计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2018,27(7):199−204 [doi: 10.15888/j.cnki.csa.006416]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于网络搜索数据的游客量组合预测模型①

谢天保, 赵　萌

(西安理工大学 经济与管理学院, 西安 710054)

摘　要: 随着信息技术的不断发展, 基于网络数据对事物近期发展态势预测成为热点. 本文以北京市月度游客量预测为目标, 以相关网络关键词搜索指数为自变量建立了 BP 神经网络、支持向量回归和随机森林三种单一预测模型, 在此基础上构建组合模型以提高预测准确度. 实验结果表明: 基于 GBDT 建立的组合模型达到了较高的预测准确度, 误差仅为 3.16%, 预测结果可以为旅游管理部门提供决策支持.

关键词: 游客量预测; 网络搜索数据; 机器学习算法; 组合模型

引用格式: 谢天保,赵萌.基于网络搜索数据的游客量组合预测模型.计算机系统应用,2018,27(7):199–204. http://www.c-s-a.org.cn/1003-3254/6416.html

Multi-Models Combination Tourists Quantity Forecasting Based on Network Search Data

XIE Tian-Bao, ZHAO Meng

(School of Economics and Management, Xi'an University of Technology, Xi'an 710054, China)

Abstract: With the continuous development of information technology, the forecast of the recent development of things based on the network data has become a hotspot. In order to predict the monthly number of tourists in Beijing, this study established three kinds of single models with the search index of the relevant network keywords as independent variables: BP neural network, support vector regression, and random forest, and constructed a variety of combinatorial models to improve the prediction accuracy. The experimental results show that the combination of models based on GBDT have achieved higher prediction accuracy, the error is 3.16%. The forecast results can provide decision support for tourism management.

Key words: tourists quantity forecasting; network search data; machine learning algorithm; combinatorial model

近年来, 伴随旅游业蓬勃发展的同时, 游客普遍反映旅游体验在逐渐变差. 究其根本, 主要源于在旅游高峰期, 景点接待能力与涌入的游客量不匹配. 各地著名景区在节假日期间往往游客爆棚、人满为患, 管理难度大幅度提升导致超出了景区管理人员的可控范围, 使得游客的游玩体验受到严重影响, 游客的人身财产安全也难以保证. 因此, 如果能实现对未来一段时间尤其是旅游旺季的游客量预测, 管理者就可以结合实际的承载能力提前制定有效的防范措施, 确保服务质量和景区安全, 具有极强的现实意义.

1　研究现状分析

传统的旅游人数预测研究采用的主要方法有时间序列模型[1]、灰色系统理论[2]以及人工神经网络[3]等, 但这些研究采用的历史数据存在较大延迟性, 时间粒度也很大, 大都集中于国家或省级层面的年度入境人数预测. 随着大数据时代的到来以及基于网络数据的经济社会类行为预测研究的广泛开展, 在研究旅游行为相关问题时, 越来越多的研究人员将目光投向了网络搜索数据. 文献[4]发现我国部分 3A 级旅游景区客流量与网络关注度密度具有明显呼应的关系; 文献

① 基金项目: 陕西省重点学科资助项目 (107-00x901)
Foundation item: Funding Project for Key Disciplines of Shaanxi Province (107-00x901)
收稿时间: 2017-10-31; 修改时间: 2017-11-21; 采用时间: 2017-12-01; csa 在线出版时间: 2018-06-27

[5]证实网络关注度和旅游人数存在长期均衡关系和 Granger 因果关系; 文献[6–10]等关系研究均表明网络搜索数据包含着许多有价值的行为信息, 对现实游客量存在前兆效应, 具有一定的预测能力. 文献[11]基于谷歌趋势构建了一般的 ARIMA 模型及加入网络数据作为自变量的预测模型, 发现后者拟合效果和预测精度更高, 但关注的仍是全国入境人数这种大范围预测; 文献[12]发现加入百度关键词作为解释变量的模型相比传统的 ARMA 模型, 预测精度提高了 14.5%, 但依然存在较大误差; 文献[13]采用直接取词法选取 5 个关键词数据作为解释变量分别建立了向量自回归和 BP 神经网络模型, 发现神经网络比回归法预测精度略高, 但关键词过少, 难免会因信息遗漏使模型与实际有一定偏离.

