中国科学: 信息科学 2020 年 第 50 卷 第 7 期: 1055–1068
SCIENTIA SINICA Informationis
《中国科学》杂志社
SCIENCE CHINA PRESS
领域大数据知识图谱专题 · 论文

# 基于旅游知识图谱的可解释景点推荐

高嘉良[1,2], 仇培元[3], 余丽[4], 黄宗财[1,2], 陆锋[1,2,5*]

1. 中国科学院大学, 北京 100049
2. 中国科学院地理科学与资源研究所资源与环境信息系统国家重点实验室, 北京 100101
3. 山东建筑大学测绘地理信息学院, 济南 250101
4. 中国科学院文献情报中心, 北京 100190
5. 江苏省地理信息资源开发与利用协同创新中心, 南京 210023
* 通信作者. E-mail: luf@lreis.ac.cn

收稿日期: 2019–12–01; 修回日期: 2020–03–17; 接受日期: 2020–04–28; 网络出版日期: 2020–07–09

国家自然科学基金项目 (批准号: 41631177, 41801320) 资助

**摘要** 景点推荐系统可以帮助游客过滤大量的无关信息, 还能辅助商家发掘潜在的顾客. 然而, 现有的基于传统方法的推荐系统, 如基于内容的推荐或协同过滤系统, 虽推荐过程相对透明直观, 但由于数据稀疏性的存在, 推荐结果往往不够准确; 基于深度学习的推荐方法, 虽在一定程度上提高了推荐结果的精度, 但由于缺乏可解释性和透明度, 难以满足部分用户理解推荐依据的愿望, 也阻碍了此类方法的推广应用. 为了解决当前方法所存在的局限, 本文引入基于知识图谱的景点推荐框架, 将推荐过程与知识图谱嵌入相结合, 推断用户兴趣在知识图谱上的传播路径, 以此作为推荐依据. 此外, 本文通过对真实旅游数据的多角度时空分析, 探究旅游活动的时空规律, 并将其应用于景点推荐框架中, 提出一种面向旅游的基于知识图谱的可解释推荐方法 —— Geo-RippleNet, 并通过构建基于开放网络资源的旅游知识图谱, 对 Geo-RippleNet 进行了全面的实验验证. 结果表明, 本文提出的基于知识图谱的景点推荐方法, 不仅可以最大限度地吸收知识图谱丰富的语义信息, 从而实现可观的性能提升, 还能充分利用图谱的关系知识, 推理兴趣传播路径, 以增强推荐结果的可解释性. 此外, 将旅游活动的时空规律融入到上述推荐框架中, 能够还原用户出游和决策的时空过程, 进一步提高方法的性能表现.

**关键词** 旅游知识图谱, 景点推荐, 可解释性, 推荐系统, 旅游管理

## 1 引言

互联网在线旅游平台为用户提供了方便快捷的信息服务, 但与此同时, “信息过载” 问题也愈发严重, 导致用户往往需要花费大量时间和精力来阅读、甄别和筛选这些信息以辅助行程规划. 因此, 推荐

**引用格式:** 高嘉良, 仇培元, 余丽, 等. 基于旅游知识图谱的可解释景点推荐. 中国科学: 信息科学, 2020, 50: 1055–1068, doi: 10.1360/SSI-2019-0268
Gao J L, Qiu P Y, Yu L, et al. An interpretable attraction recommendation method based on knowledge graph (in Chinese). Sci Sin Inform, 2020, 50: 1055–1068, doi: 10.1360/SSI-2019-0268

 

系统应运而生 [1]. 它们通过分析用户的兴趣和需求, 推断与之相关的目的地或景点集, 从而过滤掉大量的无关信息, 辅助人们的出游规划 [2].

现有的景点推荐方法可分为两类: 以基于内容的推荐 (content-based, CB) [3]、协同过滤 (collaborative filtering, CF) [4] 为代表的传统推荐方法, 和基于深度学习的推荐方法 (deep learning, DL) [5]. 传统的 CB 和 CF 方法相对直观透明, 通过度量用户历史与候选景点之间的相似性来做出推荐, 但性能受限于复杂的特征工程和不可避免的数据稀疏性. DL 方法通过引入大量参数来刻画旅游活动中的未知机制, 从而获得优于传统方法的性能提升, 但存在广受诟病的 "黑箱" 问题 [6], 导致缺乏可解释性和透明度, 难以满足部分用户理解推荐依据的愿望, 以至于用户对推荐结果心存疑虑, 也阻碍了此类方法的推广应用. 推荐结果的准确性是推荐系统的核心需求. 但与此同时, 可解释性往往会影响用户对推荐结果的接受程度. 具有良好可解析性的系统不仅能够提升系统透明度, 还能够提高用户对系统的信任水平、对推荐结果的接受概率, 以及用户满意度.

为了弥补目前推荐方法存在的不足, 业界将知识图谱 (knowledge graph, KG) 引入到推荐系统中. 知识图谱将现实世界中的实体和关系表达为三元组形式, 并以实体作节点, 关系作边, 构成一个大规模的异构语义网络 [7]. 基于知识图谱的推荐方法一般将知识图谱和用户 – 景点交互矩阵拼接成一个异构信息网, 通过定义一些特殊元路径/图的方式, 利用知识图谱中的关系知识. 该类方法较为简单直接, 但在实际操作中往往受限于元路径/图的手工构建过程, 一旦场景发生变化, 元路径/图需要重新构建, 并且在元路径/图的基础上仍需进行张量分解等操作, 导致可解释性依旧较差. 因此, 业界提出了基于知识图谱的端到端的推荐框架 RippleNet [8]. 此框架将旅游场景每个用户和景点, 及其知识图谱中的实体和关系, 嵌入表示为向量或矩阵, 采用联合学习的方式, 将知识图谱嵌入和景点推荐两个任务相结合, 最大限度地挖掘推荐场景和知识图谱的语义知识. 此外, 该框架还能够在知识图谱上, 推断从用户历史到候选景点的链接路径, 以刻画用户兴趣的传播过程, 最后将该链接路径作为推荐结果的解释.

景点推荐有别于通用推荐场景, 因为旅游活动发生于真实的物理世界, 必然存在着某些特定的时空规律. 因此, 如何刻画这些时空规律对于景点推荐至关重要. 本文从地理学视角出发, 对真实的旅游数据集进行了全面的时空分析, 挖掘其中存在的地理聚集效应 [9] 和季节性规律 [10], 并将其嵌入到 RippleNet 框架, 提出了一种全新的基于旅游知识图谱的可解释景点推荐方法 —— Geo-RippleNet. 本文基于网络开放资源构建了一个大规模的旅游知识图谱, 并在真实用户出游数据集上对 Geo-RippleNet 方法进行了对比验证. 实验结果表明, 相较于传统的推荐方法, Geo-RippleNet 在 AUC (area under the curve) 和准确率上的提升达 84% 和 72%. 最后, 通过案例分析, 以用户历史到候选景点的兴趣传播路径, 说明了 Geo-RippleNet 对推荐结果的解释过程.

