计算机科学
COMPUTER SCIENCE

面向兴趣点推荐系统的自然噪声过滤算法

朱俊, 韩立新, 宗平, 徐逸卿, 夏吉安, 唐铭

引用本文

朱俊, 韩立新, 宗平, 徐逸卿, 夏吉安, 唐铭. 面向兴趣点推荐系统的自然噪声过滤算法[J]. 计算机科学, 2023, 50(11): 132-142.

ZHU Jun, HAN Lixin, ZONG Ping, XU Yiqing, XIA Ji'an, TANG Ming. Natural Noise Filtering Algorithm for Point-of-Interest Recommender Systems [J]. Computer Science, 2023, 50(11): 132-142.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.230400045

面向兴趣点推荐系统的自然噪声过滤算法

朱 俊[1,2] 韩立新[2] 宗 平[2] 徐逸卿[1] 夏吉安[1] 唐 铭[1]
1 南京工业职业技术大学计算机与软件学院 南京 210023
2 河海大学计算机与信息学院 南京 211100

摘 要 推荐系统源数据中存在着固有的自然噪声,给推荐算法带来了误差与干扰。现有研究更加关注以各类安全攻击为代表的恶意噪声,仅有少数文献针对更为隐蔽、更难处理的自然噪声进行研究,且这些研究几乎都集中在传统推荐领域。在兴趣点推荐场景中,无论是源数据特征,还是自然噪声的产生原因和表现方式,均与传统推荐领域有较大差别。针对兴趣点推荐系统中的自然噪声,提出了基于离散特征量化与聚类距离分析的自然噪声过滤算法 NFDC。该算法定义并计算用户签到数据的离散度,量化数据驱动的不确定性,利用推荐算法的准确度(F1 值)量化预测驱动的不确定性,深入挖掘两者之间的相关性,构建经验模型,推导潜在自然噪声比例;采用模糊 C 均值聚类方法分析用户行为模式的相似性,在聚类距离分析的基础上筛选可疑噪声,并自定义噪声验证规则,删除真正的自然噪声。在两个真实的位置社交网络数据集(Brightkite 和 Gowalla)中,分别采用 NFDC 算法和其他 4 种基准方法对源数据进行预处理,将处理后的数据集分别输入到 5 类代表性的兴趣点推荐算法中,对比不同的降噪技术对提升各类兴趣点推荐算法准确性的影响程度。实验结果表明,NFDC 算法能够有效降低系统源数据中的自然噪声,为后续的推荐算法提供可靠的输入。与其他降噪数据集中的最高推荐精度相比,各类推荐算法在 NFDC 处理后的 Brightkite 和 Gowalla 数据集中的准确度分别平均提高了 15.95%和 5.00%。

关键词:推荐系统;兴趣点推荐;自然噪声;不确定性;离散度;聚类
中图法分类号 TP181

Natural Noise Filtering Algorithm for Point-of-Interest Recommender Systems

ZHU Jun[1,2], HAN Lixin[2], ZONG Ping[2], XU Yiqing[1], XIA Ji'an[1] and TANG Ming[1]
1 School of Computer and Software, Nanjing Vocational University of Industry Technology, Nanjing 210023, China
2 College of Computer and Information Engineering, Hohai University, Nanjing 211100, China

Abstract The inherent natural noise in the original dataset of recommender systems(RSs) causes error and interference to recommendation algorithms. Existing studies pay more attention to the malicious noise represented by various security attacks. The natural noise which is more subtle and difficult to deal with has rarely been documented. Most researches about natural noise are conducted for conventional RSs. However, the data feature and the causes and forms of natural noise in point-of-interest(POI) RSs are all different from those in conventional RSs. To filter the natural noise for POI RSs, a novel natural noise filtering method (NFDC) based on dispersion quantification and clustering distance analysis is proposed. The dispersion of a subset of the original check-in dataset is defined and calculated to indicate the data-driven uncertainty, and the accuracy metric F1 is adopted to represent the prediction-driven uncertainty. The measures of dispersion and accuracy metric vectors are empirically categorized to identify the proportion of the potential noise. The fuzzy C-means-based denoi-sing algorithm is performed to analyze the similarity of user behavior patterns and then screen the potentially noisy points based on clustering distance analysis. A customized rule is designed to further verify and delete the natural noise. Extensive experiments are conducted on two real-world location-based social network datasets, Brightkite and Gowalla. The datasets processed by NFDC and the other four benchmark algorithms are respectively input into five representative POI recommendation algorithms for comparison. Experimental results show that NFDC effectively filters the natural noise and provides reliable input for RSs. Compared with the highest accuracy supported by other denoi-

到稿日期:2023-04-07 返修日期:2023-07-04
基金项目:国家自然科学基金(41771251);江苏省高校自然科学研究项目(21KJB520009);南京工业职业技术大学引进人才科研启动基金(YK23-05-01)
This work was supported by the National Natural Science Foundation of China(41771251), Natural Science Foundation of the Higher Education Institutions of Jiangsu Province, China(21KJB520009) and Start-up Fund for New Talented Researchers of Nanjing Vocational University of Industry Technology(YK23-05-01).
通信作者:朱俊(zj_zijin@163.com)

sing methods, the accuracy in NFDC-processed Brightkite and Gowalla datasets is respectively improved by 15.95% and 5.00% on average.

Keywords　Recommender system, Point-of-Interest recommendation, Natural noise, Uncertainty, Dispersion, Clustering

1　引言

推荐系统(Recommender System, RS)作为一种有效的信息过滤技术，既能够主动帮助用户解决大数据背景下的信息过载问题，又可以为商家吸引更多的潜在客户，是位置社交网络中重要的服务技术[1]。为了提升推荐系统的预测准确度，大多数研究人员将推荐算法的设计作为主要研究内容，而忽略了源数据集中无法避免的噪声信息[2]。早在2006年，O'Mahony等就提出推荐系统中存在着两类固有的噪声数据(恶意噪声和自然噪声)[3]。根据误差传播定律，在使用同一推荐算法的情况下，不同的噪声过滤方法会传递不同程度的不确定性给推荐算法，造成推荐结果存在差异[4]。噪声数据使推荐系统成为一个“伪专家”，极大地降低了用户满意度。因此，推荐系统的研究不仅要关注推荐算法的设计，还应重视源数据中不同性质的噪声干扰信息，为推荐算法提供可靠的输入保障。

作为传统推荐系统与位置社交网络协同发展的必然产物，兴趣点(Point-Of-Interest, POI)推荐已成为推荐系统领域的一个研究热点[5-6]。目前，与自然噪声处理相关的研究大多都集中在以非地理特征对象(如电影、音乐、笑话等)为推荐目标的传统推荐系统中，在兴趣点推荐领域几乎没有与自然噪声处理相关的文献。实际上，在两类推荐系统中，无论是源数据特征，还是自然噪声的产生原因和表现方式，均有较大差别(见表1)。

表1　兴趣点推荐系统和传统推荐系统中的自然噪声差别

Table 1　Differences of natural noise in POI RSs and conventional RSs

推荐系统类别	推荐对象	源数据	影响推荐准确度的主要因素	自然噪声产生原因	自然噪声表现方式
传统推荐系统	音乐、图书、电影等非地理特征项目	用户显式评分	用户偏好	用户在急躁、开玩笑或消遣等情况下的随意评分	不一致的评分
兴趣点推荐系统	具有地理经纬度坐标的“位置”	用户签到记录	用户偏好 社交关系 地理距离 访问时间	网络通信信号不稳定 定位技术不够精确 用户无意的签到行为	多样化、离散化的签到数据

传统推荐系统中的自然噪声主要表现为用户评分的偏差和不一致性[7]，而在兴趣点推荐系统中，引起噪声的原因以及噪声数据的表现方式则更为复杂。本文面向兴趣点推荐系统，分析用户签到行为中的多样化、离散化特征，将恶意噪声定义为商家为了某种目的而刻意制造的虚假签到记录。例如，商家经常以赠送小礼品的方式吸引大量用户进店(增加了该商店的签到记录)，或对某平台进行安全攻击(如托攻击、注入攻击等)[8]，这种在人为的、有目的的引导下产生的“伪热门”数据就属于恶意噪声[9]。将兴趣点推荐系统中的自然噪声定义为用户由于外部因素的影响而无意引入的不一致的、离散的签到信息，具体表现为：网络通信信号不稳定或定位技术不够精确而导致用户无意地访问非目标地址；个别用户由于情绪波动或个人习惯问题而产生访问地址多样化、离散化的特征。在基于签到记录生成的兴趣点评分矩阵中，这些非恶意的签到偏差分布于某行或某列中，最终表现为评分数据中的自然噪声。恶意噪声往往具有一定的规律性，能够通过分析用户的统计特征或使用攻击检测算法等技术进行处理，而自然噪声更为隐蔽，且不服从任何规律和分布，很难通过格式化、规则化的程序简单地将其过滤掉[10-11]。因此，在兴趣点推荐系统中，如何针对自然噪声数据设计有效的识别和过滤方法，以降低数据的误差传播率，是一个值得研究的难题。

