Hans Journal of Data Mining 数据挖掘, 2020, 10(1), 56-67
Published Online January 2020 in Hans. http://www.hanspub.org/journal/hjdm
https://doi.org/10.12677/hjdm.2020.101006

Hans汉斯

Construction of Tourism Scene Knowledge Graph Based on the Content of Travelogs

Xiaohong Zhang*, Biao Ma

Department of Management, Donghua University, Shanghai
Email: *hello_smallred@163.com

Received: Dec. 18th, 2019; accepted: Jan. 2nd, 2020; published: Jan. 9th, 2020

Abstract

The traditional knowledge representation based on the concept of tourism ontology, which is focusing on the description of static features and lacks dynamic features. If some dynamic information hidden in the online UGC is fully utilized, such as the inheritance relationship of tourism events, group tourism Features such as emotions, can further enhance the interpretability of the tourism knowledge graph, thereby providing a more inferred knowledge graph for intelligent applications of experiential tourism. Therefore, this paper proposes a tourism scene knowledge graph (TSKG) representation model based on Bayesian network. The model uses tourism events as unit entity to obtain people's travel emotional information from the travelogs. The sentiment difference and transition probability between destinations is used as the dynamic attribute of edges to enrich the information of edges in the graph. The experimental result shows that the LTTE recommendation model based on the TSKG performs significantly better than the benchmark algorithm TF-IDF in the tourism destination recommendation experiment.

Keywords

Knowledge Graph of Travel Scene, Bayesian Network, Tourism Emotion, Travelog Mining

基于游记文本内容的旅游场景知识图谱的构建

张小红*，马　彪

东华大学管理科学与工程专业，上海
Email: *hello_smallred@163.com

收稿日期：2019年12月18日；录用日期：2020年1月2日；发布日期：2020年1月9日

摘　要

[目的/意义]：传统以旅游本体概念为基础的知识表示形式，侧重对静态特征的描述，缺乏对动态特征的

*通讯作者。

文章引用：张小红，马彪. 基于游记文本内容的旅游场景知识图谱的构建[J]. 数据挖掘, 2020, 10(1): 56-67.
DOI: 10.12677/hjdm.2020.101006

捕捉，如若考虑融入在线UGC中的一些动态信息，如：旅游活动事件发生的顺承关系，群体旅游情绪等特征，可进一步增强旅游知识图谱的可解释性，从而为当今体验式旅游的智能应用提供更加具有可推理性的知识库。[方法/过程]：因此，本文提出一种基于贝叶斯网络的旅游场景知识图谱表示模型，该模型以旅游活动事件作为单位实体节点，从游记文本中获取人们的旅游情感信息扩充实体的动态属性，并以人们在目的地之间的转移概率和情感差值作为边的动态属性，丰富图谱中边的信息，从而提升知识图谱的可解释性。[结论]：最终实验表明，基于本文所构建的旅游场景知识图谱设计的LTTE推荐模型，在旅游目的地推荐实验中表现的效果明显优于基准算法TF-IDF。

关键词

旅游场景知识图谱，贝叶斯网络，旅游情感，游记挖掘

1. 引言

在线旅游目的地信息服务是指针对出游用户在目的地的游玩需求所提供的有关目的地景区、餐饮、当地玩乐、购物、住宿、交通等方面的在线信息参考服务。相比于地方景区所提供的官方景点介绍，游客更倾向于浏览其他游客生成的在线游记。来自《2018 年中国景区旅游消费研究报告》的数据显示，59.7%的游客会在在线旅游网站进行门票预订，近九成用户选择移动端。可见基于用户真实体验的长篇游记以及评论是游客借助群体智慧进行旅行规划的重要信息来源。而对于旅游行业的各大 OTA 平台来说，提供个性化的智能旅行服务是当前的热点研究问题，在线游记中涵盖了各类用户对目的地的评价信息，成为企业直接了解游客真实感受的关键数据源。由此可见，针对游记文本内容的挖掘是获取旅游目的地相关知识的重要来源之一。并且将大众所积累的“经验知识”进行有效地表示与定期的更新，不仅能够成为游客进行旅游行程规划与决策的依据，也是企业实现智能化客服服务的动态知识库。

然而，鉴于文本内容本身语言表达的复杂性，目前针对游记本身知识提取方法的研究比较匮乏。仅有的研究主要从游记中挖掘主题，缺少对游记中旅游路线、旅游情感、旅游时间等知识的利用。并且现有针对游记文本挖掘的知识表示形式多采用图网络的方式，缺失语义的联系，不利于应用层面的推理。因此，本文将充分挖掘游记中的情感信息，将旅游路线的顺承关系作为知识关联的方式，构建旅游场景知识图谱。

2. 相关工作

2.1. 旅游领域知识图谱研究现状

国外旅游本体的研究起步较早并日趋成熟，目前为止，已经有许多研究机构尝试开发专门的旅游本体。相对经典且应用较广的旅游本体主要有：旅游开放联盟标准规格(OTA Specification)，旅游休闲词库(The Thesaurus on Tourism and Leisure Activities) [1]，旅行本体(Onto Tour) [2]，旅行目的地本体[3]。

国内旅游本体研究起步较晚，对旅游本体的研究主要是信息系统中的应用以及信息检索。冯欣[4]提出的旅游信息系统中包含了旅游信息本体、旅客本体以及语义 web 浏览器，本体参考 OTA 分类方案和