## 2 相关工作

### 2.1 景点推荐方法

CB 方法是景点推荐中最为经典的推荐策略 [2]. 该方法一般分为两个模块, 特征工程和相似度匹配. 特征工程即为每个景点构建表示特征, 如属性或关键词等; 相似度匹配, 在特征空间上计算用户历史和候选景点之间的相似性, 以最为相似的候选作为推荐结果 [11], 或者更为复杂的方式是从用户历史中学习一个分类器, 如朴素贝叶斯 (naive Bayes) 或线性判别分析 (linear discriminant analysis, LDA), 预测候选景点会被用户访问的概率 (click through rate, CTR) [3]. 基于内容的推荐简单高效, 但实际操作易受限于特征工程的手工构建过程和冷启动问题约束.

CF 方法不需要特征工程, 而是将全体用户和景点表达为一个交互矩阵, 以用户表示景点特征 (item-based CF), 或者反之, 以景点表示用户特征 (user-based CF), 通过相似度计算, 匹配相关景点, 完成推荐. 虽然 CF 方法广泛应用于各种推荐场景中, 但其受限于数据稀疏性、“灰山羊” [12] 和可扩展性等问题约束.

DL 方法近年来相对流行. 该类方法希望通过大量参数和非线性变换来拟合真实世界中的复杂机制. 在景点推荐中, 通常将用户和景点各表示为一个向量, 每个维度对应于某种隐式特征, 通过内积等操作, 将用户和景点的隐式特征向量转化为预测概率 [13,14]. DL 相较于传统方法, 虽一定程度上提升了性能, 但也带来 “黑箱” 问题, 无法向用户合理地解释推荐过程.

### 2.2 基于知识图谱的推荐过程

在知识图谱的概念提出前, 已有许多类似的研究, 将旅游本体作为辅助信息, 用于增强 CB 或 CF 方法 [15,16]. 旅游本体是一个由旅游领域专家手工构建的, 包含了各种旅游活动和游客画像的概念体系. 基于旅游本体定义的相似性度量, 可以缓解数据稀疏性, 提升推荐性能. 但本体的构建过程极其复杂. 目前存在的旅游本体相对较小, 难以支撑大规模的推荐系统. 因此业界提出了旅游知识图谱概念 [17]. 早期基于知识图谱的景点推荐研究往往遵循异构信息网络的思想, 定义一些面向具体场景的元路径/图, 表达用户与景点之间的连通性, 以提升矩阵分解等方法的表现 [18,19].

但是基于元路径/图的方法, 仍然依赖于领域专家根据推荐场景构造图结构, 成本较高; 并且仅将知识图谱视为有向图, 损失了结构化知识的语义信息. 因此, 知识图谱嵌入技术 (knowledge graph embedding, KGE) [20] 被引入推荐场景中. KGE 将图谱中的每一个实体和关系都表示为一个独立的向量或矩阵, 最大化地获取结构化知识的语义信息, 增加推荐场景中的信息量, 提升推荐性能 [21]. 基于 KGE 的兴趣传播框架 RippleNet, 在性能提升的同时, 也通过推断兴趣传播路径为可解释推荐提供了可能.

## 3 方法

### 3.1 相关定义

$\mathcal{U}=\{u_1,u_2,\ldots\}$ 和 $\mathcal{V}=\{v_1,v_2,\ldots\}$ 分布表示景点推荐场景中的用户和景点集合, 根据用户是否访问过某个景点, 我们可以定义一个用户 – 景点的交互矩阵 $\boldsymbol{Y}$:

$$y_{uv}=\begin{cases}1, & \text{如果用户访问},\\ 0, & \text{反之}.\end{cases} \tag{1}$$

知识图谱 $\mathcal{G}$ 由大量 “头实体 – 关系 – 尾实体” 的三元组构成, 即 $(h,r,t)$, 其中 $h\in\mathcal{E}$, $r\in\mathcal{R}$ 和 $t\in\mathcal{E}$, $\mathcal{E}$ 和 $\mathcal{R}$ 分别代表知识图谱中的实体和关系集合. 景点推荐任务的形式化定义如下.

**定义1** (景点推荐任务) 给出用户 – 景点交互矩阵 $\boldsymbol{Y}$ 和知识图谱 $\mathcal{G}$, 预测用户 $u$ 未来到访候选景点 $v$ 的概率 $\hat{y}_{uv}$, 并根据此排序生成一个推荐结合:

$$\hat{y}_{uv}=\mathcal{F}(u,v;\Theta), \tag{2}$$

其中, $\mathcal{F}(\cdot)$ 代表推荐方法, $\Theta$ 代表方法中的参数集合.

本文的符号使用与原始的 RippleNet 框架一致, 其中两个关键定义如下.

**定义2** (相关实体)　给出用户 – 景点交互矩阵 $\boldsymbol{Y}$ 和知识图谱 $\mathcal{G}$, 用户 $u$ 的 $k$ 阶相关实体定义为

$$\mathcal{E}_u^k = \{t|(h,r,t) \in \mathcal{G} \text{ and } h \in \mathcal{E}_u^{k-1}\}, \quad k = 1,2,\ldots,H, \tag{3}$$

其中, $\mathcal{E}_u^0 = \mathcal{V}_u = \{v|y_{uv}{=}1\}$ 是用户 $u$ 访问过的历史景点集合, 代表用户兴趣的起始点.

**定义3** (水波集)　用户 $u$ 的 $k$ 阶水波集 (ripple set) 定义为, 以 $\mathcal{E}_u^{k-1}$ 为头实体的三元组集合:

$$S_u^k = \{(h,r,t)|(h,r,t) \in \mathcal{G} \ \text{ and } \ h \in \mathcal{E}_u^{k-1}\}, \quad k = 1,2,\ldots,H. \tag{4}$$

### 3.2　RippleNet 框架

RippleNet 以一个用户 $u$ 和一个候选景点 $v$ 为输入, 输出该用户 $u$ 访问景点 $v$ 的概率. 用户 $u$ 的历史记录 $\mathcal{V}_u$ 代表了其潜在的兴趣, 以此作为知识图谱上兴趣传播的起始点, 生成各阶兴趣水波集 $S_u^k$ $(k = 1,2,\ldots,H)$. RippleNet 的整体框架请参考文献 [8] 中的图 2.

