Computer Science and Application 计算机科学与应用, 2021, 11(8), 2126-2136
Published Online August 2021 in Hans. http://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2021.118218

Hans汉斯

基于主题分类的旅游路线推荐规划模型

——以北京市为例

韩天祎，白千雪，李霈雯

中央民族大学理学院，北京
Email: muchty@163.com

收稿日期：2021年7月23日；录用日期：2021年8月19日；发布日期：2021年8月26日

摘　要

随着经济的发展，旅游逐渐成为人们生活的刚需，但计划行程、旅途的疲惫常常牵绊住人们外出的步伐。因此，本文基于北京市景点的文本评论运用LDA模型、K均值聚类进行主题提取、运用TF-IDF值进行评价打分为用户推荐最适宜的景点，节省了用户阅读攻略、规划行程的时间。不同于以往的数据分析，文本评论可以更直接反映用户的想法、更接近实际。除此之外，对于被选出来的景点，通过转化为旅行商问题，运用运筹学的蚁群算法为用户合理规划路线，减少步行时间以及交通时间。

关键词

文本挖掘，LDA主题模型，TF-IDF，K均值聚类，蚁群算法

Tourism Route Recommendation and Planning Model Based on Topic Classification

—Taking Beijing as an Example

Tianyi Han, Qianxue Bai, Peiwen Li

College of Science, Minzu University of China, Beijing
Email: muchty@163.com

Received: Jul. 23rd, 2021; accepted: Aug. 19th, 2021; published: Aug. 26th, 2021

文章引用: 韩天祎, 白千雪, 李霈雯. 基于主题分类的旅游路线推荐规划模型[J]. 计算机科学与应用, 2021, 11(8): 2126-2136. DOI: 10.12677/csa.2021.118218

Abstract

With the development of the economy, travel has gradually become the need of people's life, but the fatigue of planning trips and the exhaustion of the journey have hampered the pace of people going out. Therefore, based on the text review of scenic spots in Beijing, this paper uses LDA model, K-means clustering to extract topics, and TF-IDF value to evaluate the most suitable scenic spots, recommend for users, in order to save the user's time of reading strategy and planning the trip time. Unlike previous data analysis, text comments can reflect users' ideas more directly and be closer to reality. In addition, for the selected scenic spots, by transforming into a travelling salesman problem, the ant colony algorithm of operations research is used to plan the route reasonably for users to reduce walking time and traffic time.

Keywords

Text Mining, LDA Theme Model, TF-IDF, K-Means Clustering, Ant Colony Algorithm

1. 引言

随着旅游业的快速发展，旅游逐渐成为推动我国经济发展的主要动力，同时随着人们生活水平的逐渐提高，假期出游成为人们放松身心和缓解工作疲劳的主要方式。然而在旅游过程中，旅游路线的规划一直是令人头疼的问题，需要考虑时间、费用、偏好等等因素，同时还要在网上查阅许多的资料。面对这一问题，有许多的学者都进行探究，通常将旅游路线规划问题简记为 TSP 问题[1] [2]。张宇菲，彭旭等人[1]利用 0-1 数学规划模型，建立以总旅游时间最短为目标函数的模型，对 201 个国家 5A 级景点为目的地，进行探究得到结果；同时，还利用 TOPSIS 模型，对各景点的决策指标进行赋权，对前文的模型进行改进，并得到优化后的结果。刘忠花，李宪印等人[3]同样是利用 201 个国家 5A 景点为旅游目的地，将 TSP 问题划分为 3 个阶段，利用遗传算法进行路线规划和调整新的赋权图，得到路线结果。徐锋，杜军平[4]利用蚁群算法进行旅游路线规划，还提出了相关的改进算法——偶遇算法，提高了蚁群算法中的蚂蚁一次周游的质量。在实际应用中，考虑了现实景区载荷的均衡性。