目前，已有的自然噪声过滤研究大多存在着以下几点不足：

1)相关研究都集中在传统推荐领域，几乎没有文献在兴趣点推荐场景下研究自然噪声。传统推荐系统中的自然噪声处理技术在兴趣点推荐领域是否适用尚有待验证。如何面向兴趣点推荐系统，针对用户签到行为的离散化特征设计合理的离散度量化方法，是兴趣点推荐系统中自然噪声过滤的关键问题。

2)一些自然噪声过滤方法[12-14]除了需要原始评分之外，还需要额外收集大量的项目和用户信息，大大地影响了推荐系统的执行效率。另一些文献[4,15]基于严格的二值逻辑识别自然噪声，忽略了自然噪声固有的不确定性和模糊性。

3)推荐系统中存在着能够引发自然噪声的各类不确定性，这些不确定性可以概括为两类：数据驱动的不确定性和预测驱动的不确定性。前者来源于用户签到偏好以及兴趣点特征的多样性[16]，后者则是在推荐过程中由推荐算法的预测机制导致的[17]。已有一些研究工作围绕不确定性展开研究，但未深入分析两类不确定性之间的潜在关联，降低了自然噪声过滤的可解释性。

为了解决以上问题，本文面向兴趣点推荐系统设计了一种基于离散特征量化与聚类距离分析的自然噪声过滤方法，主要贡献如下：

1)定义并计算用户访问位置的一般频次距离与频次矫正距离，分析用户签到数据的离散度，量化数据驱动的不确定性；利用推荐算法的准确度量化预测驱动的不确定性。通过构建经验模型，深入挖掘两类不确定性之间的相关性。

2)采用模糊C均值聚类方法分析用户签到行为的相似

性，分别定义数据点到聚类中心的相对距离和绝对距离，在聚类距离分析的基础上定义自然噪声的验证规则，删除自然噪声。

3）在两个真实的位置社交网络数据集中开展了大量的离线实验，首先利用5种降噪方法对两个源数据集进行处理，然后将处理后的数据分别输入至5类经典的兴趣点推荐算法中，比较50组推荐场景中的预测精度。实验结果表明，本文算法能够有效降低数据干扰和误差传播，在面向兴趣点推荐系统进行自然噪声处理时具有明显优势。

2 相关工作

自然噪声指用户由于外部因素的影响而无意引入的不一致的评分数据[10]。Cosley等[18]曾经基于MovieLens网站做过一个经典的二次评分实验，证实了自然噪声的存在。Amatriain等[12]提出让用户对已评价过的项目重新打分，并将两次评分之差与预先设定好的门限值进行比较，以确定原评分是否为自然噪声。尽管这种方法能够显著提高推荐精度，但重新评分的过程对于用户来说是一项繁重的任务，实际上也难以完成。Pham等[13]设计了一个交互式推荐系统，先根据一部分评分为用户建立个人偏好模型，再根据该模型判断用户的其他原始评分是否合理，若某项目与之前构建的用户偏好模型相似度较低，而用户对该项目的评分却偏高，或者项目的主属性符合用户偏好模型却获得了较低的评分，则将该类评分识别为自然噪声。Yu等[14]制定了噪声分级指标，并根据分级结果将用户分成3组，针对不同的组别采取相应的评分修正方法。Li等[19]基于评分信息和用户评论，提出了一种基于情感的多指标集成评分方法，采用加权平均法生成更能反映用户偏好的综合评分，解决了用户评分和评论之间的不一致问题。

以上方法除了原始评分之外，还需要另外收集与项目和用户相关的其他信息，大大地影响了降噪算法的执行效率。因此，只需要利用评分矩阵而不依赖额外信息的降噪模型被相继提出。基于同一用户对高相关性的项目评分应该相似这一理论假设，Li等[20]认为自然噪声是违反了该理论的、自我矛盾的用户行为，并将捕获这种“自我矛盾”的机制形式化为约束二次优化问题。Toledo等[15]首先对用户、项目和评分进行分类，提取评分背后的隐含知识，进而识别潜在噪声。在检测到噪声评分后，采用协同过滤推荐算法进行预测，基于原始值和预测值之间的差异来纠正噪声。Bag等[4]用Bhattacharya系数代替Pearson相关系数，用二次分类方法代替再预测方法来校正自然噪声。

尽管以上文献针对自然噪声提出了各类处理方法，但其对噪声的管理是严格且死板的，这与自然噪声固有的不确定性和模糊性相违背[21]。因此，Sharon等[11]提出了一种基于模糊语言方法的自然噪声检测算法，缓解了非恶意噪声检测中的不确定性问题。Yera等[21-22]利用模糊逻辑方法灵活地处理电影推荐系统中的自然噪声。Castro等[23]首次提出面向群组推荐系统过滤自然噪声，随后又提出利用模糊学习方法改进群组推荐中的自然噪声处理效果[10]。Wang等[24]对电影和音乐评分进行模糊分类，根据用户和项目的模糊属性检测自然噪声，并根据最大隶属度原则对自然噪声进行更新替换。

现有的自然噪声研究几乎都集中在传统推荐领域，很少有文献在兴趣点推荐场景中研究自然噪声。然而，两类推荐系统中的源数据特征、自然噪声的产生原因和表现方式均有较大差别，兴趣点推荐系统中的自然噪声主要存在于用户多样的、离散的签到行为中，因此如何面向兴趣点推荐系统设计合理的签到行为离散度量化方法，是兴趣点推荐系统中自然噪声过滤的关键问题。

3 问题描述及概念定义

3.1 问题描述

兴趣点推荐系统中的自然噪声指由于外部因素的影响而无意引入的不一致的、离散的用户签到信息。设原始签到数据集 $\boldsymbol{C}$ 中包含 n 条签到记录，记作 $\boldsymbol{C}=\{c_1,c_2,\cdots,c_n\}$，每一条签到记录均是由用户ID、签到时间、地理纬度、地理经度以及兴趣点ID五元组表示。基于签到数据集 $\boldsymbol{C}$，采用FIF方法[25]将签到记录转换为用户-兴趣点评分矩阵 $\boldsymbol{R}=\{r_{ij}\}$：

$$r_{ij}=\frac{n_{ij}}{lc_j}\times\log_{10}\frac{NL}{ln_i}+\frac{n_{ij}}{uc_i}\times\log_{10}\frac{NU}{un_j} \tag{1}$$

其中，$i\in[1,NU]$，$j\in[1,NL]$，NL 和 NU 分别表示签到数据集中的兴趣点数量和用户数量，评分 r_{ij} 代表用户 u_i 对某个兴趣点 l_j 的历史喜爱程度，n_{ij} 是用户 u_i 在兴趣点 l_j 上的访问次数，lc_j 表示兴趣点 l_j 被访问的总次数，uc_i 表示用户 u_i 的总签到次数，ln_i 表示用户 u_i 访问过的兴趣点个数，un_j 表示访问过兴趣点 l_j 的用户总数。

兴趣点推荐系统中的自然噪声过滤的主要任务是：根据签到数据集 $\boldsymbol{C}$ 量化用户签到行为的离散化特征，基于用户-兴趣点评分矩阵 $\boldsymbol{R}$ 分析用户签到行为模式的差异性，筛选签到数据中的自然噪声并将其删除。完成自然噪声过滤后的数据集可作为后续推荐算法的输入，以提升推荐性能。

3.2 算法框架

本文提出的基于离散特征量化与聚类距离分析的自然噪声过滤（Natural Noise Filtering based on Dispersion Quantification and Clustering Distance Analysis，NFDC）算法主要包含以下两个步骤（见图1）。

步骤1 潜在自然噪声比例估计。首先通过无放回的随机重采样方法从原始签到数据集中随机选取一部分数据构建若干个子数据集，定义并计算用户访问位置的一般频次距离与频次矫正距离，量化用户访问兴趣点的离散度；然后，在各个子数据集中运行推荐算法，对推荐结果进行准确性评估；利用签到数据的离散度量化数据不确定性，利用推荐准确度量化预测不确定性，探索两类不确定性之间的相关性，拟合经验模型，推导潜在自然噪声的比例。

步骤2 自然噪声筛选、验证和删除。利用模糊C均值聚类算法对用户评分进行聚类，分析用户行为模式的相似性，分别定义数据点到聚类中心的相对距离和绝对距离。在聚类距离分析的基础上，结合步骤1输出的潜在噪声比例筛选可疑噪声，并定义自然噪声的验证规则，删除真正的自然噪声。

