基于噪声处理的网络搜索和 QCR – HHT 模型的九寨沟客流量预测

李晓炫[1],吴　奇[2]
(1. 阜阳师范大学 经济学院,安徽 阜阳 236037;2. 阜阳师范大学
物理与电子工程学院,安徽 阜阳 236037)

摘　　要:客流量预测可以弥补强周期性和波动性客流冲击给景区和游客造成的影响,使有限的旅游资源提前得到合理调度和配置. 在考虑网络搜索噪声的基础上,建立 QCR(Query Chain Retrieve)搜索词链和 HHT 的网络搜索数据预测模型,对九寨沟旅游日客流量进行预测. 通过对比时间序列模型、未经噪声处理的网络搜索预测模型和 BP 神经网络发现,QCR – HHT 拟合效果最佳,对九寨沟客流量的预测精度显著提高. 使用考虑噪声的 QCR – HHT 网络搜索预测模型,能够更准确地对旅游客流量进行预测,便于景区和管理部门制定更加高效准确的决策.

关　键　词:网络搜索数据; 搜索词链; 噪声处理; Hilbert 频率谱分析
中图分类号:TP393.4;F592.3　　**文献标识码**:A　　**文章编号**:1673 – 1670(2021)02 – 0053 – 07

0　引言

随着我国社会经济的日益发展,我国旅游行业发展速度逐步加快,各大旅游地的客流量均呈快速上升趋势. 旅游产业的周期性强,不同月份之间的客流量差距大,休假制度设计不甚完善导致的节假日旅游需求集中释放现象较为明显. 这种强周期性和波动性的客流量波动给景区和旅游目的地造成较大冲击,因景区超载、游客滞留等问题带来的安全隐患对游客出行体验和旅游产业的健康发展产生了诸多不良影响. 对客流量的精准预测能够使得旅游经营和管理者提前通过合理调度和配置有限旅游资源的方式最大限度避免这种混乱局面的发生. 搜索引擎作为网民信息搜寻的主要工具,捕捉并记录了海量线上信息搜集行为. 已有大量研究基于网络搜索实现客流量预测[1-2]、股市价格监控[3]和汇率浮动[4]等. 然而不管是旅游客流量还是网络搜索数据,数据统计和搜寻信息过程中都会受到外界噪声信号干扰,导致数据序列存在大量的噪声. 这些噪声会影响对于真实游客行为和决策过程的理解并且降低网络搜索数据对社会经济的预测能力,因此噪声是阻碍精准预测旅游客流量的一大根源.

针对信号中噪声的处理方法,在物理学中常用频谱分析技术实现,如小波变化和傅里叶变换. 傅里叶变换要求所处理的信号必须是平稳的,并且具有明显区别于噪声的谱特性. 然而社会经济中序列信号大多是非平稳序列,因此傅里叶变换受到较大使用限制. 小波变换基于傅里叶变换之上,凭借多分辨特性在合适的尺度下,对非平稳信号的有效成分也能识别出与噪声截然不同的谱特性,在获得信噪比增益的同时能够保持对突变信息的良好识辨,因此在处理非平稳信号时具有明显优势[5]. HHT 变换(Hilbert – Huang Transform)是最新发展起来的用于处理非线性非平稳信号的时频分析方法. 它保留了小波变换多分辨的优势,同时克服了小波变换选择小波基的问题,近年来被广泛用于非平稳信号的滤波和去噪. 笔者拟利用网络搜索实现对旅游客流量的预测,并具体分析滤噪技术的使用对于提高网络搜索预测精度方面的作用.

对搜索引擎数据的挖掘和分析能够最直接地得到人们的行为模式和决策过程. 继 2009 年某搜索引擎预测美国流感爆发之后,网络搜索数据的预测能力被研究者广泛分析. 王长琼等人[6]将神经

收稿日期:2020 – 10 – 14
基金项目:安徽省自然科学基金青年项目(1908085QG305);安徽省高校人文社会科学研究重点项目(SK2020A0341,SK2020A0330);阜阳师范大学青年人才重点项目(rcxm202008)
作者简介:李晓炫(1987—),女,安徽省阜阳市人,管理学博士,阜阳师范大学经济学院讲师,主要从事网络经济、大数据预测研究.

