北京大学学报(自然科学版) 第 57 卷 第 5 期 2021 年 9 月
Acta Scientiarum Naturalium Universitatis Pekinensis, Vol. 57, No. 5 (Sept. 2021)
doi: 10.13209/j.0479-8023.2021.020

基于统计年鉴和网络大数据的房屋竣工面积估算

原雯[1] 王君[1] 申鸿怡[1] 王新民[2,†]

1. 北京大学前沿交叉学科研究院, 北京 100871; 2. 北京大学数学科学学院, 北京 100871;
† 通信作者, E-mail: wangxinmin@pku.edu.cn

摘要 选择北京市年鉴中的若干数据指标, 构建经济社会因子体系, 采用偏最小二乘回归、LASSO 回归和 RBF 神经网络 3 种模型, 对 2017 和 2018 年北京市房屋竣工面积进行预测。由于各年鉴数据统计渠道和指标粒度不同, 且 2019 年建筑业部分指标数据的公布存在延迟, 难以用模型拟合的方式对该年度竣工面积做出估计。因此, 利用爬虫技术获取高质量数据, 并深入挖掘网络数据中的信息, 通过互联网大数据估算北京市房屋竣工面积。首先, 建立基于网络大数据的建筑数据获取框架, 通过调用服务接口和关键字搜索等技术, 爬取北京地区 8 类建筑物的属性数据; 然后, 利用正则表达式和条件过滤, 对网页返回的 HTML 非结构化数据进行抽取和清洗; 最后, 对 2019 年北京市房屋竣工面积及各功能分区的竣工面积做出估算。

关键词 竣工面积; 回归分析; 网络爬虫; 模板抽取

Estimation of Area of Completed Houses Based on Statistical Yearbooks and Online Big Data

YUAN Wen[1], WANG Jun[1], SHEN Hongyi[1], WANG Xinmin[2,†]

1. Academy for Advanced Interdisciplinary Studies, Peking University, Beijing 100871; 2. School of Mathematical Sciences, Peking University, Beijing 100871; † Corresponding author, E-mail: wangxinmin@pku.edu.cn

Abstract The authors select several indicators in the Beijing Yearbook to construct an economic and social factor system, and uses partial least squares regression, LASSO regression and RBF neural network models to predict the area of completed buildings in Beijing in 2017 and 2018. However, considering the difference of statistical channels and granularity of the yearbook indicators, and the delay in the release of some indicator data for the construction industry in 2019, it is hard to estimate the area of the year by model fitting. Therefore, crawler technology is used to obtain high-quality data and dig deep to obtain information of online big data to estimate the completed area. Firstly, a web-based building data acquisition framework is established, to crawl the attribute data of eight types of buildings in Beijing by calling service interface, keyword search and other technologies. Secondly, regular expressions and conditional filtering are used to extract and clean the HTML data returned by web pages. Finally, the area of completed houses in Beijing and the area of each functional partition in 2019 are estimated.

Key words area of completed houses; regression analysis; web crawler; template extraction

作为国民经济中重要的物质生产部门, 建筑业与国家经济的发展、人民生活的改善息息相关。作为建筑业的关键指标之一, 房屋竣工面积[1]常用于能源消耗和房地产价格估计等课题。曹爱丽等[2]采用趋势拟合与相关分析, 研究城郊温差与城市人口、GDP、能源消耗量、建成区面积和房屋竣工面积等各项城市发展指标的关系。王北星等[3]用人均生产总值、商品房竣工面积和竣工房屋造价等定义经济发展因子, 建立吉林省商品房价格影响因素的因子分析决策模型。因此, 对房屋竣工面积的估计

国家重点研发计划(2018YFC0704300)资助
收稿日期: 2020–06–29; 修回日期: 2021–01–18

和测算具有重要的现实意义。

除通过国家统计报表得到官方数据外，有研究者采用其他技术手段对建筑面积进行估算。匡文慧等[4]采用基于知识规则的遥感影像分类方法以及空间网格技术，通过建立网格内城市建筑用地面积所含的阴影面积比例与容积率的关系，计算每个网格的建筑容积率，进而估算城市不同土地利用类型的建筑面积。

信息技术的高速发展使得互联网承载着大量多领域、多维度和多粒度的数据。随着大数据时代的到来，人们希望深入挖掘网络数据中的信息，为决策提供支持，因此大规模异构数据的高效采集及获取方法受到广泛关注。在互联网领域，爬虫一般指在众多公开网站或网页上抓取数据的相关技术，能够按照一定的规则，自动地抓取万维网信息，能够在信息超载时有效地提高获取效率。聚焦爬虫[5]是一种定向抓取相关网页资源的技术，与通用网络爬虫[6]不同，聚焦爬虫不追求覆盖度，而将目标定为抓取与某一特定主题内容相关的网页，为面向主题的用户查询提供数据资源。针对与主题相关的网络资源的特点，研究者们设计多种针对网页的爬取策略，以期提升爬取效率及质量。常见的爬取策略有深度优先搜索(Depth-First Search)策略[7]、广度优先搜索(Breadth-First Search)策略[7]、最佳优先搜索(Best-First Search)策略[8]和 PageRank 策略[9]等。

网络数据采集指通过网络爬虫或网站公开 API 等方式，从网站上获取数据信息，可将非结构化数据或半结构化数据从网页中提取出来，并以结构化的方式存储为统一的本地数据文件。这一技术在国内外很多行业广泛使用。周中华等[10]为快速获取微博中的数据，开发一款支持并行的微博数据抓取工具，并应用于流感问题分析。范超等[11]通过网络爬虫和文本挖掘技术，探索 P2P 网络借贷这一重要新经济业态的风险甄别问题。Shemshadi 等[12]创建一组工具，用于从给定的数据源中采集物联网数据，并为人机用户提供动态物联网数据的实时搜索服务。Young 等[13]利用网络数据采集的样本，评估特定地区是否有农业活动。

本文基于统计年鉴数据和网络大数据，研究房屋竣工面积的估算方法。在处理统计年鉴数据时，我们发现从国家统计报表采集的官方数据存在延迟。截至目前，北京市房屋竣工面积这一指标仅更新到 2016 年。因此，本文选择国家统计年鉴中反映城市用地需求、房地产行业发展水平及社会发展水平的若干指标，对比分析多种模型，对 2017 和 2018 年北京市房屋竣工面积做出预测。同时，考虑到各年鉴数据统计渠道及指标粒度不同，且国家统计局发布的主要城市年度数据中部分指标的公布存在延迟(如截至目前 2019 年北京市城市建设用地面积等数据暂未公布)，无法用模型拟合的方式对该年度竣工面积进行预测。因此，我们希望借助互联网数据，通过网络爬虫技术获取北京地区 8 类建筑物的属性数据，从而对 2019 年北京地区房屋竣工面积做出估算。