本文基于以上学者的研究思路和结果，提出一种基于主题分类的旅游路线推荐规划模型，具有创新性，并以北京市为例阐述该模型的建立过程。首先，运用熵权法筛选得到 52 个最受欢迎的景点，并爬取其文本评论数据。然后，通过 LDA 主题模型和 K-means 聚类方法确定各景点对应的可供用户选择的标签。接着，运用 TF-IDF 值确定同一个标签下各景点的推荐顺序，并综合用户自定义的标签权重，确定多标签下景点的推荐顺序。最后，设定“多日游”和“一日游”场景，分别用蚁群算法解决路线规划的 TSP 问题。

2. 数据的选取与文本预处理

2.1. 景点的选取——熵权法

对携程网[1]公布的北京景点名单进行爬取，得到 5491 个景点的评分及评论数。对评论数小于 400 的

[1]携程网北京景点 https://you.ctrip.com/sight/beijing1.html。

景点以及大景点下的小景点(例如故宫里的太和殿)进行剔除。通过高德地图 APP，得到截至 2021 年 4 月 1 日的高德用户近一个月导航去过的游览数据。

由于景点评分的差异性较小、部分景点很受欢迎但其评论数较少，近一个月内导航去过数据可以反映潮流热门的情况，考虑将这三者结合起来得出“受欢迎度”综合得分，将评分较低的景点进行剔除。熵权法可以通过使用信息论中信息熵这个工具，来计算权重，给出综合得分。故运用熵权法计算三个数据的权重，其具体步骤如下：

步骤一：对 n 个样本，m 个指标，则 x_{ij} 为第 i 个样本的第 j 个指标的数值。

$$\left(i=1,2,\cdots,n;\ j=1,2,\cdots,m\right)$$

步骤二：对指标进行归一化处理，归一化后的数据 x'_{ij} 仍记为 x_{ij}。

正向指标：$x'_{ij}=\dfrac{x_{ij}-\min\left\{x_{1j},\cdots,x_{nj}\right\}}{\max\left\{x_{1j},\cdots,x_{nj}\right\}-\min\left\{x_{1j},\cdots,x_{nj}\right\}}$

负向指标：$x'_{ij}=\dfrac{\max\left\{x_{1j},\cdots,x_{nj}\right\}-x_{ij}}{\max\left\{x_{1j},\cdots,x_{nj}\right\}-\min\left\{x_{1j},\cdots,x_{nj}\right\}}$

步骤三：计算第 j 项指标下第 i 个样本值占该指标的比重：

$$p_{ij}=\frac{x_{ij}}{\sum_{i=1}^{n}x_{ij}}$$

步骤四：计算第 j 项指标的熵值

$$e_j=-k\cdot\sum_{i=1}^{n}p_{ij}\cdot\ln\left(p_{ij}\right)，\text{其中 } k=\frac{1}{\ln(n)}>0，\text{满足 } e_j\geq 0$$

步骤五：计算信息熵冗余度(差异)：$d_j=1-e_j$

步骤六：计算各项指标权重：$w_i=\dfrac{d_j}{\sum_{j=1}^{m}w_j\cdot x_{ij}}$

步骤七：加权计算得到综合评分

$$s_i=\sum_{j=1}^{m}w_j\cdot x_{ij},i=1,\cdots,n$$

由熵权法计算结果得：

综合得分 $=0.3208*$ 高德近一个月人次 $+0.0127*$ 携程评分 $+0.6665*$ 评论数

综合得分最高为 100 分，最低为 0 分，分数越高说明该景点越受欢迎。经过筛选得到 52 个景点，其评分结果的热力图[2]展示见图 1。

2.2. 景点文本数据的选取与预处理

首先，通过后羿采集器对 52 个景点在携程网、美团网、穷游网截至到 2021 年 4 月 14 日的评论数据进行爬取，得到 88,486 条数据。然后用 python 对数据进行去重、去除无效评论的处理，并去除评论数据中的特殊符号与表情等。最后，运用 python 的 jieba 程序包进行停用词处理及中文分词，为后续的建立模型做准备。