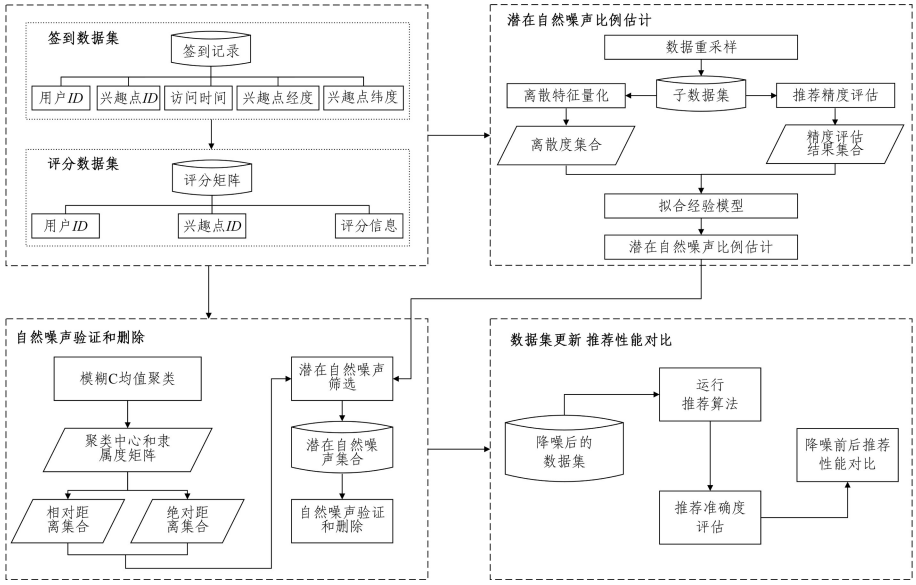

图 1　NFDC 算法的执行流程

Fig. 1　Workflow of NFDC algorithm

3.3　概念定义

定义 1(一般频次距离)　设用户 u 已访问过的所有位置的地理中心点为 P,则用户 u 访问兴趣点 l 的一般频次距离为中心点 P 到位置 l 的地理距离乘以用户 u 在位置 l 上的总签到次数。

定义 2(频次矫正距离)　设用户 u 访问兴趣点 l 的次数为 c,设置 c 个与签到频次成正比的矫正系数。用户 u 访问兴趣点 l 的频次矫正距离为 c 项之和,其中每一项为用户已访问位置的中心点 P 到位置 l 的地理距离乘以该次访问的矫正系数。

定义 3(用户签到数据的离散度)　基于用户 u 已访问的兴趣点集合 $\boldsymbol{L}$,用户 u 签到数据的离散度为用户 u 在集合 $\boldsymbol{L}$ 中的频次矫正距离之和与一般频次距离之和的比值。

定义 4(数据点与聚类中心的绝对距离)　数据点 r_i 与聚类中心 cen_k 之间的绝对距离为:

$$dis_i=\sqrt{\sum_{l=1}^{NL}(r_{il}-cen_{kl})^2} \tag{2}$$

定义 5(数据点与聚类中心的相对距离)　依据模糊 C 均值聚类的隶属度矩阵 $\boldsymbol{Mem}$,确定最大隶属度为聚类中心 cen_k 的数据点集合,计算这些数据点距离 cen_k 的绝对距离的平均值,将其记作 $dism_k$。数据点 r_i 与聚类中心点 cen_k 之间的相对距离为绝对距离 dis_i 与平均距离 $dism_k$ 的比值。

4　基于离散特征量化与聚类距离分析的自然噪声过滤算法

4.1　潜在自然噪声比例估计

4.1.1　面向数据不确定性的离散特征量化

与传统推荐系统不同,兴趣点推荐系统中的自然噪声主要表现为用户签到行为的多样化和离散化特征。这些非恶意的签到偏差分布于兴趣点评分矩阵的某行或某列中,造成评分数据的不确定性。NFDC 算法通过计算用户签到数据的离散度来量化数据驱动的不确定性。

首先,在原始签到数据集 $\boldsymbol{C}$ 中采用无放回的随机重采样方法选择部分用户及其历史签到记录构成子数据集,每次被选中的用户数量占比为 $p(0<p<1)$。独立执行 $iter$ 次随机重采样后,得到签到子数据集的集合 $\boldsymbol{C}_s=\{\boldsymbol{C}_{s\_1},\boldsymbol{C}_{s\_2},\cdots,\boldsymbol{C}_{s\_iter}\}$。设某个子数据集 $\boldsymbol{C}_{s\_r}(1\leqslant r\leqslant iter)$ 中被选中的用户集合为 $\boldsymbol{U}_{s\_r}=\{u_1,u_2,\cdots,u_{p\times NU}\}$,各用户的访问位置个数为 $\{m_1,m_2,\cdots,m_{p\times NU}\}$。

设用户 $u_i(1\leqslant i\leqslant p\times NU)$ 已访问的兴趣点集合为 $\boldsymbol{L}_i=\{l_1,l_2,\cdots,l_{mi}\}$,在各位置上的签到次数分别为 $\{c_1,c_2,\cdots,c_{mi}\}$。计算 $\boldsymbol{L}_i$ 集合中所有位置的地理中心点并将其记作 P_i。

按照定义 1 计算用户 u_i 访问兴趣点 l_j 的一般频次距离 dp_{ij}:

$$dp_{ij}=dis(P_i,l_j)\times c_j \tag{3}$$

其中,可根据中心点 P_i 的经纬度$\langle lon_p,lat_p\rangle$和位置 l_j 的经纬度$\langle lon_j,lat_j\rangle$计算两者之间的地理距离:

$$dis(P_i,l_j)=R*\arccos[\sin lat_p*\sin lat_j+\cos lat_p*\cos lat_j*\cos(lon_p-lon_j)] \tag{4}$$

其中,R 为地球半径,$R=6\,371\,\mathrm{km}$。

按照定义 2 计算用户 u_i 访问兴趣点 l_j 的频次矫正距离 $\tilde{d}p_{ij}$:

$$\tilde{d}p_{ij}=\begin{cases}dis(P_i,l_j), & c_j=1\\ dis(P_i,l_j)+\sum_{k=2}^{c_j}dis(P_i,l_j)\times\log_{cmax}k, & c_j\geqslant 2\end{cases} \tag{5}$$

其中,$cmax$ 为子数据集中所有用户签到次数的最大值。可以看出,矫正系数随着用户访问次数的增加而逐渐增大,能够反映出用户重复访问某兴趣点的频次。

基于用户 u_i 已访问的兴趣点集合 $\boldsymbol{L}_i=\{l_1,l_2,\cdots,l_{mi}\}$,按照定义 3 计算用户 u_i 签到数据的离散度:

$$disper_i=\begin{cases}0, & m_i=1\\ \dfrac{\sum_{j=1}^{m_i}\tilde{d}p_{ij}}{\sum_{j=1}^{m_i}dp_{ij}}, & m_i\geqslant 2\end{cases} \tag{6}$$

值得注意的是,用户 u_i 重复访问兴趣点 l_j 的频次越高,频次矫正距离 $\tilde{d}p_{ij}$ 的项数越多,$\tilde{d}p_{ij}$ 与一般频次距离 dp_{ij} 的差距

就越大，两者之和的比值则越小。因此，离散度 $disper_i \in [0, 1]$，其值越小代表用户的签到行为越集中，当 $disper_i = 0$ 时，表示用户只在 1 个地址上访问过若干次；反之，$disper_i$ 越大则表示用户签到行为的离散特征越明显，当 $disper_i = 1$ 时，表示用户在所有地址上均只访问过一次，签到行为的离散度非常高。

在求出子数据集 $\boldsymbol{C}_{s\_r}$ 中所有用户的离散度 $\{disper_1, disper_2, \cdots, disper_{p\times NU}\}$ 后，计算本次子数据集中的平均离散度：

$$AD\_r = \frac{\sum_{i=1}^{p\times NU} disper_i}{p \times NU} \tag{7}$$

在 $iter$ 个子数据集中按以上方法独立计算各个子数据集的平均离散度，记录 $iter$ 次随机重采样的离散度结果集合 $\boldsymbol{AD} = \{AD\_1, AD\_2, \cdots, AD\_iter\}$。

4.1.2 面向预测不确定性的推荐准确度评估

为了量化预测驱动的不确定性，在各个签到记录子数据集对应的评分子数据集 $\boldsymbol{R}_{s\_r}$ ($1 \leqslant r \leqslant iter$) 中独立运行推荐算法 Ars，并对推荐结果进行精度评估，计算推荐算法在当前子数据集中的精确率(Precision)、召回率(Recall)和综合指标 F1 值。

推荐算法 Ars 在子数据集 $\boldsymbol{R}_{s\_r}$ 中的精确率和召回率分别为：

$$precision\_r = \frac{\sum_{i=1}^{p\times NU} \frac{|\boldsymbol{Rec}(\boldsymbol{u}_i) \cap \boldsymbol{Like}(\boldsymbol{u}_i)|}{N_{\text{train}}}}{p \times NU} \tag{8}$$

$$recall\_r = \frac{\sum_{i=1}^{p\times NU} \frac{|\boldsymbol{Rec}(\boldsymbol{u}_i) \cap \boldsymbol{Like}(\boldsymbol{u}_i)|}{|Like(u_i)|}}{p \times NU} \tag{9}$$

其中，N_{train} 是在推导潜在自然噪声比例时为了评估推荐模型的准确性而事先设定的推荐列表长度，$\boldsymbol{Rec}(\boldsymbol{u}_i)$ 是提供给用户 u_i 的推荐列表，$\boldsymbol{Like}(\boldsymbol{u}_i)$ 是用户 u_i 真正喜欢的兴趣点集合。

为了衡量精确率和召回率之间的平衡度，采用 F1 值量化子数据集中预测驱动的不确定性。

$$F1\_r = \frac{2 \times precision\_r \times recall\_r}{precision\_r + recall\_r} \tag{10}$$

