doi:10.3969/j.issn.1673-9833.2018.04.011

基于 Tukey 怀疑度模型旅游线路 M 估计协同推荐

冯 莉

（厦门城市职业学院 人文社科与艺术系，福建 厦门 361008）

摘 要：为提高旅游线路推荐结果的有效性，降低干扰数据对推荐结果的影响，提出一种基于图基（Tukey）检验的 M 估计游客怀疑度模型的旅游线路协同推荐方法。首先，基于游客兴趣偏好构建旅游景点路线推荐算法框架，并基于 *k*- 距离的游客怀疑度构建可靠邻居模型；其次，针对提出的模型，提出一种基于图基检验 M 估计的鲁棒矩阵分解算法，构建游客特征矩阵和项目特征矩阵，通过调整游客间的相似性，减少干扰配置项目对特征矩阵鲁棒估计的影响；最后，在网爬数据集上进行仿真测试。测试结果显示，所提算法具有更高的游客整体满意度、更低的游客痛苦度，并且旅游景点路线多样化效果更好。

关键词：图基检验；M 估计；旅游线路；协同推荐；游客怀疑度

中图分类号：F590 **文献标志码**：A **文章编号**：1673-9833(2018)04-0067-07

Tukey Skeptical Model Based on M Estimation of Collaborative Recommendation Algorithm for Tourist Routes

FENG Li

（Department of Humanities & Social Sciences，Xiamen City University，Xiamen Fujian 361008，China）

Abstract：In order to improve the effectiveness of tourist routes recommended results, and reduce the influence of interference data on the recommendation results, a proposal has been made of Tukey skeptical model based on M estimation of collaborative recommendation algorithm for tourist routes. Firstly, a framework of scenic route recommendation algorithm is to be constructed based on interest preferences of tourists, and a reliable neighbor model is to be constructed based on k-distance skepticism. Secondly, according to the proposed model, we propose a decomposition algorithm of matrix estimation based on robust Tukey M, thus constructing the user feature matrix and project feature matrix. By adjusting the similarity between tourists, the influence has been reduced of interference configuration items on the robust estimation of the characteristic matrix. Finally, the simulation test on the web crawl data set shows that the proposed algorithm is characterized with such advantages as with a higher overall tourist satisfaction, lower tourist pains, and better diversity of tourist attractions.

Keywords：Tukey test；M estimation；tourist route；collaborative recommendation；tourist skepticism

0 引言

随着社会的发展，旅游业得到快速发展，游客数量激增，导致多数主流旅游平台出现信息饱和问题，如何解决信息过载，提供更加合理的旅游景点线路推荐方法，具有重要应用价值[1-2]。旅游景点线路的

收稿日期：2018-03-12

基金项目：国家开放大学科研基金资助项目（G16A2001Z），厦门城市职业学院第五批校企合作课程基金资助项目（xqkc2017119）

作者简介：冯 莉（1972-），女，湖北襄阳人，福建厦门城市职业学院副教授，硕士，主要研究方向为图像处理，数据挖掘，系统控制，E-mail：fengli@xmcu.cn

规划是在多条线路中，推荐出更加合理并且符合游客期望的线路。当前，已有多种旅游景点线路推荐方法：文献 [3] 通过网站游客的旅游共享照片进行旅游景点线路的分析推荐，取得较好的效果；文献 [4] 针对游客旅游景点的标签和主题，获得不同旅游景点主题类型的访问频繁度，用来指导旅游线路推荐；文献 [5] 基于聚类方法将游客对旅游景点的评价进行分类，并按照时间序列构建游客的旅游轨迹，然后利用 Markov 模型构建概率行为特征，对将来景点路线进行预测；文献 [6] 基于游客对景点社区评分价值的贡献度，利用 Bayes 模型构建游客特征和旅游方式的学习模型，实行游客旅游景点线路的个性化定制。

近年来，基于目标优化的方法得到了广泛研究，文献 [7] 基于定向策略对旅游景点路线进行规划推荐，利用已知和未知的旅游兴趣点进行旅游景点线路的目标得分最大化优化；文献 [8] 基于旅行约束限制，进行最优化的旅游景点推荐；文献 [9] 借鉴旅行商分析策略进行旅行线路推荐，获得较好的推荐效果；文献 [10] 基于旅行兴趣点对定向问题进行改进，并基于多目标准则进行旅游景点的线路推荐。

鲁棒推荐系统是指在推荐数据库中存在攻击或噪声污染时，仍能提供稳定的推荐结果。虽然上述文献中提出了一些鲁棒的协同推荐算法，但它们仍然有以下局限性：1）上述算法对于攻击的鲁棒性很差，导致其预测结果存在较大的不稳定性；2）推荐算法的鲁棒性以推荐精度为代价，导致推荐质量较差。并且上述算法在进行旅游线路推荐过程中，虽然充分利用了已有景点的游客评分值，进行新的路线推荐，但是并未考虑到当前游客的想法，存在人为操作或者不具有个性化的问题。

对此，本文在进行旅游线路设计过程中，着重考虑了旅游线路推荐过程中游客怀疑度对旅游线路选择的影响，降低干扰数据对于旅游景点路线推荐结果的影响，提高推荐算法对于干扰和攻击过程的鲁棒性，实验结果验证了所提方法的有效性。