1 城市房屋竣工面积预测研究

对城市房屋竣工面积进行预测，有助于推断未来房屋用地规模，为城市发展规划提供重要的决策支持。因此，对城市房屋竣工面积进行科学的预测成为城市建筑发展的核心问题之一。

政府每年公布的年鉴中会提供本行政区经济、人口和行业的综合数据。房屋竣工面积受城市经济、社会和环境等多方面因素的综合影响，相互之间不是简单的线性关系，而是一种复杂的非线性关系。为容纳多种社会经济因素的影响，我们使用年鉴提供的多种影响因子对房屋竣工面积进行预测，并对结果进行综合分析。

1.1 地区概况

选取北京市作为实验区。北京市地处中纬度地带，是国家的政治、经济和文化中心，是京津冀地区城市群的核心城市。20 世纪末以来，北京城市化进程大大加速，城市空间布局在现有中心城区基础上向东南西北 4 个方向拓展，城市建成区规模迅速扩张。北京市是中国快速城市化区域的典型代表，以北京市作为实验区，可为其他大型城市的建筑竣工面积预测提供参考借鉴。

1.2 指标选择及数据来源

竣工指房屋建筑工程已按工程承包合同和设计要求全部完工，达到居住和使用条件，经验收鉴定合格并正式交付使用的状态。民用房屋一般是将房屋的土建工程及其附属水、暖、电、卫工程和通风、电梯等设备安装全部完成视为竣工。工业及科研等生产性房屋建筑，在厂房和作为其组成部分的生活间、操作间和烟囱等土建工程以及水、暖、电、卫、通风等工程(不包括生产设备安装和工艺

管线敷设)全部完成, 经验收鉴定合格后, 作为竣工房屋。房屋竣工面积指报告期内房屋建筑按照设计要求已全部完工, 达到居住和使用条件, 经验收鉴定合格或达到竣工验收标准, 可正式移交使用的各栋房屋建筑面积的总和。

结合城市发展规律, 以北京市 2007—2018 年的序列资料为数据基础, 构建影响北京市房屋竣工面积的经济社会因子体系。从宏观层面理解, 北京市房屋竣工面积的影响因素有城市用地需求、房地产行业发展水平及社会发展水平, 其中社会发展水平包括经济发展水平和人民生活水平。本文从国家统计年鉴中选取若干指标(表 1), 分析北京市房屋竣工面积影响因素。

从年鉴中获取的数据存在部分缺失, 其中需要填补的数据指标为城市建设用地面积(km^2)、房地产业增加值(亿元)、居民消费水平(元)和居民人均可支配收入(元), 根据指标数据特征, 可以选择不同的方法填补缺失值。

1.2.1 利用多项式拟合填充

指标X_1(城市建设用地面积)中 2010 年的数据值缺失, 因此在 2006—2014 年的数据中, 利用多项式拟合填补缺失值, 图 1 为多项式拟合结果示意图。所用一次、二次多项式的均方根误差分别为 0.23 和 0.17, 故二次多项式拟合效果好于一次多项式, 该指标的填充结果为 1377.11 km^2。

表 1 影响因子及符号表示

Table 1 Impact factors and symbol representation

影响因素	影响因子(指标)	符号
城市用地需求	城市建设用地面积(km^2)	X_1
	城市人口密度(人/km^2)	X_2
	非农业人口(万人)	X_3
房地产行业发展水平	房地产开发住宅投资额(亿元)	X_4
	住宅商品房平均销售价格(元/m^2)	X_5
	房地产业增加值(亿元)	X_6
	全社会固定资产投资(亿元)	X_7
社会发展水平	第三产业增加值(亿元)	X_8
	人均地区生产总值(元/人)	X_9
	居民人均可支配收入(元)	X_{10}
	居民消费水平(元)	X_{11}
	地方生产总值(亿元)	X_{12}
	地方财政一般预算收入(亿元)	X_{13}
	地方财政税收收入(亿元)	X_{14}
被解释变量	房屋竣工面积(万m^2)	Y

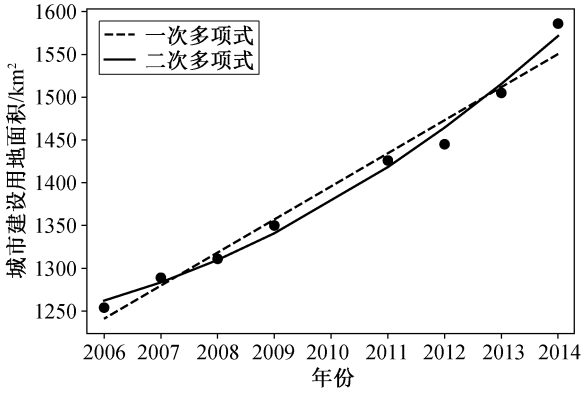

图 1 数据填充

Fig. 1 Data padding

1.2.2 利用线性拟合填充

指标X_{10}(居民人均可支配收入)只有 2013 年及以后的数据, 由于年鉴中城镇居民人均可支配收入指标值在 2002 年以后的数据均可获得, 由此可计算得到城镇居民人均可支配收入与居民人均可支配收入的比例在 1.09 左右。将城镇居民人均可支配收入除以相应的比例, 可得居民人均可支配收入在 2013 年之前各年份的估计值。

指标X_6(房地产业增加值)和X_{11}(居民消费水平)皆缺失 2018 年的数据, 本文也采用类似的方法, 拟合线性模型, 并填补缺失值。

填补缺失值后, 得到北京市 2007—2018 年各指标的真实值。为消除指标的不同量纲对数据的影响, 对每项指标按式(1)进行归一化处理, 得到各指标的相对值, 并基于该数据集进行模型拟合。

$$X' = \frac{X_{\text{real}} - X_{\min}}{X_{\max} - X_{\min}}。\tag{1}$$

式中, X'表示归一化后指标X的相对值, $X_{\max}$和$X_{\min}$分别表示指标X的最大值和最小值, X_{real}表示X的真实值。在归一化的结果中, 指标X相对值的范围在 0~1 之间, 取值为 0 表示此项数据为$X_{\min}$, 取值为 1 表示此项数据为$X_{\max}$。表 2 列出填补后 2011—2018 年各指标的相对值X'。