按以上方法独立计算推荐算法在 $iter$ 个评分子数据集中的推荐精度，记录 $iter$ 次 F1 值结果集合 $\boldsymbol{AF} = \{F1\_1, F1\_2, \cdots, F1\_iter\}$。

4.1.3 两类不确定性的相关性挖掘

NFDC 算法系统地整合数据驱动和预测驱动的不确定性，着重研究离散度集合 $\boldsymbol{AD} = \{AD\_1, AD\_2, \cdots, AD\_iter\}$ 以及准确度集合 $AF = \{F1\_1, F1\_2, \cdots, F1\_iter\}$ 之间的相关性，构建经验模型，推导潜在自然噪声的比例。

首先，将离散度集合 $\boldsymbol{AD}$ 和准确度集合 $\boldsymbol{AF}$ 中的元素按从高到低的顺序排序，将排序后的 $\boldsymbol{AD}'$ 集合划分为 $\boldsymbol{HAD}$，$\boldsymbol{MAD}$ 和 $\boldsymbol{LAD}$ 这 3 组，将排序后的准确度集合 $\boldsymbol{AF}'$ 划分为 $\boldsymbol{GAF}$，$\boldsymbol{MAF}$ 和 $\boldsymbol{BAF}$ 这 3 类(见表 2)。$\boldsymbol{HAD}$ 对应的子数据集的离散度较大，表示由用户签到行为的离散特征引起的数据不确定较高，容易引起较大的推荐误差。$\boldsymbol{MAD}$ 和 $\boldsymbol{LAD}$ 则分别对应包含中等不确定性和低不确定性的子数据集，基于这些数据集生成的推荐结果分别会产生中度误差和低推荐误差。

表 2 离散度的计算场景和推荐准确度的分类情况

Table 2 Computing scenarios for dispersion quantification and classification for prediction accuracy

	符号定义	表示含义	所取元素
离散度计算场景	$\boldsymbol{HAD}$	离散度高	$\boldsymbol{AD}'$ 集合前 1/3 元素
	$\boldsymbol{MAD}$	离散度中等	$\boldsymbol{AD}'$ 集合中间 1/3 元素
	$\boldsymbol{LAD}$	离散度低	$\boldsymbol{AD}'$ 集合后 1/3 元素
推荐准确度分类	$\boldsymbol{GAF}$	预测准确性好	$\boldsymbol{AF}'$ 集合前 1/3 元素
	$\boldsymbol{MAF}$	预测准确性中等	$\boldsymbol{AF}'$ 集合中间 1/3 元素
	$\boldsymbol{BAF}$	预测准确性差	$\boldsymbol{AF}'$ 集合后 1/3 元素

基于以上半定量描述的分类方式，将离散度计算场景和推荐准确度组合，进一步量化潜在自然噪声的存在程度，定义潜在自然噪声数量的 3 个级别 *Much*，*Moderate* 和 *Little*(见表 3)。为了确定 3 个数量级别对应的具体比例值，对处理自然噪声的相关文献进行经验总结，发现在数据集中一般存在着约 4%～10% 的潜在自然噪声[15,23]。基于此，NFDC 算法将 *Much* 级别数据集的自然噪声比例设定为 6%，将 *Moderate* 级别数据集的自然噪声比例设定为 4%，将 *Little* 级别数据集的自然噪声比例设定为 2%。

表 3 子数据集中潜在自然噪声存在程度的量化

Table 3 Quantification of potential natural noise in subsets

离散度所属集合	推荐精度所属集合	潜在自然噪声数量级别	含义
$\boldsymbol{HAD}$ 数据不确定性高	$\boldsymbol{BAF}$ 推荐准确性差	*Much*	潜在自然噪声较多
$\boldsymbol{MAD}$ 数据不确定性中等	$\boldsymbol{MAF}$ 推荐准确性中等	*Moderate*	潜在自然噪声中等
$\boldsymbol{LAD}$ 数据不确定性低	$\boldsymbol{GAF}$ 推荐准确性好	*Little*	潜在自然噪声较少

由于自然噪声数据是随机分布的，当执行足够多次的随机重采样后(例如 $iter \geqslant 100$)，子数据集中潜在噪声数据的比例结果也适用于总数据集。因此，将所有子数据集的潜在噪声比例集合 $\{\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_{iter}\}$ 的平均值作为源数据集最终的潜在噪声比例结果 ε。

$$\varepsilon = \frac{\varepsilon_1 + \varepsilon_2 + \cdots + \varepsilon_{iter}}{iter} \tag{11}$$

4.2 自然噪声的筛选、验证和删除

考虑到兴趣点项目特征的描述不可避免地具有模糊性，且用户的行为偏好也具有一定的多样性，用户对兴趣点的反馈结果可能是不精确的、主观的，NFDC 算法采用模糊 C 均值聚类方法甄别潜在自然噪声。

4.2.1 潜在自然噪声筛选

对用户-兴趣点评分数据集 $\boldsymbol{R} = \{r_{ij}\}$ ($i \in [1, NU]$，$j \in [1, NL]$) 中的数据进行模糊 C 均值聚类，生成 c 个聚类中心 $\boldsymbol{Cen} = \{cen_1, cen_2, \cdots, cen_c\}$ 和最终的隶属度矩阵 $\boldsymbol{Mem}$。

假设 r_i 为第 i 个数据点，在隶属度矩阵 $\boldsymbol{Mem}$ 中，第 i 行的最大值为 mem_{ik}。按照定义 4 和式(2)计算数据点 r_i 与聚类中心 cen_k 之间的绝对距离 dis_i，同时，按照定义 5 计算数据点 r_i 与聚类中心点 cen_k 之间的相对距离 dis_i'：

$$dis_i' = \frac{dis_i}{dism_k} \tag{12}$$

与文献[26]在处理噪声数据时将噪声点聚为一类不同，兴趣点推荐系统中用户和地址特征的多样性导致了噪声数据可能不会收敛于某一集群中[27]。因此，NFDC 算法针对各个

聚类，将聚类内的数据点按相对距离从大到小排序，依据潜在自然噪声比例 ε，将每个聚类中距离聚类中心相对距离较高的若干个数据点加入到潜在自然噪声数据集 ***Up*** 中。

4.2.2　自然噪声验证和删除

潜在自然噪声数据集 ***Up*** 中的数据点仅是疑似噪声，是否为真正的自然噪声数据还有待于进一步验证。假设 r_i 为第 i 个数据点，除了其最大隶属度聚类中心点 cen_k 之外，数据点 r_i 与其他聚类中心点之间的相对距离定义为 $\boldsymbol{Diso}=\{diso_1,\cdots,diso_{c-1}\}$。针对所有聚类，逐个计算各聚类中的平均绝对距离，形成集合 $\boldsymbol{Dism}=\{dism_1,\cdots,dism_c\}$。根据聚类的距离分析模式[26]，NFDC 算法定义了潜在自然噪声数据的验证规则。

针对 ***Up*** 中的每个潜在自然噪声数据，如果同时满足以下两个条件：

1）数据点 r_i 与其最大隶属度聚类中心点 cen_k 的相对距离 dis_i' 小于该数据点的 ***Diso*** 集合中的所有元素；

2）数据点 r_i 与其最大隶属度聚类中心点 cen_k 的相对距离 dis_i' 大于 ***Dism*** 集合中的所有元素。

则认为该数据点为真正的自然噪声数据，将其加入到自然噪声数据集 ***Un*** 中。

最后，从原始签到数据集 ***C*** 和用户-兴趣点评分数据集 ***R*** 中删除自然噪声集合 ***Un***，分别得到过滤自然噪声后的签到数据集 $\boldsymbol{C}'$ 和评分数据集 $\boldsymbol{R}'$。

5　实验结果与分析

5.1　实验数据集

本文在两个公开的位置社交网络数据集 Brightkite[1] 和 Gowalla[2] 中进行离线实验。Brightkite 数据集包含在 2008 年 4 月－2010 年 10 月期间 58 228 名用户在该网站上的社交关系和签到记录。Brightkite 数据集中的兴趣点数量为 693 362，用户在兴趣点上的签到记录共 4 747 281 条，用户之间形成了 214 078 条社交关系。Gowalla 数据集包含在 2009 年 2 月－2010 年 10 月期间 196 591 名用户在该网站上的社交关系和签到信息。Gowalla 数据集中的兴趣点数量为 1 256 379，用户在兴趣点上的签到记录共 6 442 892 条，用户之间形成了 950 327 条社交关系。

5.2　数据预处理

由于 Brightkite 和 Gowalla 数据集中的兴趣点和签到记录主要分布在美国地区，其他地区的数据分布非常零散，而过分稀疏的源数据和不足量的数据规模可能会导致实验结果的偏差[6]。因此，为了便于对数据进行比较和分析，本文实验首先将观测数据范围限定在美国地区，然后根据兴趣点的经纬度，筛选出美国地区的签到记录，并标注每个美国地点的县（county）名。在对各个县的兴趣点数量和签到次数进行统计和排序后，选择美国数据量最丰富的县（Brightkite 数据集中的 Los Angeles 县和 Gowalla 数据集中的 Travis 县）作为实验观测对象。