1 旅游景点路线推荐的可靠邻居模型

1.1 算法框架描述

旅游景点线路的推荐算法框架如图 1 所示。

在本文旅游景点路线推荐算法中主要包含两种算法：路线推荐过程和兴趣偏好学习过程。首先，建立游客的兴趣喜好矩阵，并基于游客的兴趣景点分析，获得当前分析旅游景点的兴趣偏好和流行度分析，并在定向问题上，基于兴趣偏好综合分析过程，将评价分值最高的旅游定向线路向游客推荐。

游客对于旅游景点的偏好度指标可定义为：对游客在某时段的旅游行程进行特征提取和统计分析，并将其旅游景点的类型选择数量与旅游总次数之比作为偏好度指标 [11~12]

$$C_i = m_i \Big/ \sum_{j=1}^{n} m_j , \tag{1}$$

式（1）中：m 为游客在某时段行程选择的景点类别数量；n 为游客在某时段行程选择的总数。

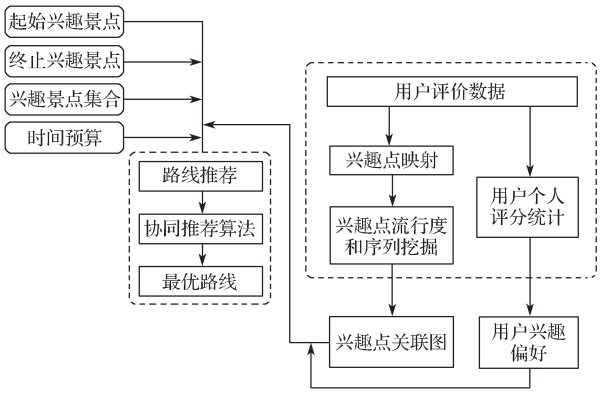


图 1 旅游景点路线推荐算法框架

Fig. 1 Framework of route recommendation algorithm for tourist attractions

如果游客在旅游线路中未曾选择某类型的景点，则该游客对于该景点类型的偏好度指标值为 0。基于式（1）定义的偏好度指标，可得游客对于旅游景点的评价信息矩阵 $\boldsymbol{R}$ 为

$$\boldsymbol{R}_{p,q} = \begin{bmatrix} r_{11} & r_{12} & \cdots & r_{1q} \\ r_{21} & r_{22} & \cdots & r_{2q} \\ \vdots & \vdots & \ddots & \vdots \\ r_{p1} & r_{p2} & \cdots & r_{pq} \end{bmatrix}。 \tag{2}$$

式（2）中：r_{ij} 为游客 i 对旅游景点 j 的评价值；q 为旅游景点数量；p 为游客数量；评价信息矩阵 $\boldsymbol{R}_{p,q}$ 为游客 p 对于旅游景点 q 的评分数值。同上，如果游客对于旅游景点未到访或者未进行评价，则该旅游景点的评分值为 0。

可通过游客怀疑度计算及构建怀疑度邻居模型，减少干扰数据对推荐结果的影响。

1.2 基于 $k-$ 距离的游客怀疑度计算

基于 k 最近邻距离的离群点挖掘算法，提出基于 $k-$ 距离的游客怀疑度计算方法。

定义 1 对象 p 的邻域删除。对于一个给定的数据集 D，$\exists p \in D$，删除对象 p 的邻域可表示为 $\overset{\circ}{U}(p,\delta)$：

$$\overset{\circ}{U}(p,\delta) = \{o \mid 0 < |op| < \delta, o \in D\}。 \tag{3}$$

定义 2 对象 p 的 $k-$ 距离。对 $\overset{\circ}{U}(p,\delta)$ 中的所有

对象与对象 p 的距离进行排序获得$\{|o_n p|\}$，其满足$\{|o_1,p|\}\leqslant\{|o_2,p|\}\leqslant\cdots\leqslant\{|o_k,p|\}\leqslant\{|o_n,p|\}$。若 k 为与对象 p 距离最近邻居节点，则对象 p 的 k- 距离可定义为

$$k\_dist(p)=|o_k,p|。\tag{4}$$

根据上述定义可知，对象 p 的 k- 距离是其 k 个邻居距离的最大值。如果对象的 k- 距离较大，则意味着其对象的邻居数量较少，这表明对象偏离总体数据，反之亦然。因此，该指标可用来评价对象的离群度。

定义 3　欧氏距离。对于任意的游客 $u_a\in U$，$u_b\in U$，令 $R(u_a)$ 和 $R(u_b)$ 分别为 u_a 和 u_b 的评级项目集，令 $I(u_a, u_b)$ 为 u_a 和 u_b 的共同额定项目集，则 u_a 和 u_b 之间的欧氏距离可表示为 [13~14]

$$dist(u_a,u_b)=\begin{cases}(|R(u_a)|+|R(u_b)|)\sqrt{\sum_{i=1}^{n}(r_{ai}-r_{bi})^2}, \\ \qquad R(u_a)\cap R(u_b)=\varnothing; \\ \dfrac{(|R(u_a)|+|R(u_b)|)}{|I(u_a,u_b)|}\sqrt{\sum_{j=1}^{n}(r_{ai}-r_{bi})^2}, \\ \qquad R(u_a)\cap R(u_b)\neq\varnothing。\end{cases}\tag{5}$$

定义 4　游客怀疑度。对于任意游客 $u\in U$，其 k- 距离可表示为 $k\_dist(u)$，则怀疑度 S_u 计算形式为

$$S_u=\left(e^{\beta}-1\right)\Big/\alpha,\tag{6}$$

$$\beta=\left(k\_dist(u)-\min_{k\_\text{dist}}\right)\Big/\left(\max_{k\_\text{dist}},-\min_{k\_\text{dist}}\right)。\tag{7}$$

