 Top-N 个用户感兴趣的事项推荐给目标用户.

除此之外,还有很多其他改进的推荐方法,如基于内容的推荐、基于模型的推荐、基于矩阵分解的推荐、基于情境感知的推荐以及基于各种方法混合的推荐.例如,Zheng 等人[6]综合考虑用户相似性和服务相似性,通过两者加权进行 QoS(Quality of Service)预测,从而实现服务推荐;Yao 等人[7]综合考虑用户评分相似性与语义内容相似性,构建基于用户协同过滤与基于内容的统一推荐模型;Zhang 等人[8]在各种时间情境(temporal context)下提出一种非负张量分解的 QoS 预测算法;Xie 等人[9]使用一个不对称矩阵表示用户和服务之间的隐藏关联,来缓解数据稀疏问题,从而提高服务推荐质量;Zheng 等人[10]集成基于用户的协同过滤与矩阵分解方法来预测服务 QoS 值.

上述工作主要聚焦于如何改进用户(事项)相似度算法、引入更丰富的用户信息或综合各种推荐算法,但以上所涉及的相似度计算有效的前提是所有用户信息均为真实可靠.然而,在现实中,难免会存在一些异常用户,他们出于各种目的,有意无意地对使用过的服务进行信息误导,影响推荐结果.例如,

电商平台上用户“刷”单行为、虚假好评和恶意差评，等等.另外，在一些开源服务平台上，用户评价信息也存在随意编写的情况.据统计[11]，Stack Overflow (https://stackoverflow.com)上有约 14.7% 的问题(issue)在发布后，用户对其标注的标签与问题本身关系不大，需要重新编辑.可想而知，如果不排除这部分不可靠用户信息的干扰，不论模型如何优化，服务推荐结果的准确性都会受到影响.

为此，Qiu 等人[12]定义了用户的信誉度量，优先使用信誉高的用户数据进行分析.但该方法中用户信誉指标容易受诸多因素影响，且不同情况的参数选取对结果的影响很大.Noorian 等人[13]根据环境条件、信息可获取性、参与人的行为信息构建了一个去中心化的自适应信用模型 zTrust，该模型虽取得了较好的服务推荐效果，但其实验条件要求较高，导致方法的通用性受限.Kim[14]根据用户在特定主题上分享的服务使用经历来度量用户的可信度.该方法也定义了一个具体的可信指标，但用户分享的信息非常有限，且依赖其他用户的点评来衡量用户的可信与否，计算工作量较大.为克服上述问题，Wu 等人[15]从聚类的角度，利用 K-means 方法对每个服务上的用户进行有效划分，并将离群用户视为该服务上的不可信用户，从而省去复杂的指标定义，使得影响因素明显减少.

然而，尽管 Wu 等人的方法取得了一定的效果，但 K-means 算法在使用前须事先设置 k 个聚类中心，并且该值在整个运算过程中无法更改.对于关注用户较少的服务与关注用户较多的服务，如采用相同的 k 值进行聚类，效果都会受到影响.例如，有两个服务 s_1 和 s_2，关注服务 s_1 的用户有 3 人，而关注服务 s_2 的用户有 50 人.若 K-means 聚类时 k 值取 2，对服务 s_2 而言，用户聚类效果仍较为粗糙；若 k 取值大于 3，则对服务 s_1 而言意义不大.因此，针对该问题，本文引入改进的 ISODATA(Iterative Self-Organizing Data Analysis Technique)聚类方法[16]，使得在聚类过程中，当属于某个分组的样本数过少时，可考虑把这个分组去除；当属于某个分组的样本数过多、分散程度较大时，可将该分组拆分为两个子分组，从而实现在每个服务上进行自适应的用户聚类.

本文应用无监督的 ISODATA 算法对服务的关注用户进行聚类，将离群用户视为不可信对象进行过滤，从而提出一种基于可信用户的服务推荐方法(Trustworthy User based Service Recommendation，Truser).本方法将采用两次 ISODATA 聚类来识别不可信用户，可概括如下：首先，针对每个服务，对所有关注(表现为使用标签对服务进行了标注)过它的用户进行聚类，并将聚类后规模最小的分组中的用户视为离群用户，即候选不可信用户；其次，基于用户被标记为候选不可信的次数再对用户进行第二次聚类，并将平均候选不可信值最大的分组中的用户视为最终的不可信用户；最后，在可信用户集中为目标用户推荐 Top-N 个最为相似的用户，并按照这些相似用户的偏好进行协同推荐.

本文的主要贡献可归纳如下：

(1)采用两次无监督的改进 K-means 聚类方法(ISODATA 算法)隐式地识别不可信用户，提出了一种基于可信用户的服务推荐方法 Truser，减少不可信用户数据对推荐结果的影响.

(2)在两个公共数据集上验证了所提方法的有效性，且相比已有的基准方法，在推荐精度上可分别提高 16.1% 和 4.5%.

本文第 2 节归纳聚类算法应用于服务推荐的一些相关工作；第 3 节重点介绍 Truser 服务推荐方法及相关算法；第 4 节陈述实验数据和实验结果分析；第 5 节主要讨论文中涉及的一些实验处理问题、应用价值与不足之处；最后一节为全文总结.

2　相关工作

2.1　服务推荐

一直以来，服务推荐作为服务计算领域的一个重要主题备受关注，其中基于协同过滤的推荐方法应用最广，其大致思路为采用相似用户(事项)的结果进行推荐，可划分为基于记忆的(memory-based)推荐、基于模型的(model-based)推荐、基于情境的(context-based)推荐等等.

基于记忆的推荐方法包括基于用户、事项或两者混合的协同过滤.Shao 等人[17]考虑用户的 QoS 相似性，提出了基于用户的协同过滤方法进行个性化 QoS 预测.Wang 等人[18]利用用户的定性定量偏好信息，提出了一种新的基于用户相似度的服务推荐方法，结果表明该方法不仅能够识别相似用户，还能提高推荐精度.Hu 等人[19]将多视图学习(multi-view learning)应用于商品推荐，通过将每个用户的购买记录作为一个视图，利用梯度下降算法获取商品之间的关系和用户之间的打分差异，从而将商品推荐给最可能购买的用户.Zheng 等人[6]综

合考虑用户相似性和服务相似性，通过线性组合加权的方式进行 QoS 预测，从而实现服务推荐.

基于模型的推荐旨在从已有数据中学习出特有的模式用于推荐. Chen 等人[20]基于事项的协同过滤和隐语义模型，提出了一种面向排序的服务推荐方法，通过服务之间排序结果的相关系数来度量服务相似度. Ren 等人[21]针对每个用户，利用 SVM 算法(Support Vector Machine)从其历史评分数据中学习一个超平面，再通过计算每个服务到该超平面的距离判断用户对服务的偏好程度，从而省去对服务的 QoS 或评分预测. Deng 等人[22]利用可扩展的随机游走算法进行服务推荐. Cao 等人[23]利用关系主体模型刻画 Mashup 和服务之间的关系以及挖掘潜在主题，再通过因子分解机训练潜在主题，从而实现 QoS 感知的服务推荐. 此外，还有研究者通过各种矩阵分解的方法进行服务推荐[5,9,24].

影响用户偏好和服务 QoS 的情境因素很多，比如时间、位置、日期等. Chen 等人[25]通过位置信息和服务历史 QoS 值对用户和服务进行聚类，再利用聚类后的结果做个性化服务推荐. 考虑到用户的移动性，Wang 等人[26]利用移动边缘计算构建了一种改进的服务推荐算法. Xie 等人[27]基于 SLOPE ONE 算法设计了一个情境感知的协同过滤算法，考虑用户和服务的位置、日期和时间等信息. Wei 等人[2]根据用户的评分信息与社会化标签信息，通过构建社会服务网络与偏好主题模型，提出了一种融合以上两类信息的推荐模型，以提升服务推荐质量.