为实现更准确、更具有时效性、地域针对性更强的预测, 本文拟基于网络搜索数据, 结合多种机器学习算法建立游客量预测模型, 时间粒度选取为月度, 以提高预测的及时性和实用性, 同时考虑到组合预测法的思想, 即在诸种单一预测模型各异的情况下, 组合预测模型可能会得到比任何一个独立预测值更好的预测值, 显著改进预测效果[14], 进一步构建组合模型以优化预测结果.

2 网络搜索关键词的选取

首都北京在我国旅游城市排行中首屈一指, 本文选取北京市游客量作为研究对象, 收集了 2011 年 1 月至 2016 年 12 月期间, 每个月北京市所有旅游景区、景点接待的全部游客总量, 但模型也可推广应用至其他地区和省市.

搜索引擎能够帮助游客从数以亿计的网页中快速定位到所需要的信息, 而关键词搜索是游客在线信息搜索时最常用的策略[15], 所以基于网络搜索数据的预测研究的第一步就是选取相关搜索关键词. 本文中所用到的关键词网络搜索量来源于国内应用最为官方的搜索引擎的百度指数.

2.1 选定核心关键词

本文采用文本挖掘的方法, 结合旅游六要素, 即食、住、行、游、购、娱, 对网络上与北京旅游相关的新闻、文章、点评、分享交流等信息进行查找收集, 剔除掉一些无用信息后, 再使用 NLPIR 汉语分词系统对原始文本集合进行处理, 得到关键词列表及其权重, 权重越高, 越应被选为核心关键词. 最终选定了 6 个核心关键词: “北京小吃”、“北京住宿”、“北京旅游地图”、“北京旅游”、“北京特产”及“北京景点”.

2.2 核心关键词搜索指数的预测能力分析

显然网络搜索数据和实际游客量数据都属于时间序列, 平稳性是时间序列数据统计推断的基础. 检查序列平稳性的标准方法是各种单位根检验, 本文采用 ADF (Augmented Dickey-Fuller Test) 检验对 6 个核心关键词的搜索指数序列和实际游客人数序列的平稳性进行检验, 结果表明原序列中部分为非平稳序列, 但在一阶差分下所有变量均在 1% 的显著性水平上拒绝原假设 (原假设为序列至少有一个单位根, 即不平稳), 即均为一阶单整序列, 符合协整检验的前提条件.

本文通过 Johansen 协整检验来考察变量间的协整关系, 检验结果如图 1 所示, 可以发现特征根迹检验和最大特征值检验在 5% 的显著性水平上都是拒绝原假设的, 说明协整关系存在, 依据现代协整理论, 对于非平稳时间序列, 只要各变量之间存在协整关系, 就可以直接建立 VAR 模型[16].

Unrestricted Cointegration Rank Test (Trace)

Hypothesized No. of CE(s)	Eigenvalue	Trace Statistic	0.05 Critical Value	Prob.**
None *	0.589824	173.8108	125.6154	0.0000
At most 1 *	0.530342	123.0142	95.75366	0.0002
At most 2 *	0.471578	79.93636	69.81889	0.0063
At most 3	0.321658	43.57837	47.85613	0.1191
At most 4	0.224385	21.45649	29.79707	0.3298
At most 5	0.112382	6.972835	15.49471	0.5808
At most 6	0.003112	0.177660	3.841466	0.6734

Trace test indicates 3 cointegrating eqn(s) at the 0.05 level
* denotes rejection of the hypothesis at the 0.05 level
**MacKinnon-Haug-Michelis (1999) p-values

(a) 特征根迹检验

Unrestricted Cointegration Rank Test (Maximum Eigenvalue)

Hypothesized No. of CE(s)	Eigenvalue	Max-Eigen Statistic	0.05 Critical Value	Prob.**
None *	0.589824	50.79664	46.23142	0.0152
At most 1 *	0.530342	43.07779	40.07757	0.0223
At most 2 *	0.471578	36.35798	33.87687	0.0248
At most 3	0.321658	22.12188	27.58434	0.2142
At most 4	0.224385	14.48366	21.13162	0.3267
At most 5	0.112382	6.795175	14.26460	0.5137
At most 6	0.003112	0.177660	3.841466	0.6734

Max-eigenvalue test indicates 3 cointegrating eqn(s) at the 0.05 level
* denotes rejection of the hypothesis at the 0.05 level
**MacKinnon-Haug-Michelis (1999) p-values

(b) 最大特征值检验

图 1 Johansen 协整检验结果

实验收集了 2011 年至 2016 年共计 72 个月的月度数据, 选取前 5 年 (即前 60 个) 数据作为样本集用于建模, 2016 年 1 月至 12 月的数据则作为测试集用于模型验证. 建立 VAR 模型需要确定滞后阶数, 本文结合似然比 LR、AIC、SC 准则等多种检验方法, 最终确定建立 VAR(3) 模型. 如图 2 所示, 该 VAR 模型所有特征根的倒数均落于单位圆内, 即均小于 1, 模型稳定. 应用该模型预测样本集外数据, 结果如图 3 所示.