[2] 本文地图均来自于高德开放平台 https://lbs.amap.com/tools/picker。

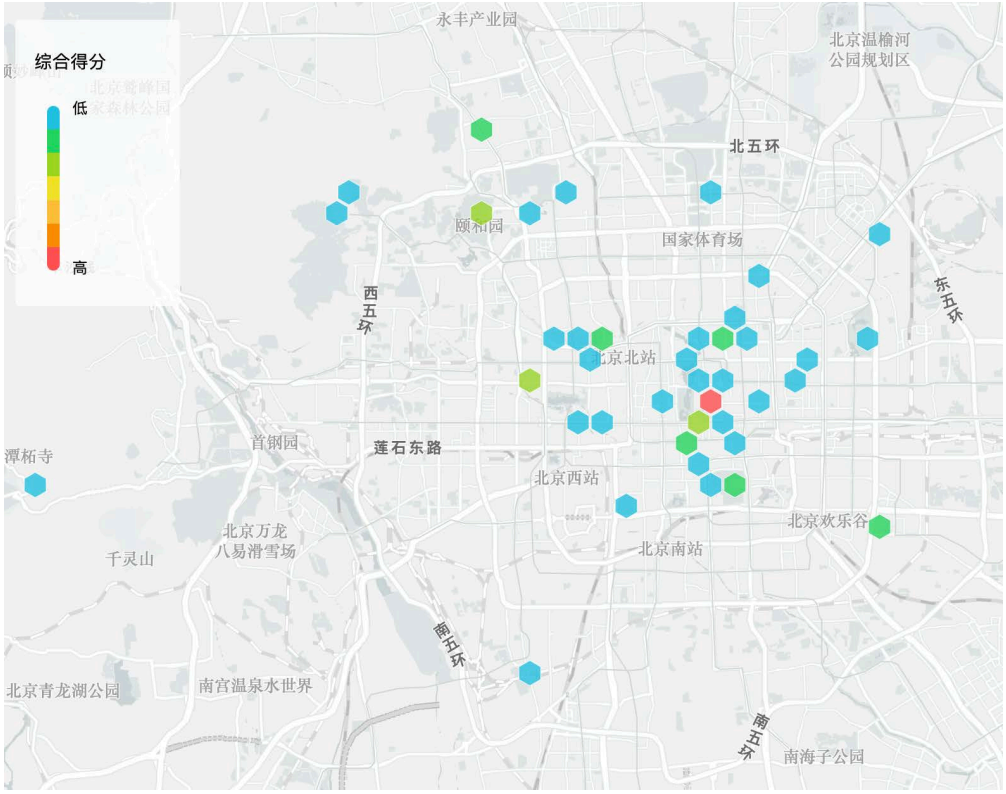

Figure 1. Heat map of attractions in Beijing urban area
图 1. 北京城区景点得分热力图

3. 景点的主题词分析

在该部分主要通过文本挖掘的方法得出可供用户选择的标签。首先，通过 LDA 主题模型提取各景点的分主题及其主题词；然后，通过 K-means 聚类方法根据主题词将分主题聚为六类，并通过总结每个类别的景点特点确定景点的标签。

3.1. 景点的主题划分——LDA 模型

LDA [5]是一种文档主题生成模型，也是一种非监督的机器学习技术，常称为一个三层贝叶斯概率模型，包含词、主题和文档三层结构。它认为一篇文章的每个词都是通过“以一定概率选择了某个主题，并从这个主题中以一定概率选择某个词语”这样一个过程得到，其中文档到主题服从多项式分布，主题到词服从多项式分布。

步骤一：基于各景点的中文分词，运用 LDA 模型，提取各景区的三个主题和十个关键词。

步骤二：通过逐一分析各景点的主题词提取结果，修改逐个景点的主题数，但不修改主题词数目，得到更符合实际的各景点分主题。其中北海公园和朝阳公园的主题划分结果见表 1：

Table 1. Topic division results (Partial)
表 1. 主题划分结果(部分)

主题一	主题二	主题一	主题二
白塔	休闲	建筑	项目
皇家	白塔	休闲	很大

Continued

园林	双桨	很大	玩
九龙壁	荡起	设施	游乐
建筑	开心	建筑	孩子
游玩	玩	城市公园	游玩
琼岛	拍	面积	设施
完整	荷花	拍照	娱乐
划船	小船	娱乐	小朋友
双桨	女儿	游玩	面积