1.3 回归分析

1.3.1 初步分析

以北京市房屋竣工面积Y为因变量, 选择部分影响因子构建房屋竣工面积与各影响因子的一元线性回归预测方程, 表 3 和图 2 分别展示各方程拟合结果。从图 2 可以看出, 线性回归预测方程的拟合

表 2　2011—2018 年填补后数据

Table 2　Data after filling in 2011–2018

年份	2018	2017	2016	2015	2014	2013	2012	2011
X_1	0.29	0.24	0.24	0.18	1.00	0.49	0.12	0.00
X_2	0.00	0.02	0.02	1.00	0.96	0.89	0.81	0.72
X_3	0.88	0.99	1.00	0.99	0.86	0.62	0.32	0.00
X_4	1.00	0.17	0.75	0.66	0.55	0.24	0.00	0.38
X_5	1.00	0.85	0.59	0.31	0.14	0.11	0.05	0.00
X_6	1.00	0.88	0.76	0.46	0.32	0.34	0.22	0.00
X_7	1.00	0.83	0.71	0.57	0.40	0.38	0.16	0.00
X_8	1.00	0.84	0.68	0.49	0.35	0.24	0.11	0.00
X_9	1.00	0.81	0.62	0.42	0.31	0.22	0.10	0.00
X_{10}	1.00	0.84	0.69	0.57	0.45	0.33	0.10	0.00
X_{11}	1.00	0.86	0.72	0.39	0.28	0.19	0.09	0.00
X_{12}	1.00	0.84	0.67	0.48	0.36	0.25	0.12	0.00
X_{13}	1.00	0.87	0.75	0.62	0.37	0.24	0.11	0.00
X_{14}	1.00	0.85	0.75	0.66	0.47	0.31	0.13	0.00
Y	—	—	0.00	0.48	1.00	0.36	0.18	0.39

表 3　部分影响因子线性回归结果

Table 3　Linear regression results of some factors

自变量	因变量	拟合方程
北京市非农业人口(万人)	北京市房屋竣工面积 Y (万 m^2)	Y=–0.506X_3+0.6634
北京市全社会固定资产投资(亿元)		Y=–0.627X_7+ 0.6189
北京市地方生产总值(亿元)		Y= –0.685X_{12}+ 0.6054

优度均不理想，说明房屋竣工面积受多方面因素综合影响。

1.3.2 拟合与预测

本文使用的数据集为表 1 中各个指标在 2005—2016 年补全后的值。通过计算表 1 中各因子组成矩阵 $\boldsymbol{X}$ 的 $\boldsymbol{X}'\boldsymbol{X}$ 行列式，发现结果非常接近 0，我们认为解释变量间可能存在多重共线性。因此，建立偏最小二乘回归模型、LASSO 回归模型和 RBF 神经网络来拟合房屋竣工面积及各个解释变量之间的关系。

模型 1：偏最小二乘回归(Partial Least Squares Regression，PLS)[14]。此模型是近年来应实际需要而产生的一种具有广泛适用性的多元统计分析方法，具有主成分分析、典型相关分析和线性回归分析等方法的特点，能有效地解决变量间存在多重共线性的问题[15]。偏最小二乘回归分析使用成分提取的方法，对输入特征进行重组而不是剔除，考虑因变量与自变量之间的线性关系，并选择对自变量和因变量解释性最强的综合变量，排除噪声干扰，既能保证多重共线性问题的消除，又能保证模型的稳定。图 3 展示偏最小二乘回归模型的拟合效果。

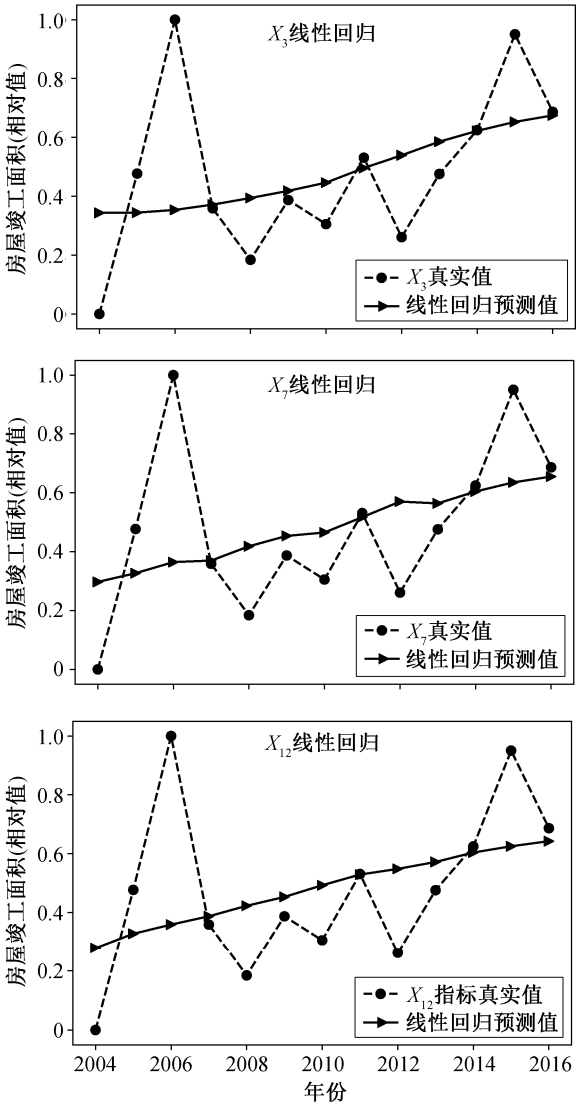

图 2　部分影响因子一元回归拟合结果

Fig. 2　Regression fitting results of some factors

模型 2：LASSO (Least Absolute Shrinkage and Selection Operator)[16]。LASSO 回归是一种压缩估计方法，通过构造一个惩罚函数，得到一个较为精炼的模型，使得它可以压缩一些回归系数，即强制系数的绝对值之和小于某个固定值；同时，设定一些回归系数为零。该方法保留了子集收缩的优点，是一种适用于复共线性数据的有偏估计模型。LASSO 回归的过程可视为变量选择，最终将不重要的变量系数值取值为 0，保留的变量为对房屋竣工面积有实际影响的变量。