此外，为了减少数据集中的异常数据给推荐结果带来的影响，进一步过滤无社交关系的、不活跃的用户 u 和不受欢迎的兴趣点 l[28]。设定的过滤规则如下。

1）删除孤立用户：只考虑存在至少 1 条社交关系的用户。

2）删除不活跃的用户：在 Brightkite 的 Los Angeles 数据集中，只考虑访问过至少 2 个位置的用户；对于 Gowalla 中的 Travis 数据集，将阈值调整为 10 以缓解稀疏性，即只考虑访问过至少 10 个位置的用户。

3）删除冷门的位置：若某位置被访问的用户数少于 NV，则认为该地点是冷门的、不受欢迎的，可以考虑将其删除；对于 Brightkite 的 Los Angeles 数据集，$NV=2$；对于 Gowalla 中的 Travis 数据集，$NV=10$。

过滤后的社交关系和签到数据按照式（1）转化为用户-兴趣点二维评分矩阵。最终，Brightkite 中的 Los Angeles 数据集（下文中简称为“Brightkite”）包含 1 233 名用户和 2 951 个兴趣点，Gowalla 中的 Travis 数据集（下文中简称为“Gowalla”）包含 3 280 名用户和 3 335 个兴趣点。数据集的其他统计信息如表 4 所列。

表 4　数据预处理后 Brightkite 和 Gowalla 数据集的统计

Table 4　Statistics of Brightkite and Gowalla datasets after preprocessing

统计项	Brightkite (Los Angeles)	Gowalla (Travis)
用户数量	1 233	3 280
兴趣点数量	2 951	3 335
社交关系数量	8 432	36 050
签到记录数量	61 710	200 817
平均每个用户的签到记录数量	50.05	61.22
平均每个用户的社交关系数量	6.84	10.99
平均每个兴趣点被访问的次数	20.91	60.21
用户-兴趣点评分矩阵的稀疏度	0.983	0.982

5.3　基准方法选择

5.3.1　基准预处理方法选择

本文分别采用 NFDC 算法和以下 4 种基准方法对推荐系统源数据进行预处理，4 种方法的基本原理如下。

1）Base：按照第 4.2 小节中的流程对签到数据集进行评分矩阵转换，在生成评分矩阵后不进行任何噪声处理。

2）Diff-based[3]：该方法首次面向推荐系统提出了自然噪声和恶意噪声的定义，是能够同时处理恶意噪声和自然噪声的经典算法，已成为推荐系统领域噪声研究的重要基础。在自然噪声的识别过程中，该方法借助推荐算法生成的预测评分，衡量原始评分与预测评分之间的一致性，通过比较一致性程度与自定义门限值来确定原始评分是否为自然噪声。

3）NN-Crisp[15]：该方法专门针对协同过滤推荐系统中的自然噪声进行识别和纠正。首先，利用评分情况对用户、项目和评分进行分类，根据分类结果识别出可疑噪声数据；然后，采用协同过滤推荐算法进行评分预测，基于原始值和预测值之间的差异确定并纠正自然噪声。

4）PCA-based[29]：利用在数据压缩、冗余消除和噪声消除等领域都有着广泛应用的主成分分析（Principal Components Analysis，PCA）方法进行噪声数据的判断，主要针对恶意配置的属性进行检测和删除。

以上 4 类方法中，Base 方法仅过滤无社交关系的、不

1) https://snap.stanford.edu/data/loc-Brightkite.html

2) http://snap.stanford.edu/data/loc-Gowalla.html

活跃的用户和不受欢迎的兴趣点；其余方法均是在此基础上进一步过滤不同种类的噪声数据：PCA-based 方法是恶意噪声消除的经典方法；NN-Crisp 方法主要针对自然噪声进行数据清洗；Diff-based 方法则兼顾恶意噪声和自然噪声的识别。因此，从研究对象和研究方法来看，以上 4 种基准方法均具有一定的代表性。

5.3.2 推荐算法选择

完成数据预处理之后，需要将不同的数据处理结果输入至推荐算法中，以比较不同的降噪技术对各类推荐算法准确性提升的影响程度。考虑到兴趣点推荐作为推荐系统的一个重要代表，无论是发展历程还是关键技术，都与传统推荐系统一脉相承。因此，在推荐算法选择方面，首先选择 3 种经典的协同过滤算法，即基于用户的协同过滤算法 UBCF[30]、基于项目的协同过滤算法 IBCF[31] 和奇异值分解算法 SVD[32]，然后考虑到位置社交网络中关系上下文和位置上下文对用户签到行为的重要影响，分别选择基于社交关系的兴趣点推荐算法 FCF[33] 和基于地理特征的核密度估计算法 KDE[34] 作为推荐模型。

5.4 两类不确定性的相关性研究

在进行潜在自然噪声比例估计时，采用无放回的随机重采样方法选择子数据集，设定每次被选中的用户比例为 $p=0.7$，随机重采样次数为 200（$iter=200$）。图 2 和图 3 给出了在 Brightkite 和 Gowalla 中 200 个子数据集的离散度 $\{AD_{\_1}, AD_{\_2}, \cdots, AD_{\_200}\}$ 和 $F1$ 值 $\{F1_{\_1}, F1_{\_2}, \cdots, F1_{\_200}\}$ 的频率分布直方图，图中虚线表示离散度和 F1 值的平均值。

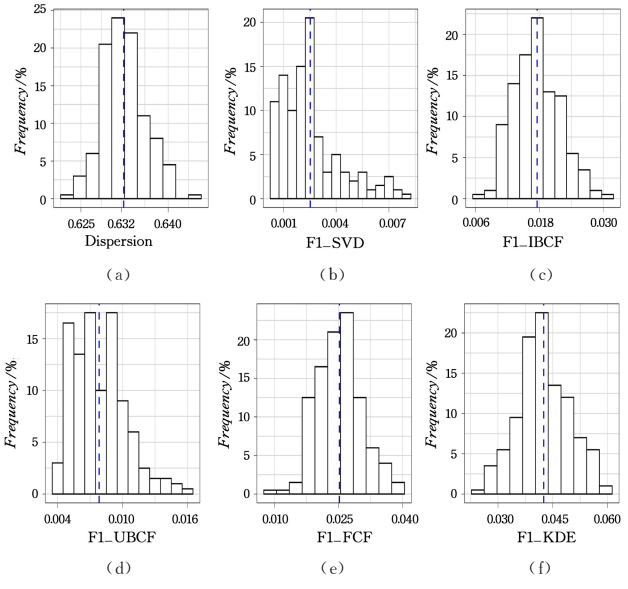

图 2 Brightkite 子数据集中离散度和 F1 值的频率分布直方图

Fig. 2 Histogram of frequency distribution of dispersion and F1 values in Brightkite subsets

基于图 2 和图 3 中的 200 组离散度和各类推荐算法的 F1 值，选择常见的皮尔逊相关系数（Pearson Correlation Coefficient）作为反映离散度和推荐准确度之间相关性的统计指标，皮尔逊相关系数 r 及其对应的显著性水平 p 值（p value）如表 5 所列。

从表 5 可以看出，不管是哪种推荐算法，其 F1 值都与离散度具有显著相关性（p 值均小于 0.05），且均呈现负相关关系（皮尔逊相关系数 r 均小于 0），即数据驱动的不确定性越高（离散度越大），推荐算法的预测准确度就越低。由此可见，本文将离散度和推荐准确度分类组合在一起，从而计算潜在自然噪声比例是可行且符合实验发现的。最终，Brightkite 和 Gowalla 数据集在各推荐场景中的潜在自然噪声比例如表 5 所列。

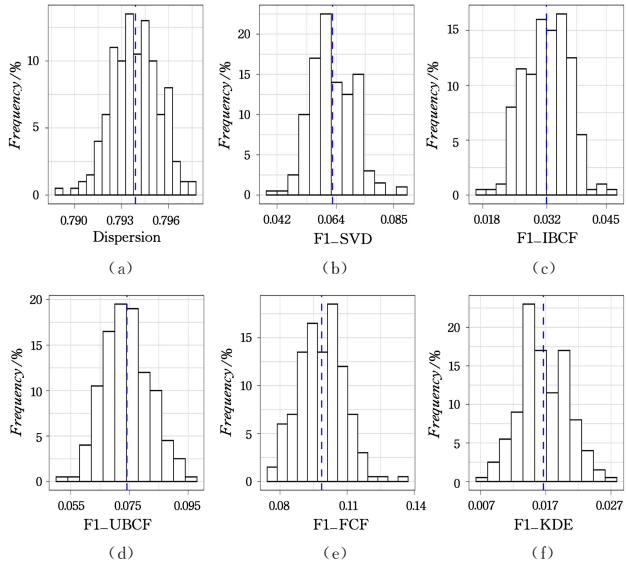

图 3 Gowalla 子数据集中离散度和 F1 值的频率分布直方图

Fig. 3 Histogram of frequency distribution of dispersion and F1 values in Gowalla subsets

表 5 离散度与各推荐算法 F1 值之间的相关性

Table 5 Correlation between dispersion and F1 values of each recommendation algorithm