网络和支持向量机引入,提出了结合网络搜索数据的组合预测模型,将电商网络订单量的预测精度提高了 2.67%. 李凤岐、李光明[7]在百度指数基础上利用 PS 方法自动挖掘网络指数与经济指标间的关系,分析了多源搜索数据的预测能力,并实现了网络搜索数据对 CPI 和 CCI 这类宏观经济指标的预测. Bangwayo – Skeete 和 Skeete[8]利用搜索词"酒店"和"机票"的搜索量,结合 AR – MIDAS 模型对旅游客流量进行预测,对比时间序列基准模型发现,加入网络搜索后的 AR – MIDAS 模型的预测精度显著提高. Höpken 等人[9]预测瑞士山旅游客流量时,提取关键词环节改进了仅利用搜索引擎推荐的方法,选择在关注搜索词多语种表达基础上提取与目标词最相关的搜索词方法,再利用网络数据进行预测时发现预测效果明显优于时间序列模型. 然而网络数据的预测并非总是有效的,Volchek 等人[10]利用某搜索指数从微观层面对英国 5 个博物馆旅游客流量预测,分别利用简单线性回归、季节性回归,SARMAX 等模型,发现在不同的数据频率上,没有一个模型可以优于其他模型的预测精度. Li 等人[11]利用百度指数预测九寨沟旅游客流量时得到类似的结论,发现网络搜索数据的预测效果甚至与简单线性回归模型相差不大. 这说明仅简单利用网络搜索数据进行预测,可能受到多种干扰,这些可能会导致网络数据的应用受限.

网络搜索数据的预测能力被很多学者证实,然而不管是搜索数据还是研究对象的统计数据都存在一定的噪声. 这些序列信号中的噪声不可避免,而如何处理信号噪声则是在预测中面对的另外一大问题. 考虑序列非平稳性非线性特征,HHT 变换在处理社会经济数据时具有明显优势. HHT 变换是由 Hilbert 和 Huang 提出的,主要包括经验模态分解(EMD)和希伯特变换(Hilbert Transform). 该方法一经提出就开始在机械故障[12]和地球物理[13]等领域得到广泛的应用. 最近开始有学者试图将 HHT 应用到经济管理中并且发现效果显著. 李晓炫等人[14]利用百度指数,将 EMD – BP 组合预测模型应用于预测九寨沟旅游客流量时,发现 EMD 分解可以有效识别网络搜索数据中的噪声,相比于时间序列模型、网络搜索预测模型和 BP 神经网络模型,经过 EMD 分解后搜索数据预测能力显著提高. 陆利军[15]利用张家界旅游客流量数据再一次验证了 EMD 对分离网络搜索数据噪声的有效性.

笔者拟在之前学者基础上,首先提出基于网络搜索用户搜索路径的关键词提取方法,结合使用 HHT 方法对搜索数据进行序列的过滤噪声处理,对九寨沟旅游客流量日度数据进行预测,比较滤噪对客流量预测精度的影响效果.

1 Hilbert – Huang 变换

HHT 变换包括两个部分:Huang 提出的经验模态分解(EMD)和 Hilbert 谱分析. EMD 是针对非线性非平稳序列进行的一种自适应时间序列分解方法,原始序列经过 EMD 算法分解后,可以得到若干条彼此影响甚微的本征模函数(IMF)向量,这些分量具有不同的尺度,代表了不同频率的序列变动,从而简化了原始序列中不同尺度的特征信息之间的干涉或耦合. 各条 IMF 分量较好满足 Hilbert 变换对于窄带条件的要求,可以进一步求得 IMF 分量的 Hilbert 瞬时频率和 Hilbert 功率谱,据此识辨出序列中的噪声.

EMD 将一条时间序列分解为若干 IMF 分量和一条残波,每一条 IMF 分量包含着不同的时间尺度局部特征,并且必须满足两个基本条件:1) 在整个事件序列范围内,局部极值个数(极大值和极小值)与过零点个数相差不超过 1; 2)在任何时间序列范围内,局部均值为 0. 分解步骤为:1)计算原序列的所有局部极值点. 将所有极大值点连接生成上包络线 $U(t)$,所有极小值点连接得到下包络线 $L(t)$,求得上下包络线之间每一点的均值. 用原序列减去均值,得到第一条分量 IMF_1;2)将上述步骤中求得的 IMF_1 从原序列中分离出来得到剩余序列,重复整个步骤 1 并逐次得到剩余的各本征模函数 IMF_s,直到最终剩余的序列是常数或单调函数.