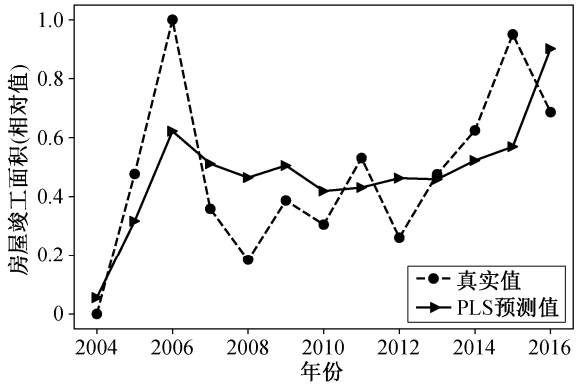

图 3 偏最小二乘回归模型拟合结果

Fig. 3 Partial least squares regression model fitting result

使用表 1 中 14 个影响因子, 建立 LASSO 回归模型, 对 Y 进行拟合, 拟合效果如图 4 所示。

模型 3: RBF(Radial Basis Function)神经网络[17]。RBF 神经网络具有全局逼近性质和最佳逼近性能, 可以较好地挖掘和揭示复杂非线性系统的实际结构。图 5 展示 RBF 神经网络的网络结构, 包括输入层、隐含层和输出层。其中, 输入层将输入矢量直接映射到隐空间, 起到传输信号的作用; 隐含层含若干隐单元节点, 节点数量视求解的具体问题而定, 隐含层可对网络输入做非线性映射, RBF 是一个径向对称、双方向衰减的非负非线性函数; 输出层则对隐含层的输出采用线性加权求和的映射模式。

由此可见, RBF 神经网络是线性和非线性的有机统一, 即从输入层到隐含层是非线性映射, 采用非线性优化策略, 学习速度较慢; 从隐含层到输出层是线性变换, 采用线性优化策略, 学习速度较快。与 BP 网络相比, RBF 神经网络的学习速度大大

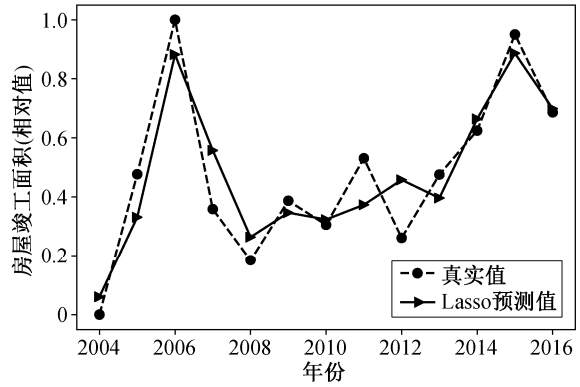

图 4 多元回归模型拟合结果

Fig. 4 LASSO regression model fitting result

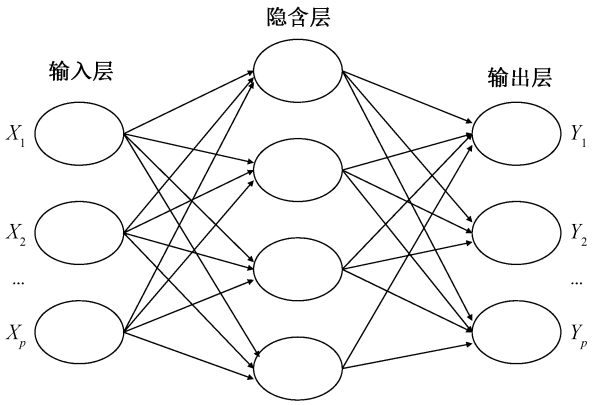

图 5 RBF 神经网络结构

Fig. 5 RBF neural network structure

加快, 并能够有效地避免陷入局部最小的不足。图 6 为 RBF 神经网络拟合结果。

1.4 模型预测及评估

将各解释变量 2017—2018 年的指标值代入模型, 可得到 2017 和 2018 年北京市房屋竣工面积的预测值, 如表 4 所示。由于 2019 年指标值缺失过多, 本文不对其进行预测。

选择均方根误差(RMSE)作为模型评估指标:

$$\mathrm{RMSE}=\sqrt{\frac{1}{m}\sum_{i=1}^{m}\left(y^{(i)}-\widehat{y^{(i)}}\right)^2}\ 。\tag{2}$$

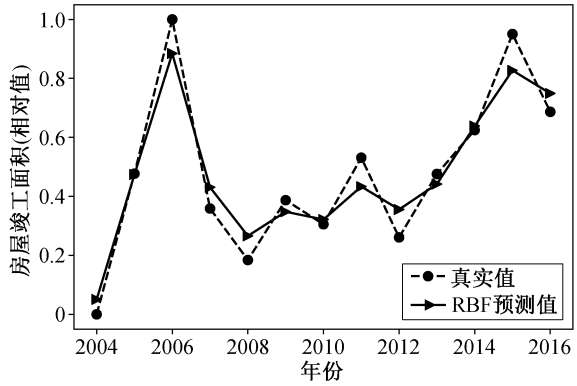

图 6 RBF 神经网络拟合结果

Fig. 6 RBF neural network fitting result

表 4 北京市房屋竣工面积预测值

Table 4 Predictions of area of completed buildings in Beijing

模型	2017 年预测值/万 m^2	2018 年预测值/万 m^2
偏最小二乘回归模型	2803.64	2609.23
LASSO 模型	2792.67	2583.42
RBF 神经网络	2731.31	2535.71

计算标准化还原后预测值与真实值的均方根误差(RMSE), 结果如表 5 所示。

结合各个模型的拟合图像及均方根误差可以看出, 多元线性回归模型、偏最小二乘回归模型和RBF 神经网络均能较好地拟合房屋竣工面积的变化趋势, 其中 RBF 神经网络的拟合效果最好。据此, 选择 RBF 神经网络给出的 2731.31 和 2535.71 万 m^2 分别作为 2017 和 2018 年北京市房屋竣工面积的预测值。

2 基于网络大数据的建筑数据获取与研究

基于网络大数据的建筑数据获取流程包括建筑名称、建筑面积和建成时间的获取。通过调用百度地图的服务获取建筑名称数据; 建筑面积和建成时间的获取依赖于建筑名称的获取结果, 数据来源于百度百科基本信息页、百度关键字搜索以及谷歌关键字搜索。图 7 展示基于网络大数据的建筑数据获取流程。