	相关性指标	实验数据集	
		Brightkite	Gowalla
相关性指标	r(Dispersion, F1_SVD)	−0.3899	−0.5072
	p value(Dispersion, F1_SVD)	0.0024	0.0095
	r(Dispersion, F1_IBCF)	−0.4078	−0.5217
	p value(Dispersion, F1_ IBCF)	0.0016	0.0015
	r(Dispersion, F1_UBCF)	−0.4060	−0.5384
	p value(Dispersion, F1_ UBCF)	0.0004	5.24×10^{-7}
	r(Dispersion, F1_FCF)	−0.3702	−0.4146
	p value(Dispersion, F1_ FCF)	0.0063	0.0118
	r(Dispersion, F1_KDE)	−0.3514	−0.4513
	p value(Dispersion, F1_ KDE)	0.0094	4.29×10^{-6}
潜在噪声比例	潜在噪声比例 $\varepsilon_{\_SVD}$	0.0421	0.0391
	潜在噪声比例 $\varepsilon_{\_IBCF}$	0.0393	0.0406
	潜在噪声比例 $\varepsilon_{\_UBCF}$	0.0395	0.0419
	潜在噪声比例 $\varepsilon_{\_FCF}$	0.0400	0.0424
	潜在噪声比例 $\varepsilon_{\_KDE}$	0.0411	0.0397

5.5 不同降噪方法对推荐准确性的影响对比

为了比较同一推荐算法在不同的数据处理结果中的推荐精度，首先分别利用 NFDC 算法和 4 种基准方法对 Brightkite 和 Gowalla 数据集进行数据预处理。然后，将生成的 10 个数据处理结果分别输入到 5 类代表性的兴趣点推荐算法中，进行 50 组仿真实验。在每组实验中，所有推荐算法都独立执行 100 次，最终的推荐准确度是 100 次评估结果的平均值。在推荐过程中，采用离线实验的方法将数据集随机划分为训练集和测试集，其中训练集中包含 90% 的数据，测试集中包含 10% 的数据。将预测评分排名前 N 位的兴趣点推荐给用户，本文实验测试当推荐列表长度 $N=5, 10, 15, 20$ 时各推荐

算法的准确度表现[35-36]。

图4—图7列出了5类推荐算法在采用5种不同方法预处理后的Brightkite和Gowalla数据集中的精确率Precision和召回率Recall。

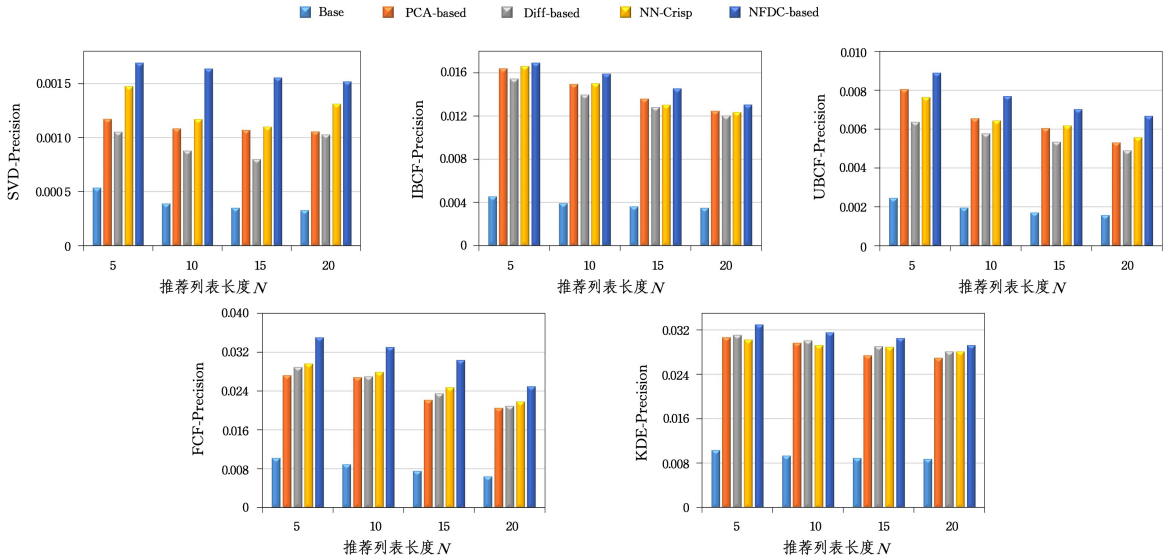

图4　各推荐算法在采用不同方法预处理后的Brightkite数据集中的Precision值

Fig. 4　Precision values of each recommendation algorithm on Brightkite preprocessed by different methods

从图4—图7所示的结果中可以看出:

1)精确率Precision的值一般随着推荐列表长度N的增大而降低,但召回率Recall的值一般都与推荐列表长度N成正比。这是因为精确率计算的是,在提供给用户的推荐列表中用户真正喜欢的项目所占的比例。因此,当推荐列表中符合用户兴趣偏好的项目数量跟不上推荐列表长度N的增长趋势时,精确率的值就会随之下降。召回率表示在用户喜欢的所有项目中,推荐系统真正把它推荐给用户的比例,当用户喜欢的项目总数一定时,推荐列表的长度越大,推荐结果中包含符合用户兴趣的项目几率就越大,因此召回率的值也越大。

2)不管是基于Brightkite数据集(见图4和图5),还是基于Gowalla数据集(见图6和图7),5种推荐算法均在NFDC算法预处理后的数据中获得了最高的精确率和召回率。由此可见,在这5类不同类型的推荐算法中,NFDC算法对自然噪声的识别和过滤能力最强。

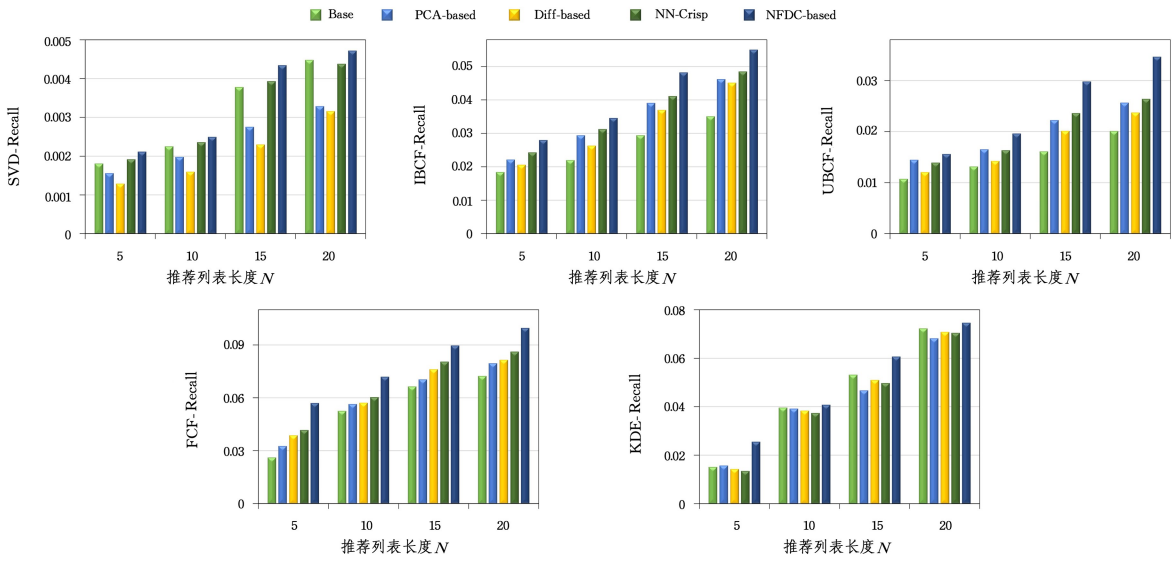

图5　各推荐算法在采用不同方法预处理后的Brightkite数据集中的Recall值

Fig. 5　Recall values of each recommendation algorithm on Brightkite preprocessed by different methods

3)在Brightkite数据集中,SVD和KDE推荐算法在NN-Crisp,Diff-based,PCA-based数据中的召回率反而低于其在未经任何降噪处理的Base数据中的值,这说明在兴趣点推荐场景中,NN-Crisp,Diff-based,PCA-based方法未能有效提升SVD和KDE这两类推荐模型输入数据的质量,进而影响了推荐结果的全面性。SVD是经典的矩阵分解算法,可解决数据稀疏问题,KDE算法则常用于挖掘兴趣点的地理特征对用户签到行为的个性化影响,这两类算法均是兴趣点推荐领域常用的推荐模型。由此可见,现有的面向传统推荐系统设计的降噪方法有时并不适用于兴趣点推荐场景,难以提升一些

经典的兴趣点推荐算法的服务质量。

4)在 Gowalla 数据集中,5 种预处理方法对不同推荐算法预测精度的影响较为一致,按准确度提升能力从高到低排序,分别为 NFDC, NN-Crisp, Diff-based, PCA-based 以及 Base 方法。由于排名靠前的两种降噪方法均专门针对自然噪声进行过滤,PCA-based 算法则是针对恶意噪声进行过滤,因此,可以推断出在 Gowalla 数据集中,自然噪声对推荐性能的影响比恶意噪声更大。