式（6）~（7）中：β 为游客 u 归一化后的 k- 距离值；α 为常值；$\max_{k\_\text{dist}}$ 为所有游客的最大 k- 距离值；$\min_{k\_\text{dist}}$ 为所有游客的最小 k- 距离值。参数 α 的设定方式如下。

对于 $\beta\in[0,1]$ 和 $S_u\in[0,1]$，可得 $0\leqslant\left(e^{\beta}-1\right)\Big/\alpha\leqslant 1$，$\alpha^{\beta}-1\geqslant 0$ 和 $\alpha>0$。令 $f(\beta)=e^{\beta}-1$，可得 $f'(\beta)=e^{\beta}>0$。因此，$f(\beta)$ 是单调递增的，并且可得 $0\leqslant\left(e^{\beta}-1\right)\leqslant\alpha$，$0=e^0-1\leqslant e^{\beta}-1\leqslant e^1-1=1.713\,8$，$\alpha\geqslant e^{\beta}-1$ 和 $\alpha\geqslant 1.718\,3$。图 2 所示为选取 6 种 α 情况下，游客 u 的怀疑度随 β 变化情况。

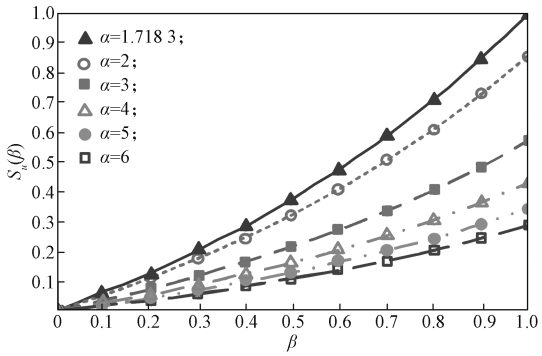


图 2　怀疑度随 β 的变化曲线

Fig. 2　Skepticism curves

如果 $\Delta\beta$ 变化相同，$\Delta S_u(\beta)$ 随着 α 的增加而逐渐降低。因此，当 α 为 1.718 3 时，$S_u(\beta)$ 随 β 增大的效果日益突出。因此，这里选取 1.718 3 作为 α 的最佳取值。

1.3　怀疑度邻居模型

基于游客的推荐算法是协同过滤算法中的一种常见算法，其主要思想是根据目标游客的目标项目最近邻评价预测目标游客的目标项目评价。在游客评价数据库中，有游客偏好特征和项目偏好特征。前者意味着一些游客给出的项目评级高于其他游客；后者意味着某些项目的评级高于其他项目。为了减少两种情况对推荐精度的影响，定义基线估计如下：

$$b_{ui}=\mu+b_u+b_i,\tag{8}$$

式中：μ 为评级数据库中所有评级的均值；b_u 和 b_i 分别为从均值观测到的游客 u 和项目 i 的偏差，且

$$b_u=\frac{1}{|I_u|}\sum_{i\in I_u}(r_{ui}-\mu),\tag{9}$$

$$b_i=\frac{1}{|U_i|}\sum_{u\in U_i}(r_{ui}-b_u-\mu);\tag{10}$$

式（9）~（10）中，I_u 为游客 u 进行评分的项目，U_i 为项目 i 的游客组。

为了降低干扰数据对推荐结果的影响，结合游客的怀疑度可靠邻居模型和基线估计方法实现协同过滤推荐。预测评级计算如下：

$$\hat{r}_{ui}=b_{ui}+\left|R(u)\right|^{-\frac{1}{2}}\sum_{v\in R_u}(r_{vi}-\overline{r}_v)*sim_{uv}(1-S_v)。\tag{11}$$

式中：$R(u)$ 为目标游客 u 的相似游客集；$\overline{r}_v$ 为游客 u 的评级均值；sim_{uv} 为游客 u 和游客 v 之间的相似性；S_v 为游客 v 的怀疑度。

根据式（9）可知，对于任意游客 u 的邻居 v，如果怀疑度 S_v 很大，邻居 v 和目标游客 u 之间的相似性可以通过 $sim_{uv}(1-S_v)$ 降低。因此，怀疑度邻居模型减少了干扰数据对推荐结果的影响。

2　基于 M 估计的鲁棒矩阵分解模型

2.1　基本矩阵分解模型

矩阵分解模型可以揭示游客和项目在评级数据中的隐藏特征，其可由游客特征矩阵 $\boldsymbol{P}$ 和项目特征矩阵 $\boldsymbol{Q}$ 表示。令 $\hat{\boldsymbol{R}}$ 是预测评级的矩阵，则基本矩阵分解模型定义如下：

$$\hat{\boldsymbol{R}}=\boldsymbol{Q}^{\mathrm{T}}\boldsymbol{P}。\tag{12}$$

式中：$\boldsymbol{P}=(\boldsymbol{p}_1,\boldsymbol{p}_2,\cdots,\boldsymbol{p}_m)$ 为 $f\times m$ 矩阵，$\boldsymbol{p}_u$ 为游客 u 的 f 维特征向量；$\boldsymbol{Q}=(\boldsymbol{q}_1,\boldsymbol{q}_2,\cdots,\boldsymbol{q}_n)$ 为 $f\times n$ 矩阵，$\boldsymbol{q}_i$ 为项目 i 的 f 维特征向量。