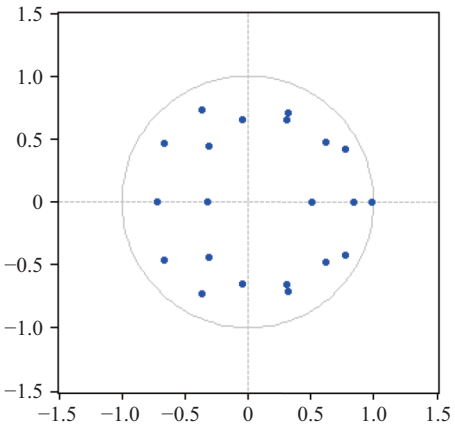

图 2 VAR 模型特征根位置图

总体来说, 预测值与实际值的趋势基本保持了一致, 说明模型具有一定的预测能力, 关键词指数的前期变化的确有助于解释实际游客量的变化. 但是预测误差明显较大, 平均绝对百分比误差 (MAPE) 高达 12.24%, 具体到每一个月的相对误差基本在几百万人次 (图 3 中游客人数单位为万人次), 显然达不到精准预测的要求.

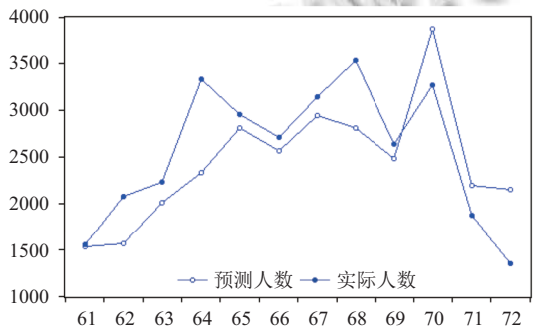

图 3 2016 年北京市实际旅游人数和预测人数的对比图

因此, 仅仅基于这 6 个核心关键词对游客人数进行预测是不科学的, 选取核心关键词的方法不完善或是核心关键词的数量过少, 都会导致信息覆盖不全面从而影响研究结果. 为了提高研究结果的准确性, 应该对核心关键词进行大范围拓展和进一步择优, 才能保障模型中所加入的自变量能尽可能的涵盖会影响到因变量变化的所有信息.

2.3 关键词的拓展与择优

拓展的目标是围绕少数的核心关键词, 拓展出数量更多的相关关键词. 拓展的依据和方法有多种, 本文综合使用了长尾关键词拓展法、百度需求图谱以及网页相关搜索推荐, 建立了一个包含 79 个关键词的初始词库.

通过判定各个关键词与研究对象的关联关系, 筛选出合适数目的最优关键词是提升模型预测准确度的关键. 因为并不是每个关键词都与实际游客量存在相关关系, 多个词之间也可能存在共线性, 导致信息重叠, 不利于模型建立. 本文首先根据 Spearman 秩相关检验筛选出相关系数大于 0.6 的搜索关键词, 共计 38 个. 然后采用时差相关分析确定上一步筛选出的关键词搜索指数与北京市游客量的时滞阶数, 并选取同行关键词指标 (网络搜索作为一种即时性行为, 游客一般都会在出行当月搜索相关的旅游信息). 最后筛选出的同行关键词及其相关系数, 共计 25 个. 如表 1 所示.

表 1 同行关键词 spearman 秩相关系数

关键词	相关系数	关键词	相关系数
八达岭长城	0.64	北京世界公园	0.90
北海公园	0.84	北京住宿哪里便宜	0.62
北京动物园	0.83	故宫	0.61
北京海洋馆	0.68	故宫门票	0.73
北京欢乐谷	0.85	北京小吃	0.66
天安门升旗时间	0.65	北京小吃街	0.78
北京景点	0.79	北京一日游	0.80
北京景点地图	0.76	北京周边好玩的地方	0.77
北京好玩的地方	0.82	北京住宿攻略	0.64
北京旅游	0.66	北京酒店	0.60
北京旅游攻略	0.74	颐和园	0.60
北京旅游景点大全	0.79	毛主席纪念堂	0.63
北京美食攻略	0.61		

VAR 模型本质就是把系统中每一个变量描述为系统中所有变量的滞后值的线性函数, 当变量多达 25 个时, 难以保证各变量之间仅仅存在线性关系. 因此, 对于解释变量众多、平稳性和协整关系难以保证、可能存在非线性关系等情况, 应用适应性更为广

泛的机器学习算法建立预测模型比传统的 VAR 模型更为合适.