表 5　模型评估

Table 5　Model evaluation

模型	RMSE
偏最小二乘回归	260.56
LASSO 回归	232.79
RBF 神经网络	162.39

2.1 建筑名称获取

百度地图提供多种服务, 如定位、地图、搜索和鹰眼轨迹等基础服务(调用网址为 http://api.map.baidu.com/place/v2/search)。用户通过调用百度地图提供的服务接口, 发起检索请求, 可得到格式化的检索结果。本文使用百度地图的地点检索服务, 获取地点(POI)的基本信息, 如名称、省、市、区、具体地址和经纬度等。地点检索服务涉及两个必要的请求参数——关键字和区域, 其中关键字决定检索请求的地点类别, 区域参数限定检索的地理范围。

根据检索区域请求参数的不同, 该服务有行政区划检索、圆形区域检索、矩形区域检索 3 种。其中行政区划检索对应的区域请求参数为行政区划(如北京市), 圆形区域检索的区域请求参数为中心点经纬度以及检索区域的半径, 矩形区域检索要求输入待检索矩形区域的左上角和右下角经纬度来发出请求。由于在同一关键字请求时, 行政区划检索最多召回 400 个 POI, 使用矩形检索的方式, 因此将北京市切分成多个小矩形区域, 对每个小矩形区域进行检索, 以便扩大 POI 召回。

百度地图对地点按行业进行分类, 地点检索服务请求的关键字为行业标签, 因此地点检索服务的返回结果是一类地点在指定检索区域的信息。本文按照功能将建筑划分为住宅、写字楼、政府机关、商场、宾馆饭店、医院、学校及其他共 8 类。为获取大量且多样的 8 类建筑功能的建筑名称, 使用一

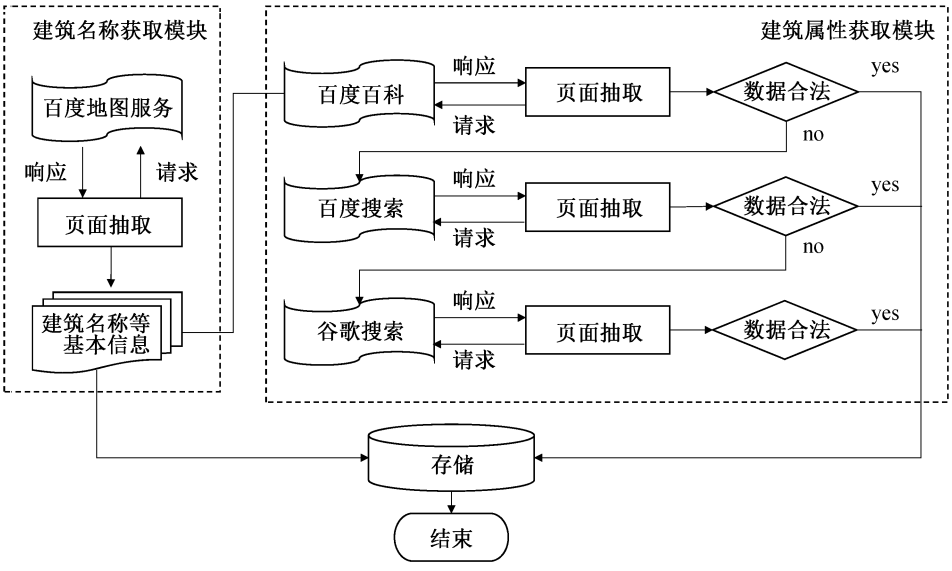

图 7　基于网络大数据的建筑数据获取流程

Fig. 7　Building data acquisition process based on online big data

表 6 建筑功能与检索请求关键字参数对应表

Table 6 Correspondence table of building function and retrieval request keyword parameters

建筑功能分类	检索请求关键字
住宅	住宅区
写字楼	写字楼
政府机关	政府机构
商场	购物, 购物中心, 百货商场
宾馆饭店	酒店
医院	医疗, 综合医院, 专科医院
学校	学校, 高等院校
其他	公司, 公司企业

级行业与二级行业分类为关键字, 以矩形检索的方式多次请求调用服务, 建筑功能与请求的关键字参数对应关系见表 6。

2.2 建筑属性获取

根据 2.1 节得到的建筑名称信息, 利用搜索引擎的搜索功能和网络数据爬取技术, 获取建筑面积和建成时间等建筑属性。

2.2.1 相关原理

网络数据爬取技术是从网页中提取所需信息的技术, 首先发送 HTTP 请求, 得到网页响应的结果, 该结果是非结构化的 HTML 源码; 然后使用网页解析技术、数据分析和抽取方法, 对非结构化数据解析, 得到所需的结构化的信息。

爬取的主要网页包括百度百科搜索、百度搜索及谷歌搜索。百度百科是全球最大的中文百科全书, 同样也是众人可协作的百科全书, 百度百科词条的内容是由对词条信息熟悉的领域专家在符合百科收录规范的前提下, 进行词条创建、内容编辑和提交。因此, 百科页面的内容权威性高, 规范性强, 从百科页面爬取的信息置信度高, 清洗难度小, 是爬取建筑物相关信息的首选网页。

当百科页面相关信息不全时, 便从搜索引擎关键字搜索返回的页面中, 提取建筑物相关信息。搜索引擎能够根据用户输入的查询关键字检索网页资源, 使用算法(如 Google 提出的 PageRank 算法[18])对网页进行排序, 并展示网页内容的摘要信息。不同的查询关键字返回的结果不同, 甚至查询关键字的微小差别也能导致搜索结果的巨大不同。因此, 对搜索引擎爬取的查询关键字进行限制, 查询关键字需要包含建筑物的名称信息以及想要爬取的建筑属性(如建筑面积和时间等)。另外, 同名建筑会造成查询结果的歧义性, 比如同一建筑开发商在不同城市开发的建筑, 由百度地图搜索的建筑名称相同, 但是区域信息、建筑面积和时间等其他建筑物信息都不相同, 因此还需在查询关键字中加入“地区”信息, 从而对查询内容进行消歧。

最终确定的用于搜索引擎爬取的查询关键字为“地区 建筑名称 建筑属性”。爬虫目标为建筑物的相关属性内容抓取, 搜索引擎返回结果为与查询关键字相关的摘要信息, 百度搜索返回的 HTML 数据中, 会将命中的关键字标红, 从而更准确地定位到待抽取位置。对于返回结果的处理方法是使用关键字提取技术对搜索引擎返回的摘要信息进行抽取, 而不是访问搜索引擎返回的 url 网页链接, 从而避免 url 跳转带来的未知风险。爬取时的请求网址和主要参数设置见表 7。本文合规地使用爬虫技术, 避免对相关网站的正常业务造成冲击, 同时也无意收集敏感数据, 并承诺对收集的数据保密。