Google 网页目录以及旅游网站等。李艳[5]参考《综合电子政务词表》构建旅游政务系统中的本体。奚凡研究了旅游活动与推荐系统中构建了旅游情景本体模型[6]。杨青云提出通过构建旅游应用本体[7]，开发旅游信息服务平台。

从上述对旅游目的地本体构建的研究现状梳理中发现，目前旅游本体中的知识起初包含静态的信息，例如景点的地理位置、主要的景点、酒店的位置。后来为了支持智能应用服务的需求，加入动态的旅游信息，例如活动的举办时间、当地的天气，住宿的优惠信息等。但由于旅游本体本身大都是以静态概念及其关系为研究对象，加入动态信息丰富语义内容并不能从根本上解决语义推理困难的问题。因此，考虑从事件(场景)角度对本体进行建模，有利于解决旅游智能应用的智能化发展。

2.2. 游记文本挖掘研究现状

游客在线生成的海量 UGC 内容，是知识挖掘领域的研究重点。近十年，从游客生成的图片集中挖掘信息兴起一阵热潮，学者[8]通过处理大量带有地点标签的图片，如 Flicker 中的游客照，识别热门目的地，精选具有代表性的图片，直观反应目的地形象与特色。Xin Lu 等人[9]，从带有地理标签的游客照中挖掘用户旅游足迹，结合游记中关于目的地的特点描述，构建了一个交互式旅行规划系统“Photo2Trip”。

相比之下，针对游记文本挖掘的相关研究较少。Zhao Zhenbin 等人运[10]用内容分析法，对太白山相关网站论坛上的游记文本内容进行词频统计分析，研究太白山背包族的行为特征。吴恒等人[11]同样基于内容分析法，运用高频词聚类分析携程网中蜜月游的基本特征以及游客的选择行为特征。上述针对旅游游记的内容分析法研究，从统计学角度出发，关注于游客对目的地形象感知、以及游客的行为特征[12]，有助于识别目的地形象塑造与人们认知的偏差，促进旅游景区改善服务。但该方法通过简单的词频统计和高频词聚类等方法处理数据，会造成语义丢失或理解上的歧义。因此，有学者开始运用 LDA 主题建模方法提取游记文本中的目的地相关信息。胡乔楠等人[13]为弥补层级热门旅游景点推荐的不足，利用文本挖掘和地理名词提取技术，从八万余篇旅游游记中得到相关的地理名词和景点名词，然后基于具有层级关系的地理本体树实现热门旅游景点推荐。Rui Cai 等人[14]提出了一个针对游记文本挖掘的框架，运用 LT (Location Topic Model)模型，将游记中的主题分为“全局主题(Global topic)”和“局部主题(Local Topic)”，从而更好地识别游记中具有地点代表性的知识。

综上所述，仅有的针对游记挖掘研究主要从游记中挖掘主题，缺少对游记中旅游路线、旅游情感、旅游时间等知识的利用，并且未对挖掘出的知识进行合理的知识表示。

3. 旅游场景知识图谱的构建

3.1. 旅游场景知识图谱的概念模型

单纯运用基于本体的语义网知识表示法[15] [16]不能够完全满足旅游智能应用的知识需求，无法充分表达不确定性知识，所以本文将贝叶斯条件概[17]与传统语义网相结合，提出了旅游场景下基于贝叶斯概率的语义网模型。该模型能够将人类的先验知识和后验概率进行结合，可以克服语义网等模型仅能表达确定性知识的弱点。该模型主要思想是将贝叶斯条件概率与传统语义网相结合，在语义网的节点之间添加置信概率，能有效解决游客 POI 判断中不确定性知识表示的选择难、多义性和确定权重难等问题，增强计算机对知识的可理解性。

因此，本文提出旅游场景知识图谱 TSKG (Tourism Scene Knowledge Graph)概念模型，通过一个二元组对其进行描述，即 $TSKG = \langle N, R \rangle$ (3.1)，其中 N (Node)表示图谱中围绕某一 POI 所发生的旅游活动事件节点，R (Relation)表示图谱中节点之间的边的语义关系。

1) 旅游场景图谱中的实体节点 N 定义见如下公式：

$$N = \langle Nid, Where, What, How, Score, Text, Rid \rangle \tag{3.2}$$

旅游活动事件节点 N 通过一个七元组来描述，其中 Nid 表示节点的唯一标识符，Where 视为图谱的基本实体单元，即旅游目的地，What (旅游活动)，How (旅游体验)均为该节点的属性特征，分别表示发生该旅游活动事件的旅游活动及旅游感受。Score (情感评分)是根据 What (旅游活动)和 How (旅游感受)综合计算得到的对于该旅游目的地的综合情感评分。Text 是围绕这一 POI 描述旅游活动事件发生全过程的相应游记片段，Rid 是相连节点的边的编号。

2) 旅游场景图谱中的旅游活动事件节点关系 R 定义见如下公式：

$$R = \langle Rid, RSematic, RProbility, RScore \rangle \tag{3.3}$$

旅游活动事件节点间关系，通过一个三元组来描述，其中 Rid 表示边的唯一标识符，RSematic 表示边的语义关系(顺承、因果、属性)，RProbility 和 RScore 是边的两个属性，其中 RProbility 表示两个节点之间的转移概率，RScore 表示两个节点的情感差值。