另外，为有效度量用户的可信程度，有研究者提出构建对应的信誉系统，即通过打分方式衡量用户的信誉水平. 例如，文献[28]中作者通过引入第三方的用户信誉管理信息来排除不可信用户，并提出了一种改进的服务选择和排序方法. 考虑到现实中很难找到具有公信力的第三方，所以该方法的通用性受限. Qiu 等人[12]通过排除低信誉用户的数据，提出了一种信誉感知的方法，但该方法中信誉计算较为复杂，且受参数设置差异的影响显著. Wang 等人[18]结合定性与定量的用户偏好可信度，利用多目标优化的方法实现服务推荐. Wei[29]利用用户共现(co-occurrence)网络计算用户的定性定量权威性，优先根据权威用户的标签信息进行服务分类与个性化推荐. 不难发现，以上方法在判断用户可信与否时，均是通过一个给定的特定指标，且侧重于从某一个角度对用户的可信性进行度量.

2.2 聚 类

聚类(Clustering)的本质是寻找联系紧密的事物，把它们区分出来，使得同一分组中的事物比不同分组中的事物更具有共性，因而是数据分析中常用的一种技术. K-means 聚类旨在将一组对象划分为 k 个分组. 关于聚类算法，Xu 在文献[30]中做了详细综述，包括启发式聚类、基于 Kernel 的聚类、模糊聚类等. 聚类算法常用于离群值(outlier)的查找，如 Yoon 等人[31]曾使用 K-means 聚类算法来检测软件度量中数据偏小且与其他数据相距较远的离群值. Chen 等人[25]利用聚类算法识别区域敏感的服务，并提出了一种新的区域模型 RegionKNN. Wu 等人[32]则基于用户的相似度进行聚类来预测服务的 QoS 值. 潘等人[33]将 Mashup 与 API 关系抽象为一个面向服务的软件网络模型，利用复杂网络社区发现算法对软件网络进行聚类，实现服务的自动分类与推荐. Kim 等人[34]利用动态 K-means 聚类对用户的音乐列表进行聚类，根据聚类结果实现用户的个性化推荐.

以上工作中的聚类都是采用传统方法，其中使用较多的是 K-means 方法. 该算法聚类中心 k 为预先设置，且聚类过程中固定不变，这使得该参数的设置至关重要. 当遇到高维度、海量的数据集时，人们往往很难准确地给出 k 值. 为此，本文利用迭代自组织数据分析法(ISODATA)改进 K-means 方法，实现聚类过程中分组的灵活合并与分裂，从而更好地发现离群用户，提高服务推荐质量.

3 基于可信用户的服务推荐

以 Delicious(https://del.icio.us)公开的数据为例，根据我们的统计，80%的用户为书签(bookmark)附加的标签(tag)数不超过 5 个，90%的用户使用的标签数不超过 7 个. 这说明大部分用户的标注行为较为统一，只有极少数用户在给书签贴标签时的行为不同，但不排除存在某个正常用户对某一书签具有明显的个性化特征，导致标注的标签特别多. 若直接按照用户对书签的标注行为来判断用户的可信与否，也不够合理.

有鉴于此，我们采用聚类方式对用户进行划分，将在一个服务上标签标注行为相似的用户分为一组(即用户的标签相似性)，而把标注行为明显不同于其他用户的用户暂时标记为候选不可信用户，并

用候选不可信指数 *CUI*(Candidate Untrustworthy Index)记录每个用户在所有服务中被标记为候选不可信用户的次数. 一个用户的 *CUI* 值越大，表示该用户为服务添加标签的行为与大众差异越大，其不可信的可能性越高.

本文所提 Truser 方法的基本框架如图 1 所示，主要包括 3 部分：

(1) 用户标签加权向量化. 根据用户在服务上的标注信息，利用 TF-IDF(Term Frequency-Inverse Document Frequency)方法对用户的标签进行加权量化，形成表征用户偏好的标签向量.

(2) 两阶段的 ISODATA 聚类. 首先，根据(1)中得到的用户标签向量，对每个服务的关注用户采用 ISODATA 方法进行聚类，得到每个服务对应的候选不可信用户集 *UG*；其次，根据每个用户在所有服务上被标记为候选不可信用户的次数，计算每个用户的候选不可信指数 *CUI*；最后，根据用户的 *CUI* 值再次进行聚类，将用户划分为可信与不可信两类.

(3) 基于用户的协同推荐. 给定一个目标用户，从(2)中得到的可信用户集 *RU* 中为目标用户返回与其最相似的 Top-*N* 个用户，并根据这 *N* 个用户的偏好情况进行服务推荐.

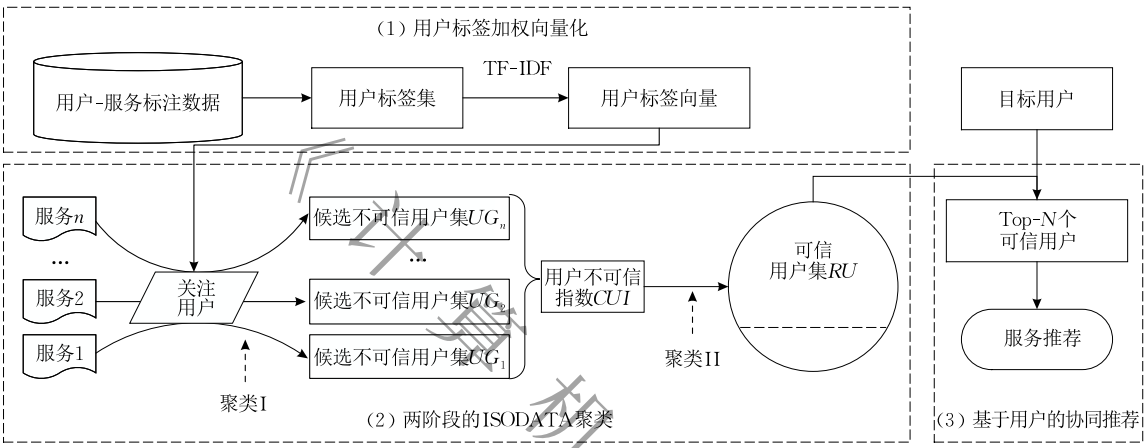

图 1　Truser 方法的框架

3.1　用户标签向量

考虑到在用户的标签集中，不同标签在代表用户兴趣偏好的权重上存在差异，如一个用户在标记不同服务时重复使用某一个标签，则该标签更倾向于代表此用户的兴趣偏好. 为有效度量用户标签的权重，我们利用经典的 TF-IDF 方法构建表示每个用户标签集的加权标签向量.

用 $\boldsymbol{L}(u)=(w_{u1},w_{u2},\cdots,w_{un'})$ 表示用户 u 的加权标签向量，其中 w 为该用户在每个标签上的权重，n' 为标签维度. 假设用户 u 的标签集中有 k 个标签，在所有用户(♯*users*)中有 ♯$user_i$ 个用户的标签集中包含标签 i，则标签 i 在用户 u 上的 tf_{ui} 和 idf_i 的计算公式为

$$tf_{ui}=\frac{f_{ui}}{\sum_k f_{uk}},\ idf_i=\log\frac{\sharp users}{\sharp user_i} \tag{1}$$

考虑到用户实际使用的标签非常有限，类似于文献[35]的处理，本文对 tf 和 idf 进行相应变换. 因此，修改后对应的 tf_{ui} 和 idf_i 为

$$tf_{ui}=\log(f_{ui}+1),\ idf_i=\log\frac{\sharp users}{\sharp user_i+1} \tag{2}$$

在式(2)中，tf_{ui} 表示用户 u 使用标签 i 标注服务的频率，即该用户在标记服务时使用标签 i 的次数 f_{ui} 越多，则 tf_{ui} 越大. 如果标签 i 被越多的用户所使用，表示该标签代表用户的兴趣独特性越低. 因此，用户 u 在标签 i 上的权值 w_i 的表达式为

$$w_{ui}=\log(f_{ui}+1)\cdot\log\frac{\sharp users}{\sharp user_i+1} \tag{3}$$

于是，可得到每个用户的标签向量 $\boldsymbol{L}$.