令 $\hat{r}_{ui}$ 为预测评级，可表达如下：

$$\hat{r}_{ui}=\boldsymbol{q}_i^{\mathrm{T}}\boldsymbol{p}_u。\tag{13}$$

为了求解 $\boldsymbol{p}_u$ 和 $\boldsymbol{q}_i$，定义最小二乘问题如下：

$$\boldsymbol{p}^*,\boldsymbol{q}^*=\arg\min\sum_{r_{ui}\neq\phi}\left(r_{ui}-\boldsymbol{q}_i^{\mathrm{T}}\boldsymbol{p}_u\right)^2+\lambda\left(\|\boldsymbol{q}_i\|^2+\|\boldsymbol{p}_u\|^2\right),\tag{14}$$

式中 $\lambda\left(\|\boldsymbol{q}_i\|^2+\|\boldsymbol{p}_u\|^2\right)$ 是一个正则化项，可以避免过拟合问题，λ 为一个常数。

定义 5 残差。残差是观测值和回归估计之间的差异，可表示为

$$e_{ui}=r_{ui}-\hat{r}_{ui}。\tag{15}$$

式中：r_{ui} 为真实的评级；$\hat{r}_{ui}$ 为预测评级。

2.2 基于 Tukey M 估计的鲁棒参数估计

为了减少参数估计的干扰影响，引入基于 Tukey 的 M 估计对游客特征矩阵和项目特征矩阵进行鲁棒估计。

定义 6 基于 Tukey 的 M 估计。令 e_{ui} 为真正的评级和预测评级在推荐系统中的差异，则利用 Tukey 的 M 估计器实现损耗函数的最小化，以实现对参数 $\boldsymbol{p}^*,\boldsymbol{q}^*$ 的求解：

$$\boldsymbol{p}^*,\boldsymbol{q}^*=\arg\min\sum_{r_{ui}>0}\rho\left(e_{ui}\right),\tag{16}$$

式中 $\rho(e_{ui})$ 为利用 Tukey 函数得到的有界函数，可表示为

$$\rho\left(e_{ui}\right)=1-\left[1-\left(e_{ui}/c\right)^2\right]^3,\tag{17}$$

式中 c 是固定值，设定 c=4.68。

定义 7 影响函数。令 $T(F)$ 为功能函数，对于 $\forall x\in R$，退化点 x 的概率分布可表示为 δ_x，当存在以下限制，并且种群分布为 F 时，功能 T 的影响函数形式如下：

$$IF\left(x,F,T\right)=\lim_{\varepsilon\to0^+}\frac{T\left[\left(1-\varepsilon\right)F+\varepsilon\delta_xT\left(F\right)\right]}{\varepsilon}=\frac{\partial}{\partial\varepsilon}T\left[\left(1-\varepsilon\right)F+\varepsilon\delta_x\right]\Big|_{\varepsilon\to0^+}。\tag{18}$$

直观地说，影响函数是概率分布空间中鲁棒估计的一阶导数，并描述了孤立点数据对鲁棒估计量的影响能力。因此，在鲁棒统计中，估计的鲁棒性可以通过影响函数反映。Tukey 的 M 估计器影响函数为

$$\varphi\left(e_{ui}\right)=\rho'\left(e_{ui}\right)=e_{ui}\left[1-\left(e_{ui}/c\right)^2\right]^2。\tag{19}$$

定义 8 总误差敏感性。总误差敏感性是基于显性鲁棒测量，定义为 γ^*，可表示如下：

$$\gamma^*=\sup\left|IF\left(x,F,T\right)\right|。\tag{20}$$

式中 sup 是指上限。如果它是有限的，其对于分布函数 F 具有鲁棒性。

由以上论述可知，如果一个估计的影响函数是无界的，那么残差将趋向于无限大，而影响函数的值往往也是无限大的。因此，估计将是敏感的异常值。接下来，证明基于 Tukey 的 M 估计方式，通过对损失函数的有界性和唯一性的解决可以实现影响函数的鲁棒性估计。

性质 1 有界性。基于 Tukey 的 M 估计的影响函数是有界函数。

证明 对于影响函数 $\varphi(e_{ui})$，可得，

$$\varphi\left(e_{ui}\right)=e_{ui}\left[1-\left(e_{ui}/c\right)^2\right]^2,\ e_{ui}\in[-5,5],$$

令

$$\left|\varphi\left(e_{ui}\right)\right|=\left|e_{ui}\left[1-\left(e_{ui}/c\right)^2\right]^2\right|=\left|e_{ui}\right|\left\|\left[1-\left(e_{ui}/c\right)^2\right]\right\|^2\leqslant\left|e_{ui}\right|\left(1+\left(e_{ui}/c\right)^2\right)^2\leqslant5\left(1+25/c^2\right)^2。$$

由此可以得知，$\exists M>0$，$M=5\left(1+25/c^2\right)^2$，亦即 $\left|\varphi\left(e_{ui}\right)\right|\leqslant M$。因此，基于 Tukey 的 M 估计的影响函数是有界函数。证毕。

性质 2 唯一解。基于 Tukey 的 M 估计具有唯一解。

证明 对于损耗函数 $\rho\left(e_{ui}\right)=1-\left[1-\left(e_{ui}/c\right)^2\right]^3$，其一阶导数形式为

$$\rho'\left(e_{ui}\right)=e_{ui}\left[1-\left(e_{ui}/c\right)^2\right]^2,\tag{21}$$

二阶导数形式为

$$\rho''\left(e_{ui}\right)=\left[1-\left(e_{ui}/c\right)^2\right]\left[1-5\left(e_{ui}/c\right)^2\right]。\tag{22}$$