综上，基于贝叶斯概率语义网模型进行知识表示的旅游场景知识图谱可以抽象成简单的三种基本单元：

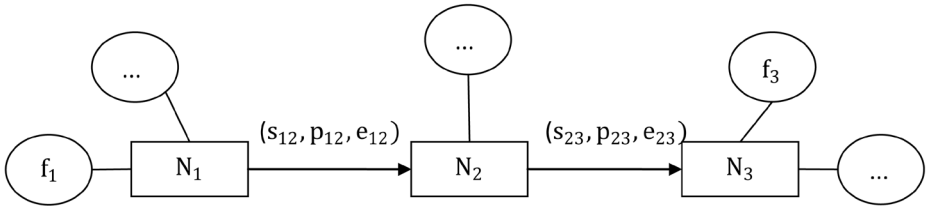

Figure 1. Knowledge unit in single travel scene
图 1. 单个旅游场景下的知识单元

图 1 为单个旅游场景下的知识单元，表示单个旅游场景下活动事件 N_1, N_2, N_3 之间的顺承关系走向为 $N_1 \rightarrow N_2 \rightarrow N_3$，每个旅游活动事件具有上述定义的属性特征 f_n，每条边具有上述定义的三个属性，分别用 s (语义关系)，p (转移概率)，e (情感评分)表示。在实际旅游活动路线，可能会产生闭环现象。如从 N_3 返回到 N_2 或 N_1，本文不考虑该闭环情况。一方面是因为通过实际数据研究发现，该情况十分稀少，不具有群体代表性。另一方面是因为一天行程结束后返回最初节点的情况，大多为返回酒店入住，不具有深入挖掘的价值。

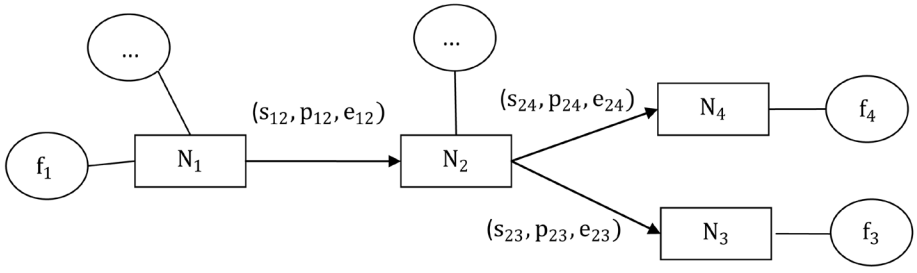

Figure 2. Knowledge unit in many travel scenes
图 2. 多个旅游场景下的知识单元

图 2 为多个旅游场景下的一种知识单元类型，即由两种旅游路线组合而成，$N_1 \rightarrow N_2 \rightarrow N_3 / N_4$。

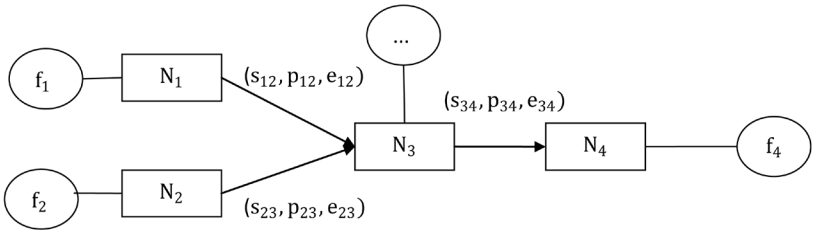

Figure 3. Knowledge unit in many travel scenes
图 3. 多个旅游场景下的知识单元

图 3 为多个旅游场景下的一种知识单元类型，即由两种旅游路线组合而成，$N_1 / N_2 \rightarrow N_3 \rightarrow N_4$。

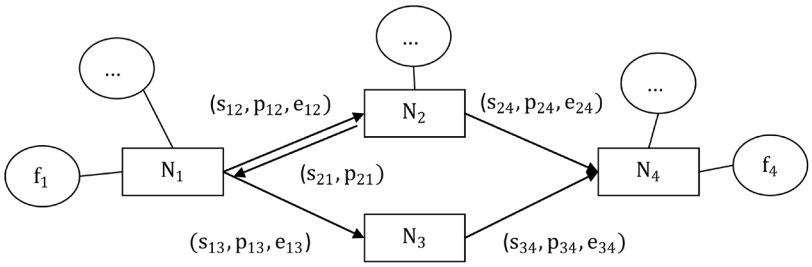

Figure 4. Knowledge unit in combined travel scene
图 4. 融合后的旅游场景知识图谱

图 4 为多个旅游场景融合之后的旅游场景知识图谱，由于多条旅游路线组合的缘故，从图中可见 N_1 与 N_2 之间可能存在双向箭头，但据上述分析，本文不考虑闭环情况，因此双向箭头一定不属于同一旅游场景。

3.2. 旅游场景知识图谱的构建

本文所构建的旅游场景知识图谱，其实体、实体属性值及边属性值均来源于群体贡献的在线游记。因此，本章将针对游记文本设计知识挖掘框架，实现对旅游目的地(实体节点)，旅游路线(实体间的顺承关系)，旅游目的地相关主题活动及情感(实体属性)，及旅游目的地之间转移概率(边属性)的抽取与计算，最终以旅游场景知识图谱的知识表示形式组织旅游活动事件及其关系。总体实现逻辑如图 5 所示：