2.2.2 基于半监督学习的模板抽取

一栋建筑涉及的指标有占地面积和建筑面积等, 不同的表达对应的面积含义也不同。本文关注的是建筑面积, 是建筑物的楼地面面积, 因此在百科 infobox 信息页和搜索引擎查询关键字时, 建筑属性即为建筑面积, 获取时的模板表述明确且一致。一栋建筑物的建成时间则相对复杂, 比如“竣工时间”、“建成时间”和“建成年代”等均对应该栋建筑物的建成时间。因此, 在挖掘建筑物建成时间的信息时, 使用半监督学习策略, 明确建成时间的提取模板。

半监督学习只使用有限的标记数据, 对大量的未标记数据进行学习。自举法[19]是半监督学习的一种, 采用少量的种子数据, 轮流发现学习。本文基于自举法的思想, 利用种子数据和网络资源进行建成时间表述的模板抽取。这里, 种子数据即为(建筑物地区+建筑名称, 对应的建成时间)构成的二元组数据。种子数据要具有较高的数据质量, 本文

表 7 建筑属性爬虫主要设置

Table 7 Crawler settings of building properties

抽取策略	请求网址	查询关键字
百科页面	https://baike.baidu.com/item/	建筑名称
百度搜索	https://baidu.com/s	地区 建筑名称 建筑属性
谷歌搜索	https://www.google.com/search	地区 建筑名称 建筑属性

利用网络上经相关专家整理发布的北京市住宅小区的详细信息(这些数据大多来源于专业的房屋信息网站), 从中随机采样少量样本作为模板抽取的种子数据。利用自举法抽取建成时间模板的算法执行步骤如下。

1) 构造数据质量较高的少量种子数据。

2) 利用百度百科搜索引擎, 将种子数据二元组作为查询关键字, 到网页中回标包含二元组的句子。

3) 由于搜索结果是按照匹配度排序的, 取前5条命中结果, 并利用百度搜索对摘要信息页中命中关键字标红的特点, 直接定位到回标的句子。

4) 对回标结果进行相关文本筛选, 并统计建成时间表述模板。

算法运行结果如图8所示, 将竣工时间、建筑年代、建造年代和建成时间作为建筑物时间属性的抽取模板。

2.2.3 正则表达式数据抽取

由于网络数据量巨大, 数据质量参差不齐, 所以存在建筑面积与建成时间表述形式不统一的现象。在获取HTML非结构化数据后, 使用正则表达式对数据进行抽取, 完成对建筑面积和建成时间数据的初步清洗。正则表达式[20]描述文本中隐含信息的共性, 常用于模板数据的抽取。

对建筑名称来说, 调用百度地图服务时, 得到的返回值具有随机性和具体性, 如XX商场-A座、XX小区-Y号楼、XX学校(YY校区), 返回的是更具体的父类+子类名称, 而将这些表述应用于百度(或谷歌)搜索建筑面积或建成时间时, 得到的是其父类的相关基本信息。因此, 将百度地图返回的建筑名称进行对齐处理, 将子类的名称转化为其父类的名称, 并对多个相同父类进行去重处理。

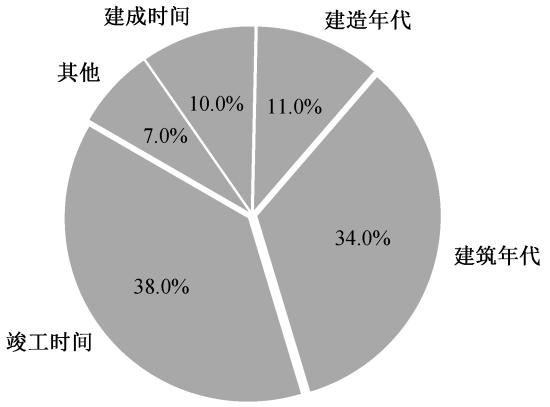

图 8　建成时间属性模板抽取比例

Fig. 8　Proportion of extracted time attribute template

对建筑面积来说, 抽取数据的模板固定, 而不同的网页有不同的数字表示。我们主要考虑以下情况: 模板“建筑面积”前后有非数字字符或其他汉字, 如“建筑面积:”, “建筑面积为”; 数字使用逗号进行百位和千位的分隔; 数字具有小数的情况; 数字小数点使用中文字符的“.”; 阿拉伯数字使用汉字数学单位“百、千、万”; 面积后的单位表达不同, 如“平方米、平米、m2”。

对建成时间来说, 难点在于模板的获取, 而数字的表达, 不同数据源也有所不同, 比如“2020年1月1日”, “2020-01-01”, “2020/01/01”。但是, 年份数字的表述形式是统一的, 因此只考虑阿拉伯数字的表达, 即将年份表述为4位阿拉伯数字。另外, 若抽取的数字与现实不符(如“3000-00-00”), 则需要进行清洗筛除。

我们使用正则表达式和条件过滤, 对网页返回的HTML非结构化数据进行抽取和清洗。用表8中正则表达式来匹配建筑名称的子类片段或HTML网页数据中“抽取属性+属性值”的片段; 通过将匹配到的片段替换为空字符完成建筑名称中子类名称的过滤; 通过使用去除中文字符的正则表达式再次匹配和转换, 完成属性数值的抽取。如果抽取结果为空或不满足属性值的基本特性, 则抽取下一条网页摘要页, 或进入下一步召回策略。

2.3 建筑数据分析

2.3.1 数据结果展示

建筑名称的获取基于百度地图。由于网络数据内容丰富, 关于建筑面积的描述具有多种形式。另外, 建筑名称的获取和建筑面积的获取是两个具有先后顺序的子任务, 且各自的数据源不同。上述现

表 8　抽取字段的正则表达式

Table 8　Regular expression for extracting fields

抽取字段	正则表达式
建筑名称	r' [-(]\D*'
建筑面积	r'建筑面积\D*\d+[,,]?\d*[,,]?\d*[,,]?\d*[\..]?\d*\D*[平方米㎡ m]'
建成时间	r'竣工时间\D*\d{4}'
	r'建筑年代\D*\d{4}'
	r'建造年代\D*\d{4}'
	r'建成时间\D*\d{4}'