对 EMD 分解后的 IMF 分量(设为 c_i)构造解析信号:

$$z_i(t) = c_i(t) + \mathrm{j}\hat{c}_i(t) = a_i(t)\mathrm{e}^{\mathrm{j}\Phi_i(t)}. \quad (1)$$

其幅值函数、相位函数和瞬时频率分别为:

$$a_i(t) = \sqrt{c_i^2(t) + \hat{c}_i^2(t)}, \quad (2)$$

$$\Phi_i(t) = \arctan(\hat{c}_i(t)/c_i(t)), \quad (3)$$

$$f_i(t) = \frac{1}{2\pi}\omega_i(t) = \frac{1}{2\pi}\frac{\mathrm{d}\Phi_i(t)}{\mathrm{d}t}. \quad (4)$$

若省略残余分量 R_n,原信号 $Y(t)$ 可表示为:

$$Y(t) = \mathrm{Re}\sum_{i=1}^{n} a_i(t)\mathrm{e}^{\mathrm{j}\Phi_i(t)} = \mathrm{Re}\sum_{i=1}^{n} a_i(t)\mathrm{e}^{\mathrm{j}\int\omega_i(t)\mathrm{d}t}, \quad (5)$$

$$H(\omega,t) = \mathrm{Re}\sum_{i=1}^{n} a_i(t)\mathrm{e}^{\mathrm{j}\int\omega_i(t)\mathrm{d}t}. \qquad (6)$$

其中,Re 表示取信号的实部,式(6)为信号的 Hilbert 谱. Hilbert 谱很好地反映了信号增幅在整个频率轴上随频率和时间变化的规律.

2　实证分析

2.1　数据来源

我国九寨沟在景区网络搜索排名中遥遥领先,成为游客关注的最著名景点之首. 加上该地区不存在其他搜索信息干扰,搜索热词几乎跟九寨沟旅游本身密切相关,排除了当地人搜索和与旅游无关搜索的问题. 九寨沟的旅游客流量数据来自其官方旅游局网站发布的定期数据,因考虑到 2017 年 8 月 6 日地震灾害导致九寨沟景区损毁重建至今,景区每日将客流量限制在 2 000 人,笔者截选 2012 年 6 月 1 日至 2017 年 8 月 6 日共 1 893 天旅游客流量数据. 网络搜索数据取自百度指数网站,考虑到搜索的先行性,笔者取搜索数据时间提前于客流量一个月,截选 2012 年 5 月 1 日至 2017 年 8 月 6 日,共计 1 924 天数据. 笔者综合考虑模型稳定性和九寨沟旅游淡旺季的特点,选择 2012 年 6 月 1 日至 2016 年 10 月 31 日为训练集(共 1 614 天,占样本量 85%),2016 年 11 月 1 日至 2017 年 8 月 6 日为预测集(共 279 天,占样本量 15%).

2.2　QCR(Query Chain Retrieve)——合成搜索指数

百度指数发布的每一个搜索关键词搜索指数都是实际搜索量,笔者采用 QCR 搜索词条链对搜索数据进行指数合成. 网民利用搜索引擎进行信息搜寻时,无法一次性完成所需信息的搜索和收集,往往需要进行一系列的点击和一连串的关联搜索来达到信息收集的目的. 从搜索词条链中不仅可以看出网民的搜索轨迹,而且可以通过分析和文本挖掘发现网民的搜索意图和未来时期的决策方向,从而为分析和预测提供一个很好的思路和可能性. 一般来说,最初的搜索词条是较为广义的搜索关键词或中心词,在此基础上逐渐通过扩展和修改搜索关键词的方式一步步确定和缩小搜索范围,进而满足自己对于多样化信息的收集. 使用搜索引擎的网民数量庞大,而若以每个网民为单位研究搜索词条链,个体样本的选择极易造成样本有偏,进而导致研究和预测的偏差. 因此本文的 QCR 方法是基于所有网民的群体智慧,利用搜索引擎中的全体搜索的来源相关词和去向相关词提取搜索词条链.

根据“百度指数”的定义,“来源相关词”用于反映用户在当前搜索词之前的搜索需求,通过过滤当前搜索词上一步搜索行为来源的相关词,按照相关程度排序得到;“去向相关词”类似,反映用户在当前搜索词之后的搜索需求. 因此搜索词条链中每一轮搜索词的选择只依赖于与当前搜索词高度相关的来源词和去向词,即只选择与搜索词相关性最高的前 10 个来源相关词和去向相关词. 具体步骤包括:

1)根据游客的决策目标,选择“九寨沟”为预测目标词,提取出其排名前 10 的来源搜索词和去向搜索词. 网民在搜索“九寨沟”之前,搜索最多相关性最高的词条包括“旅游”“成都”和“九寨沟天气”等,之后搜索最多相关性最高的词条包括“九寨沟旅游”和“九寨沟旅游攻略”. 表 1 所示为以“九寨沟”为中心词的来源和去向相关词.