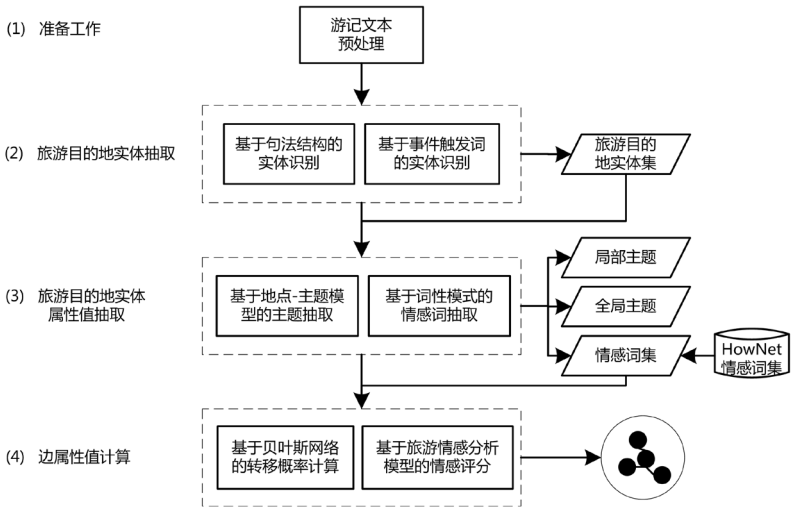

Figure 5. Construction logic of travel scene knowledge graph
图 5. 旅游场景知识图谱的构建逻辑图

1) 旅游目的地实体的抽取及判别

本文提取 OTA 平台中 TOP 级景点信息构成基本的旅游目的地词典 POI_dict，基于 POI_dict 词典精确匹配游记 l 处理后的每个词语 w，观察发现每个自然段可能提及多个 POI，除了作者在本段中实际描述的景点以外，还存在其他干扰地点。为解决该问题，本文主要从 POI 附近的触发词及上下文联系两个方面排除干扰，判定自然段实际描述的实体对象。

a) 基于触发词判别段落的实际描述对象

触发词作为发现 POI 的重要特征，可以用于进行 POI 判别。如例句 A，通过词典匹配将识别出“中环”和“海洋公园”两个地点，依据依存句法分析，可以得到关键触发词“到达”，从而判定该段接下来将会描述“到达”该词后面指示的地点“海洋公园”，而非“中环”。同样，例句 B 中依据依存句法分析，可以得到触发词“去”，其指示的旅游目的地为“尖沙咀星光大道”。

例句 A：“双层巴士装载客人的速度很快。八达通一刷，10.6 HKD。大概半个小时就从中环到达海洋公园了。”

例句 B：“去香港一定要去一趟尖沙咀星光大道。透过维多利亚港隔海欣赏对岸香港岛的璀璨灯光。”

b) 结合上下文过滤干扰实体

作者在书写游记的过程中，行文逻辑一般以发生旅游活动事件的先后顺序展开，描述完一个 POI 后转而切换到下一个 POI 进行描述，很少穿插论述。因此，每个自然段对 POI 的描述是上下文相关联的，若在连续自然段中均对 POI_1 进行描述，中间匹配出与上下多个连续自然段不符的 POI_2，则对这类 POI_2 进行过滤。并且，若一个段落中同时匹配出三个以上的旅游目的地实体，且实体之间的距离相邻近，则可以基本认定作者在罗列地点名词(如例句 C)，可能是在概述行程，也可以在回忆过往，因此对这类并非针对某一 POI 进行描述的段落予以剔除。

例句 C：从小就看港剧、港片长大，却是第一次去南丫岛、大澳、大屿山，电视剧中看过无数次的景色真实出现在眼前的感觉真实难以言喻。

综上，可以从游记文本中识别出旅游目的地实体，并判别自然段实际描述的 POI，从而依照实体出现段落次序形成旅游顺承路线。

2) 旅游目的地实体属性值的抽取

a) 旅游主题活动的抽取

本文将传统 LDA 模型的主题分为局部主题(Local Topic)和全局主题(Global Topic)两类，并将旅游目的地视为局部主题的概率分布。若游记集 L 所构成的词袋容量为 W，旅游目的地词集容量为 K，则游记集可以表示为一个 $L \times W$ 的[词语 - 文档]稀疏矩阵(Word-TravelLog matrix)，如图 6 最左侧所示，其中第 m 篇文章在 W 向量上的词语概率分布即为第 m 列；第 m 列第 x 行的元素表示条件概率(第 x 个词语|第 m 篇文章)。于是，[词语 - 文档]矩阵可以分解为 I 与 II 两部分。

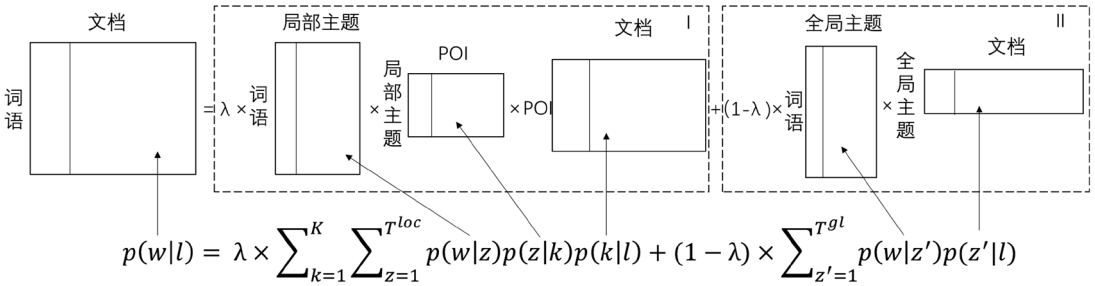

Figure 6. Graph of Location-topic model
图 6. 地点 - 主题概率模型建模示意图

从上图可以直观地看到一篇游记中每个词语的产生源于条件概率分布下做出的一系列随机选择，通过将传统 LDA 模型中[词语－主题]矩阵拆分为[词语－局部主题]和[词语－全局主题]两类，两类矩阵分别表示局部主题和全局主题在词语上的概率分布 $\left\{p\left(\text{词语}w\,|\,\text{主题}z\right)\right\}_{w=1}^{W}$，因而可以对每个词语所描述的主题类型进行区分，从而筛选出专门刻画旅游目的地特色的代表性知识，其中[局部主题－地点]矩阵则反映了地点与局部主题之间的概率关系 $\left\{p\left(\text{局部主题}z\,|\,\text{地点}l\right)\right\}_{z=1}^{T^{loc}}$。对于[地点－文档]矩阵的填充，本文依照上一节中的 POI 识别及判别规则对每篇游记实际描述的 POI 进行提取，并映射到相应的文章片段。