RippleNet 将每个候选景点嵌入表示为一个 $d$ 维向量 $\boldsymbol{v} \in \mathbb{R}^d$. 给出用户 $u$ 的访问历史 $\mathcal{V}_u$, 一阶水波集上的三元组 $(h_i, r_i, t_i)$ 与候选景点 $v$ 将计算一个相关性概率 $p_i$ 如下:

$$p_i = \text{softmax}(\boldsymbol{v}^{\mathrm{T}}\boldsymbol{R}_i\boldsymbol{h}_i) = \frac{\exp(\boldsymbol{v}^{\mathrm{T}}\boldsymbol{R}_i\boldsymbol{h}_i)}{\sum_{(h,r,t)\in S_u^1}\exp(\boldsymbol{v}^{\mathrm{T}}\boldsymbol{R}_i\boldsymbol{h}_i)}, \tag{5}$$

其中, $\boldsymbol{R}_i \in \mathbb{R}^{d\times d}$ 是一阶水波集上的三元组关系的 $d \times d$ 维矩阵表示, $\boldsymbol{h}_i \in \mathbb{R}^d$ 是三元组的头实体 $d$ 维向量表示. 相关性概率 $p_i$ 可以理解为候选景点 $v$ 与用户历史 $\mathcal{V}_u$, 即头实体 $h_i$, 在三元组关系 $r_i$ 上的语义相关性. 然后, 潜在兴趣将在一阶 RippleNet 上由头实体 $h_i$ 向尾实体 $t_i$ 进行传播, 并在尾实体处积累. 因此, 以相关性概率 $p_i$ 为权重, 将所有尾实体 $t_i$ 进行求和, 得到一阶 RippleNet 上的潜在兴趣表示 $\boldsymbol{o}_u^1$:

$$\boldsymbol{o}_u^1 = \sum_{(h_i,r_i,t_i)\in S_u^1} p_i\boldsymbol{t}_i, \tag{6}$$

其中, $\boldsymbol{t}_i \in \mathbb{R}^d$ 是尾实体 $t_i$ 的 $d$ 维向量表示.

对于二阶水波集的兴趣表示 $\boldsymbol{o}_u^2$, 只需用 $\boldsymbol{o}_u^1$ 替换式 (5) 中 $\boldsymbol{v}$, 其他保持不变. 对于更高阶的水波集兴趣, $\boldsymbol{o}_u^3,\ldots,\boldsymbol{o}_u^H$, 重复以上迭代过程即可. 用户 $u$ 的向量表示定义为该用户在各阶水波集上兴趣的累加:

$$\boldsymbol{u} = \boldsymbol{o}_u^1 + \boldsymbol{o}_u^2 + \cdots + \boldsymbol{o}_u^H, \tag{7}$$

最终, 用户的向量表示 $\boldsymbol{u}$ 和候选景点的向量表示 $\boldsymbol{v}$ 通过内积相互结合, 得到访问概率预测值 $\hat{y}_{uv}$:

$$\hat{y}_{uv} = \text{sigmoid}(\boldsymbol{u}^{\mathrm{T}}\boldsymbol{v}). \tag{8}$$

### 3.3　旅游活动时空规律挖掘

旅游活动所具有的时空规律对于推荐过程具有重要影响. 因此, 本文从地理学视角对旅游行为进行时空分析, 挖掘其中蕴含的地理聚集效应和季节性规律. 本文时空分析的数据集来自旅游评论网站 —— 马蜂窝, 共有 72638 名用户关于 20952 个全国景点的 114724 条出游记录.

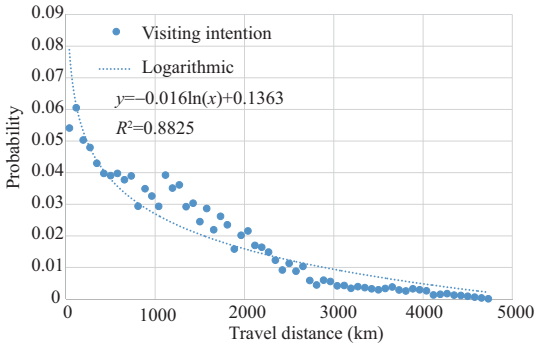


**图 1 (网络版彩图) 通行距离对景点访问意愿的影响**

**Figure 1** (Color online) The influence of travel distance on the visiting intention of attractions

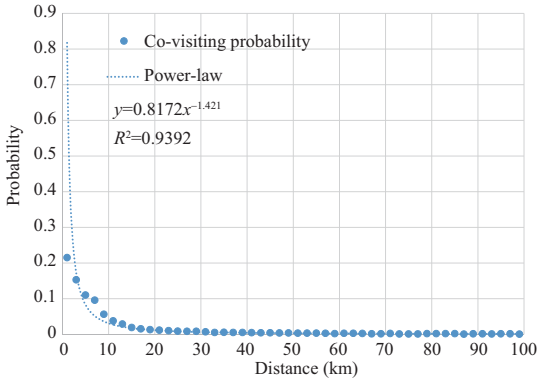


**图 2 (网络版彩图) 景点共现概率与相对距离的关系**

**Figure 2** (Color online) The relationship between the co-occurrence probability and relative distance of attractions

### 3.3.1 地理聚集效应

首先, 我们分析通行距离对景点访问意愿的影响, 以用户居住地到目的地的距离作为通行距离, 绘制如图 1 所示的概率分布直方图. 从中可以发现, 随着通行距离的增加, 游客访问意愿服从对数式衰减. 此外, 在 1000~2000 km 的距离范围, 直方图略高于拟合的对数式曲线, 说明存在 “近景易嫌” 现象, 即人们对可达范围内且新颖的景点更感兴趣.

此外, 大量研究证明人类出行服从幂律分布. 由于该空间模式类似于动物觅食的行为, 所以又被称为 “Levy-flight” 模式 [22]. 本文分析了景点被同一用户共同访问的概率与景点之间相对距离的关系, 发现旅游活动也呈现显著的幂率分布模式, $R^2$ 达到 0.939, 如图 2 所示.

上述分析结果表明, 旅游活动体现出两种现象: (1) 用户倾向于访问居住地附近的景点; (2) 用户倾向于历史记录附近的景点. 该现象说明了旅游活动中存在着地理聚类效应.

### 3.3.2 季节性规律

时间效应同样会对旅游行为产生极大影响. 首先, 我们对 2014 年 1 月 ~ 2019 年 4 月的游客出行数据按月份进行统计, 并通过 STL 算法 [23] 对其进行周期性分解. 如图 3 所示, 旅游活动体现出明显的季节性和平稳的周期性. 在新年到来时, 由于春节团聚, 景点的访问量减少至最低, 而暑假和 “十一” 黄金周, 景点的访问量分别达到峰值.