3.2　ISODATA 聚类算法

ISODATA 聚类算法在运行过程中能够根据各个聚类分组的实际情况采取两种操作来调整聚类中心数 K：一种是分裂操作，对应着增加聚类中心数；另一种是合并操作，对应着减少聚类中心数. 在使用该方法时，需要涉及以下几类参数的设置：

(1) 预期的聚类中心数目 K_0. ISODATA 算法在运行过程中聚类中心数目是可变的，但仍需要事先指定一个聚类中心的参考标准，然后算法针对参考标准 K_0，在$[K_0/2,2K_0]$范围动态调节.

(2) 每个聚类分组所要求的最少样本数目 $N_{\min}$. 用于判断当某个分组所包含样本分散程度较大时是否可以进行分裂操作. 如果分裂后会导致某个子分组包含的样本数偏小，则不对该分组进行分裂操作.

(3)最大方差 $\sigma_{\max}$. 用于衡量某个分组中样本的分散程度. 当样本的分散程度超过这个值时,则有可能进行分裂操作(注意分裂时需满足(2)中的条件).

(4)两个分组对应的聚类中心之间所允许的最小距离 $d_{\min}$. 如果两个分组靠得非常近(即这两个分组对应的聚类中心之间的距离小于 $d_{\min}$),则需对这两个分组进行合并操作.

ISODATA 算法的主要描述如算法 1,其中与 K-means 算法最大的不同为 Step5 的分裂操作和 Step6 的合并操作.

算法 1. ISODATA 算法的主要流程.

输入:样本集$\{x_i\}$

输出:聚类分组情况 $C=(C_0,C_1,\cdots,C_K)$

Step1. 从数据集中随机选取 K_0 个样本作为初始聚类中心 $C=(C_0,C_1,\cdots,C_{K_0})$;

Step2. 针对每个样本 x_i,计算它到 K_0 个聚类中心的距离,并将其分到距离最近的聚类中心所对应分组中;

Step3. 判断上述每个分组中的样本数是否小于 $N_{\min}$,如果小于则放弃该分组,令 $K=K-1$,并将该分组中的样本重新分配到下一个距离最小的分组中;

Step4. 对每个分组 C_i,根据其所有样本重新计算它的聚类中心 $c_i=\dfrac{1}{|C_i|}\sum_{x\in C_i}x$;

Step5. 如果当前 $K\leq\dfrac{K_0}{2}$,进行分裂操作;

(1)计算每个分组中所有样本在每个维度下的方差 σ;

(2)针对每个分组的所有方差选取最大值 $\sigma'_{\max}$;

(3)如果某个分组的 $\sigma'_{\max}>\sigma_{\max}$ 并且该分组的样本数量 $n\geq N_{\min}$,则进行分裂操作,前往步骤(4),否则退出分裂操作;

(4)$K=K+1$,分裂后的分组的聚类中心位置为 $c_i^+=c_i+\sigma_{\max}$, $c_i^-=c_i-\sigma_{\max}$;

Step6. 如果当前 $K\geq 2K_0$,进行合并操作;

(1)计算当前所有分组聚类中心之间的距离,用矩阵 $\boldsymbol{D}$ 表示,其中 $\boldsymbol{D}(i,i)=0$;

(2)对于 $\boldsymbol{D}(i,j)<d_{\min}(i\neq j)$ 的两个分组进行合并,合并后新分组聚类中心位置为

$$c_{\text{new}}=\frac{1}{n_i+n_j}(n_ic_i+n_jc_j);$$

Step7. 当迭代运行到聚类分组不再发生改变时终止,否则回到 Step2 继续执行.

3.3 不可信用户过滤

令 $\boldsymbol{M}_{m\times n}$ 表示用户与服务的关注矩阵,m 为用户总数,n 为服务总数,如果用户 i 对服务 j 用标签进行过标注,则表示用户关注该服务,矩阵元素 $q_{i,j}=1$,否则 $q_{i,j}=0$. 考虑到现实中大部分用户参与标记的服务相对有限,所以矩阵 $\boldsymbol{M}$ 其实为一个稀疏矩阵.

阶段一. 基于用户标签向量的聚类. 针对每个服务,对所有关注过它的用户根据 3.1 小节中得到的标签向量进行 ISODATA 聚类. 利用 ISODATA 算法对用户进行聚类,ISODATA 算法以欧式距离作为相似度计算方法,用户 u 到质心 c 的距离如式(4)所示.

$$d(u,c)=\sqrt{\sum_{i=1}^{n'}(w_{ui}-w_{ci})^2} \tag{4}$$

在 ISODATA 聚类中,需给定一个预期聚类分组参考标准 K_0,为了区分两次聚类中的预期聚类参数,第一次聚类的参数标记为 K_{01},第二次为 K_{02}. 根据文献[32]的处理方法(将少数与其他值分布差异较大的数据视为离群值)以及我们的前期调研情况,本文将每个服务上聚类后所得规模最小的那个分组中的用户标记为候选不可信用户.

在上述对每个服务的关注用户聚类之前,需要初始化 $\boldsymbol{A}_{m\times n}$ 和矩阵 $\boldsymbol{B}_{m\times m}$,它们分别用于记录该阶段聚类后每个服务上候选不可信用户和两个用户被聚类到同一个分组的次数. 换言之,如果用户 u 在服务 s 上聚类后被标记为候选不可信用户,则 $a_{u,s}=1$;如果用户 u 和用户 v 在 4 个不同的服务上都被聚类在同一个分组中,则 $b_{u,v}=4$.

聚类后,服务 s 上各分组的大小($|C_s^k|$)会有所差异,使用 C_s^k 表示服务 s 上第 k 个分组($K_{01}/2\leq k\leq 2K_{01}$),将规模最小的分组 $\arg\min_{1\leq k\leq K_1}|C_s^k|$ 中的用户标记为该服务上的候选不可信用户 $UG_s^{\min}$,如式(5)所示.

$$UG_s^{\min}=\{u|u\in C_s^t,t=\arg\min_{1\leq k\leq K_1}|C_s^k|\} \tag{5}$$

更新矩阵 $\boldsymbol{A}$ 中所有用户 $u(u\in UG_s^{\min})$ 的 $a_{u,s}=1$. 如果存在多个最小分组,则需将这些分组中的所有用户都标记为候选不可信用户. 同时,每当完成一个服务上用户的聚类,根据分组情况更新矩阵 $\boldsymbol{B}$;然后,对每个服务重复上述步骤,并更新矩阵 $\boldsymbol{A}$ 和矩阵 $\boldsymbol{B}$.