表 1　“九寨沟”的高度来源相关词和去向相关词

来源相关词		去向相关词	
旅游	成都到九寨沟	成都天气	攻略
成都	黄龙	成都	成都到九寨沟
天气	门票	天气	九寨沟天气
九寨沟天气	九寨沟天气预报 15 天	九寨沟旅游	九寨沟国家级自然保护区
九寨沟门票	九寨沟图片	九寨沟旅游攻略	预报

2)基于第一轮的选择,将来源相关词和去向相关词中明显与九寨沟旅游无关的搜索词剔除,剩余的搜索词在第二轮中作为中心词,提取出相应的前 10 个来源相关词和去向相关词,将已经存在的和重复的删除,最后剩余 74 个有效关键词.

3)分别计算每个搜索词与九寨沟客流量之间提前 0 ~ 31 期的 Pearson 相关系数. 观察每个搜索词与九寨沟客流量相关系数最大的时期,确定为该搜索词相对于九寨沟客流量的提前期. 将所有搜索词中提前期为 0 的剔除. 考虑到相关系数阈值的确定对预测结果和噪声产生较大影响,阈值过低会使约束过松,降低搜索指数与客流量相关性并包含过多噪声,阈值过高会使约束条件过于节俭,有遗漏关键搜索词的风险. 综合考虑,本文相关系数阈值确定为 0.7,共保留 28 个关键词搜索词条,若将阈值提高到 0.8 则只剩下 5 个关键词搜索词条,而若将阈值降低至 0.6,则有效关键词多达 53 个. 由此

可见过高或过低的阈值选择都不利于对搜索指数的有效筛选和提取.

4)将筛选后保留的 28 个搜索关键词进行加总,得到合成搜索指数 Index7. 图 1 中两个序列之间的变化趋势总体较为一致,吻合度较高,大略显示了利用合成搜索指数对旅游客流量预测的可能性.

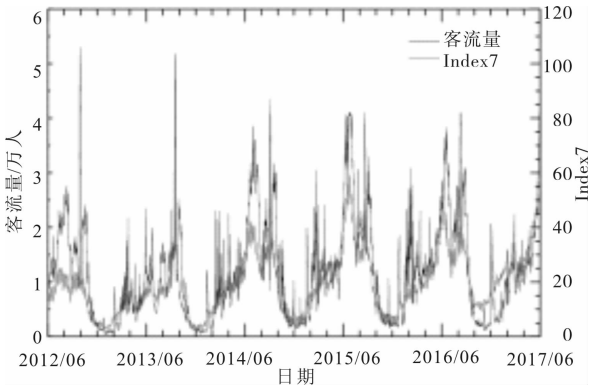

图 1 九寨沟旅游客流量与 Index7 序列图

2.3 HHT 变换噪声处理

根据 HHT 变换,首先利用 R 语言对合成搜索指数序列 Index7 进行 EMD 分解,该软件的 EMD 程序包主要根据 Huang 等的 EMD 算法编程. 经过分解获得 9 条 IMF 函数序列和一条残差,如图 2 所示.