象导致建筑面积获取与建筑名称描述之间存在偏差与歧义性，因此基于网络大数据获取建筑面积具有一定的挑战性。本文爬取了北京地区 8 类建筑物的主要建筑，每类建筑的建筑名称、建筑面积及建成时间的爬取数量见图 9。对于住宅数据，除使用将上述方法爬取外，还融合 2.2.2 节介绍的种子数据集，因此其数量远超过其他建筑类型。图 9 的横坐标表示建筑数目统计的 3 个维度：建筑名称、建筑面积以及建成年份，并对同时含有建筑面积与建成年份的建筑物进行统计(“完整数据”一列)。图 10 统计北京市每个行政区的建筑数据爬取情况，共爬取到北京市 16 个市辖区的建筑数据。

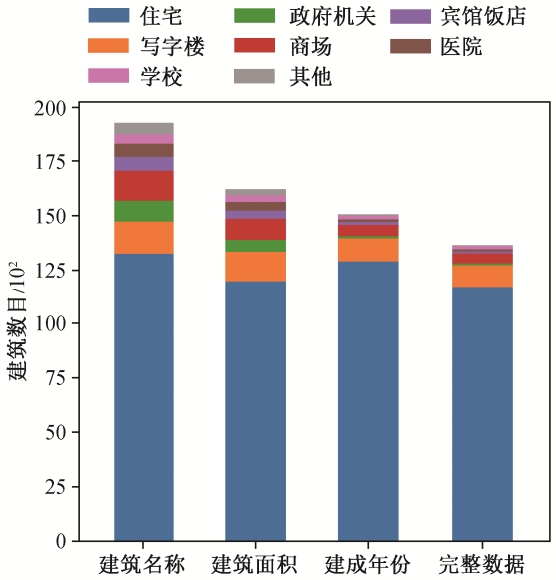

图 9 基于网络大数据爬取的北京市建筑数据获取结果

Fig. 9 Results of building data in Beijing based on online big data crawling

2.3.2 基于网络大数据的北京地区房屋竣工面积预估

考虑到搜索引擎排序算法或网页内容的误差，在利用爬虫数据进行建筑面积预估前，使用箱形图算法对建筑面积数据进行逐年清洗。同时，整理北京市统计年鉴中房屋竣工面积 2005—2018 年的数据，统计基于互联网大数据的获取比例，对 2019 年北京房屋竣工面积进行估算(至成文之时，该数据未公布)。基于网络大数据的获取比例见图 11，其中首都功能核心区包括东城区和西城区，城市功能拓展区包括朝阳区、丰台区、海淀区和石景山区，城市发展新区包括房山区、顺义区、昌平区、通州区和大兴区，生态涵养发展区包括门头沟区、怀柔区、密云区、延庆区和平谷区。我们发现，爬取到的首都功能核心区的建筑面积总和大于年鉴的房屋竣工面积，可能是由于建成时间的爬取存在误差或建筑物所属市辖区的基本信息错误。造成建成时间数据错误的原因可能有以下两点：一是利用搜索引擎抽取建成时间时，由于搜索引擎的页面排序算法的误差，导致与查询关键字匹配度并不高的网页优先返回，从而在错误的摘要页上抽取；二是查询关键字的消歧力度不够，导致抽取在歧义性网址的摘要页上进行。

设数据获取比例为 r_i^j，i 的取值范围为{0，1，2，3}，分别表示首都功能核心区、城市功能拓展区、

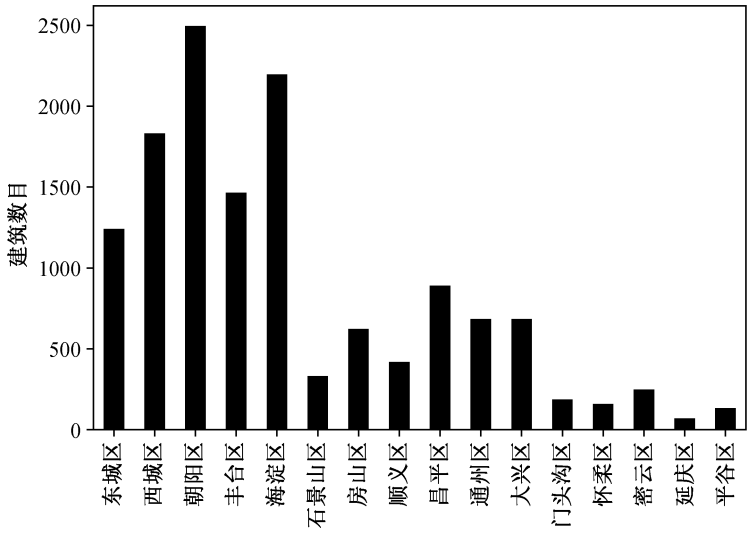

图 10 基于网络大数据爬取的北京市分区数据获取结果

Fig. 10 Results of Beijing regional district data based on online big data crawling

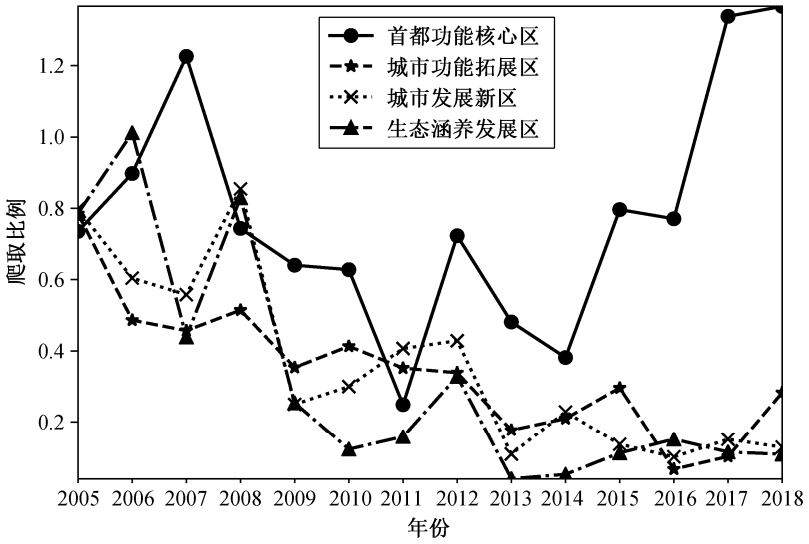

图 11　各功能区数据爬取获取比例

Fig. 11　Acquisition ratio of crawl data of each functional partition

城市发展新区和生态涵养发展区; j 表示年份, 取值范围为{2005, …, 2018}。从图 11 可以看出, 部分年份部分功能区的获取比例超过 1, 而超过 1 的获取比例是不合理的。因此, 需把不合理的获取比例删除, 平均获取比例的计算公式为

$$\overline{r_i} = \frac{1}{\sum_{j=2005}^{2018} I(r_i^j < 1)} \sum_{j=2005}^{2018} I\left(r_i^j < 1\right) r_i^j , \qquad (3)$$