3.2. 主题词的提取——K-Means 文本聚类

步骤一：根据从景点提取出来的分主题的主题词，运用 R 语言的 tm 包生成词频权重矩阵，再进行降维处理得到标准数值矩阵

步骤二：通过 R 语言的 k-means 函数，运用无监督的 K 均值聚类法[6]将分主题进行聚类。

1) 适当选择 c 个类的初始中心；

2) 在第 k 次迭代中，对任意一个样本，根据词频求其到 c 各中心的欧式距离，将该样本归到距离最短的中心所在的类；

3) 利用均值等方法更新该类的中心值；

4) 对于所有的 c 个聚类中心，如果利用 2) 3)的迭代法更新后，值保持不变，则迭代结束，否则继续迭代。

步骤三：通过聚类分析，将前面某景点分主题进行了合并，得到 24 个景点具有两个主题，28 个景点仅有一个主题。通过生活经验，确定聚为六类符合客观实际，并通过人为微调以及根据景点名称的聚类结果，进一步地得到景点标签，并命名为休闲－胡同、休闲－公园、必去－新景点、必去－名胜古迹、亲子、博物馆类。部分聚类结果展示见表 2：

Table 2. Top 7 attractions under each tag (In random order)
表 2. 各标签下前 7 个景点(顺序随机)

休闲－胡同	休闲－公园	必去－新景点	必去－名胜古迹	亲子	博物馆类
大栅栏	798 艺术区	玉渊潭	天安门广场	北京动物园	自然博物馆
东交民巷	北海公园	南锣鼓巷	八达岭长城	北京植物园	中国科学技术馆
南锣鼓巷	北京大学	古北水镇	故宫博物院	自然博物馆	北京天文馆
前门大街	北京欢乐谷	香山公园	北海公园	北京欢乐谷	军事博物馆
什刹海	北京植物园	鸟巢水立方	天坛公园	军事博物馆	国家图书馆
五道营胡同	朝阳公园	国家大剧院	恭王府	朝阳公园	国家博物馆
烟袋斜街	地坛	三里屯	圆明园	北京天文馆	大观园

4. 景点的推荐

当用户选择完满意的标签后，存在着两个问题：同一个标签下各景点的推荐顺序问题以及多标签下景点的推荐顺序问题。

4.1. 分类别的景点排名

在前面选择景点的分析中，用熵权法计算了景点的“受欢迎度”，但并没有考虑到景点与标签的贴合度问题。即在选择景点时很多景点虽然被分为一类，但其主题与该类别标签的贴合度并不一定是等权重的。通过查阅文献，TF-IDF 值的大小能够体现它在文本集合中的某一个文档里的重要性。故建立模型如下：

步骤一：基于中文分词数据，运用 ROSTCM6 计算各景点各中文分词的 TF-IDF 值。

$$\text{词频 TF} = \text{某个词在文章中的出现次数}$$

$$\text{逆文本频率指数 IDF} = \lg\frac{\text{语料库的文档总数}}{\text{包含改词的文档数}+1}$$

$$\text{TF-IDF} = \text{TF} * \text{IDF}$$

其中词频可以反映出景点的受欢迎度，IDF 可以体现出关键词的特征，故 TF-IDF 值可以作为某标签下景点的排名依据。

步骤二：对于某一个标签而言，选择景点与该标签主题词相关的 TF-IDF 值最大的一项作为该景点在该标签下的 TF-IDF 值。例如大栅栏，与“名胜古迹”这一标签的值最高的为“老字号”，与“胡同”最贴合的是“逛逛”，从而得出大栅栏在这两项标签下的 TF-IDF 值。