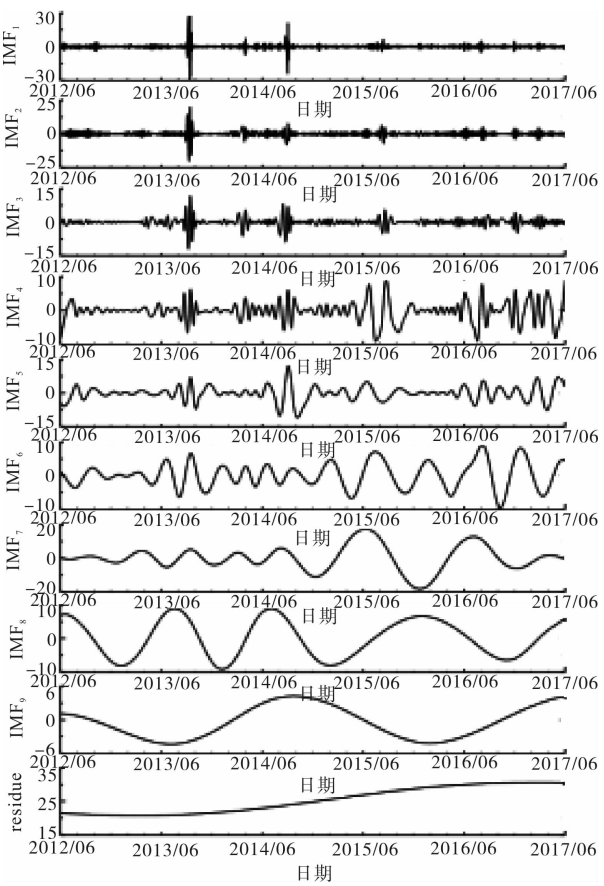

图 2 合成搜索指数各 IMF 分量图

各分量中 IMF_1 具有最高波动频率,接下来各分量的波动依次下降,残差则接近单调函数,表示合成搜索指数的长期变动趋势. 在 EMD 分量分解基础上对各 IMF 分量作 Hilbert 谱分析,得到各分量的波动频率,如图 3 所示.

如图 3 所示,IMF_1 分量的瞬时频率在时间轴范围内无法变现出明显频率带,各数据点较为均匀地分布在整个高频区域,将该分量辨别为高频噪声. 得到其余各分量与残波的有效信号部分,其加总之和作为 HHT 去除噪声后的有效序列,表示为 Index7_hht.

2.4 模型设定和训练

笔者在考虑高频随机噪声干扰的基础上,经过 HHT 变换将高频随机噪声分离后,得到合成搜索指数的有效序列 Index7_hht,对九寨沟旅游客流量进行预测. 便于对噪声处理的有效性进行评价,笔者分别选择时间序列 ARMA 模型,未经噪声处理的网络搜索数据预测 ARMAX 模型和 BP 神经网络模型作为基准模型.

$$visitor_t = c + \alpha_1 visitor_{t-1} + \alpha_2 visitor_{t-7} + \mu_t, \quad (7)$$

$$visitor_t = c + \beta_1 Index7_{t-2} + \mu_t, \quad (8)$$

$$visitor_t = c + \gamma_1 Index7\_hht_{t-2} + \mu_t. \quad (9)$$

其中:$visitor_t$ 代表 t 时期九寨沟旅游客流量,$Index7_{t-2}$代表利用 QCR 搜索词条链合成的搜索指数,它相对于客流量有 2 天的提前期,$Index7_hht_{t-2}$代表经过 HHT 变换噪声处理后的合成搜索指数序列部分. ADF 平稳性检验显示,九寨沟客流量和合成搜索指数序列在 5% 显著性水平下平稳,且 Grager 非因果检验显示两序列互为 Granger 原因.

从表 2 可以看出,基于网络搜索的 M(2)拟合效果略优于时间序列模型,而经过 HHT 分解去噪后的有效网络搜索序列的 M(3)拟合效果则比未分解的模型效果有进一步改善. 所有的模型系数均在 5% 水平下显著. 具体的模型拟合结果如式(10)~(12)所示:

$$\left.\begin{aligned} visitor_t &= 12.673 + 0.784\ visitor_{t-1} + \\ &\quad 0.178\ visitor_{t-7} + \mu_t, \\ \mu_t &= \varepsilon_t + 0.072\ \varepsilon_{t-1}. \end{aligned}\right\} \quad (10)$$

$$\left.\begin{aligned} visitor_t &= 5.891 + 1.275\ Index7_{t-2} + \mu_t, \\ \mu_t &= 0.787\ \mu_{t-1} + 0.135\ \mu_{t-7} + \varepsilon_t + \\ &\quad 0.081\ \varepsilon_{t-7}. \end{aligned}\right\} \quad (11)$$