考虑 $\rho'(e_{ui})$，可得停滞点 e_{ui}=0 和 $e_{ui}=\pm c$，将其分别代入 $\rho''(e_{ui})$，可得到 $\rho''(0)$=1>0，$\rho''(\pm c)$=0。对于 $\rho''(e_{ui})$>0，损失函数为严格凸函数，在 e_{ui}=0 位置具有唯一解，避免了在参数估计过程中，局部最优收敛问题。

因为对于大多数的 M 估计问题获得其解析解很困难，通常采用迭代方案，利用如下最小二乘问题进行求解：

$$\boldsymbol{p}^*,\boldsymbol{q}^*=\arg\min w\left(e_{ui}\right)e_{ui}^2,\tag{23}$$

式中 $w(e_{ui})$ 为通过基于 Tukey 的 M 估计的影响函数得到的权重函数。可表示为

$$w\left(e_{ui}\right)=\varphi\left(e_{ui}\right)/e_{ui}=\left[1-\left(e_{ui}/c\right)^2\right]^2。\tag{24}$$

根据式（23）~（24），可通过权重函数对于每个残差设置不同的权重。图 3 所示为本文采用的 Tukey 权重函数图和 Huber 权重函数图对比情况。

由图 3 可以看出，随着残差的逐渐增加，两个估计量的权重函数值逐渐减小。此外，基于 Tukey 的 M 估计器的权重函数的估计值明显低于基于 Huber

的 M 估计器。因此，对于基于 Tukey 的 M 估计器，较大的残差将获得较小的权重。较大的残差意味着干扰数据的可能性较大。

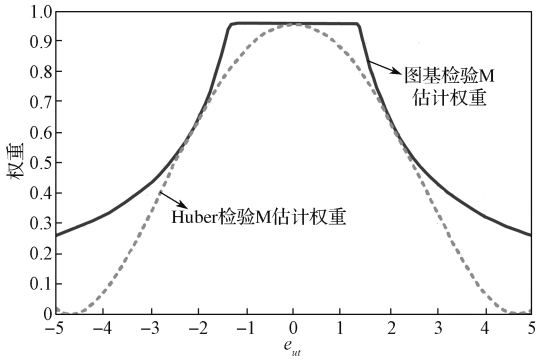


图 3　权重函数图对比情况

Fig. 3　Comparison of weight function graphs

2.3　推荐算法步骤

在所设计推荐算法过程中，预测评级的计算过程如下：

$$\hat{r}_{ui}=b_{ui}+\boldsymbol{q}_i^{\mathrm{T}}\boldsymbol{p}_u+|R(u)|^{-\frac{1}{2}}\sum_{v\in R(u)}(r_{vi}-\bar{r}_v)*sim_{uv}*(1-S_v)。\tag{25}$$

式（18）可以通过随机梯度下降法求解，迭代公式为

$$\begin{cases}\boldsymbol{q}_i\leftarrow\boldsymbol{q}_i+\gamma w(e_{ui})\boldsymbol{p}_u,\\ \boldsymbol{p}_u\leftarrow\boldsymbol{p}_u+\gamma w(e_{ui})e_{ni}\boldsymbol{q}_i,\\ b_u\leftarrow b_u+\gamma w(e_{ui})e_{ui},\\ b_i\leftarrow b_i+\gamma w(e_{ui})e_{ui}。\end{cases}\tag{26}$$

所提算法的计算过程如下所示。

算法 1：推荐算法过程；

输入：游客项目评分矩阵为 $\boldsymbol{R}$，游客偏好为 b_u，项目偏差为 b_i；

输出：目标游客 v 在项目 j 上的预测评级。

Begin

初始化特征矩阵，$\boldsymbol{P}=(\boldsymbol{p}_1,\boldsymbol{p}_2,\cdots,\boldsymbol{p}_m)$，

$\boldsymbol{Q}=(\boldsymbol{q}_1,\boldsymbol{q}_2,\cdots,\boldsymbol{q}_n)$；

for eachdo $u_a\in U$ do

for each $u_b\in U$ && $u_a\neq u_b$ do

$sim(u_a,u_b)\leftarrow similarity(u_a,u_b)$；// 相似度计算

endfor

邻居模型构建S_{u_a}；

endfor

repeat

for each $u\in U$ do

for each $i\in I$ do

if $r_{ui}\neq 0$ then

$$\hat{r}_{ui}=b_{ui}+\boldsymbol{q}_i^{\mathrm{T}}\boldsymbol{p}_u+|R(u)|^{-\frac{1}{2}}\sum_{v\in R(u)}(r_{vi}-\bar{r}_v)*sim_{uv}*(1-S_v)；$$

$e_{ui}\leftarrow r_{vi}-\hat{r}_{ui}$；

for k=1:f do

$q_{ik}\leftarrow q_{ik}+\gamma w(e_{ui})e_{ui}p_{uk}$；

$p_{uk}\leftarrow p_{uk}+\gamma w(e_{ui})e_{ui}q_{ik}$

endfor

$b_u\leftarrow b_u+\gamma w(e_{ui})e_{ui}$；

$b_i\leftarrow b_i+\gamma w(e_{ui})e_{ui}$；

endif

endfor

endfor

until P, Q 保持不变；

$$\hat{r}_{vj}=b_{vj}+\boldsymbol{q}_i^{\mathrm{T}}\boldsymbol{p}_v+|R(v)|^{-\frac{1}{2}}\sum_{v\in R(v)}(r_{vj}-\bar{r}_v)*sim_v*(1-S_v)；$$