接下来, 我们将访问量占比超过 15% 的月份设为旺季, 低于 5% 的月份设为淡季, 所有景点的按月淡旺季分布差异如图 4 所示. 经过更细的景点统计, 我们发现不同景点具有完全不同的淡旺季分布, 这意味着在不同时间段, 不同景点的吸引力完全不一致.

## 3.4 Geo-RippleNet 推荐方法

我们将旅游行为的时空规律嵌入到知识图谱兴趣传播框架中, 提出了一种针对旅游场景的新的推荐方法 —— Geo-RippleNet.

对于地理聚集效应, 我们分别修改模型的兴趣激活和输出层访问概率的计算过程. 根据 3.3.1 小节, 由于用户对历史记录附近的景点更感兴趣, 所以在计算兴趣相关概率时, 我们考虑历史景点与候

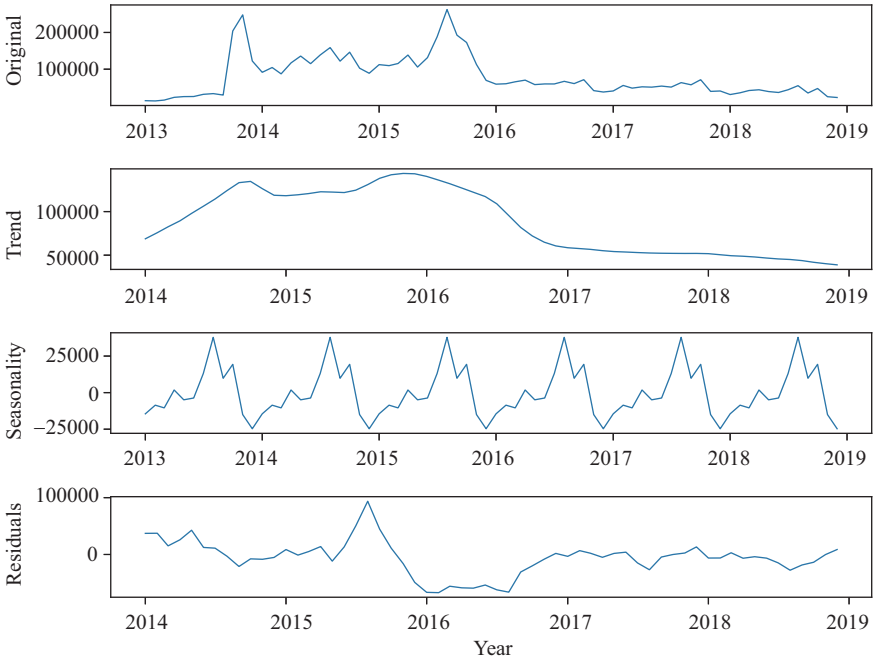


**图 3** (网络版彩图) 旅游行为的周期性分解

**Figure 3** (Color online) The periodic decomposition of travel behavior

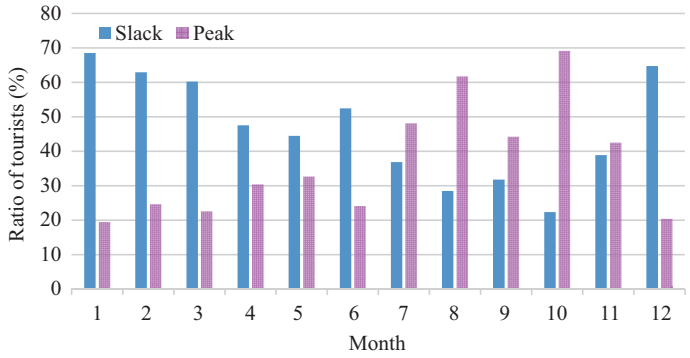


**图 4** (网络版彩图) 景点的淡旺季分布

**Figure 4** (Color online) The off-peak season distribution of tourists to attractions

选景点相对距离对兴趣概率的幂率式衰减, 将式 (5) 修改如下:

$$p_i = \mathrm{softmax}(\boldsymbol{v}^{\mathrm{T}}\boldsymbol{R}_i\boldsymbol{h}_i \cdot \mathrm{dec}_{a2a}) = \frac{\exp(\boldsymbol{v}^{\mathrm{T}}\boldsymbol{R}_i\boldsymbol{h}_i \cdot \mathrm{dec}_{a2a})}{\sum_{(h,r,t)\in S_u^1}\exp(\boldsymbol{v}^{\mathrm{T}}\boldsymbol{R}_i\boldsymbol{h} \cdot \mathrm{dec}_{a2a})}, \tag{9}$$

$$\mathrm{dec}_{a2a} = a_1 \cdot \mathrm{dist}_{a2a}^{b_1} + c_1, \tag{10}$$

其中, $\mathrm{dec}_{a2a}$ 是景点间相对距离 $\mathrm{dist}_{a2a}^{b_1}$ 的幂率衰减项, $a_1$, $b_1$, $c_1$ 为衰减项的参数.

用户出行意愿也受到居住地到目的地之间通行距离的影响. 因此, 在访问概率计算的输出层, 我们同时考虑了通行距离所致的对数式衰减.

$$\hat{y}_{uv} = \mathrm{sigmoid}(\boldsymbol{u}^{\mathrm{T}}\boldsymbol{v} \cdot \mathrm{inf}_{\mathrm{space}}), \tag{11}$$

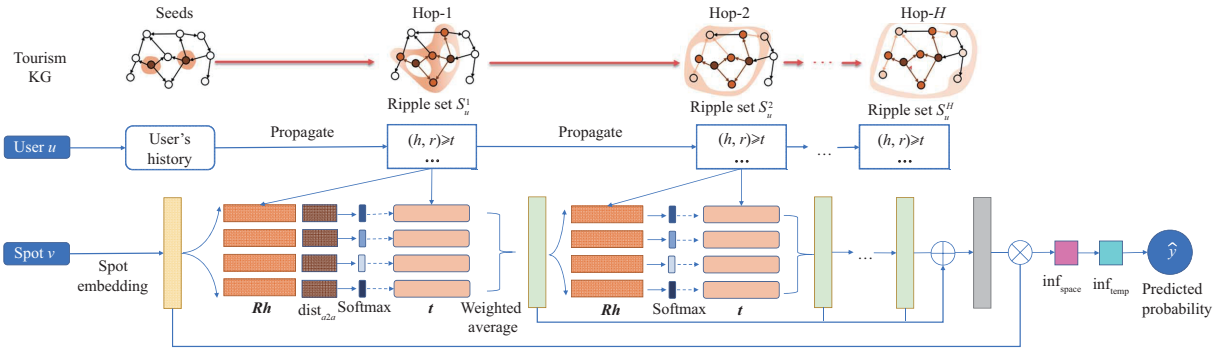


**图 5 (网络版彩图) Geo-RippleNet 模型框架**

**Figure 5** (Color online) The framework of Geo-RippleNet

$$\text{inf}_{\text{space}} = a_2 \cdot \ln(\text{dist}_{r2d}) + b_2, \tag{12}$$

其中, $\text{inf}_{\text{space}}$ 为通行距离 $\text{dist}_{r2d}$ 的对数衰减项, $a_2$, $b_2$ 为对数函数的参数.