表 1 为一个具有 5 个用户和 4 个服务的样例数据集,l_{ij} 为用户在服务 i 上标注的第 j 个标签,对应的关注矩阵如图 2 所示. 值得注意的是,现实中不是所有的用户都会关注某个服务,而且同一个服务上两个用户也可以标注相同的标签. 例如,用户 u_4 和 u_5 没有关注服务 s_1,而用户 u_2 和 u_4 在服务 s_2 上拥有相同的标签集$\{l_{22},l_{23}\}$. 根据前述方法,若取 $K_{01}=2$ 时,对 4 个服务上的用户进行聚类,结果如图 3 所示. 每个服务上的用户被划分为 2 个分组,规模最小的分组被标记为 $UG_s^{\min}$. 服务 s_1 上 $C_{s_1}^1$ 分组包含用户

u_1 和 u_2，$C_{s_1}^2$ 分组只包含 u_3 一个用户，所以用户 u_3 将被记为服务 s_1 上的候选不可信用户（$UG_{s_1}^{\min}=\{u_3\}$），矩阵 $\boldsymbol{A}$ 中对应的 $a_{3,1}=1$. 类似地，其他三个服务的最小聚类分组分别为：$UG_{s_2}^{\min}=\{u_1\}$，$UG_{s_3}^{\min}=\{u_4,u_5\}$，$UG_{s_4}^{\min}=\{u_1,u_2,u_4,u_5\}$. 另外，如果两个用户在聚类过程中被聚到同一分组，则存在一次 $b_{i,j}$ 增加 1，最终更新后的矩阵 $\boldsymbol{A}$，$\boldsymbol{B}$ 如图 4 所示. 因用户 u_2 和用户 u_3 在服务 s_2 和服务 s_3 上都划分在同一分组，所以 $b_{2,3}$ 和 $b_{3,2}$ 的值为 2.

表 1　一个简单的样例数据

	s_1	s_2	s_3	s_4
u_1	$\{l_{11}\}$	$\{l_{21}\}$	$\{l_{31},l_{32}\}$	$\{l_{41},l_{45}\}$
u_2	$\{l_{11},l_{12}\}$	$\{l_{22},l_{23}\}$	$\{l_{32},l_{33},l_{34}\}$	$\{l_{42},l_{43}\}$
u_3	$\{l_{13}\}$	$\{l_{22},l_{23},l_{24}\}$	$\{l_{31},l_{32},l_{34}\}$	—
u_4	—	$\{l_{22},l_{23}\}$	$\{l_{35}\}$	$\{l_{42},l_{44}\}$
u_5	—	$\{l_{23},l_{24}\}$	$\{l_{35},l_{36}\}$	$\{l_{41}\}$

	s_1	s_2	s_3	s_4
u_1	1	1	1	1
u_2	1	1	1	1
u_3	1	1	1	0
u_4	0	1	1	1
u_5	0	1	1	1

关注矩阵 $\boldsymbol{M}_{5\times4}$

图 2　用户-服务关注矩阵

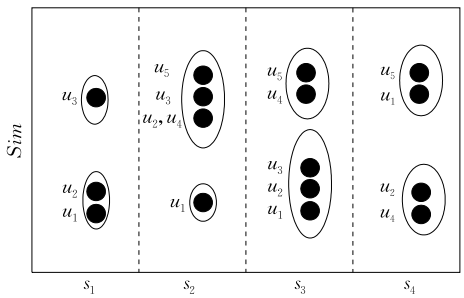

图 3　基于用户标签相似度的聚类

	s_1	s_2	s_3	s_4
u_1	0	1	0	1
u_2	0	0	0	1
u_3	1	0	0	0
u_4	0	0	1	1
u_5	0	0	1	1

$\boldsymbol{A}_{5\times4}$

	u_1	u_2	u_3	u_4	u_5
u_1	0	2	1	0	1
u_2	2	0	2	2	1
u_3	1	2	0	1	1
u_4	0	2	1	0	2
u_5	1	1	1	2	0

$\boldsymbol{B}_{5\times5}$

图 4　更新后的矩阵 $\boldsymbol{A}$ 和 $\boldsymbol{B}$

阶段二. 基于用户 *CUI* 指标聚类. 通过前面对每个服务上用户标签集的聚类，可得到每个用户被标记为候选不可信用户的次数，即 CUI_u 为用户 u 的候选不可信指数，表示如下：

$$CUI_u=\sum_{j=1}^{m}a_{u,j},\ 1\le u\le m \tag{6}$$

如果一个用户的 *CUI* 值明显高于其他用户，则该用户不可信的概率越大.

基于用户的 *CUI* 指标值，再次使用 ISODATA 聚类方法，将用户进一步划分为 K 个分组. 如文献[15]所述，用户可大致分为可信、不确定、不可信三种，所以第二阶段的聚类预期参数 K_{02} 设为 3，实际聚类分组数 K 的范围为[2,6]. 其中，用户整体 *CUI* 指标最高的分组，即被多次标记为候选不可信的用户，记为不可信用户集 *Hgroup*；用户整体 *CUI* 指标值最低的分组，记为可信用户集 *Lgroup*，其他分组用户均记为 *Mgroup*.

最后，再结合矩阵 $\boldsymbol{B}$ 中记录的两个用户被聚类到同一个分组的次数信息，去除 *Lgroup* 中与用户 $u(u\in Hgroup)$ 的 $b_{u,v}\ge\rho$ 的用户 $v(v\in Lgroup)$；同时，增加 *Mgroup* 中与用户 $u(u\in Lgroup)$ 的 $b_{u,v}\ge\rho$ 的用户 $v(v\in Mgroup)$，得到最终的可信用户集 *RU*.

$$RU=\{u\mid u\in Lgroup,v\in Hgroup,b_{u,v}<\rho \text{ 或 } u\in Mgroup,v\in Lgroup,b_{u,v}\ge\rho\} \tag{7}$$

对应算法的核心实现如算法 2 所示.

算法 2.　Truser 获取可信用户集.

输入：用户集 U、服务集 S、用户服务-标签信息

参数 K_{01}、K_{02}、$N_{\min}$、$\sigma_{\max}$、ρ

输出：可信用户集 *RU*

Step1. 初始化矩阵 $\boldsymbol{M}$，$\boldsymbol{A}$，$\boldsymbol{B}$，每个服务的不可信候选用户 $UG^{\min}$，每个用户的 *CUI*，*Hgroup*、*Lgroup*、可信用户集 *RU*；

Step2. 对每个用户 u，根据用户的标签集使用 TF-IDF 技术构建用户的加权标签向量 $\boldsymbol{L}(u)$；

Step3. 对每个服务 s，对 $\boldsymbol{M}$ 中所有 $q_{i,s}=1$ 的用户 i 的标签向量 $\boldsymbol{L}$ 进行 ISODATA 聚类；

Step4. 将每个服务上聚类后规模最小的分组 $\arg\min\limits_{1\le k\le K_1}|C_s^k|$ 中的用户标记为该服务 s 上的不可信候选用户 $UG_s^{\min}$，并更新矩阵 $\boldsymbol{A}$ 和 $\boldsymbol{B}$；

Step5. 根据矩阵 $\boldsymbol{A}$，计算每个用户的 $CUI_i=\sum\limits_{j=1}^{m}a_{i,j}$；

Step6. 再对所有用户的 *CUI* 值使用 ISODATA 算法进行聚类；

Step7. 获取 *Hgroup*、*Mgroup* 和 *Lgroup* 用户集；

Step8. 根据矩阵 $\boldsymbol{B}$，从 *Lgroup* 中排除与 *Hgroup* 中用户的 b 值超过 ρ 的用户，增加 *Mgroup* 中与 *Lgroup* 中用户的 b 值超过 ρ 的用户；

Step9. 返回最终的可信用户集 *RU*.