b) 旅游对象－情感词对的抽取

对于每篇游记文本 l，经过主题挖掘后，将包含主题项 T^{loc} 的句子作为主题句，然后利用词性模式识别主题句中的情感词，构成主题(旅游评价对象)－情感词组。用 T 表示主题词(topic word)，S 表示情感词(sentimental word)，主题情感词组用 T-S 表示。表 1 为反复实验所总结的词性模式，其中各字母代表含义为：n：名词、vn：动名词、v：动词、a：形容词、d：副词、CT：候选主题词、CS：候选情感/情绪词。

Table 1. Extract part-of-speech patterns of topic-emotional phrases
表 1. 抽取主题－情感词组的词性模式

序号	词性模式	输出	子句词性示例	评价对象－情感组示例
1	n + d + a	(CT = n，CA = a)	“过山车/n 非常/d 刺激/a”	(过山车/n，刺激/a)
2	n + d + v	(CT = n，CA = v)	“价格/n 很/d 划算/v”	(价格/n，划算/v)
3	n + d + n	(CT = n，CA = n)	“卫生/n 干净/n”	(卫生/n，干净/n)
4	n + d + vn	(CT = n，CA = vn)	“价格/n 实惠/vn”	(价格/n，实惠/vn)
5	vn + a	(CT = vn，CA = a)	“服务/vn 好/a”	(服务/vn，好/a)

3) 旅游目的地之间转移概率和情感差值的计算

a) 基于贝叶斯网络的转移概率的计算

设旅游活动事件 $N=\left\{N_1,N_2,\cdots,N_n\right\}$ 为随机变量，每个变量都只有两个取值，去——1/不去——0 (旅游目的地)。设样本数为 m，可以从样本集中得到旅游活动事件之间的关系，即样本集中有多少人去了 A 地后又去了 B 地，通过词频共现统计 $f\left(N_1\to N_2\right)$ 可以构建出 CPT 表，用于参数计算。表达式 $p\left(X_1=x_1,X_2=x_2,\cdots,X_n=x_n\right)$ 表示一个联合概率，即变量 $X_1,X_2,\cdots,X_n$ 的值分别是 $x_1,x_2,\cdots,x_n$ 时的概率。其一般形式为：

$$P\left(X_1,X_2,\cdots,X_n\right)=\prod_{i=1}^{n}p\left(X_1\,|\,X_i,X_2,\cdots,X_{i-1}\right) \tag{3.4}$$

b) 基于情感分析模型的情感差值计算

本文构建了一个六元组(公式 3.5)用于存储对情感极性产生明显影响的词性或标志，并制定了一系列规则用于 POI 情感值的计算。

$$Score_k=\lambda\sum_n E\left\langle\text{tword},\text{sword}\right\rangle\times\alpha\left(\text{nword}\right)\times\beta\left(\text{dword}\right)\times\gamma\left(\text{position}\right)\times\delta\left(\text{rword}\right)/m \tag{3.5}$$

其中，$Score_k$ 表示第 k 个 POI 的情感评价总分，m 指描述该 POI 的游记篇数，n 指一篇游记中的候选主题句数目，tword 指旅游评价对象，sword 指情感/情绪词，nword 指否定词，dword 指程度副词，position 指否定词和程度副词的相对位置，$\alpha,\beta,\gamma,\delta$ 指上述四个元素情感强度的影响系数，λ 指正负向情感的矫正系数。

规则 1：对于程度副词的权值，采用 HowNet(知网)程度副词文档进行处理。文档中根据程度副词的强烈程度，将其划分成五个倍数级别，对其分别给予 0.5~3 的倍数。

规则 2：对于否定词，当句子中出现否定词，将对其修饰的情感词取相反含义。而对于多重否定句式的归纳，奇数重否定表示否定义，偶数重否定表示肯定义。

规则 3：若句中不仅存在否定词，还存在程度副词时，需要考虑两者的相对位置。当否定词在程度副词之前，则整体情感强度降低 0.5，相反，当否定词位于程度副词之后，则整体情感强度增加 0.5。