return $\hat{r}_{vj}$

3　实验分析

3.1　实验测试集

本实验所采用的测试集，是基于网爬技术在旅游网站服务器上获取的数据，该测试集所具有的数据结构见图 4。

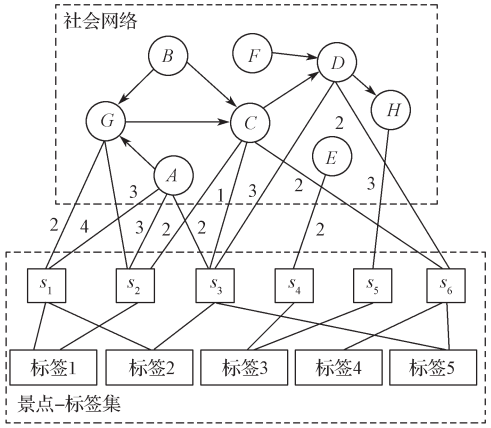


图 4　测试集数据结构

Fig. 4　Test set data structure

图 4 中，长方形对象为游客进行标记的旅游景点标签，正方形对象为与游客产生关联的旅游景点，圆圈对象为游客；游客间通过关注方式可构成社会关系网络，其为一个有向图网络。如果存在由游客 *C* 指向游客 *A* 的网络连接边，则说明游客 *C* 关注了游客 *A*；图 4 中所示边的权值为游客对旅游景点的评价值。

所获取的实验测试集，含有 5 个不同的城市（上海、桂林、杭州、香港和厦门），共含有 992 组旅游景点，以及这些旅游景点的评价游客数量 9 208 人，

评分记录 22 516 条，景点标签 138 个。测试集中的社会网络由 26 948 条边和 15 384 个游客组成。

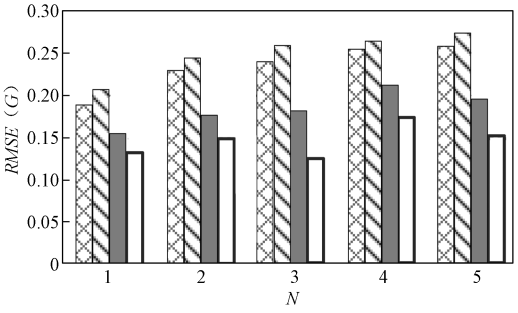
3.2 满意度指标

游客群体对于景点路线的满意度指标可采取均方误差（root mean square error，RMSE）方式进行评价，具体形式为 [15]：

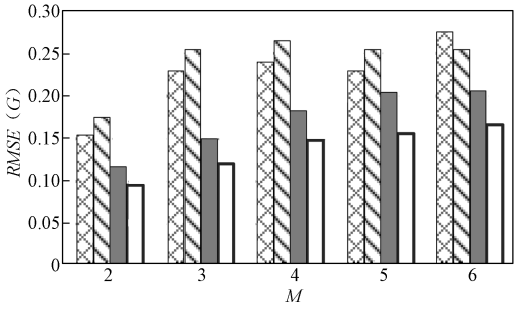
$$RMSE(G)=\sqrt{\frac{1}{M}\sum_{n=1}^{M}\left(Profit(u_n)'-Profit(u_n)''\right)^2}\ 。（27）$$

式中：$Profit(u_n)''$ 为游客 u_n 在最优旅游路线中个人偏好；$Profit(u_n)'$ 为游客 u_n 在最优旅游路线中景点收益均值（满意度）；M 为游客的总体数量。

$RMSE(G)$ 指标越小，表示游客在所设计的群体路线内的均值收益和个人喜好路线收益越接近，此时的群体游客满意度较高。对比算法选取 Least Misrry（最小痛苦策略）、Multiply（多目标满意度乘策略）和 Average（满意度均值化策略）3 种对比策略，对比上述几种情况的 $RMSE(G)$ 指标，实验结果如图 5 所示。



a）参数 N 对算法满意度指标影响



b）参数 M 对算法满意度指标影响

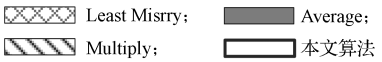


图 5 游客满意度对比

Fig. 5 Comparison of tourists' satisfaction

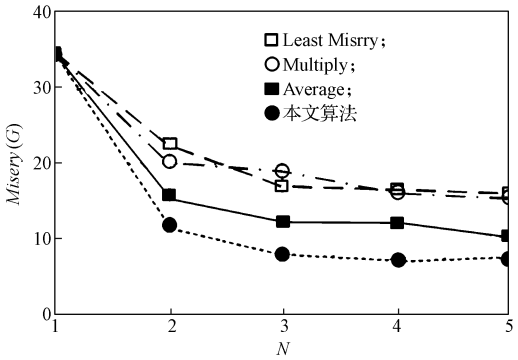
图 5a 给出满意度指标 $RMSE(G)$ 随所设计的旅游景点线路内景点数量 N 的变化趋势。随数量 N 的增大，几种对比策略的满意度指标 $RMSE(G)$ 均增大，这说明景点数量 N 越大，线路推荐的满意度越低。图 5b 给出满意度指标 $RMSE(G)$ 随游客数量变化趋势，可见随旅行团内游客数量 M 增大，几种规划策略的满意度指标 $RMSE(G)$ 值均随之增大，说明旅游团规模越大，线路推荐结果满意度越低。而对比上述选取的几种算法，本文策略相对于选取的对比策略能使群体游客整体上获得相对较高的满意度。