3.4 基于可信用户的协同推荐

根据聚类理论,同一聚类分组中的用户会倾向于具有相似的偏好特性.对于给定的目标用户 u,从矩阵 $\boldsymbol{B}$ 中获取与用户 u 的 b 值不低于 b_u^n 值的所有可信用户,组成 SU_u,其中 b_u^n 为用户 u 与所有其他用户的 b 值按从大到小排序后的第 n 个值,即为目标用户返回 Top-N 个最相似的可信用户.

$$SU_u=\{u'\mid b_{u,u'}\geq b_u^n,u'\neq u,u'\in RU\} \tag{8}$$

对于用户 u 未关注的服务 s',可根据用户 u 的可信用户 $u'(u'\in SU_u)$ 对服务 s' 的关注情况,计算该用户在该服务上的关注概率 $P_{u,s'}$.当 $P_{u,s'}$ 大于等于阈值 $\theta(0.5\leq\theta\leq1)$ 时,则 $q_{u,s'}=1$,即将服务 s' 推荐给用户 u.

$$P_{u,s'}=\frac{1}{|SU_u|}\sum_{u'=1}^{|SU_u|}q_{u',s'} \tag{9}$$

4 实验分析

4.1 实验数据

实验数据(https://grouplens.org/datasets/hetrec-2011)为两个具有代表性的公开数据集 Last.FM 和 Delicious.Delicious 数据集包含用户 1867 个,服务(书签)104 799 个,标签 53 388 个;Last.FM 数据集包含用户 1892 个,服务(音乐)17 632 个,标签 11 946 个,如表 2 所示.实验过程中,为减少偏差,我们仅保留数据集中关注服务数不少于 2 个的用户,以及关注用户不少于 5 个的服务.

表 2 实验数据的统计情况

数据集	用户数	服务数	标签数
Last.FM	1892	17 632	11 946
Delicious	1867	104 799	53 388

4.2 评价指标与基准方法

评价推荐结果质量的方法有很多,常用的有准确率、召回率、平均绝对误差、均方根误差等,本文采用均方根误差 *RMSE*(Root Mean Square Error)作为评价推荐质量的标准,*RMSE* 值越小,推荐质量越高.通过对比目标用户 u 在推荐服务 s 上的关注情况 $q_{u,s}$ 与用户实际关注情况 $\hat{q}_{u,s}$ 的差异,具体评价指标表示如下:

$$RMSE=\sqrt{\frac{\sum_{u,s}(q_{u,s}-\hat{q}_{u,s})^2}{N}} \tag{10}$$

此外,本文还选取 UPCC[17]、IPCC[36]、WSRec[6] 和 CAP[15] 四种已有方法作为实验的基准方法:

(1) UPCC 是一种使用皮尔逊相关性系数(Pearson Correlation Coefficient,PCC)的基于用户的协同过滤推荐方法.该方法的基本思想为,基于用户对服务的偏好找到相似用户,然后将相似用户喜欢的服务推荐给当前用户.

(2) IPCC 是一种使用 PCC 的基于事项的协同过滤推荐方法.该方法本思想为,基于用户对服务的偏好找到相似的服务,然后根据用户的历史偏好,推荐相似的服务给当前用户.

(3) WSRec 是一种综合了 UPCC 与 IPCC 的协同过滤(混合)推荐方法.

(4) CAP 是一种利用两阶段 K-means 聚类的推荐方法.

4.3 实验结果

4.3.1 推荐结果评价

根据 ISODATA 算法的基本要求,$N_{\min}$ 至少为 1,最多为 5,方差 σ 通常设置为 0.01.考虑到实验数据中每个服务至少有 5 个关注用户,那么对于只含有 5 个关注用户的服务,聚类结果至少为 1 个分组,最多为 5 个分组,即 $0\leq K_0/2\leq5$,$1\leq2K_0$,所以 $1\leq K_0\leq10$.同时,根据前面介绍,在第二阶段的聚类过程中,预期聚类分组数为 3,即 $K_{02}=3$.

为验证本文所提 Truser 方法的有效性,我们随机从用户-服务关注矩阵 $\boldsymbol{M}$ 中去除比例 r 为 1%~5%的 $q=1$ 元素.设置第一阶段聚类时参数 $K_{01}=5$,$N_{\min}=1$、$\sigma_{\max}=0.01$,最终为目标用户返回 Top-5 个相似可信用户,以及关注概率阈值 $\theta=0.5$.根据统计结果发现,聚类后两个用户之间的平均 b 值为 2.14,所以我们将 ρ 固定为 7(详细说明见讨论部分).推荐结果的 *RMSE* 值如表 3 所示.

首先,实验结果显示在不同比例 r 范围内,本文所提方法 Truser 的 *RMSE* 值比已有的四种方法都要小,即推荐准确度更高,尤其是与常规的 UPCC、IPCC 和 WSRec 相比,改进较为明显.对于 CAP 方法,在 Last.FM 数据集上,随着 r 的变化,Truser 的 *RMSE* 降低幅度在 7%~16%之间,其中 $r=1\%$ 时,所提方法改进幅度最大为 16.1%(以斜体表示);而在 Delicious 数据集上,Truser 方法的改进幅度整体较小,随着 r 的变化,*RMSE* 降低幅度在 5%以内,其中 $r=3.5\%$ 时,改进幅度达到最大 4.5%(以斜体表示).

其次,根据 Truser 在两个数据集上的 *RMSE* 值,不难发现 r 在 1%~5%范围内变化时,*RMSE* 值的变化幅度不大,在 Last.FM 数据集上,*RMSE* 平均变化幅度仅为 1.6%;在 Delicious 数据集上,*RMSE* 值平均变化幅度仅为 0.9%.由此可见,比例

表 3　推荐结果对比情况（表中数值为推荐结果的 *RMSE* 值，该值越小表示推荐越准确）

数据集	方法	去除关注矩阵 $\boldsymbol{M}$ 中 $q=1$ 元素的比例 r(%)								
		1	1.5	2	2.5	3	3.5	4	4.5	5
Last.FM	UPCC	0.868	0.875	0.878	0.884	0.890	0.893	0.899	0.904	0.907
	IPCC	0.809	0.831	0.841	0.857	0.870	0.875	0.885	0.894	0.898
	WSRec	0.801	0.801	0.797	0.794	0.775	0.787	0.782	0.778	0.791
	CAP	0.767	0.747	0.744	0.745	0.745	0.746	0.749	0.749	0.749
	Truser	**0.643**	**0.648**	**0.655**	**0.641**	**0.637**	**0.648**	**0.671**	**0.677**	**0.692**
	$\left(\frac{\text{CAP-Truser}}{\text{CAP}}\right)\times100\%$	***16.1*%**	13.3%	12.0%	14.0%	14.5%	13.1%	10.4%	9.6%	7.6%
Delicious	UPCC	0.789	0.792	0.794	0.800	0.806	0.808	0.814	0.819	0.822
	IPCC	0.709	0.733	0.744	0.764	0.778	0.785	0.797	0.806	0.811
	WSRec	0.676	0.677	0.676	0.677	0.677	0.678	0.678	0.680	0.694
	CAP	0.592	0.592	0.593	0.595	0.596	0.597	0.597	0.598	0.601
	Truser	**0.584**	**0.576**	**0.577**	**0.576**	**0.579**	**0.570**	**0.576**	**0.577**	**0.589**
	$\left(\frac{\text{CAP-Truser}}{\text{CAP}}\right)\times100\%$	1.3%	2.8%	2.7%	3.2%	2.8%	***4.5*%**	3.6%	3.6%	2.0%

r 对 Truser 的影响非常有限，说明本文所提方法具有较好的鲁棒性.

4.3.2　参数影响分析

(1) 聚类过程中参数 K_{01}、$N_{\min}$ 的影响分析

Truser 方法关键的一步是过滤不可信用户，即两阶段的 ISODATA 聚类过程，其中第二阶段的 ISODATA 聚类中 K_{02} 固定为 3，而第一阶段聚类的参数 K_{01} 是动态可调的，而且不同的取值会对最后可信用户集 RU 的生成有影响. 理论上，参数 K_{01} 的值越大，聚类时分裂操作会更多，聚类分组会更精细. 为此，我们进一步分析了第一阶段聚类时，参数 K_{01} 对推荐结果的影响情况. 根据 4.3.1 小节中结果显示，实验中用于测试的元素比例 r 对推荐结果的影响不明显. 考虑到测试元素比例在 2.5%时，结果在两个数据集上的综合效果是最好的. 因此，本次实验我们设置 $r=2.5\%$，并保持为目标用户返回 Top-5 个相似可信用户和关注概率阈值 $\theta=0.5$，$N_{\min}=1$、$\sigma_{\max}=0.01$，分析参数 K_{01} 从 1 增至 10 时的推荐结果.