规则 4：若主题句中出现关联词/转折词“但是”、“却”、“仍然”等，则句子整体情感倾向取决于关联词/转折词之后的子句。

规则 5：情感矫正系数为 3，当游记中正面情感超出负面情感的 3 倍时，才将总体情感倾向判定为正向。

基于上述规则计算每个 POI 的情感评分，POI 之间的情感差值计算公式如下：

$$\Delta E = Score_k - Score_{k-1} \tag{3.6}$$

其中 $Score_k$ 是 $Score_{k-1}$ 相邻的父节点， $Score_k$ 计算公式见(3.5)。

4. 基于旅游场景知识图谱的应用

4.1. 旅游目的地实体相似度度量

基于 LT 模型识别游记中旅游目的地的相关局部主题，然后可以利用 W 维的词语空间或者 T^{loc} 维的局部主题空间作为每个 POI 的代表性知识。基于 LT 模型训练的参数 φ_k，即地点 k 在局部主题上的多项分布，可用于 POI 在局部空间的知识分布表示。而 POI 在词语空间的分布，可以基于吉布斯样本，统计游记集中各个词语被分配到地点 k 的次数，推导出地点 k 在 W 个词语上的概率分布 $p(w|k)$ (见如下公式 4.1)，作为其在词语空间中的知识分布表示。

$$p(w|k) = \frac{n_k^w}{\sum_{w'=1}^{W} n_k^{w'}} \tag{4.1}$$

其中，n_k^w 指游记集中词语 w 出现在描述地点 k 的片段中的次数，$p(w|k)$ 指在所有描述地点 POI_k 的词语中，词语 w 的占比。

因而，两个 POI_k 和 POI_{k+1} 之间的相似度 $LocSim(POI_k, POI_{k+1})$ 可以通过计算它们在局部主题空间的距离导出，本文采用 JS 散度(Jensen-Shannon divergence)度量该相似度，公式如下：

$$LocSim(POI_k, POI_{k+1}) \overset{\text{def}}{=} \exp\left[-\tau \cdot D_{JS}\left(\varphi_{k_1}, \varphi_{k_2}\right)\right] \in (0,1] \tag{4.2}$$

$$D_{JS}\left(\varphi_{k_1}, \varphi_{k_2}\right) = \frac{1}{2} D_{KL}\left(\varphi_{k_1} \| \frac{\varphi_{k_1} + \varphi_{k_2}}{2}\right) + \frac{1}{2} D_{KL}\left(\varphi_{k_2} \| \frac{\varphi_{k_1} + \varphi_{k_2}}{2}\right) \geq 0 \tag{4.3}$$

$$D_{KL}\left(\varphi_{k_1} \| \varphi_{k_2}\right) = \sum_i \varphi_{k_1 i} \log \frac{\varphi_{k_1}}{\varphi_{k_2}} \geq 0 \tag{4.4}$$

其中，φ_{k_1} 和 φ_{k_2} 分别是地点 k_1 和 k_2 在局部主题上的多项分布，$D_{JS}\left(\varphi_{k_1}, \varphi_{k_2}\right)$ 表示两个概率分布 φ_{k_1} 和 φ_{k_2} 之间的 JS 散度，$D_{KL}\left(\varphi_{k_1} \| \varphi_{k_2}\right)$ 表示从 φ_{k_1} 和 φ_{k_2} 的 KL 散度，因子 $\tau > 0$ 用于控制由 JS 散度到地点相似度的映射。

基于旅游目的地相似性进行推荐的问题，一般指给定候选地点集 K 和目标地点 $k_{\varphi_{k_1}}$，通过计算每个候选地点 $k \in K$ 与目标地点 k_{φ_k} 的相似度，对候选地点进行降序排序，从而推荐相似度排名靠前的旅游目的地。本文将考虑结合旅游场景知识图谱中的边属性值，即旅游目的地间的转移概率 P (公式)和情感差值 ΔE (公式)，对候选地点集中各地点推荐值进行打分，具体计算公式如下：

$$Score(k)=\log LocSim\left(k,k_{\varphi_k}\right)+\alpha\cdot\log\Delta E\left(k,k_{\varphi_k}\right)+\beta\cdot\log P\left(k,k_{\varphi_k}\right) \tag{4.5}$$

其中，非负系数 α 和 β 是两个目的地之间转移概率和情感差值在推荐评价打分时的权重，当两个旅游目的地之间不存在顺承关系时，则视 β 为 0，推荐分值则依据相似度和情感评分决定。

4.2. 基于旅游场景知识图谱的推荐

基于旅游目的地相似性进行推荐的问题，一般指给定候选地点集 K 和目标地点 $k_{\varphi_{k_1}}$，通过计算每个候选地点 $k\in K$ 与目标地点 k_{φ_k} 的相似度，对候选地点进行降序排序，从而推荐相似度排名靠前的旅游目的地。本文将考虑结合旅游场景知识图谱中的边属性值，即旅游目的地间的转移概率 P (公式 3.4)和情感差值 ΔE (公式)，对候选地点集中各地点推荐值进行打分，具体计算公式如下：

$$Score(k)=\log LocSim\left(k,k_{\varphi_k}\right)+\alpha\cdot\log\Delta E\left(k,k_{\varphi_k}\right)+\beta\cdot P\left(k,k_{\varphi_k}\right) \tag{4.6}$$

其中，非负系数 α 和 β 是两个目的地之间转移概率和情感差值在推荐评价打分时的权重，当两个旅游目的地之间不存在顺承关系时，则视 β 为 0，推荐分值则依据相似度和情感评分决定。

5. 实验设计与结果分析

通过网络查询发现，携程旅行网是中国领先的在线旅行服务公司，目前占据中国在线旅游 50%以上市场份额，是绝对的市场领导者，因此笔者选定“携程旅行网”中“游记攻略”版块里的游记作为本研究的样本数据。利用网络爬虫工具采集发布时间在 2019 年 3 月 31 日前十年游记(共有 5341 余篇香港游记)作为本文的研究数据，采集的内容包括每篇游记的链接、文本内容、标题等。