根据 3.3.2 小节所述的旅游行为的季节性, 景点在不同月份具有不同的吸引力. 考虑到各个景点均有着不同的变化模式, 且难以确定刻画其分布的数学模型, 因此, 在本研究中, 我们尝试通过简单多项式方法近似拟合其分布. 通过实验发现, 三次多项式表现最优. 因此, 我们采用三次多项式来刻画景点这种随月份变化的吸引力, 并将其嵌入到推荐框架中:

$$\hat{y}_{uv} = \text{sigmoid}(\boldsymbol{u}^{\mathrm{T}}\boldsymbol{v} \cdot \text{inf}_{\text{temp}}), \tag{13}$$

$$\text{inf}_{\text{temp}} = a_3 \cdot (\text{level}_{vj})^3 + b_3 \cdot (\text{level}_{vj})^2 + c_3 \cdot \text{level}_{vj} + d_3, \tag{14}$$

$$\text{level}_{vj} = \frac{\text{num}_{vj}}{\max_{i\in\{1,2,\ldots,12\}}(\text{num}_{vi})}, \tag{15}$$

其中, $\text{inf}_{\text{temp}}$ 是景点的季节性吸引力作用项, $\text{level}_{vj}$ 是候选景点 $v$ 在 $j$ 月的吸引力大小, 由游客访问量的全年占比计算得到, $a_3$, $b_3$, $c_3$, $d_3$ 是三次多项式函数的参数.

综上, 我们得出了知识图谱兴趣传播框架的地理版本 —— Geo-RippleNet, 模型框架如图 5 所示.

## 4 实验

### 4.1 知识图谱构建

本文利用网络开放资源构建了一个大规模的旅游知识图谱. 信源来自 3 种类型的数据: 结构化、半结构化和非结构化数据. 结构化数据指已构建的知识库. 本文选择 CN-DBpedia 作为结构化数据源. CN-DBpedia 构建自中文百科语料, 已经成功应用于智能问答系统和智能医疗等任务 [24]. 以头实体属于的地点和旅游两个类别下的三元组作为我们的旅游知识图谱的结构化部分. 半结构化数据来自马蜂窝网站的景点主页中的网络表格, 本文共爬取了超过 10 万个景点的首页介绍, 并通过网页解析和正则表达式提取出景点的基本属性, 如门票、交通、开放时间等, 最后将这些属性以三元组形式纳入旅游知识图谱中. 非结构化数据来自马蜂窝或百度百科中的旅游景点介绍和评论. 对于非结构化的文本数据, 我们采用 TF-IDF 算法从中获取关键字来表达景点的活动信息.

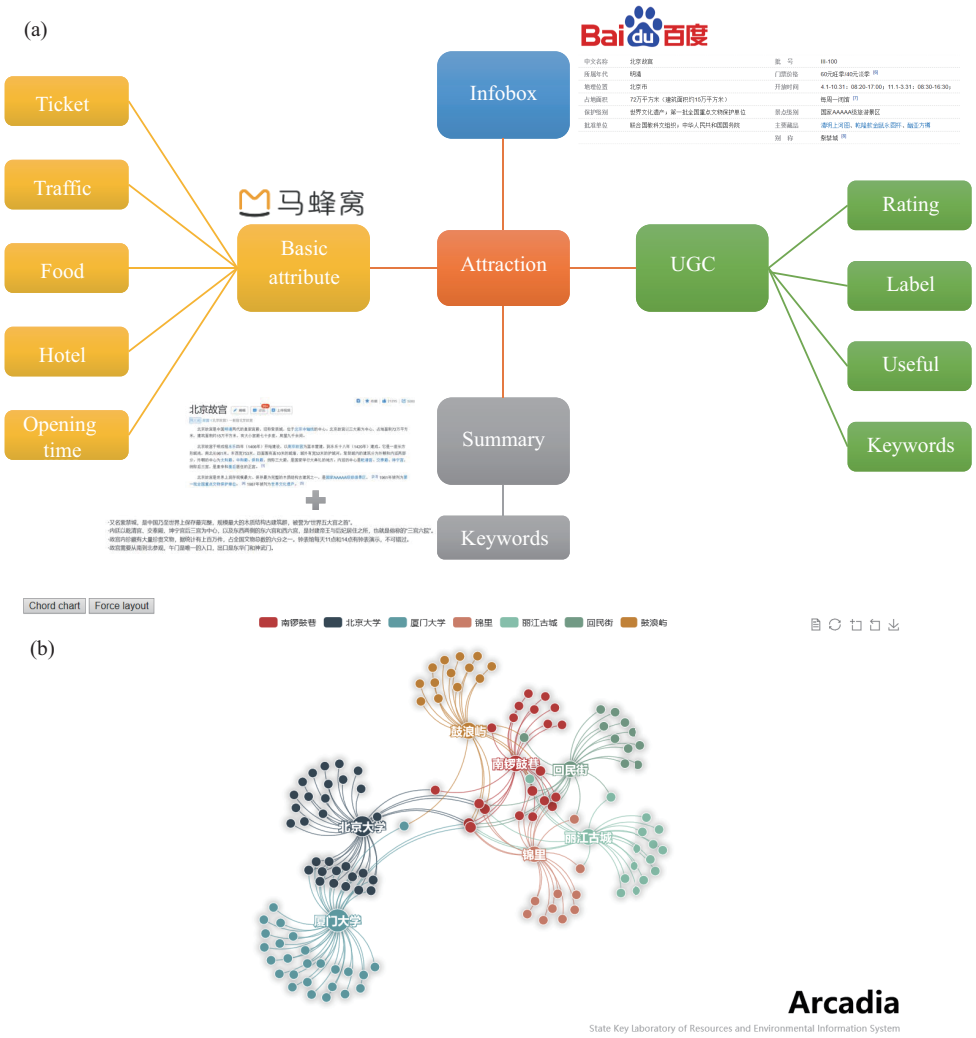


**图 6** (网络版彩图) 旅游知识图谱的 (a) 框架示意图和 (b) 示例

**Figure 6** (Color online) (a) The framework diagram and (b) the demo of tourism knowledge graph

综合以上 3 种数据源, 本文所构建的旅游知识图谱如图 6 所示[1]. 该旅游知识图谱覆盖了全国 700 多个城市的 20952 个景点, 共包含 369599 个三元组, 123261 个实体和 323042 个关系.