如图 5 所示，$K_{01}=1$ 时，*RMSE* 最大，因为此时采用 ISODATA 算法将使得每个服务上用户聚类成 1 或 2 个分组. 这对于用户个数较多的服务而言，区分用户可信与不可信的效果并不明显，而且也容易出现所有用户聚成一个分组的情况，等价于对可信与不可信用户没有进行区分. 随着 K_{01} 值的增大，在两组数据集上，*RMSE* 值均呈先下降再有所上升，最后趋于稳定. 一种解释是 K_{01} 值增大，意味着初始分组增多，聚类过程中一些分组之间将进行有效的合并与分裂操作，使得每个服务上用户聚类的效果更精确；但随着 K_{01} 值的继续增大，将更多的是分组间进行分裂，将导致部分原被标记为不可信的候选用户将变为可信用户或不确定性用户，从而增加推荐结果的误差. 纵观实验结果，大体上 K_{01} 在等于 6 时，两个数据集下得到的 *RMSE* 值最小，分别为 0.625 和 0.560，即为推荐效果最佳.

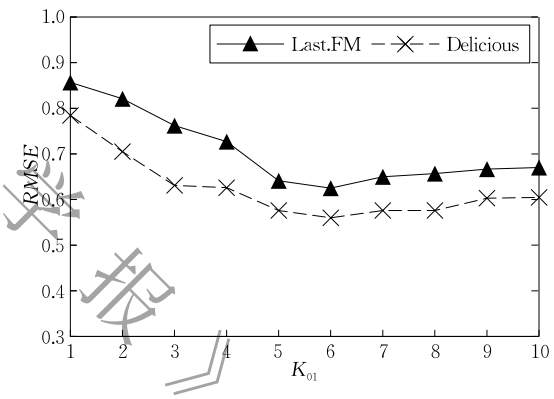

图 5　第一阶段 *K*-means 参数对推荐结果的影响

同样，设置 $K_{01}=6$，其他参数保持不变，令参数 $N_{\min}$ 从 1 到 5 变化. 结果如图 6 所示. $N_{\min}$ 越小越好，当 $N_{\min}$ 超过 3 时，推荐结果的 *RMSE* 开始显著增长，理由是数据集中拥有 10 个用户以内的服务占比

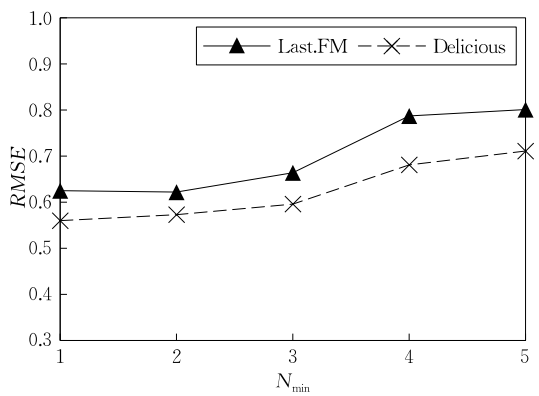

图 6　$N_{\min}$ 参数对推荐结果的影响

在 60%以上，当 N_{min} 增大时，聚类分组将越少，导致用户之间的区分度更低，使得通过聚类来识别离群用户的意义不显著.

(2) 协同推荐过程中参数 Top-N、θ 的影响分析

在推荐过程中，最终为目标用户返回的相似可信用户个数 N 也会对推荐结果的质量产生影响. 因此，我们在设置 $K_{01}=6$，且其他条件与实验(1)相似的条件下，对比分析为目标用户返回不同数量的可信用户用于协同推荐的情况，如图 7 所示.

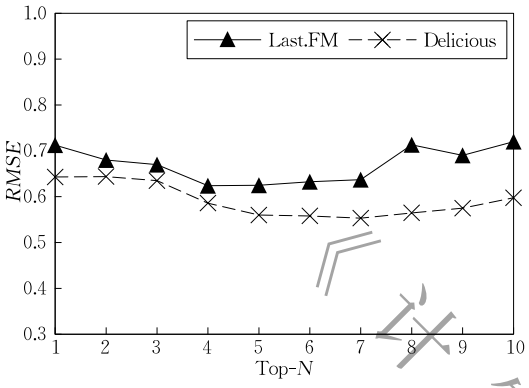

图 7 Top-N 参数对推荐结果的影响

图 7 结果显示，当只为目标用户返回最相似的一个可信用户(Top-$N=1$)时，由于可利用的信息受限且不排除随机性因素影响，推荐结果并不乐观. 在一般推荐情况中，Top-N 值太大也常会因引入的用户提供价值不高的信息反而增加推荐误差. 该问题在本次实验中也得到进一步验证，表现为当 Top-N 值在 4～7 之间时，$RMSE$ 值更小. 对于 Last.FM 数据集，Top-$N=4$ 时，$RMSE$ 最小为 0.624，对于 Delicious 数据集，则 Top-$N=7$ 时，$RMSE$ 值最小，为 0.553. 总之，为目标用户返回的 Top-N 个相似可信用户不宜太少或太多.

最后，对于用于判断目标用户是否会关注某一个服务的阈值 θ，在保持其他参数固定且 Top-$N=5$ 时，分析该参数对推荐结果的影响. 如图 8 所示，对

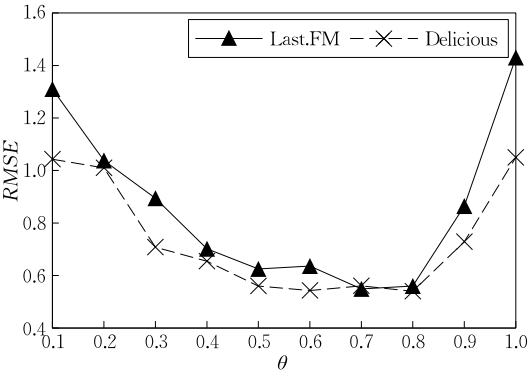

图 8 Top-$N=5$ 时阈值 θ 对推荐结果的影响

于数据集 Last.FM，阈值 θ 在[0.5，0.8]区间时，$RMSE$ 相对偏低，最低为 0.549($\theta=0.7$)，而其他范围下的 $RMSE$ 值增长幅度较大；对于数据集 Delicious，结果趋势类似，阈值 θ 的合适范围也在[0.5，0.8]区间时，$RMSE$ 值相对偏低，最低为 0.540($\theta=0.8$). 结果说明：在向目标用户推荐服务时，该用户的 Top-5 个相似可信用户中至少有 70%的用户关注了某服务，该服务才适合被推荐.

因此，本文所提的 Truser 方法在使用过程中，第一阶段聚类 K_{01} 取值 6，为目标用户推荐最相似的前 5 个可信用户，以及关注概率阈值 θ 取值为 0.7 或 0.8 时，推荐效果整体上会更好.

5 讨 论

本节主要讨论文中涉及的一些实验处理问题、应用价值与不足之处.

首先，在数据预处理阶段中，我们仅选取关注服务数不少于 2 个的用户，以及关注用户不少于 5 个的服务. 只关注一个服务的用户，其兴趣偏好可能存在一定随机性因素；而一个服务如果关注它的用户过少，在第一阶段的聚类过程中，随着参数 K_{01} 的增大，会出现分组个数超过用户数的情况，从而导致聚类失去意义. 在正常情况下，一个服务的关注用户越多，通过聚类划分用户的效果越明显.