精确率 Precision 体现了推荐系统与用户兴趣的契合程度,召回率 Recall 体现了推荐系统的全面性,两者往往无法兼顾。为了衡量两者之间的平衡度,本文进一步采用 F1 值来评价推荐系统的综合准确度。表 6 和表 7 详细列出了 5 类推荐算法在不同数据集中的 F1 值,其中每一行加粗的数值表示该行的最大值。

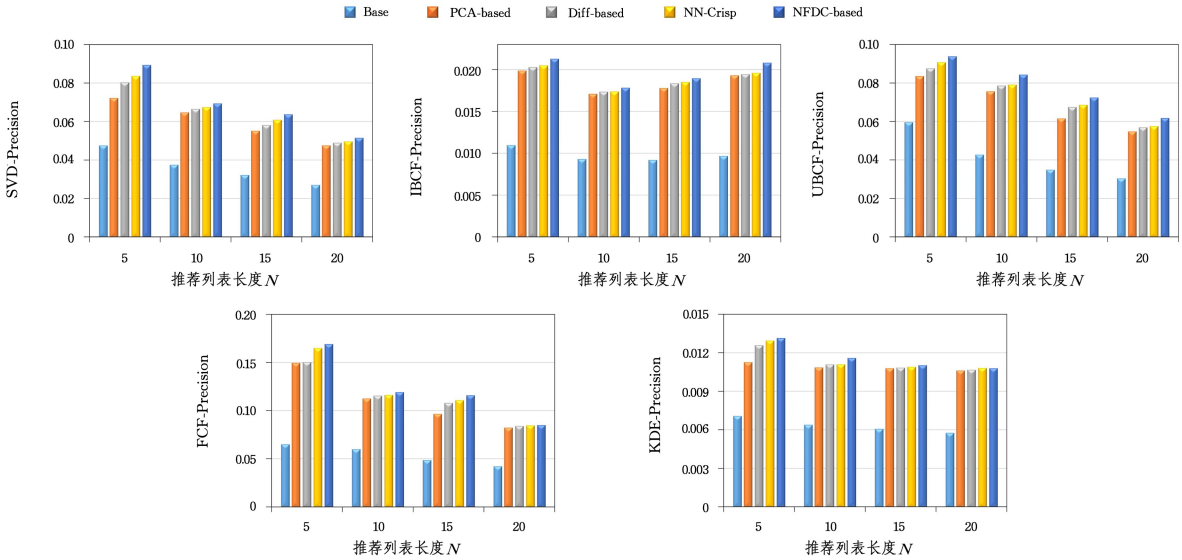

图 6 各推荐算法在采用不同方法预处理后的 Gowalla 数据集中的 Precision 值

Fig. 6 Precision values of each recommendation algorithm on Gowalla preprocessed by different methods

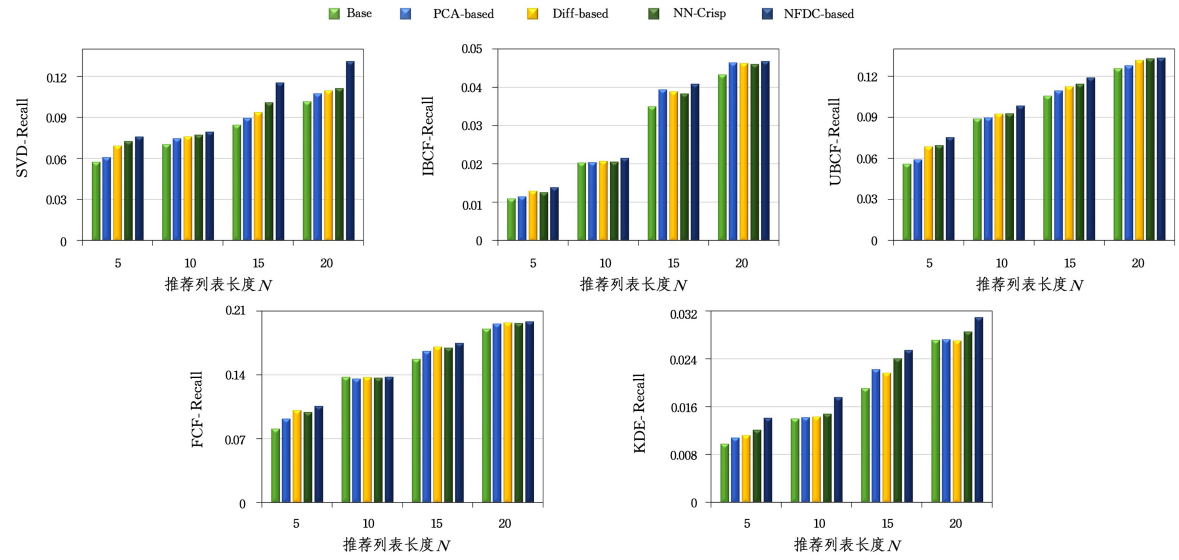

图 7 各推荐算法在采用不同方法预处理后的 Gowalla 数据集中的 Recall 值

Fig. 7 Recall values of each recommendation algorithm on Gowalla preprocessed by different methods

从表 6 和表 7 的展示结果中可以看出:

1)F1 值与推荐列表的长度 N 没有直接的线性关系,这是因为精确率一般随着 N 的增大而降低,但召回率一般与 N 成正比。

2)所有推荐算法在降噪之后的数据集中的推荐性能都优于在 Base 数据集中的表现,这说明每个位置社交网络数据集中都存在着噪声信息,且这些噪声会对推荐算法的预测性能产生较大的影响。

3)各类推荐算法在 NFDC 算法处理后的 Brightkite 数据集和 Gowalla 数据集中都获得了最高的推荐精度 F1 值。与各推荐算法在 PCA-based, Diff-based 和 NN-Crisp 数据中的最高准确度(表 6 中每一行第 3—6 列的最大值)相比,SVD, IBCF, UBCF, FCF 和 KDE 推荐算法在 NFDC 算法处理后的 Brightkite 数据集中的 F1 值在 N 的 4 种取值下仍分别平均提高了 21.52%, 8.06%, 16.22%, 19.53% 和 14.44%;在 NFDC 算法处理后的 Gowalla 数据集中的 F1 值则分别平均

提高了 6.11%，4.56%，5.67%，2.83%和 5.85%。

表 6　各推荐算法在采用不同方法预处理后的 Brightkite 数据集中的 F1 值

Table 6　F1 values of each recommendation algorithm on Brightkite preprocessed by different methods

推荐算法	推荐列表长度	Base	PCA-based	Diff-based	NN-Crisp	NFDC
SVD	5	0.0008	0.0013	0.0012	0.0017	**0.0019**
	10	0.0007	0.0014	0.0011	0.0016	**0.0020**
	15	0.0006	0.0015	0.0012	0.0017	**0.0023**
	20	0.0004	0.0017	0.0016	0.0020	**0.0025**
IBCF	5	0.0073	0.0189	0.0177	0.0198	**0.0211**
	10	0.0066	0.0198	0.0183	0.0203	**0.0218**
	15	0.0064	0.0202	0.0191	0.0198	**0.0224**
	20	0.0063	0.0196	0.0190	0.0197	**0.0212**
UBCF	5	0.0040	0.0104	0.0083	0.0099	**0.0113**
	10	0.0034	0.0094	0.0082	0.0092	**0.0111**
	15	0.0031	0.0095	0.0084	0.0098	**0.0114**
	20	0.0029	0.0088	0.0081	0.0092	**0.0112**
FCF	5	0.0147	0.0297	0.0331	0.0347	**0.0434**
	10	0.0152	0.0364	0.0367	0.0382	**0.0453**
	15	0.0135	0.0337	0.0359	0.0379	**0.0454**
	20	0.0117	0.0326	0.0333	0.0349	**0.0399**
KDE	5	0.0123	0.0208	0.0195	0.0186	**0.0288**
	10	0.0151	0.0338	0.0337	0.0328	**0.0356**
	15	0.0153	0.0345	0.0370	0.0366	**0.0406**
	20	0.0156	0.0387	0.0402	0.0400	**0.0420**

表 7　各推荐算法在采用不同方法预处理后的 Gowalla 数据集中的 F1 值

Table 7　F1 values of each recommendation algorithm on Gowalla preprocessed by different methods

推荐算法	推荐列表长度	Base	PCA-based	Diff-based	NN-Crisp	NFDC
SVD	5	0.0520	0.0661	0.0744	0.0777	**0.0821**
	10	0.0489	0.0693	0.0709	0.0720	**0.0741**
	15	0.0466	0.0683	0.0718	0.0760	**0.0821**
	20	0.0427	0.0659	0.0677	0.0686	**0.0740**
IBCF	5	0.0110	0.0146	0.0158	0.0156	**0.0168**
	10	0.0128	0.0186	0.0189	0.0188	**0.0195**
	15	0.0146	0.0245	0.0248	0.0250	**0.0259**
	20	0.0158	0.0273	0.0274	0.0275	**0.0288**
UBCF	5	0.0578	0.0693	0.0770	0.0788	**0.0837**
	10	0.0578	0.0821	0.0851	0.0855	**0.0909**
	15	0.0525	0.0789	0.0845	0.0859	**0.0901**
	20	0.0490	0.0768	0.0795	0.0804	**0.0845**
FCF	5	0.0721	0.1138	0.1209	0.1239	**0.1301**
	10	0.0834	0.1231	0.1255	0.1257	**0.1279**
	15	0.0742	0.1221	0.1323	0.1341	**0.1394**
	20	0.0690	0.1159	0.1178	0.1183	**0.1190**
KDE	5	0.0082	0.0111	0.0119	0.0125	**0.0136**
	10	0.0088	0.0123	0.0125	0.0127	**0.0140**
	15	0.0092	0.0145	0.0144	0.0150	**0.0154**
	20	0.0095	0.0150	0.0153	0.0157	**0.0160**