3.3 痛苦度指标对比

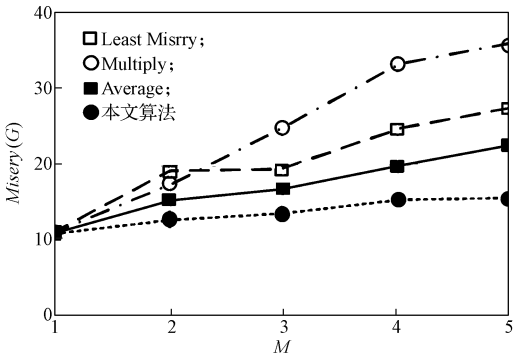
采用在群体旅游线路中的最小收益游客，在景点的收益均值作为痛苦度指标，可定义如下：

$$Misery(G)=\frac{1}{\arg_{u_n\in G}\min\left(Profit(u_n)'\right)}。\quad（28）$$

根据公式可知，指标 $Misery(G)$ 越大，表明旅行团内游客的个别偏好被忽略的可能性更大，旅行团内的游客满意度存在越大的差异。实验结果如图 6 所示。



a）参数 N 对群体痛苦度指标影响



b）参数 M 对群体痛苦度指标影响

图 6 群体痛苦程度对比

Fig. 6 Comparison of group pains

图 6a 给出痛苦度指标 $Misery(G)$ 随所设计的旅游景点线路内景点数量的变化趋势。随着数量 N 的增大，几种对比策略的 $Misery(G)$ 指标均降低，总体上，本文策略的痛苦度指标始终低于选取的几种对比策略。图 6b 给出痛苦度指标 $Misery(G)$ 随游客数量变化趋势，可见，随着旅行团内游客数量增大，几种规划策略的痛苦度指标 $Misery(G)$ 越大，说明此时旅行团内个别游客具有较低的满意度。根据图 6 所示实验结果，本文策略的痛苦度指标要小于选取的几种对比策略，可见本文策略综合考虑了旅行团内所有成员

的个性喜好，使得旅行团内每名游客可获得相对满意的推荐结果。

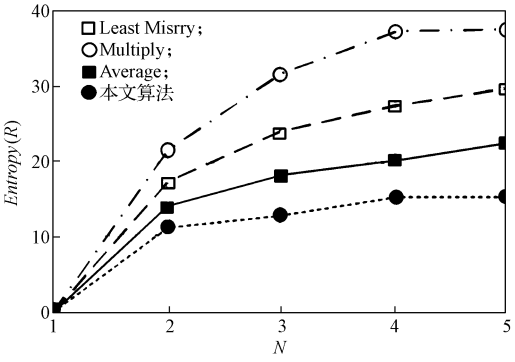
3.4　路线景点多样化指标

在旅行过程中，游客往往希望线路内可以包含更多类型的景点。这里选取旅行路线景点类别熵作为景点多样化指标，该值越大表明旅行线路内的景点多样化程度越大，具体计算形式为

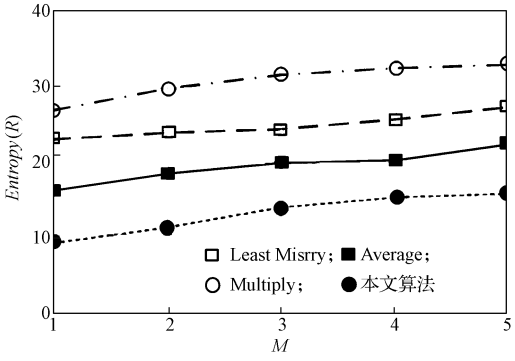
$$Entropy(R)=-\sum_{c\in S(R)}\left(\frac{n(c)}{|R|}\log_{|C|}\frac{n(c)}{|R|}\right)。（29）$$

式（29）中：C 为所有景点的类别集；$|C|$ 为景点的所属类别数；$S(R)=v_1.c\cup v_2.c\cup\cdots\cup v_N.c$ 为景点线路 R 内所有景点的所属类别集；$n(c)$ 为线路内类别是 c 的景点数；$|R|$ 为线路内景点数。

图 7 为线路内景点多样化指标对比结果。



a）参数 N 对地点多样性指标影响



b）参数 M 对地点多样性指标影响

图 7　线路内景点多样性指标对比

Fig. 7　Comparision of diversity index of scenic spots

图 7a 给出旅行线路熵值多样性指标随线路内景点数量 N 的变化趋势。随着数量 N 的增大，几种对比策略的熵值显著增大，这表明此时的旅行线路的景点类别多样性增加。图 7b 显示了给出旅行线路熵值多样度指标随随旅行团内游客数目 M 的变化趋势。可见随旅行团内游客数量 M 的增大，线路熵值多样度指标增大。根据图 7 可知，本文策略所获得的多样性指标结果要优于选取的对比策略，推荐效果较好。

4　结语

本文提出一种基于图基检验（Tukey）的 M 估计游客怀疑度模型的旅游线路协同推荐方法，基于 $k-$距离的游客怀疑度构建可靠邻居模型构建旅游景点路线推荐算法框架，并利用图基检验 M 估计的鲁棒矩阵分解算法，实现旅游线路的有效推荐。本文的主要贡献如下：

1）为了减少攻击对推荐结果的影响，通过利用用户的怀疑度提出了一个可靠的邻居模型。

2）采用 Tukey M 估计，提出了一个强大的矩阵分解模型来实现用户特征矩阵和项目特征矩阵估计，减少攻击对于项目特征矩阵的影响。

3）结合可靠邻居模型和鲁棒矩阵分解模型，提出了一种鲁棒的协同推荐算法。

本文算法具有较广的应用范围，例如与具有节点关联性的网络数据库的信息挖掘算法中。未来的工作主要集中在，如何获取和分析更多种类旅游景点数据，例如旅游照片等，以完善游客相似度计算，提高分析精度。

参考文献：

[1] GAVALAS D，KONSTANTOPOULOS C，MASTAKAS K，et al. A Survey on Algorithmic Approaches for Solving Tourist Trip Design Problems[J]. Journal of Heuristics，2014，20(3)：291−328.