实验流程主要包括三部分，首先，对游记文本进行预处理；然后，训练 TSKG 旅游场景知识图谱；最后，将 TSKG 应用于旅游目的地推荐，进行对比实验。

5.1. 实验语料预处理

编写爬虫代码从携程网获取游记后，对游记文本进行分词、去除停用词及低频词过滤等处理。另外，本文构建了两个词库，一个是根据携程旅游网及蚂蜂窝旅行网提供的香港旅游目的地地名，构建的香港景点词典，用于识别旅游目的地实体，共包括 340 个旅游目的地。另一个是基于游记中的高频形容词集合知网情感词典，构建的旅游情感词库，共包含 1545 个正面词汇，1045 个中性词汇以及 1240 个负面词汇。

5.2. 基于旅游场景知识图谱的推荐

1) 实验设计及评估指标

为对比验证基于旅游场景知识图谱的推荐方法有效性，本文将爬取的 5341 条数据分为 3341 训练集和 2000 测试集，利用训练集样本构建旅游场景知识图谱，然后基于测试集进行实验验证，将实际用户到达的旅游目的地作为基准，判定推荐列表前十的景点中是否包含用户实际前往的旅游目的地。本文采用推荐系统常用的评测指标召回率(Recall)及 Mean Reciprocal Rank (MRR)对基于旅游场景知识图谱的推荐方法的有效性进行评价。将本文提出的推荐算法(LTTE)与基于词语 TF-IDF 向量的推荐算法(TCS)进行对比。

TCS 方法：对于每个地点，将所有提及它的游记连接成一偏“大”文档，然后利用 TF-IDF 表示这些对应于地点的文档，最后将两个地点的相似度定义为二者对应文档的 TF-IDF 向量之间的余弦相似度。

a) 召回率(Recall)

召回率是评判推荐系统准确性的通用指标，对于每个预测结果而言，指标 $R(A_u)$ 计算用户实际到达的旅游目的地在推荐列表中的比例。

用户 u 的推荐召回率的计算公式如下：

$$R(A_u)=\frac{A_u \cap B_u}{B_u} \tag{5.1}$$

推荐系统整体召回率：

$$R(A)=\frac{1}{n}\sum_{u\in U} R(A_u) \tag{5.2}$$

其中，A_u 是用户u的推荐列表，B_u 是测试集中用户 u 实际前往的旅游目的地，n 是推荐总次数。

b) Mean Reciprocal Rank (MRR)

MRR 指标是在国际上对推荐算法进行评判的通用机制，用于评判推荐算法质量的高低。其评估思想是，若第一个推荐结果即与实际旅游目的地相符，则计分为 1，若第二个推荐结果是正确的则计分为 1/2，由此推导则第 N 个推荐结果与实际相符，则计分为 1/n，若推荐列表中不存在，则计分为 0，将所有分数进行加和即得到最终分数，计算公式如下：

$$\text{MRR}=\frac{\sum_{n=1}^{N}\frac{1}{rank_n}}{N} \tag{5.3}$$

其中，N 表示推荐总次数，$rank_n$ 表示实际旅游目的地存在于推荐列表中的排名，若不存在与列表中，则 $rank_n=0$。

基于 Recall 和 MRR 两个评估指标，对本文提出的基于旅游场景知识图谱的推荐算法(LTTE)与推荐算法(TCS)进行比较。另外，推荐列表中推荐数目的变化，也是影响推荐效果的重要因素，所以后续本文将改变推荐数目，分析其对推荐效果的影响。

2) 实验结果分析

a) 推荐数目对推荐效果的影响

基于旅游场景知识图谱进行推荐的实验中，推荐列表中推荐数目的多少会对实验结果产生影响，所以接下来将推荐数目分别设置为 1 到 11，依次计算 R 值和 MRR 值，得到结果如图 7 所示：

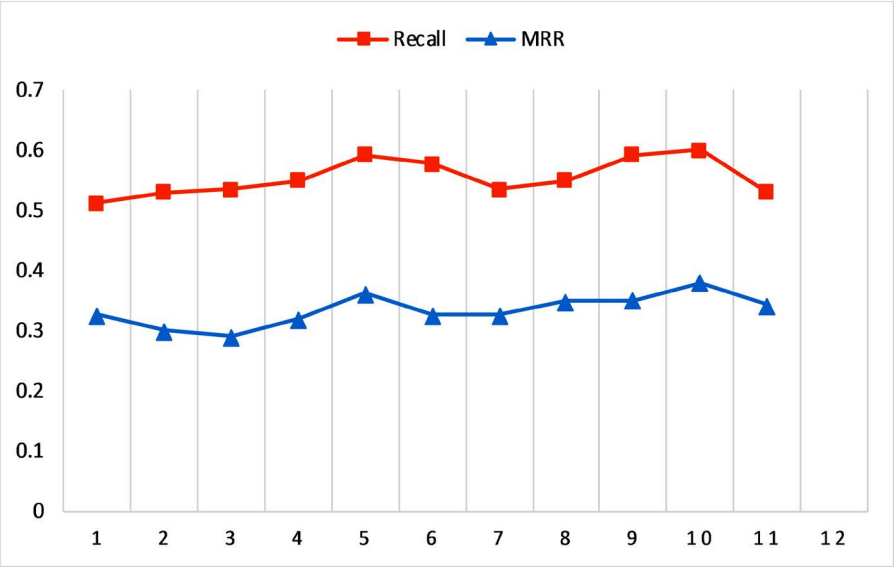

Figure 7. Recall and MRR value of different recommended list numbers
图 7. 不同推荐列表数目的 R 值与 MRR 值