3 单一预测模型的构建

3.1 BP 神经网络模型

理论上已经证明三层神经网络可以无限逼近任意连续函数, 本文建立单隐藏层的 BP 神经网络模型, 再对模型隐藏层的节点数目和迭代次数进行优化, 以确定出最优的模型误判率.

实验发现, 训练集误差跟随隐藏层节点数的增加而下降, 但测试集误差先下降后面反而上升, 这是由于模型中隐藏层节点数增加而引起的模型过度拟合导致的, 考虑到预测模型应注重模型的推广能力, 当隐藏层节点数为 4 时, 测试集 MAE 值最小且训练集误差也在接受范围内, 因此确定最优的隐藏层节点数为 4. 同时, 当训练周期达到 300 以后, 训练集和测试集的 MAE 均趋于平稳且已经达到了较小的值, 因此最终确定出一个隐藏层节点数为 4, 训练周期为 300 的单隐藏层 BP 神经网络模型.

3.2 支持向量回归模型

支持向量机最初是根据分类问题发展起来的, 但也可应用于回归问题. 建立 SVR(支持向量回归机) 模型, 需要确定分类方式和核函数的组合方式, 针对数值型变量的分类方式主要有两种 (eps-regression 和 nu-regression), 核函数则有四类 (linear, polynomial, radial 和 sigmoid).

实验发现, 按照 *MAE* 值最小原则无论是测试集预测还是训练集拟合均应选择 eps-regression 和 radial 的组合. 在此基础上对惩罚因子 *cost* 和 *gamma* 参数进行优化, 同样按照 *MAE* 值最小原则确定出测试集 *cost* 取 1, *gamma* 取 0.1, 训练集则 *cost* 取 10, *gamma* 取 1.

3.3 随机森林模型

在构建随机森林模型的过程中有两个重要参数: 一是树节点预选的变量个数 *mtry*, 决定着单棵决策树的情况; 二是随机森林中树的个数 *ntree*, 决定着整片森林的总体规模.

实验发现当 *mtry*= 5 时, 模型对变量的解释率最高, 为 86.05%, 残差平方均值最小, 所以节点上变量个数确定为 5. 接着确定整片森林的规模, 实验发现模型误差随决策树数量的增多逐渐降低并趋于平稳, 当决策树数量约大于 1300 之后, 模型误差基本稳定, 因此将 *ntree* 值确定为 1300.

以上三种模型预测误差如表 2(见 4.2 节) 所示, 从 *MAPE* 值来看, 支持向量回归最优, 随机森林次之, BP 神经网络则相对较差. 但总体来说, 这三种单一模型的预测准确度和稳定性都优于前述的 VAR 模型, 这一方面说明了关键词拓展的必要性, 另一方面也说明网络搜索指数与实际游客量之间存在部分非线性关系, 因此机器学习在这种预测方面更具优势.

4 基于机器学习算法的组合预测模型

4.1 建立 GBDT 组合预测模型

以往研究中使用频率较高的是简单便捷的定权组合法 (如等权平均法、方差倒数法), 但其实笼统的赋予定值权重, 对于提高预测准确度是不理想的, 因为不同单一模型在不同时刻的预测误差是不一样的, 如果按照时刻和预测误差的变化赋予各个模型动态变化的权值, 效果会更佳, 本文提出基于 GBDT 的组合预测模型.

GBDT (Gradient Boosting Decision Tree) 是一种梯度提升的决策树算法, 核心思想是将损失函数的负梯度在当前模型的值作为回归问题提升树算法中的残差的近似值, 拟合一个回归数. 将三种单一模型训练集的拟合序列作为新的训练集, 将单一模型测试集的预测序列作为新的测试集建立 GBDT 模型, 模型中赋予各个单一模型的权重系数应是随时间点不同而变化的. 算法流程如下文.