## 4.2 实验数据集

本文以马蜂窝网站 2014~2019 年的用户访问记录作为景点推荐任务的实验数据集. 为了避免冷启动问题, 我们清洗掉历史记录不足 5 个景点的用户和访客量不足 100 个的景点. 该数据集共含有 10000 个用户, 分布在全国 700 多个城市的 10000 个景点和 3759921 条访问记录. 如果用户访问某景点, 两者之间的交互设为 1, 否则设为 0. 在实验过程中, 将数据集按照 6:2:2 划分为训练集、验证集和测试集.

1) https://gaojialianglreis.github.io/KG_Demo/.

### 4.3 对比方法

对于景点推荐场景, 本文选择以下方法进行性能比较.

• 基于内容的推荐 (CB): 以景点作头实体的三元组作为独热特征, 选择朴素贝叶斯分类器预测用户对候选景点的访问概率.

• 基于协同过滤的推荐 (CF): 基于物品的协同过滤算法 (item-based CF).

• Wide&Deep: 基于深度学习的强基线推荐方法, 将线性与非线性特征通道相结合. Wide&Deep 的景点特征输入与 CB 一致.

• Rank-GeoFM: POI 推荐场景中的强基线方法, 将签到行为的时空特性融合于张量分解框架.

• RippleNet: 本文方法所依据的基础框架.

### 4.4 评价指标

本文采用两种评价体系, 对比各方法的性能表现:

(1) 在点击率预测中 (CTR), 使用训练好的模型对测试集中每个用户 – 景点交互进行预测, 以 AUC 和准确率 (accuracy) 来评估各方法的性能.

(2) 在最优推荐集中 (Top-K), 使用训练好的模型为测试集中的每个用户在所有景点中, 筛选预测概率最高的前 $K$ 个候选景点作为推荐结果, 以 Precision@K, Recall@K 和 F1@K 来评估推荐结果.

## 5 结果

### 5.1 方法对比

各方法在点击率预测 (CTR) 和最优推荐集 (Top-K) 中的性能表现如表 1 和图 7 所示. 实验结果分析与讨论如下:

• CB 和 CF 方法表现最差. 两者准确率在 0.5 附近, 相当于我们对测试集都预测为正也能达到近似的效果, 并且 CB 的 AUC 接近于 0.5, 意味着 CB 预测结果接近于完全随机分布. 造成 CB 完全失效的原因是, 以知识图谱中的三元组为高维稀疏特征, 难以获取知识图谱的语义信息, 更无法捕捉用户的兴趣分布. CF 表现较差的原因是用户 – 景点交互数据的稀疏性. 实验数据集中各景点平均到访游客数仅为 21, 难以通过所有访客行为代表该景点的特征. 该实验结果与之前较多研究不一致, 原因是后者往往采取在线测试的方法, 将推荐结果呈现给用户, 让用户主观判断, 而我们采取的是 CTR 测试, 即预测用户是否会在未来真正到访所推荐的景点. 两者造成差异的原因是用户感兴趣的景点较多, 但现实中有机会到访的景点却很少. 我们在进行误差分析中, 也发现概率最高的假阳性景点均为非常著名的景点, 概率最高的假阴性景点均为较冷门的景点. 说明传统的 CB 和 CF 方法倾向于推荐知名景点, 而对冷门景点的关注不足.

• Wide&Deep 相较于 CB 方法, 具有显著的性能提升, 说明深度学习框架的特征表示能力优于传统的特征工程. 相较于 CF 方法, Wide&Deep 方法的实验结果也证明了只要合理使用知识图谱中的三元组, 知识图谱可以作为一种可靠的辅助信息, 缓解用户 – 景点交互数据的稀疏性.

• Geo-RankFM 的表现性能接近于 Wide&Deep 方法. Geo-RankFM 方法属于基于模型的 CF 方法, 没有利用知识图谱等其他辅助信息, 仅考虑旅游行为的时空特性便可以达到接近于 Wide&Deep 的推荐效果, 说明旅游行为时空效应的刻画对景点推荐至关重要.

表 1　各推荐方法的 AUC 和准确率表现

**Table 1**　AUC and accuracy performance of each recommended method

| Model | AUC | Accuracy |
|---|---|---|
| CB | 0.504 | 0.499 |
| CF | 0.782 | 0.501 |
| Wide&Deep | 0.855 | 0.773 |
| Geo-RankFM | 0.837 | 0.754 |
| RippleNet | 0.912 | 0.821 |
| Geo-RippleNet | 0.928 | 0.852 |