其次，式(7)中阈值 ρ 是用于进一步过滤与不可信用户聚类到同一分组且达到一定次数的可信用户. 我们对用于存储用户被聚类到同一个分组次数的矩阵 $\boldsymbol{B}$ 进行了分析，发现绝大部分用户被聚类到同一个分组的次数不超过 7. 其中，在数据集 Last.FM 中，89%以上的用户对的 b 值不超过 7，而 Delicious 上接近 80%的用户对的 b 值不超过 7. 需要注意的是，统计结果不包含 $b=0$ 的情况. 因此，本文在两次聚类后生成可信用户集 RU 时，对阈值 ρ 采取设置为固定值 7.

然后，实验结果显示 K_{01}、Top-N 的取值都应适中，推荐效果才更好. 以 Last.FM 数据集为例，分析发现数据集中 80%的音乐的关注用户在 15 人以下，如果参数 K_{01} 偏大，容易使聚类过程中分裂出很多规模为 1 的用户分组，从而增加候选不可信用户比例，甚至使大部分用户被标记为候选不可信. 另外，图 7 实验比较结果显示，为目标用户返回 Top-5 个可信用户时，结果相对更好. 我们进一步根据聚类后用户矩阵的元素分布也可得到(如图 9 所示)，平均每个用户只存在 4.25 个 $b>0$ 的可信用户. 因此，

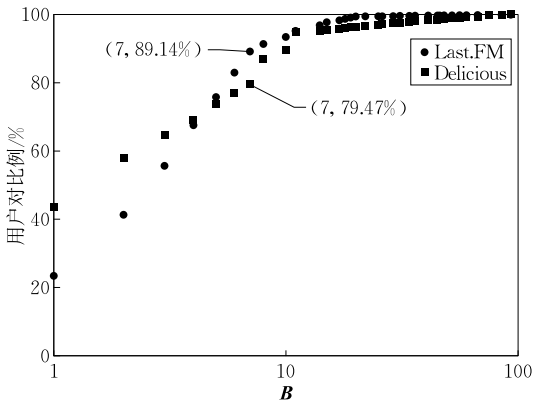

图 9　两个数据集聚类后矩阵 $\boldsymbol{B}$ 中元素值分布情况

返回的可信用户偏多容易使部分参考价值不大的用户被用于协同推荐，降低推荐质量.

最后，考虑到原始数据中，并没有给定用户可信与否的确切类别，在事先并不知道任何样本的类别的情况下，本文选用了聚类而非分类的方法进行不可信用户过滤，且并不需要关心具体是哪类不可信用户，只需把相似的东西自动聚到一起，属于无监督学习(unsupervised learning).

本文也存在以下一些不足的地方待改进：

(1) 我们只使用了 Last.FM 和 Delicious 两个数据集，目前可供使用的数据还包括 Movielens、WS-DREAM 等. 考虑到本次使用的两个数据集的质量已在大量前人研究工作中得以验证，我们选用它们作为我们的实验对象，相信我们的方法在其他满足所需信息的数据集中会同样适用.

(2) 众所周知，聚类方法很多，本文我们采用了 K-means 算法的一种改进方法 ISODATA，目前流行的改进方法还有 K-means++ 和 Kernel K-means，且除参数 K_{01} 外，整个算法中的方差 σ 设置为默认值. ISODATA 算法在聚类中心位置的选取延续了普通 K-means 方法，而在 K-means++ 中，聚类中心的选取采取互相离得越远越好的策略，这种策略不仅直观简单，而且也在某些数据集上被证实非常有效，后期工作可在此思路上进一步优化.

(3) 除本文选取的 UPCC、IPCC、WSRec 和 CAP 四种对比方法外，目前还有一些相关推荐方法. 表 4 统计了部分代表性推荐方法及其推荐过程中所利用的信息. 整体上，除了本文所利用的用户与服务之间关系矩阵和标签信息外，用户评分、服务描述内容、用户评论、社交关系和时间/位置的情境信息也常被用于服务推荐. 由此可见，每种方法所利用的信息种类不同，导致很难在一个公共的框架下进行横向对比. 因此，本文后续将考虑引入更多的信息来提高聚类质量，从而提升推荐效果，再与类似的方法进行比较.

表 4　部分推荐方法统计(第 1 列为各方法所利用的信息)

信息	Truser	CTR[37]	PMF[38]	HR[2]	Sorec[23]	MVIR[19]	WSPred[8]	AWSR[39]	VFDSR[40]
用户-服务关系	√	√	√	√	√	√	√	√	√
标签	√			√					
评分		√	√	√	√	√			
服务描述		√						√	
评论									√
社交关系					√				
时间/位置							√		

6　结　论

为排除不可信用户信息带来的干扰，提高服务推荐质量、确保用户的服务体验，本文从用户聚类的角度，采用两阶段的 ISODATA 算法过滤不可信用户，提出了一种基于可信用户的服务推荐方法 Truser. 所提方法在两个公开数据集 Last.FM 和 Delicious 上进行了实证分析. 实验结果表明：相比已有方法，本文所提方法在两个数据集上的推荐精度相比最好的基准方法 CAP 分别提高了 16.1% 和 4.5%，且发现第一阶段聚类中预期聚类分组个数为 6 时，推荐效果最好；同时，在推荐过程中为目标用户返回 Top-5 个可信用户最为适宜.

致　谢　在此向对本文实验过程中的算法设计工作给予支持的老师表示感谢，并向对本文工作提出宝贵评审意见的审稿专家表示衷心的感谢！

参　考　文　献

[1] Linden G, Smith B, York J. Amazon.com recommendations: Item-to-item collaborative filtering. IEEE Internet Computing, 2003, 7(1): 76-80

[2] Wei S, Zheng X, Chen D, et al. A hybrid approach for movie recommendation via tags and ratings. Electronic Commerce Research & Applications, 2016, 18(C): 83-94

[3] Lawrence E. Everything is a recommendation netflix, altgenres and the construction of taste. Knowledge Organization, 2015, 42(5): 358-364

[4] Xia Li-Xin, Li Chong-Yang, Cheng Xiu-Feng, et al. Multi dimension recommendation model of heterogeneous network resources—A case study of douban. Library & Information Service, 2017, 61(3): 6-13(in Chinese)
(夏立新, 李重阳, 程秀峰等. 异质网络资源多维度推荐模式研究——以豆瓣网为例. 图书情报工作, 2017, 61(3): 6-13)

[5] Benzi K, Kalofolias V, Bresson X, et al. Song recommendation with non-negative matrix factorization and graph total variation //Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing. New Orleans, USA, 2016: 2439-2443

[6] Zheng Z, Ma H, Lyu M R, et al. WSRec: A collaborative filtering based Web service recommender system//Proceedings of the IEEE International Conference on Web Services. Los Angeles, USA, 2009: 437-444

[7] Yao L, Sheng Q Z, Ngu A H H, et al. Unified collaborative and content-based Web service recommendation. IEEE Transactions on Services Computing, 2015, 8(3): 453-466

[8] Zhang Y, Zheng Z, Lyu M R. WSPred: A time-aware personalized QoS prediction framework for Web services// Proceedings of the IEEE International Symposium on Software Reliability Engineering. Dallas, USA, 2012: 210-219

[9] Xie Q, Zhao S, Zheng Z, et al. Asymmetric correlation regularized matrix factorization for Web service recommendation //Proceedings of the IEEE International Conference on Web Services. San Francisco, USA, 2016: 204-211

[10] Zheng Z, Ma H, Lyu M R, et al. Collaborative Web service QoS prediction via neighborhood integrated matrix factorization. IEEE Transactions on Services Computing, 2013, 6(3): 289-299