步骤三：对于某一个标签通过各景点的 TF-IDF 值降序排序的排名即为推荐顺序。其中名胜古迹类别情况见表 3：

Table 3. The ranking under the places of interest tag
表 3. 名胜古迹标签下的排名

名胜古迹	TF-IDF 值	排名	名胜古迹	TF-IDF 值	排名
故宫博物院	1664	1	景山公园	437	10
八达岭长城	1083	2	前门大街	408	11
明十三陵	934	3	什刹海	370	12
北海公园	928	4	雍和宫	350	13
颐和园	819	5	大栅栏	299	14
天坛公园	575	6	王府井步行街	295	15
恭王府	537	7	东交民巷	169	16
圆明园	468	8	钟鼓楼	124	17
天安门广场	451	9	孔庙国子监	218	18

4.2. 多标签下的景点推荐

由于在多标签下，用户基于自己的兴趣选择标签，但选择的顺序不同，其重要性也不同，因此，考虑让用户自定义所选择第 i 个标签的权重 w_i，$i=1,2,\cdots,6$，其中未被选中的标签其权重为 0；第 i 个标签下第 j 个景点的 TF-IDF 值记为 t_{ij}，$j=1,2,\cdots,52$，若第 j 个景点没有出现在第 i 个标签下，记其 $t_{ij}=0$，则经过加权计算，在 n 个标签下各景点的得分为 $S_k=\sum_{i=1}^{6} w_i \cdot t_{ik}$，$k=1,2,\cdots,52$。

- 实例展示：用户选择的标签为名胜古迹、亲子，其权重分别为 0.7 和 0.3。根据上述公式计算可得打分排名结果见图 2：

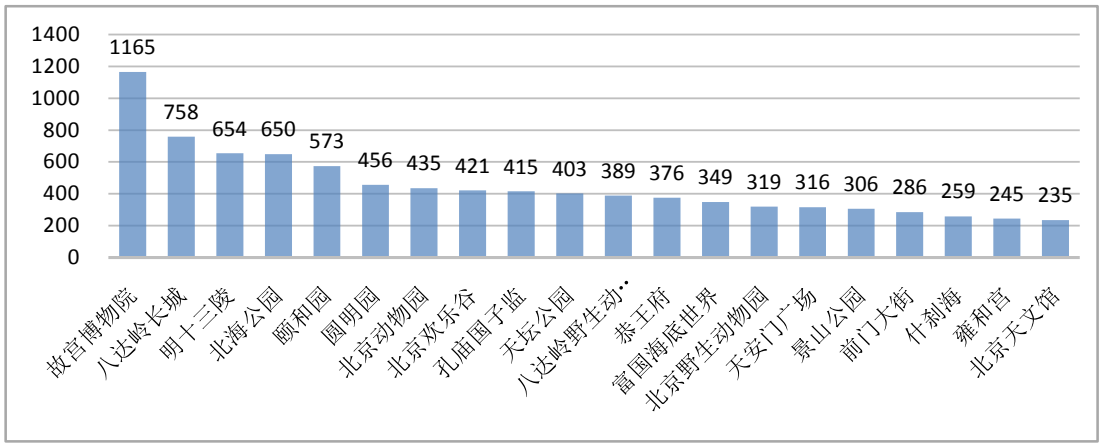

Figure 2. The results of attractions recommendation
图 2. 景点推荐顺序结果

通过推荐顺序结果可知，该结果较符合客观常识，由于天安门广场的数据在网页上并没有显示，存在天安门广场排名不精准的情况。

5. 旅游路线推荐

在选出最适合用户的景点群后，考虑为用户设计最优的游览路线，该部分通过设定两种不同的场景进行展示。

5.1. 多日游总体路线规划(大景区内的路线规划)

- 情境设定

不考虑机场、火车站、酒店、餐厅的位置，不考虑旅行天数限制，仅考虑游完十个景点的最短路径路线，游客选择的标签为名胜古迹、亲子。主要为了展示多个标签下的景点选择问题以及景点确定后的路线规划。

- 模型建立

步骤一：得到推荐的景点

根据 4.2 的结果，被推荐的景点依次为故宫博物院、八达岭长城、明十三陵、北海公园、颐和园、圆明园、北京动物园、北京欢乐谷、孔庙国子监、天坛公园。

步骤二：获取景点的经纬度数据并计算距离矩阵

首先，运用高德地图的坐标拾取器[3]获取景点的经纬度。设第 i 个景点与第 j 个景点的经纬度坐标分别为 (x_i, y_i) 和 (x_j, y_j)，则这两点的距离为：

$$d_{ij} = R \cdot \arccos\left[\cos(x_1 - x_2) \cdot \cos y_1 \cdot \cos y_2 + \sin y_1 \cdot \sin y_2\right]$$