从图 7 可看出，推荐算法的 Recall 值和 MRR 值总体波动幅度不大，当推荐列表数目分别为 5 和 10 的时候，会有一个向上波动的趋势，当推荐列表数目达到 10 时，召回率 Recall 和 MRR 值均达到最大，此时推荐效果最佳，召回率 Recall = 60.03%，MRR = 37.99%。因此，选择推荐数目为 10 能够较好体现本算法的有效性。

b)不同推荐算法的对比实验

基于上个实验分析推荐列表数目对算法的影响结果，下面实验设定推荐列表数目为 10，然后分别进行指标计算，得到结果如表 2：

Table 2. Evaluation index results of recommended algorithm
表 2. 推荐算法评估指标结果

推荐模型	Recall	MRR
基于 TCS 的推荐算法	0.325	0.256
基于 LT 模型的推荐算法	0.589	0.368
基于 LTTE 模型的推荐算法	0.667	0.434

从上表 2 的实验结果可以看出，基于 LT 模型和 LTTE 模型的推荐结果无论在 Recall 还是 MRR 值上都优于 TCS 算法，原因在于 TCS 计算 POI 之间的相似度时，未区分局部主题与全局主题，与 POI 无关的全局主题混入相似度度量的推荐中，从而对推荐结果产生了影响。而基于 LT 模型的推荐能够利用局部主题更好地刻画旅游目的地在特色主题活动，从而得到的结果与真实结果的相似度一致性较好。相比基于 LT 模型的推荐算法，基于 LTTE 模型的推荐结果略有提升，原因在于除了考虑旅游目的地之间的相似度之外，基于群体智慧，融入旅游场景知识图谱中的边属性值，即转移概率和情感差值，分别从旅游目的地的热度和满意度两方面修正推荐结果，可见本文提出的基于旅游场景知识图谱的 LTTE 模型推荐算法效果更佳。

6. 结论

本文提出基于贝叶斯网络的语义网知识表示模型，基于游记抽取旅游目的地、旅游路线及旅游情感知识，最后构建旅游场景知识图谱并基于该图谱进行旅游目的地的推荐应用。实验表明本文基于游记所构建的旅游场景知识图谱能够在推荐应用中达到较好的效果。未来可进一步考虑旅游路线出现闭环的情况，以及融入用户兴趣特征进行推荐应用。

参考文献

[1] http://www.world-tourism.org

[2] DERI, on Tour Ontology.

[3] Mili, H., Valtchev, P., Charif, Y., *et al*. (2011) E-Tourism Portal: A Case Study in Ontology-Driven Development. In: *E-Technologies*: *Transformation in a Connected World*, Springer, Berlin Heidelberg, 76-99. https://doi.org/10.1007/978-3-642-20862-1_6

[4] 冯欣, 王成良. 本体在旅游信息系统中的应用研究[J]. 计算机与现代化, 2010(3): 128-132.

[5] 李艳, 王重英, 屈正庚. 基于主题词表的旅游政务系统本体构建研究[J]. 信息技术, 2015(3): 53-56.

[6] 奚凡. 基于情景感知的自适应旅游活动与推荐系统研究[D]: [硕士学位论文]. 上海: 东华大学, 2013.

[7] 杨青云, 尹鹏飞. 基于语义网的旅游信息服务平台研究[J]. 邵阳学院学报(自然科学版), 2014, 11(2): 19-24.

[8] Yuan, H., Qian, Y., Yang, R. and Ren, M. (2014) Human Mobility Discovering and Movement Intention Detection with GPS Trajectories. *Decision Support Systems*, **63**, 39-51. https://doi.org/10.1016/j.dss.2013.09.010

[9] Lu, X., Wang, C., Yang, J.M., *et al*. (2010) Photo2Trip: Generating Travel Routes from Geo-Tagged Photos for Trip Planning. In: *ACM International Conference on Multimedia*, ACM, New York, 143-152. https://doi.org/10.1145/1873951.1873972

[10] 赵振斌, 党娇. 基于网络文本内容分析的太白山背包旅游行为研究[J]. 人文地理, 2011, 26(1): 134-139.

[11] 吴恒, 陈燕翎. 基于 UGC 文本挖掘的游客目的地选择信息研究——以携程蜜月游记为例[J]. 情报科学, 2017, 35(1): 101-105.

[12] Banyai, M.T.D. (2012) Evaluating Research Methods on Travel Blogs. *Journal of Travel Research*, **51**, 267-277. https://doi.org/10.1177/0047287511410323

[13] 胡乔楠. 基于旅游文记的旅游景点推荐及行程路线规划系统[D]: [硕士学位论文]. 杭州: 浙江大学, 2015.

[14] Hao, Q., Cai, R., Wang, X.J., *et al*. (2009) Generating Location Overviews with Images and Tags by Mining User-Generated Travelogues. *Proceedings of the* 17*th International Conference on Multimedia* 2009, Vancouver, British Columbia, Canada, October 19-24 2009, 801-804.

[15] 梁柱, 曾绍玮. 知识表示技术研究[J]. 科学咨询(决策管理), 2010(1): 52.

[16] 廖开际, 叶东海, 闫健峻, 等. 基于加权语义网的专家知识发现及表示方法[J]. 情报学报, 2012, 31(1): 60-64.

[17] 陈祖国, 李勇刚, 卢明, 陈超洋, 刘端. 基于贝叶斯概率语义网的铝电解槽况知识表示模型与约简方法[J/OL]. 控制与决策, 1-16.