Step 1. 输入训练集 $T=\{(x_1,y_1),(x_2,y_2),\cdots,(x_n,y_n)\}$, 损失函数 $L(y,f(x))$;

Step 2. 初始化: $f_0(x)=\arg\min_c \sum_{i=1}^{n} L(y_i,c)$.

Step 3. 对于迭代轮数 $m=1,2,\cdots,M$:

(a) 对 $i=1,2,\cdots,n$, 计算负梯度 $r_{mi}=-\left[\frac{\partial L(y_i,f(x_i))}{\partial f(x_i)}\right]_{f(x)=f_{m-1}(x)}$;

(b) 对 r_{mi} 拟合一个回归树, 得到第 m 棵树的叶节点区域为 R_{mj}, $j=1,2,\cdots,J$;

(c) 对 $j=1,2,\cdots,J$, 计算最佳拟合值 $c_{mj}=\arg\min_c \sum_{x_i\in R_{mj}} L(y_i,f_{m-1}(x_i)+c)$;

(d) 更新 $f_m(x)=f_{m-1}(x)+\sum_{j=1}^{J} c_{mj}I(x\in R_{mj})$.

Step 4. 得到回归树:

$$\hat{f}(x) = f_M(x) = \sum_{m=1}^{M}\sum_{j=1}^{J} c_{mj} I(x \in R_{mj})$$

要对各参数进行优化, 包括损失函数、学习速率、迭代次数等. 损失函数选择回归问题中最常用的 Gaussian 分布, 学习速率取 0.05, 使用交叉验证确定最佳迭代次数为 2518. 最终根据此模型得到一组新的组合预测结果.

4.2 模型预测结果评价

为了有效和直观的衡量不同模型的预测能力, 本文选取均方误差 (MSE)、平均绝对误差 (MAE)、平均绝对百分比误差 (MAPE) 这三个指标来评估预测结果, 各模型预测结果如表 2.

表 2 各模型预测结果比较

误差	BP 神经网络	支持向量回归	随机森林	GBDT
均方误差	30.56	20.33	17.61	12.04
平均绝对误差	173.63	110.41	117.73	65.16
平均绝对百分比误差 (%)	8.71	4.91	5.59	3.16

从表 2 可以看出, 无论从 MSE、MAE 还是 MAPE 来说, 组合模型的预测效果均有显著优势, 相比单一模型大幅度提高了预测准确度. 各模型的预测值与实际值对比如图 4 所示.

由图 4 可知, 其中图 4(a) 和图 4(d) 清晰直观的表现出了效果最差的单一模型与效果最好的组合模型在预测准确度上的明显差异 (由于游客量数据周期性很强, 每一年走势基本一致, 因此仅展示 2014~2016 年的数据), BP 神经网络模型通过学习训练基本能预测出游客量一年的走势, 但对峰值敏感度较低, 训练集拟合效果也较差, 而 GBDT 组合模型的训练集拟合效果很好, 峰值敏感度和测试集预测效果也更优.

5 结束语

本文以北京市游客量为研究对象, 选定核心关键词后, 对其进行数据检验和预测能力分析, 证明网络搜索数据的确有助于预测实际游客量, 为提高预测的科学性和自变量信息的完善性, 进一步拓展核心关键词并择优筛选, 基于同行相关关键词的百度搜索指数, 分别建立了三种单一预测模型, 为提高预测准确度又建立了基于 GBDT 的组合模型, 模型预测结果显著体现出了组合预测的优越性. 统计局统计数据的发布至少存在两个月的滞后期, 而本文提出的基于同行网络数据的组合预测模型可以即时预测当月人数, 具有很强的现实意义. 模型的进一步推广应用与可靠性检验是接下来的研究方向.