其中 R 为地球半径，从而得到距离矩阵 D，$(d_{ij})_{10\times10}$。

步骤三：蚁群算法得出 TSP 问题最优路线规划

蚁群算法[7] (ACA)是对自然界蚂蚁的寻径方式进行模拟而得出的一种仿生算法，由大量蚂蚁组成的蚁群集体行为便表现出一种信息正反馈现象：某一路径上走过的蚂蚁越多，则后来者选择该路径的概率就越大。本质上是进化算法中的一种启发式全局优化算法。

[3] 高德地图坐标拾取器：https://lbs.amap.com/tools/picker。

通过 MATLAB 计算哈密顿环的最短路径为：颐和园→圆明园→北京动物园→北海公园→故宫博物院→天坛公园→北京欢乐谷→孔庙国子监→明十三陵→八达岭长城→颐和园(起点可以为任意一个景点)，其最短距离为 120.5241 公里，结果见图 3。

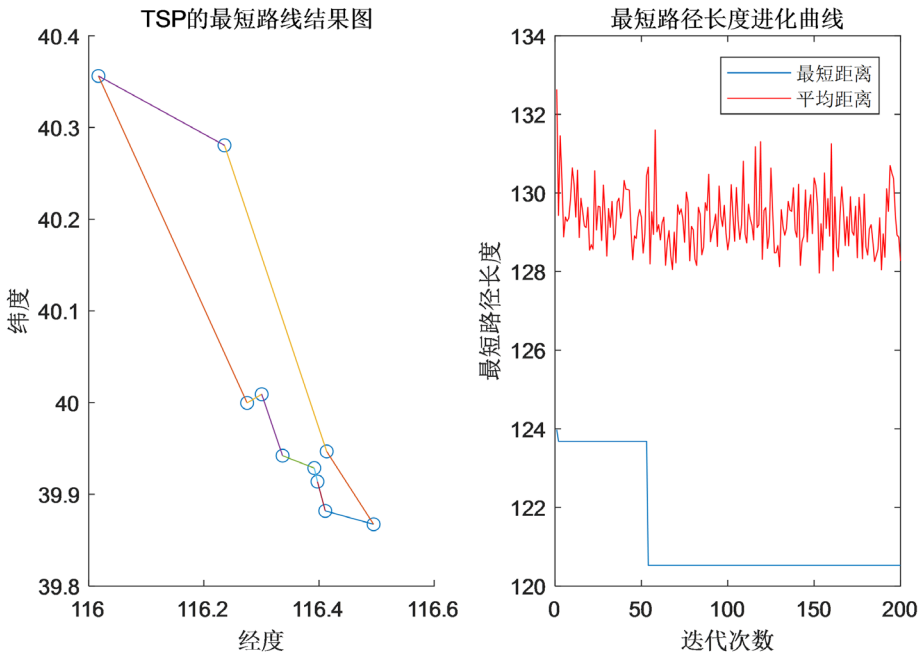

Figure 3. Optimal solution of TSP problem in ant colony algorithm
图 3. 蚁群算法下 TSP 问题的最优解

步骤四：找到该路线上最长的一条边的距离，将该路线从路线规划中剔除，从而得到最优路线，见图 4。

由距离比较得，八达岭长城与颐和园之间的距离最长，故可以考虑以八达岭长城为起点、颐和园为终点或颐和园为起点、八达岭长城为终点，其路线长度为 86.9117 公里。

Figure 4. Optimal route for multiple attractions
图 4. 多景点最优路线情况

- 模型推广

该模型还适用于大景区的旅游路线规划，即把北京市看作大景区即可，通过对小景点进行前面的文本数据挖掘，计算景点排名，得出要去的小景点，然后用上述的旅游路线规划模型即可得出最优旅游路线，减少用户在景区内的步行时间，缓解旅途的疲惫。