[2] PERBOLI G，GHIRARDI M，GOBBATO L，et al. Flights and Their Economic Impact on the Airport Catchment Area：An Application to the Italian Tourist Market[J]. Journal of Optimization Theory & Applications，2015，164(3)：1109−1133.

[3] LU H C，FANG S H，TSENG V S. Integrating Tourist Packages and Tourist Attractions for Personalized Trip Planning Based on Travel Constraints[J]. GeoInformatica，2016，20(4)：741−763.

[4] MEMON I，CHEN L，MAJID A，et al. Travel Recommendation Using Geo-Tagged Photos in Social Media for Tourist[J]. Wireless Personal Communications，2015，80(4)：1347−1362.

[5] 宋晓宇，许鸿斐，孙焕良，等 . 基于签到数据的短时间体验式路线搜索 [J]. 计算机学报，2013，36(8)：1693−1702.
SONG Xiaoyu，XU Hongfei，SUN Huanliang，et al. Short-Term Experience Route Search Based on Check-In Data[J]. Chinese Journal of Computers，2013，36(8)：1693−1702.

（下转第 78 页）

101.
CHEN Jie，CHEN Zhixiang. The EOQ Model with Multivariate Markov Demand for Deteriorating Items[J]. Journal of Industrial Engineering and Engineering Management，2016，30(4)：93−101.
[13] YANG H，LI Y，LU L，et al. First Order Multivariate Markov Chain Model for Generating Annual Weather Data for Hong Kong[J]. Energy and Buildings，2011，43(9)：2371−2377.
[14] CHING W K，FUNG E S，NG M K. Higher-Order Markov Chain Models for Categorical Data Sequences[J]. Naval Research Logistics，2004，51(4)：557−574.
[15] SIU T K，CHING W K，FUNG S E，et al. On a Multivariate Markov Chain Model for Credit Risk Measurement[J]. Quantitative Finance，2005，5(6)：543−556.
[16] 陈 杰，陈志祥，邢灵博，等 . 带有能力约束的多元马氏需求报童模型 [J]. 管理科学学报，2016，19(7)：37−49.
CHEN Jie，CHEN Zhixiang，XING Lingbo，et al. Capacitated Newsboy Model with Multivariate Markovian Demand[J]. Journal of Management Science in China，2016，19(7)：37−49.

（责任编辑：申 剑）

（上接第 73 页）

[6] CORREIA R F，BRITO C M. Mutual Influence Between Firms and Tourist Destination：A Case in the Douro Valley[J]. International Review on Public and Nonprofit Marketing，2014，11(3)：209−228.
[7] GAVALAS D，KENTERIS M，KONSTANTOPOULOS C，et al. Web Application for Recommending Personalised Mobile Tourist Routes[J]. IET Software，2012，6(4)：313−322.
[8] 吴 澎，朱家明，朱林波，等 . 基于多目标规划和智能优化算法的旅游线路设计研究 [J]. 数学的实践与认识，2016，46(15)：105−114.
WU Peng，ZHU Jiaming，ZHU Linbo，et al. Research on the Design of Travel Route Based on the Multi-Objective Programming and Intelligent Optimization Algorithm[J]. Mathematics in Practice and Theory，2016，46(15)：105−114.
[9] HE Z Q，WU Z Y，ZHOU B C，et al. Tourist Routs Recommendation Based on Latent Dirichlet Allocation Model[C]//Web Information System and Application Conference (WISA). Jinan：IEEE，2015：11−13.
[10] ETAATI L，SUNDARAM D. A Tentative Tourist Recommendation System：Conceptual Frameworks and Implementations[J]. Vietnam Journal of Computer Science，2015，2(2)：95−107.
[11] 姬鹏飞，李远刚，卢盛祺，等 . 基于语义 Web 的旅游景点路线个性化定制系统 [J]. 计算机工程，2016，42(10)：308−317.
JI Pengfei，LI Yuangang，LU Shengqi，et al. Personalized Customization System of Travel Route Based on Semantic Web[J]. Computer Engineering，2016，42(10)：308−317.
[12] GAVALAS D，KENTERIS M. A Web-Based Pervasive Recommendation System for Mobile Tourist Guides[J]. Personal & Ubiquitous Computing，2011，15(7)：759−770.
[13] NEO H F，RASIAH D，TONG D Y K，et al. Biometric Technology and Privacy：A Perspective From Tourist Satisfaction[J]. Information Technology & Tourism，2014，14(3)：219−237.
[14] SHI L，LIN F Y，YANG T C，et al. Context-Based Ontology-Driven Recommendation Strategies for Tourism in Ubiquitous Computing[J]. Wireless Personal Communications，2014，76(4)：731−745.
[15] HORNG G J. The Adaptive Recommendation Mechanism for Distributed Parking Service in Smart City[J]. Wireless Personal Communications，2015，80(1)：395−413.

（责任编辑：申 剑）