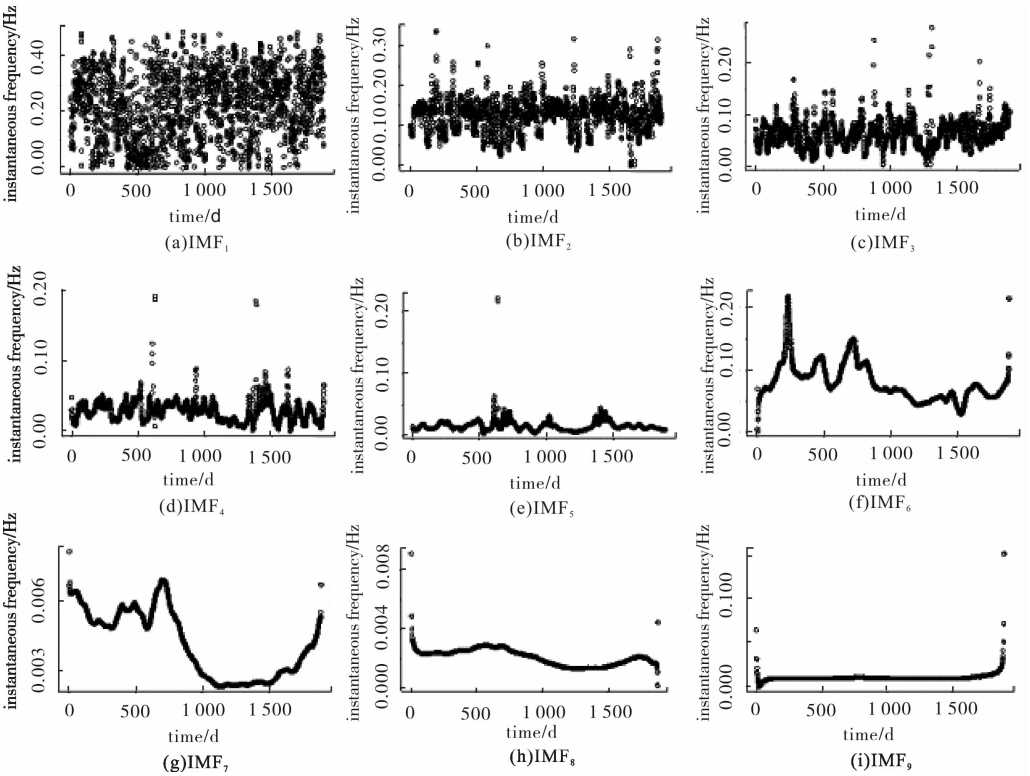

图 3　各分量 Hilbert 瞬时频率谱

表 2　模型拟合结果

数据类型	$visitor_{t-1}$	$visitor_{t-7}$	$Index7_{t-2}$	$Index7\_hht_{t-2}$	$AR(1)$	$AR(2)$	$AR(3)$	$AR(7)$	$MA(1)$	$MA(7)$
时间序列模型	0.784***	0.178***							0.072**	
Index7			1.275***		0.787***			0.135***		0.081***
HHT 变换				1.191***	2.051***	−1.746***	0.674***		0.624***	
R^2	0.875			0.884				0.994		
AIC 值	5.239			5.166				2.080		
SC 值	5.252			5.183				2.101		
DW 值	2.001			2.030				1.982		

注:***、***分别代表显著性水平 5%、10%.

$$\left.\begin{aligned} visitor_t &= 8.081 + 1.191\ Index7\_hht_{t-2} + \mu_t, \\ \mu_t &= 2.051\ \mu_{t-1} + 1.746\ \mu_{t-2} + 0.674\ \mu_{t-3} + \\ &\quad \varepsilon_t + 0.624\ \varepsilon_{t-1}. \end{aligned}\right\} \tag{12}$$

拟合结果显示,网络搜索与九寨沟客流量之间存在正向相关关系,网络搜索的增加会使未来时期内九寨沟客流量上升,这是因为搜索量的增加表明了网民对于目的地的关注增加,一方面可能是已经计划出行的游客在为即将到来的出行做准备,另一方面可能是受到外界影响开始关注九寨沟并有可能促成未来的实际出行. 结果显示网络搜索增加 1 倍将会引起未来客流量增长约 1.275 倍,而在式(12)中这一影响效果有所减弱,降低为 1.191 倍,说明在式(11)中利用网络搜索解释未来客流量增加时,会有一定的高估效应,这与多年来一直讨论的"大数据傲慢"现象是一致的,即网络搜索在预测方面的高估问题. 而经过 HHT 分解去噪后剩余的有效信号部分使得这一高估效应得到了一定的调整.

2.5　模型预测

通过上述的模型训练,发现经过 HHT 噪声过滤后的网络搜索数据预测模型效果明显优于其他两个模型. 为了验证模型的预测效果,在此基础上对预测期内九寨沟客流量进行预测,并与基准模型进行对比. 预测效果评估指数有很多,笔者选择 MAPE(Mean Absolute Percentage Error)和 RMSE(Root Mean Square Error)预测效果评估指数衡量 QCR – HHT 模型和基准模型的预测误差. 基本测

算公式如式(13)和(14)所示，相关测算值如表 3 所示，其中 P_i 代表预测值，Y_i 代表原值.

$$\mathrm{MAPE} = \frac{1}{n} \times \sum_{i=1}^{n} \left| \frac{P_i - Y_i}{Y_i} \right| \times 100\% \ , \quad (13)$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{n} \times \sum_{i=1}^{n} (P_i - Y_i)^2} \ . \quad (14)$$