5.2. 一日游情境

- 情境设定

早晨十点开始旅行，晚上六点结束，在景点沿线午餐一个小时，不涉及起终点，游客偏好胡同休闲游，由于精细景点的评论数据不容易获得，故用户在每个景点的用时为建议游览时间，景点间的交通方式为地铁。

- 优化目标的确定

为了给游客提供良好的游览体验，应当让游客的步行时间尽量缩短，应当让游客游览尽可能多的景点，即景点间的交通路径不宜过长。将地铁运行速度与步行换乘速度放在一起等价为一个综合速度，估计为 60 km/h。

- 问题的分析

根据情境，约束条件为逛景点及景点间交通的时间为八个小时，每个景点都需要满足建议游览时间的限制。由于将步行到地铁站的时间等价，为便于计算，可以直接将距离矩阵定义为景点临近地铁站之间地铁运行时间的矩阵。当用户能够走尽量多的景点时，由于总时间固定，意味着游览时间尽可能多，使得交通时间尽可能短，而排名 3~7 的景点差异并不是很大，故可以将多个优化目标确定为第一目标：景点数量尽可能地多，第二目标：景点评分尽量高。

- 模型建立

步骤一：推荐景点的确定，结果见表 4。

Table 4. Leisure attractions in Hutong
表 4. 胡同休闲游景点一栏表

景点	得分	排名	建议游览时间	用于计算的时间
南锣鼓巷	485	1	2~4 小时	2 小时
前门大街	283	2	1~2 小时	1 小时
五道营胡同	177	3	1~2 小时	1 小时
什刹海	175	4	2~4 小时	2 小时
东交民巷	169	5	1~2 小时	1 小时
大栅栏	120	6	0.5~1 小时	0.5 小时
烟袋斜街	104	7	1 小时	2 小时

其中建议游览时间数据来源于去哪儿网[4]，由于优化目标可知，选择区间下限能够使得景点数目尽量多。

步骤二：先不考虑八个小时的约束条件，对这 7 个景点运用蚁群算法计算该哈密顿环的最短路径。走完这 7 个景点需要 9.75 个小时，其结果见图 5：

[4] 去哪儿网：https://travel.qunar.com/p-oi722949-nanluoguxiang。

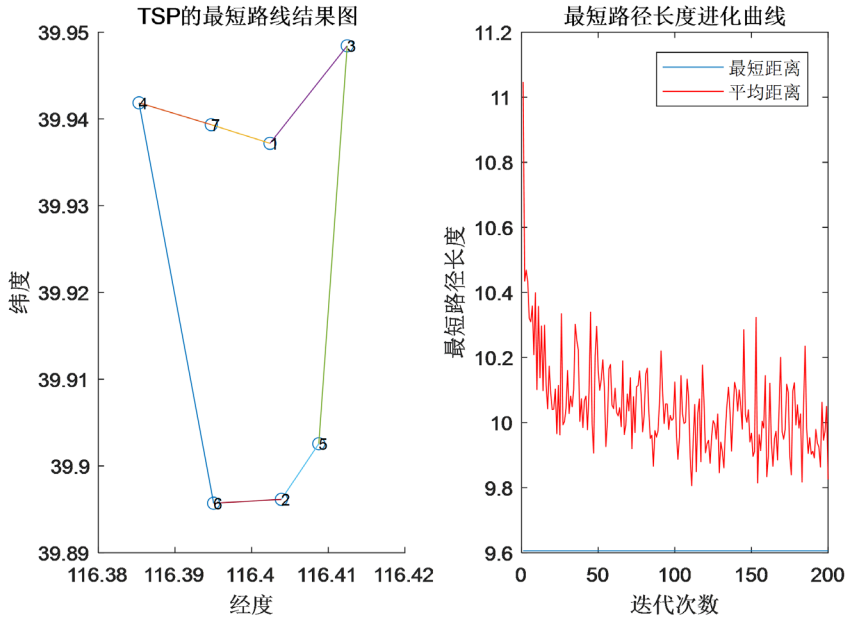

Figure 5. Optimal solution of the traveling salesman problem with seven scenic spots
图 5. 7 个景点旅行商问题的最优解情况