由此可见，本文提出的自然噪声过滤算法 NFDC 适用于位置社交网络中多种类型的兴趣点推荐算法，能够有效检测和删除自然噪声，切实提升兴趣点推荐算法的预测准确性。

结束语　针对较为隐蔽、难以处理的自然噪声问题，本文面向兴趣点推荐系统提出了一种基于离散特征量化与聚类距离分析的自然噪声过滤算法 NFDC。该算法提出了用户签到数据的离散度量化方法，并以此表示数据驱动的不确定性，同时利用推荐算法的准确度量化预测驱动的不确定性，通过构建经验模型，挖掘两类不确定性之间的相关性；基于聚类距离分析，筛选可疑噪声，并自定义噪声验证规则，删除真正的自然噪声。在两个真实的位置社交网络数据集中进行了大量的离线仿真实验，实验结果表明，NFDC 算法适用于兴趣点推荐场景，能够明显提高兴趣点推荐系统的源数据质量，降低数据干扰和误差传播，为后续的推荐算法提供可靠输入，是兴趣点推荐系统性能获得提升的基础。

考虑到自然噪声的删除会进一步加剧数据稀疏问题，给推荐算法的准确度提升带来一定的技术瓶颈，因此下一步将围绕自然噪声的矫正和更新开展更深层次的研究。

参 考 文 献

[1] ZHANG Q, YU S Y, YIN H F, et al. Neural collaborative filtering for social recommendation algorithm based on graph attention[J]. Computer Science, 2023, 50(2): 115-122.

[2] CHENG Z T, ZHONG T, ZHANG S M, et al. Survey of recommender systems based on graph learning[J]. Computer Science, 2022, 49(9): 1-13.

[3] O'MAHONY M P, HURLEY N J, SILVESTRE G. Detecting noise in recommender system databases[C]// Proceedings of the 11th International Conference on Intelligent User Interfaces. 2006: 109-115.

[4] BAG S, KUMAR S, AWASTHI A, et al. A noise correction-based approach to support a recommender system in a highly sparse rating environment[J]. Decision Support Systems, 2019, 118: 46-57.

[5] WANG Y L, JIANG C C, FENG X N, et al. Time aware point-of-interest recommendation[J]. Computer Science, 2021, 48(9): 43-49.

[6] ZHU J, HAN L X, GOU Z N, et al. A robust personalized location recommendation based on ensemble learning[J]. Expert Systems With Applications, 2021, 167: 114065.

[7] BELLOGÍN A, SAID A, DE VRIES A P. The magic barrier of recommender systems - no magic, just ratings[C]// International Conference on User Modeling, Adaptation, and Personalization. 2014: 25-36.

[8] WANG Z W, GAO M, LI J D, et al. Gray-Box shilling attack: An adversarial learning approach[J]. ACM Transactions on Intelligent Systems and Technology, 2022, 13(5): 82.

[9] LIU Z, FENG X D, WANG Y C, et al. Self-paced learning enhanced neural matrix factorization for noise-aware recommendation[J]. Knowledge-Based Systems, 2021, 213: 106660.

[10] CASTRO J, YERA R, MARTÍNEZ L. A fuzzy approach for natural noise management in group recommender systems[J]. Expert Systems With Applications, 2018, 94(15): 237-249.

[11] SHARON M J, DHINESH B L D. A fuzzy linguistic approach-based non-malicious noise detection algorithm for recommendation system[J]. International Journal of Fuzzy Systems, 2018, 20: 2368-2382.

[12] AMATRIAIN X, PUJOL J M, TINTAREV N, et al. Rate it again: increasing recommendation accuracy by user re-rating[C]// Proceedings of the Third ACM Conference on Recommender Systems. 2009: 173-180.

[13] PHAM H X, JUNG J J. Preference-based user rating correction process for interactive recommendation systems[J]. Multimedia Tools and Applications, 2013, 65(1): 119-132.
[14] YU P H, LIN L F, YAO Y G. A novel framework to process the quantity and quality of user behavior data in recommender systems[C]// International Conference on Web-Age Information Management. 2016: 3-5.
[15] TOLEDO R Y, MOTA Y C, MARTÍNEZ L. Correcting noisy ratings in collaborative recommender systems[J]. Knowledge-Based Systems, 2015, 76: 96-108.
[16] XIA B, LI T, LI Q M, et al. Noise-tolerance matrix completion for location recommendation[J]. Data Mining and Knowledge Discovery, 2018, 32: 1-24.
[17] LI D T C, LIU H, ZHANG Z L, et al. CARM: Confidence-aware recommender model via review representation learning and historical rating behavior[J]. Neurocomputing, 2021, 455: 283-296.
[18] COSLEY D, LAM S K, ALBERT I, et al. Is seeing believing? How recommender system interfaces affect users' opinions [C]// Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. 2003: 585-592.
[19] LI W H, LI X G, DENG J Z, et al. Sentiment based multi-index integrated scoring method to improve the accuracy of recommender system[J]. Expert Systems With Applications, 2021, 179: 115105.
[20] LI B, CHEN L, ZHU X Q, et al. Noisy but non-malicious user detection in social recommender systems[J]. World Wide Web, 2013, 16(5): 677-699.
[21] YERA R, CASTRO J, MARTÍNEZ L. A fuzzy model for managing natural noise in recommender systems[J]. Applied Soft Computing, 2016, 40: 187-198.
[22] YERA R, BARRANCO M J, ALZAHRANI A A, et al. Exploring fuzzy rating regularities for managing natural noise in collaborative recommendation[J]. International Journal of Computational Intelligence Systems, 2019, 12(2): 1382-1392.
[23] CASTRO J, YERA R, MARTÍNEZ L. An empirical study of natural noise management in group recommendation systems [J]. Decision Support Systems, 2017, 94: 1-17.
[24] WANG P Y, WANG Y, ZHANG L Y, et al. An effective and efficient fuzzy approach for managing natural noise in recommender systems[J]. Information Sciences, 2021, 570: 623-637.
[25] ZHOU D Q, WANG B, RAHIMI S M, et al. A study of recommending locations on location-based social network by collaborative filtering[C]// Proceedings of the 25th Canadian Conference on Advances in Artificial Intelligence. 2012: 255-266.
[26] BORGELT C, BRAUNE C, LESOT M J, et al. Handling noise and outliers in fuzzy clustering[M]// Fifty Years of Fuzzy Logic and its Applications. Springer, 2015: 315-335.
[27] SALAH A, ROGOVSCHI N, NADIF M. A dynamic collaborative filtering system via a weighted clustering approach[J]. Neurocomputing, 2016, 175: 206-215.
[28] GENG B R, JIAO L C, GONG M G, et al. A two-step personalized location recommendation based on multi-objective immune algorithm[J]. Information Sciences, 2019, 475: 161-181.
[29] CHEN K K, CHAN P P K, ZHANG F, et al. Shilling attack based on item popularity and rated item correlation against collaborative filtering[J]. International Journal of Machine Learning and Cybernetics, 2019, 10: 1833-1845.
[30] CHEN J, WANG X S, ZHAO S, et al. Deep attention user-based collaborative filtering for recommendation[J]. Neurocomputing, 2020, 383: 57-68.
[31] JIANG L C, LIU R R, JIA C X. User-location distribution serves as a useful feature in item-based collaborative filtering[J]. Physica A—Statistical Mechanics and Its Applications, 2022, 586, 126491.
[32] ZHOU X, HE J, HUANG G Y, et al. SVD-based incremental approaches for recommender systems[J]. Journal of Computer and System Sciences, 2015, 81: 717-733.
[33] LIN K H, WANG J J, ZHANG Z N, et al. Adaptive location recommendation algorithm based on location-based social networks[C]// Proceedings of International Conference on Computer Science & Education. 2015: 137-142.
[34] SI Y L, ZHANG F Z, LIU W Y. An adaptive point-of-interest recommendation method for location-based social networks based on user activity and spatial features[J]. Knowledge-Based Systems, 2019, 163: 267-282.
[35] SU C, WU P F, XIE X Z, et al. Point of interest recommendation based on user's interest and geographic factors[J]. Computer Science, 2019, 46(4): 228-234.
[36] CHEN J, ZHANG H, CAO F Y. Study on point-of-interest collaborative recommendation method fusing multi-factors[J]. Computer Science, 2019, 46(10): 77-83.

ZHU Jun, born in 1987, Ph.D, associate professor, is a member of China Computer Federation. Her main research interests include machine learning and recommender systems.

(责任编辑:喻藜)