表 3　未来 7 天预测结果和预测误差

日期	原值	预测值				MAPE/%				RMSE			
		时间序列	Index7	BP 神经网络	QCR－HHT	时间序列	Index7	BP 神经网络	QCR－HHT	时间序列	Index7	BP 神经网络	QCR－HHT
2016/11/1	11.09	12.98	12.78	13.02	11.60	17.03	15.31	17.46	4.60	1.89	1.70	1.94	0.51
2016/11/2	13.40	15.06	14.71	11.30	12.08	12.40	9.78	15.68	9.88	1.66	1.31	2.10	1.32
2016/11/3	13.96	15.98	14.85	13.55	13.22	14.48	6.38	2.97	5.33	2.02	0.89	0.42	0.74
2016/11/4	14.48	12.37	15.69	13.61	13.75	14.57	8.38	5.96	5.03	2.11	1.21	0.86	0.73
2016/11/5	14.72	16.56	15.97	14.28	14.24	12.49	8.46	2.98	3.28	1.84	1.25	0.44	0.48
2016/11/6	10.00	14.74	13.83	14.11	10.32	47.48	38.40	41.13	3.28	4.75	3.84	4.11	0.33
2016/11/7	7.79	10.00	9.64	10.90	8.61	28.49	23.75	39.99	10.54	2.22	1.85	3.11	0.82

从未来 7 期的预测结果和预测误差可知，时间序列的预测效果最差，整体预测误差较高且不稳定，加入网络搜索数据后的预测误差相对有所下降，但与 BP 神经网络相比，误差相当，说明与机器学习算法相比，网络搜索数据的预测能力并没有太大优势，这也说明了粗糙的大数据对预测应用的能力和提升作用有限. 与以上 3 个基准模型相比，经过 QCR－HHT 处理的网络搜索数据预测效果有显著提升，未来 7 期的预测误差有明显下降. 未来 15 期九寨沟客流量预测结果值如图 4 所示，基于 QCR－HHT 模型的整体预测值与原客流量序列之间的误差最小.

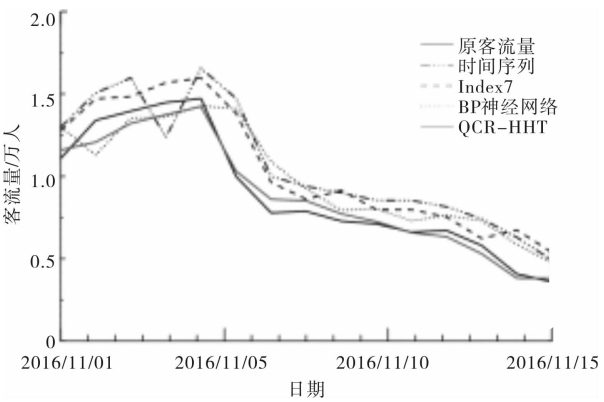

图 4　未来 15 期预测值对比

本文预测期共包含 279 天，整体预测误差如表 4 所示. 整体预测效果评价中，网络搜索数据预测的模型 MAPE 最高，其次为 BP 神经网络，说明在未来时期内搜索数据的预测效果最差，甚至劣于时间序列模型. 明显地，利用搜索链和 HHT 噪声处理后，搜索数据的预测精度最高，279 天整体预测误差控制在 7.76%，相比于未处理噪声前的预测模型，预测精度提高 64%，验证了本文方法可以较好提升网络搜索数据在九寨沟旅游客流量中的预测能力.

表 4　整体预测效果评价

模型	MAPE/%	RMSE
时间序列	19.27	2.03
Index7	21.59	1.92
BP 神经网络	20.92	2.12
QCR－HHT	7.76	0.54

3　结论

笔者针对网络搜索数据在预测中噪声干扰的问题，提出了运用 QCR 搜索链这种根据网民需求图谱来提取预测词，并利用 HHT 对网络搜索数据进行噪声处理的方法. 该方法一方面改进了传统利用推荐法提取搜索词的技术，另一方面结合 EMD 分解序列并根据 Hilbert 谱分析确定各分量序列的频率谱，识别高频噪声并分离. 以九寨沟客流量预测为例，引入时间序列、传统网络搜索预测模型和 BP 神经网络模型作为基准模型，对客流量预测效果进行了实证检验和对比分析.