步骤三：考虑八个小时的约束，采用遍历法，从排名靠后的景点开始，先尝试剔除一个景点，判断是否满足约束；若都不满足约束，则考虑剔除两个景点，以此类推。

通过计算发现，只需要裁掉一个需要游览两个小时的景点即可。这时需要考虑剔除排名最靠后的景点，即烟袋斜街。

因此，通过该模型的计算，得到最优路线为：什刹海→南锣鼓巷→五道营胡同→东交民巷→前门大街→大栅栏，其中南锣鼓巷和什刹海均可以作为起点。其线路图见图 6：

Figure 6. Optimal metro route map between attractions
图 6. 景点间的最优地铁路线图

6. 结论

本文对北京市的 52 个主要景点的 88,486 条评论文本进行主题分析，得到了“多日游”和“一日游”

场景下的景点推荐以及路线规划的结果，解决了“旅游发愁”的问题，具有很强的现实意义。

本模型不同于基于数值数据的推荐系统，采用了更容易、更真实反映用户体验的评论文本数据进行深入挖掘，更加贴合用户的实际，很好地实现景点分类和推荐的功能。结合运筹学的知识，运用启发式算法，得到旅游景点推荐以及相应的路线规划结果。

由于文本评论来源有限，携程网的评论文本占主要部分，其对景点的评论大多与导游服务等相关，也会受到节假日高峰期评论数据的影响，因此存在对主题的划分不够精确的问题。本文用到的评论文本实际是来自于偏爱评论的这部分人，以及跟团游等，没有考虑诸如北京市民、当地居民对景点的评论情况，还需要更有效的数据做进一步的完善和拓展。

7. 模型推广

7.1. 旅游行程推荐

结合用户提供的旅行消费预算、出行人数、住宿偏好、饮食偏好等，为游客推荐在景点旅游线路附近最近的满足要求的酒店住宿以及餐厅，并合理规划好游客的每日出行计划。

7.2. 带人流量实时数据的路线推荐

本文所述的旅游景点推荐仅考虑了景点的受欢迎度、主题类别以及距离的远近。但在一些旅游高峰期时，人流拥挤会大大降低游客旅行的舒适度，因此，可以考虑结合实时人流量密度数据为用户提供更适宜的景点以及旅游路线。

7.3. 带季节特色的路线推荐

以综合评分第二的玉渊潭公园为例，从主题分析可以看出“樱花”是其最大的特色，受近期樱花季的影响，高德中近一个月去过最多的景点就是玉渊潭公园，高达 218.9 万人次，说明一个景点的季节特色也是应该考虑进去的。因此，可以根据用户需要旅行的时间再景点的排名中增加季节性的指标进行推荐，使得该模型更加的贴合游客的需要。

参考文献

[1] 张宇菲, 彭旭, 邵光明, 陈华友. 旅游路线规划问题[J]. 数学的实践与认识, 2016, 46(15): 81-89.

[2] 常亮, 孙文平, 张伟涛, 宾辰忠, 古天龙. 旅游路线规划研究综述[J]. 智能系统学报, 2019, 14(1): 82-92.

[3] 刘忠花, 李宪印, 于婷, 杨博旭. 基于三阶段 TSP 算法的旅游路线规划[J]. 曲阜师范大学学报(自然科学版), 2016, 42(4): 11-16.

[4] 徐锋, 杜军平. 改进蚁群算法在旅游路线规划中的应用研究[J]. 计算机工程与应用, 2009, 45(23): 193-195+226.

[5] 徐戈, 王厚峰. 自然语言处理中主题模型的发展[J]. 计算机学报, 2011, 34(8): 1423-1436.

[6] 蒋帅. K-均值聚类算法研究[D]: [硕士学位论文]. 西安: 陕西师范大学, 2010.

[7] 杨剑峰. 蚁群算法及其应用研究[D]: [博士学位论文]. 杭州: 浙江大学, 2007.