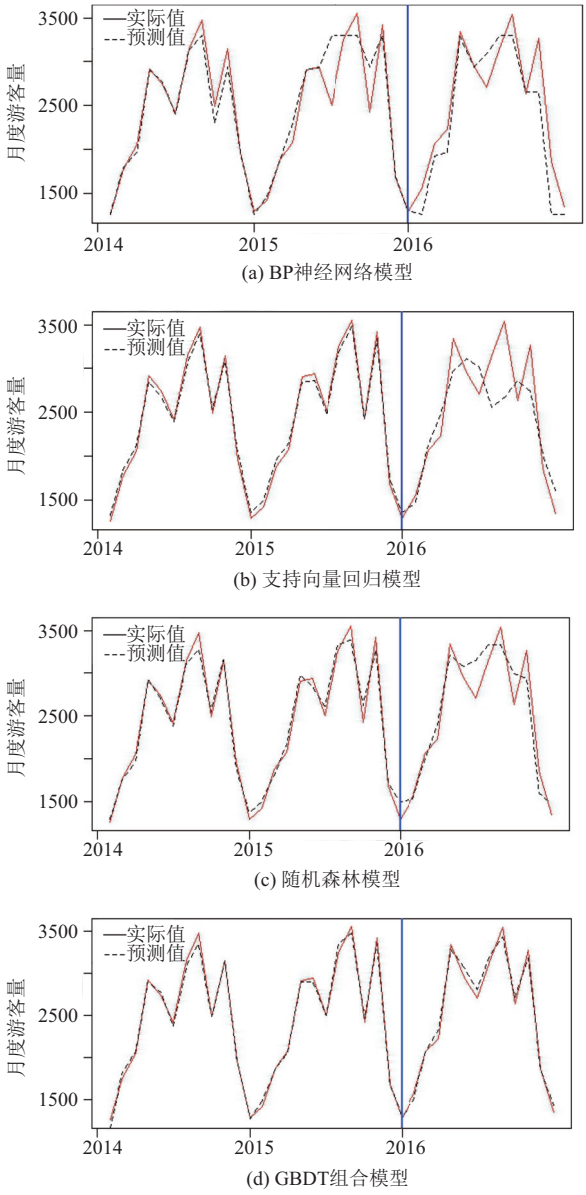

图 4 各模型预测效果图

参考文献

1 李乃文, 韩婧婧. 基于时间序列修正算法的我国入境旅游人数预测. 资源开发与市场, 2015, 31(1): 126–128.

2 曾冬玲, 喻科, 赵清俊. 基于灰色理论和马尔科夫修正的旅游需求预测——以云南省旅游市场为例. 重庆工商大学学报 (自然科学版), 2016, 33(4): 58–68.

3 Hassani H, Silva ES, Antonakakis N, *et al*. Forecasting

accuracy evaluation of tourist arrivals. Annals of Tourism Research, 2017, (63): 112–127. [doi: 10.1016/j.annals.2017.01.008]

4 汪秋菊, 黄明, 刘宇. 城市旅游客流量——网络关注度空间分布特征与耦合分析. 地理与地理信息科学, 2015, 31(5): 102–106.

5 殷杰, 郑向敏, 董斌彬. 基于 VECM 模型的景区网络关注度与旅游人数的关系研究——以鼓浪屿为例. 福建农林大学学报 (哲学社会科学版), 2015, 18(5): 68–75.

6 Miah SJ, Vu HQ, Gammack J, *et al*. A big data analytics method for tourist behaviour analysis. Information & Management, 2017, 54(6): 771–785.

7 王玉霞, 王静. 客流量与网络关注度的关系分析——以首都博物馆为例. 北京联合大学学报, 2016, 30(1): 75–80.

8 黄娟, 黄英, 张敏. 基于网络关注度构建智慧旅游公共服务体系的实证建议——以武汉为例. 现代城市研究, 2016, 31(2): 126–131.

9 邓爱民, 王瑞娟. 基于百度指数的旅游目的地关注度研究——以武汉市为例. 珞珈管理评论, 2014, (2): 143–152.

10 龙祖坤, 任红丹. 湖南省 5A 级景区网络关注度分布特征及形成机理研究. 湖南财政经济学院学报, 2016, 32(159): 141–147.

11 沈苏彦, 赵锦, 徐坚. 基于“谷歌趋势”数据的入境外国游客量预测. 资源科学, 2015, 37(11): 2111–2119.

12 黄先开, 张丽峰, 丁于思. 百度指数与旅游景区游客量的关系及预测研究——以北京故宫为例. 旅游学刊, 2013, 28(11): 93–100.

13 陈涛, 刘庆龙. 智慧旅游背景下的大数据应用研究: 以旅游需求预测为例. 电子政务, 2015, (9): 6–13.

14 徐国祥. 统计预测和决策. 上海: 上海财经大学出版社, 2005. 251–257.

15 Xiang Z, Gretzel U, Fesenmaier DR. Semantic representation of tourism on the internet. Journal of Travel Research, 2009, 47(4): 440–453. [doi: 10.1177/0047287508326650]

16 李娅, 李志鹏. 计量经济分析与 Eviews 详解. 北京: 科学出版社, 2017. 155–163.