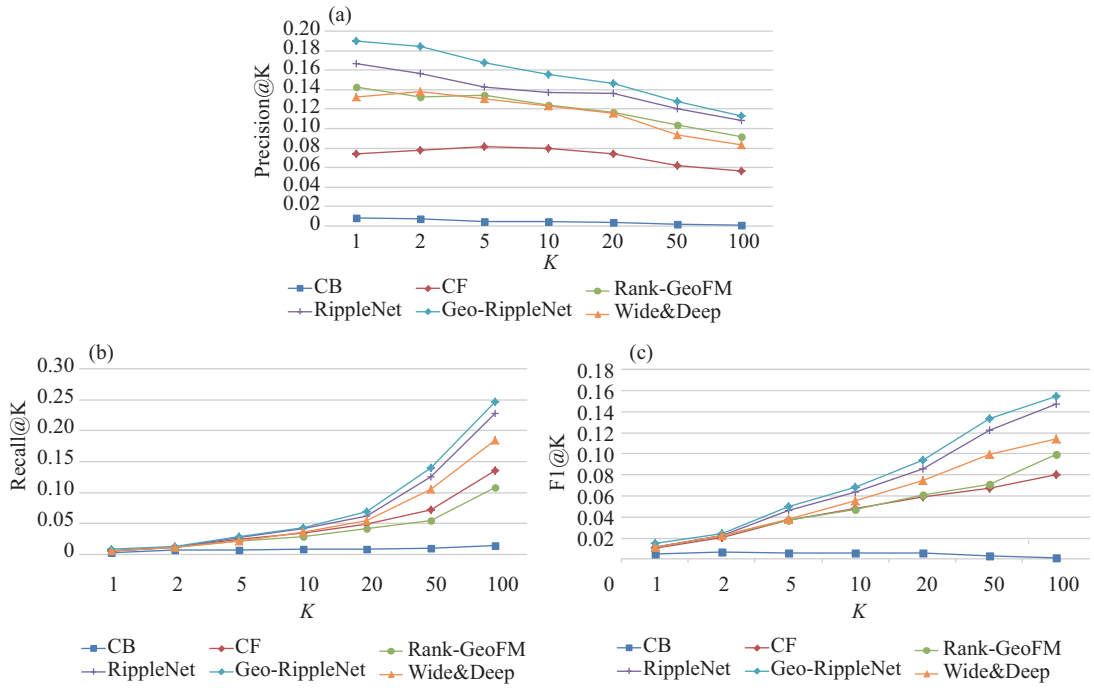


图 7　(网络版彩图) 最优推荐集的结果

**Figure 7**　(Color online) The result of the optimal recommendation set. (a) Precision@K; (b) Recall@K; (c) F1@K

• RippleNet 的表现优于以上所有的对比方法, 证明结合 KGE 技术能够最大程度获取知识图谱和推荐场景中的语义信息.

• Geo-RippleNet 的性能优于 RippleNet 和以上对比方法, 在 AUC 和准确率上分别实现了 1%∼84% 和 1%∼72% 的性能提升. 表明在对用户兴趣在知识图谱上的传播过程进行建模时, 考虑旅游活动的时空效应是非常有意义的.

## 5.2　时空效应的具体影响

我们通过点击率预测 (CTR), 对比了地理聚集效应和季节性分别对知识图谱兴趣传播框架的性能提升, 如表 2 所示. 地理聚集效应的性能提升大于季节性的提升, 说明用户在出游中更看重空间因素的影响. 两种效应共同作用的性能提升最大, 说明景点推荐中考虑时空因素是正确的.

表 2 时空效应的具体影响

**Table 2** Specific influence of the space-time effect

| Model | AUC | Accuracy |
|---|---|---|
| RippleNet | 0.912 | 0.821 |
| Space-RippleNet | 0.921 | 0.839 |
| Temp-RippleNet | 0.918 | 0.831 |
| Geo-RippleNet | 0.928 | 0.852 |

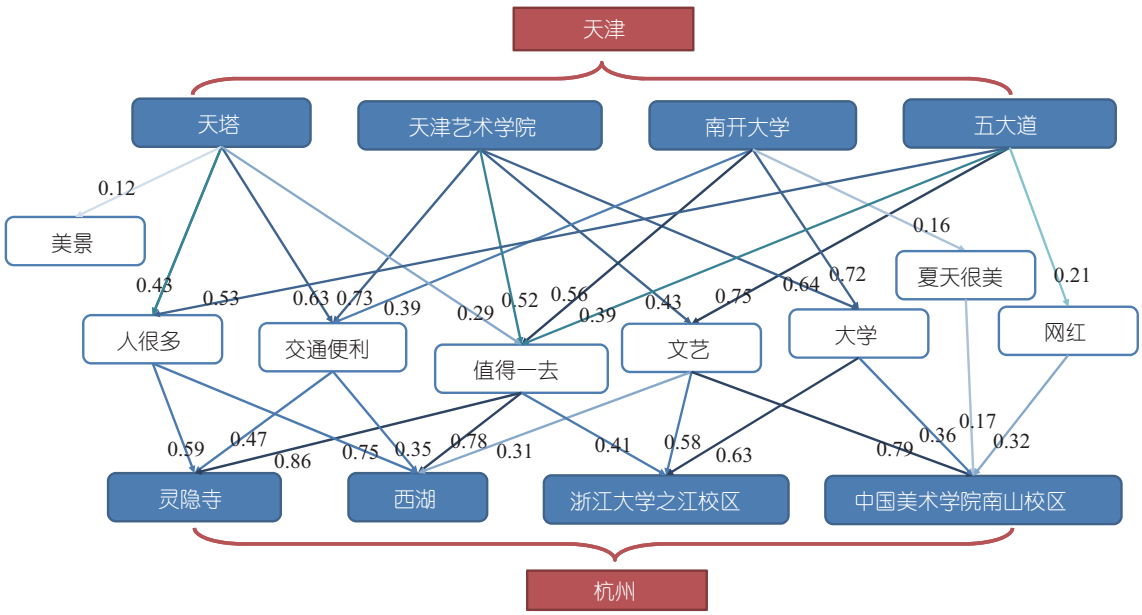


图 8 (网络版彩图) 案例展示

**Figure 8** (Color online) The case study

### 5.3 案例分析

为了直观展示 Geo-RippleNet 的良好解释性, 本文随机选取一位到访过天津 4 个景点的用户, 从其测试数据中选择杭州 4 个正确预测的候选景点, 输出杭州的 4 个候选景点与天津的历史访问景点的前 $k$ 阶兴趣相关概率 $p$, 构建从天津历史景点到杭州候选景点的兴趣传播路径, 如图 8 所示. 其中蓝色的阴影表示较大的兴趣相关概率值, 为了清楚展示, 省略了关系的名称, 同时兴趣相关概率值小于 0.1 的路径也予以省略. 图中, 杭州西湖的推荐理由是, 从用户历史记录中发现, 用户兴趣分布于 “人多”、“交通便利”、“知名”、“历史文化” 等实体, 而这些实体与杭州西湖高度关联. 另外一点值得注意的是, Geo-RippleNet 从用户的历史到候选节点构建了大量的链接路径, 从而形成了一个兴趣分布网络, 其中被多次经过的枢纽实体可以认为是用户的潜在画像.

## 6 结论

本文提出了一种面向旅游场景的基于知识图谱的景点推荐方法 Geo-RippleNet, 成功地将旅游活动的时空规律嵌入兴趣传播框架中, 实现了高精度推荐和良好的可解释性. 本文通过时空分析发现

了旅游活动具有的地理聚类效应和季节性规律, 并将景点间距离建模至兴趣激活层; 在预测访问概率时, 不仅考虑了通行距离的影响, 还表达了景点吸引力的季节性变化. 实验中, 本文整合网络开放资源, 构建了一个大规模的旅游知识图, 并在真实出游数据集上对方法进行了充分的实验验证. 结果表明, Geo-RippleNet 相较于其他对比方法具有显著的性能提升, 证明了基于知识图谱的兴趣传播框架的有效性, 和旅游活动的时空规律对推荐过程的重要意义.