[11] Wang S, Lo D, Vasilescu B, et al. EnTagRec++: An enhanced tag recommendation system for software information sites. Empirical Software Engineering, 2018, 23(3): 800-832

[12] Qiu W, Zheng Z, Wang X, et al. Reputation-aware QoS value prediction of Web services//Proceedings of the IEEE International Conference on Services Computing. Santa Clara, USA, 2013: 41-48

[13] Noorian Z, Marsh S, Fleming M. zTrust: Adaptive decentralized trust model for quality of service selection in electronic marketplaces. Computational Intelligence, 2016, 32(1): 127-164

[14] Kim Y A. A trust prediction framework in rating-based experience sharing social networks without a Web of Trust. Information Sciences, 2012, 191(9): 128-145

[15] Wu C, Qiu W, Zheng Z, et al. QoS prediction of Web services based on two-phase *K*-means clustering//Proceedings of the IEEE International Conference on Web Services. New York, USA, 2015: 161-168

[16] Pall G H. ISODATA: An interactive method of multivariate data analysis and pattern classification//Proceedings of the IEEE International Communication Conference. 1966: 116-117

[17] Shao L, Zhang J, Wei Y, et al. Personalized QoS prediction for Web services via collaborative filtering//Proceedings of the IEEE International Conference on Web Services. Salt Lake City, USA, 2007: 439-446

[18] Wang H, Tao Y, Yu Q, et al. Incorporating both qualitative and quantitative preferences for service recommendation. Journal of Parallel & Distributed Computing, 2018, 114: 46-69

[19] Hu Q Y, Zhao Z L, Wang C D, et al. An item orientated recommendation algorithm from the multi-view perspective. Neurocomputing, 2017, 269: 261-272

[20] Chen M, Ma Y, Hu B, et al. A ranking-oriented hybrid approach to QoS-aware Web service recommendation// Proceedings of the IEEE International Conference on Services Computing. New York City, USA, 2015: 578-585

[21] Ren L, Wang W. An SVM-based collaborative filtering approach for Top-*N* Web services recommendation. Future Generation Computer Systems, 2017, 78: 531-543

[22] Deng S, Huang L, Xu G. Social network-based service recommendation with trust enhancement. Expert Systems with Applications, 2014, 41(18): 8075-8084

[23] Cao B, Liu J, Tang M, et al. Mashup service recommendation based on user Interest and social network//Proceedings of the IEEE International Conference on Web Services. Santa Clara, USA, 2013: 99-106

[24] Xu J, Zheng Z, Lyu M R. Web service personalized quality of service prediction via reputation-based matrix factorization. IEEE Transactions on Reliability, 2016, 65(1): 28-37

[25] Chen X, Liu X, Huang Z, et al. RegionKNN: A scalable hybrid collaborative filtering algorithm for personalized Web service recommendation//Proceedings of the IEEE International Conference on Web Services. Miami, USA, 2010: 9-16

[26] Wang S, Zhao Y, Huang L, et al. QoS prediction for service recommendations in mobile edge computing. Journal of Parallel & Distributed Computing, 2017, DOI: https://doi.org/10.1016/j.jpdc.2017.09.014

[27] Xie Qi, Wu Kai-Gui, Xu Jie, et al. Personalized context-aware QoS prediction for Web services based on collaborative filtering//Proceedings of the International Conference on Advanced Data Mining and Applications. Brussels, Belgium, 2010: 368-375

[28] Wang H, Wang L, Yu C. Integrating trust with qualitative and quantitative preference for service selection//Proceedings of the IEEE International Conference on Services Computing. Anchorage, USA, 2014: 299-306

[29] Wei J. A personalized authoritative user-based recommendation for social tagging. Future Generation Computer Systems, 2018, 86: 355-361

[30] Xu R. Survey of Clustering Algorithms. New York, USA: IEEE Press, 2005

[31] Yoon K A, Kwon O S, Bae D H. An approach to outlier detection of software measurement data using the *K*-means clustering method//Proceedings of the International Symposium on Empirical Software Engineering and Measurement. Madrid, Spain, 2007: 443-445

[32] Wu J, Chen L, Feng Y, et al. Predicting quality of service for selection by neighborhood-based collaborative filtering. IEEE Transactions on Systems Man & Cybernetics Systems, 2013, 43(2): 428-439

[33] Pan Wei-Feng, Li Bing, Shao Bo, et al. Service classification and recommendation based on software networks. Chinese Journal of Computers, 2011, 34(12): 2355-2369(in Chinese)
(潘伟丰，李兵，邵波等. 基于软件网络的服务自动分类和推荐方法研究. 计算机学报，2011，34(12)：2355-2369)

[34] Kim D, Kim K S, Park K H, et al. A music recommendation system with a dynamic *K*-means clustering algorithm// Proceedings of the International Conference on Machine Learning and Applications. San Diego, USA, 2008: 399-403

[35] He Peng, Li Bing, Yang Xi-Hui, et al. Roster: An approach to potential peer recommendation for developers. Chinese Journal of Computers, 2014, 37(4): 859-872(in Chinese)
(何鹏，李兵，杨习辉等. Roster：一种开发者潜在同行推荐方法. 计算机学报，2014，37(4)：859-872)

[36] Sarwar B, Karypis G, Konstan J, et al. Item-based collaborative filtering recommendation algorithms//Proceedings of the International Conference on World Wide Web. Hyderabad, India, 2001: 285-295

[37] Wang C, Blei D M. Collaborative topic modeling for recommending scientific articles//Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Diego, USA, 2011: 448-456

[38] Mnih A, Salakhutdinov R R. Probabilistic matrix factorization. Advances in Neural Information Processing Systems, 2007, 20(2): 1257-1264

[39] Kang G, Liu J, Tang M, et al. AWSR: Active Web service recommendation based on usage history//Proceedings of the International Conference on Web Services. Honolulu, USA, 2012: 186-193

[40] Wang H, Chi X, Wang Z, et al. Extracting fine-grained service value features and distributions for accurate service recommendation//Proceedings of the IEEE International Conference on Web Services. Honolulu, USA, 2017: 277-284

HE Peng, born in 1988, Ph. D., associate professor. His current research interests include service-oriented software engineering, complex networks.

WU Hao, born in 1999, bachelor. His current research interest is software engineering.

ZENG Cheng, born in 1975, Ph. D., professor. His current research interest is services computing.

MA Yu-Tao, born in 1980, Ph. D., associate professor. His current research interests include software engineering and services computing.

Background

Service recommendation aims at providing users with helpful information to improve their experience of services while interacting with information service systems. The primary goal of personalized service recommendation is to recommend the most likely services for target users to meet their specific requirements. However, a critical challenge to intelligent service recommendation is how to eliminate noise information of service annotation effectively from untrusted users when using collaborative filtering.

To solve this problem, several measures have been taken in some existing studies. For example, some calculate the reputation of each user based on their contributed values, and then take advantage of reputation-based ranking to exclude the values contributed by untrustworthy users. Others calculate the trustworthiness of users according to environmental conditions, information availability, and behavioral dispositions. The quality of recommendations can indeed be improved using the above methods. However, in some context, the calculation of reputation or trustworthiness of user is challenging and not practical.

Unlike those previous studies that defined specific metrics for untrusted users, in this study, we attempt to filter out outliers, viewed as untrusted users because of their uncommon behavior compared to the public, from the perspective of clustering, and then propose a novel approach based on trusted users to recommend personalized services to users using a two-stage ISODATA clustering algorithm. Experimental results indicate that the proposed approach outperforms the other four baselines.

This work was supported by the National Key Research and Development Program of China (No. 2017YFB1400602), the National Natural Science Foundation of China (No. 61572371), the Technology Innovation Special Program of Hubei Province (No. 2018ACA13), and the Educational Commission of Hubei Province (No. Q20171008).