经过分析得到以下研究结论：1）搜索链技术以网民的搜索路径为关键词提取原则，相比于机器推荐，可以对预测目标有更高的相关性；2）网络搜索数据中由于用户本身搜索的特点，数据中包含大

量噪声,这些噪声会影响网络搜索数据的预测效果,甚至导致预测失败;3)HHT 变换可以很好地适用旅游客流量这种非线性非平稳数据,经过噪声处理,对网络搜索数据的预测效果有显著提升.

参考文献:

[1]王兰梅,陈崇成,叶晓燕,等. 网络搜索数据和 GWO – SVR 模型的旅游短期客流量预测[J]. 福州大学学报(自然科学版),2019,47(5):598 – 603.

[2]周晓丽,唐承财. 基于网络搜索大数据的 5A 级景区客流量预测分析[J]. 干旱区资源与环境,2020,34(3):204 – 208.

[3]王耀君,高扬,王耀青. 基于网络搜索指数的股票市场微观结构特征[J]. 北京理工大学学报(社会科学版),2018,20(5):54 – 62.

[4]杨超,姜昊,雷峥嵘. 基于文本挖掘和百度指数的汇率预测[J]. 统计与决策,2019,35(13):85 – 87.

[5]HUANG N E,SHEN Z,LONG S R,et al. The empirical mode decomposition and the Hilbert spectrum for nonlinear and non – stationary time series analysis[J]. Proceedings of the Royal Society Series A:mathematical,physical and engineering sciences,1998,454(1971):903 – 995.

[6]王长琼,曹乜蜻,王艳丽,等. 基于融合百度指数的电商订单量组合预测研究[J]. 计算机工程与应用,2018,54(12):219 – 225.

[7]李凤岐,李光明. 基于搜索行为的经济指标预测方法[J]. 计算机工程与应用,2017,53(6):215 – 222.

[8]BANGWAYO – SKEETE P F,SKEETE R W. Can Google data improve the forecasting performance of tourist arrivals: Mixed – data sampling approach[J]. Tourism Management,2015,46:454 – 464.

[9]HÖPKEN W,EBERLE T,FUCHS M,et al. Search engine traffic as input for predicting tourist arrivals[C]. Jönköping:Information and Communication Technologies in Tourism,2018:381 – 393.

[10]VOLCHEK K,LIU A Y,SONG H Y,et al. Forecasting tourist arrivals at attractions: Search engine empowered methodologies[J]. Tourism Economics,2019,25(3):425 – 447.

[11]LI X X,WU Q,PENG G,et al. Tourism forecasting by search engine data with noise – processing[J]. African Journal of Business Management,2016,10(6):114 – 130.

[12]KEDADOUCHE M,THOMAS M,TAHAN A. A comparative study between Empirical Wavelet Transforms and Empirical Mode Decomposition Methods:Application to bearing defect diagnosis[J]. Mechanical Systems and Signal Processing,2016,81:88 – 107.

[13]ZHANG Y,ZHANG C,SUN J,et al. Improved wind speed prediction using empirical mode decomposition[J]. Advances in Electrical and Computer Engineering,2018,18(2):3 – 10.

[14]李晓炫,吕本富,曾鹏志,等. 基于网络搜索和 CLSI – EMD – BP 的旅游客流量预测研究[J]. 系统工程理论与实践,2017,37(1):106 – 118.

[15]陆利军. 基于网络搜索指数和 EMD – ARIMA – BP 组合模型的游客量预测:以张家界为例[J]. 吉首大学学报(社会科学版),2019,40(1):138 – 150.

(责任编辑:王彦江)

Tourist Flow Forecasting in Jiuzhaigou Valley Based on the Search Engine with Denoising and QCR – HHT Model

LI Xiaoxuan[1],WU Qi[2]

(1. School of Economics,Fuyang Normal University, Fuyang, Anhui 236037, China;2. School of Physics and Electronic Engineering,Fuyang Normal University, Fuyang, Anhui 236037, China)

Abstract:The forecast of tourist flow can make up the impact from the strong periodic and volatile passenger flows on the scenic spots and tourists, so the limited tourism resources can be reasonably scheduled and allocated in advance. With the noise of the search engines data taken into consideration, this paper proposes to build a novel prediction model with QCR(Query Chain Retrieve) and HHT to forecast the daily tourist flow of Jiuzhaigou. It is found that QCR – HHT prediction model is the best and the prediction accuracy is remarkably improved, compared with the traditional regression model, ARMAX model with search engines data and BP neural network model. The use of the QCR – HHT prediction model with denoising is of great help to forecast the tourist flow more accurately, enabling the management departments of scenic spots to make more efficient and accurate decisions.

Key words:search engine data, query chain retrieve,denoising, Hilbert frequency spectrum analysis