## 参考文献


1 Ricci F, Rokach L, Shapira B. Recommender systems: introduction and challenges. In: Recommender Systems Handbook. Berlin: Springer, 2015

2 Yeh D Y, Cheng C H. Recommendation system for popular tourist attractions in Taiwan using Delphi panel and repertory grid techniques. Tourism Manage, 2015, 46: 164–176

3 Pazzani M J, Billsus D. Content-based recommendation systems. In: Proceedings of the Adaptive Web, 2007. 325–341

4 Herlocker J L, Konstan J A, Borchers A, et al. An algorithmic framework for performing collaborative filtering. In: Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 1999

5 Elkahky A, Song Y, He X D. A multi-view deep learning approach for cross domain user modeling in recommendation systems. In: Proceedings of the 24th International Conference on World Wide Web, 2015

6 Shwartz-Ziv R, Tishby N. Opening the black box of deep neural networks via information. 2017. ArXiv:1703.00810

7 Lin Y K, Liu Z Y, Sun M S, et al. Learning entity and relation embeddings for knowledge graph completion. In: Proceedings of the 29th AAAI Conference on Artificial Intelligence, 2015

8 Wang H W, Zhang F Z, Wang J L, et al. Ripplenet: propagating user preferences on the knowledge graph for recommender systems. In: Proceedings of the 27th ACM International Conference on Information and Knowledge Management, 2018

9 Petronis K R. A different kind of contextual effect: geographical clustering of cocaine incidence in the USA. J Epidemiol Commun Health, 2003, 57: 893–900

10 Soja E W. Regions in context: spatiality, periodicity, and the historical geography of the regional question. Environ Plann D, 1985, 3: 175–190

11 Mathias M, Zhou F, Torres-Moreno J M, et al. Personalized sightseeing tours: a model for visits in art museums. Int J Geogr Inf Sci, 2017, 31: 591–616

12 Gras B, Brun A, Boyer A. Identifying grey sheep users in collaborative filtering: a distribution-based technique. In: Proceedings of Conference on User Modeling Adaptation and Personalization, 2016

13 Wang R X, Fu B, Fu G, et al. Deep & cross network for ad click predictions. In: Proceedings of Knowledge Discovery and Data Mining, 2017

14 Guo H F, Tang R M, Ye Y M, et al. DeepFM: a factorization-machine based neural network for CTR prediction. 2017. ArXiv:1703.04247

15 Choi C, Cho M, Choi J, et al. Travel ontology for intelligent recommendation system. In: Proceedings of the 3rd Asia International Conference on Modelling & Simulation, 2009

16 Daramola J O, Adigun M O, Ayo C K. Building an ontology-based framework for tourism recommendation services. In: Proceedings of Information Communication Technologies in Tourism, 2009. 135–147

17 Lu C, Laublet P, Stankovic M. Travel attractions recommendation with knowledge graphs. In: Proceedings of Knowledge Acquisition, Modeling and Management, 2016

18 Sun Y Z, Han J W, Yan X F, et al, Pathsim: meta path-based top-k similarity search in heterogeneous information networks. In: Proceedings of Very Large Data Bases, 2011. 992–1003

19 Yu X, Ren X, Sun Y Z, et al. Personalized entity recommendation: a heterogeneous information network approach. In: Proceedings of the 7th ACM International Conference on Web Search and Data Mining, 2014

20 Wang Q, Mao Z D, Wang B, et al. Knowledge graph embedding: a survey of approaches and applications. IEEE Trans Knowl Data Eng, 2017, 29: 2724–2743

21 Wang H W, Zhang F Z, Zhao M, et al. Multi-task feature learning for knowledge graph enhanced recommendation.

In: Proceedings of World Wide Web Conference, 2019
22 Brockmann D, Hufnagel L, Geisel T. The scaling laws of human travel. Nature, 2006, 439: 462–465
23 Cleveland R B, Cleveland W S, Mcrae J E, et al. STL: a seasonal-trend decomposition procedure based on LOESS. J Off Stat, 1990, 6: 3–33
24 Xu B, Xu Y, Liang J Q, et al. CN-DBpedia: a never-ending chinese knowledge extraction system. In: Proceedings of International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems, 2017

# An interpretable attraction recommendation method based on knowledge graph

Jialiang GAO[1,2], Peiyuan QIU[3], Li YU[4], Zongcai HUANG[1,2] & Feng LU[1,2,5*]

1. *University of Chinese Academy of Sciences, Beijing* 100049*, China*;
2. *State Key Laboratory of Resource and Environmental Information Systems, Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing* 100101*, China*;
3. *College of Surveying and Geo-Informatics, Shandong Jianzhu University, Jinan* 250101*, China*;
4. *National Science Library, Chinese Academy of Sciences, Beijing* 100190*, China*;
5. *Jiangsu Collaborative Innovation Center for the Development and Utilization of Geographic Information Resources, Nanjing* 210023*, China*
* Corresponding author. E-mail: luf@lreis.ac.cn

**Abstract** The attraction recommendation systems not only filter out overwhelming irrelevant information for visitors but also identify potential customers for service providers. However, the current attraction recommendation methods such as content-based methods, collaborative filtering, or deep learning-based methods are either inaccurate due to data sparsity, or lack of interpretability, which results in the users' suspicion on the recommendation results. To address the limitations of the current methods, we introduce a novel framework for preference propagation on knowledge graphs (KGs), which utilizes lots of parameters to capture the abundant semantics of existing KGs more comprehensively, and meanwhile explains the results through reasoning the link paths from user's history to candidates on KGs. With a multi-view spatiotemporal analysis on real-world travel data, we investigate the geographical characteristics of human tour activities and build a tourism-oriented KG based on open web resources. Then, we propose a KG-aware attraction recommendation method named Geo-RippleNet and implement it with extensive experiments on large-scale datasets. It is argued that the framework for preference propagation on KGs not only absorb rich semantic information to achieve substantial performance gains in the attraction recommendation scenario but also enhance the interpretability of recommendation results with the support of abundant relational knowledge. Moreover, incorporating the spatiotemporal characteristics of human tour activities into the framework for preference propagation further makes the recommendation performance more aligned with the potential interests of visitors.

**Keywords** tourism knowledge graph, attraction recommendation system, interpretability, recommendation system, tourism management

**Jialiang GAO** was born in 1994. Currently, he is a Ph.D. candidate at Institute of Geographic Sciences and Natural Resources Research, CAS. His research interests include knowledge graph and tourism management.

**Lu FENG** was born in 1970. He received his Ph.D. degree from Institute of Remote Sensing Applications, Chinese Academy of Sciences, in 1999. He is currently a professor at Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences. His research interests include geo-spatial data mining, big spatial data analysis, and computational transportation science.

**Peiyuan QIU** was born in 1986. He received his Ph.D. degree from the University of Chinese Academy of Sciences, Beijing, in 2016. Currently, he is a lecturer at Shandong Jianzhu University. His research interests include geographic information extraction and geographic knowledge graph.