其中, 函数 $I(\cdot)$为指示函数, 若 r_i^j <1 成立, 则取值为 1, 否则取值为 0。利用平均获取比例及爬取数据集中 2019 年各功能分区的总建筑面积, 得到 2019 年北京市房屋竣工面积的估计值为 2511.09 万 m^2, 各功能分区的估算结果见表 9。

2.3.3 不同方法的北京地区房屋竣工面积预估结果对比

为了验证基于网络大数据爬虫估算竣工面积方法的有效性, 对 2017 和 2018 年数据, 使用同样的估算方法进行竣工面积的估算, 结果见表 10。可以看出, 基于网络大数据的北京地区房屋竣工面积估算的误差在可接受范围之内。

表 9　2019 年北京市房屋竣工总面积估计值

Table 9　Estimation of area of completed buildings in Beijing in 2019

功能分区	平均获取比例	2019 年竣工面积估算/万 m^2
首都功能核心区	0.641	154.55
城市功能拓展区	0.345	1139.29
城市发展新区	0.361	824.69
生态涵养发展区	0.269	392.56
北京市	–	2511.09

表 10　2017 和 2018 北京市房屋竣工总面积估计值及误差

Table 10　Estimation of area and error of completed buildings in Beijing in 2017 and 2018

年份	竣工面积估算/万 m^2	RMSE
2017	2101.63	160.18
2018	2444.57	

表 11　不同方法 RMSE 对比

Table 11　RMSE comparison of different methods

模型	RMSE
偏最小二乘回归	260.56
LASSO 回归	232.79
RBF 神经网络	162.39
基于网络大数据	160.18

我们将传统方法与基于网络大数据的方法对 2017 和 2018 年建筑面积数据估算的 RMSE 误差进行对比。从表 11 可以看出, 两种方法的误差相差不大, 都可以有效地估计北京市房屋竣工总面积。但是, 从运行时间的角度看, 基于网络大数据方法的爬取需要更多的时间。因此, 在年鉴数据齐全的情况下, 可以使用传统方法估计竣工面积; 当年鉴数据不完整时, 使用基于网络大数据的方法, 构建相关数据, 并对竣工面积进行合理有效的估计。

3 结论

本文考虑到城市经济、社会和环境等因素对房

屋竣工面积的综合影响, 在年鉴中选择14个影响因子指标, 分别建立偏最小二乘回归模型、LASSO回归模型和RBF神经网络, 并对比其拟合效果, 最终得到2017和2018年北京市房屋竣工面积的预测值分别为2731.31和2535.71万 m^2。但是, 2019年北京市城市建设用地面积和城市人口密度等数据暂未公布, 无法用模型拟合的方式对该年度的竣工面积进行预测。因此, 本文提出基于网络大数据的建筑数据获取流程, 根据得到的建筑名称信息, 利用搜索引擎的搜索功能和网络数据爬取技术, 获取建筑面积和建成时间, 使用正则表达式对爬取到的数据进行分析抽取, 得到2019年北京市各功能分区的建筑面积估计结果, 最终得到2019年北京市房屋竣工面积的估计值为2511.09万 m^2。

综上所述, 使用模型拟合的方法可以得到竣工面积的估算值, 但该方法依赖于年鉴的相关数据。在年鉴数据缺失的情况下, 利用本文提出的网络数据爬取方法和估算流程, 可以有效地估算房屋竣工面积。

参考文献

[1] 中国煤炭工业协会. 煤炭工业统计常用指标计算办法. 北京: 煤炭工业出版社, 2012

[2] 曹爱丽, 张浩, 张艳, 等. 上海近50年气温变化与城市化发展的关系. 地球物理学报, 2008, 51(6): 1663–1669

[3] 王北星, 陈芳怡, 卢超. 吉林省房地产价格影响因素决策模型——基于因子分析及多元回归法的实证研究. 税务与经济, 2014(3): 107–110

[4] 匡文慧, 张树文, 张养贞. 基于遥感影像的长春城市用地建筑面积估算. 土木建筑与环境工程, 2007, 29(1): 18–21

[5] 周立柱, 林玲. 聚焦爬虫技术研究综述. 计算机应用, 2005, 25(9): 1965–1969

[6] 刘金红, 陆余良. 主题网络爬虫研究综述. 计算机应用研究, 2007, 24(10): 32–35

[7] Pavalam S M, Kashmir Raja S V, Akorli F K, et al. A survey of web crawler algorithms. International Journal of Computer Science Issues, 2011, 8(6): 309–313

[8] Rawat S, Patil D R. Efficient focused crawling based on best first search // 2013 3rd IEEE International Advance Computing Conference. Ghaziabad, 2013: 908–911

[9] Zhang Ling, Qin Zheng. The improved pagerank in web crawler // 2009 First International Conference on Information Science and Engineering. Nanjing, 2009: 1889–1892

[10] 周中华, 张惠然, 谢江. 基于Python的新浪微博数据爬虫. 计算机应用, 2014, 34(11): 3131–3134

[11] 范超, 王磊, 解明明. 新经济业态P2P网络借贷的风险甄别研究. 统计研究, 2017, 34(2): 33–43

[12] Shemshadi A, Sheng Q Z, Qin Y. ThingSeek: a crawler and search engine for the internet of things // Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval. New York, 2016: 1149–1152

[13] Young L J, Hyman M, Rater B R. Exploring a big data approach to building a list frame for urban agriculture: a pilot study in the city of Baltimore. Journal of Official Statistics, 2018, 34(2): 323–340

[14] Geladi P, Kowalski B R. Partial least-squares regression: a tutorial. Analytica Chimica Acta, 1986, 185: 1–17

[15] 王惠文. 偏最小二乘回归的线性与非线性方法. 北京: 国防工业出版社, 2006

[16] Tibshirani R. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society: Series B (Methodological), 1996, 58(1): 267–288

[17] Moody J, Darken C J. Fast learning in networks of locally-tuned processing units. Neural computation, 1989, 1(2): 281–294

[18] Page L, Brin S, Motwani R, et al. The pagerank citation ranking: bringing order to the web [R]. Stanford: Stanford InfoLab, 1999

[19] Brin S. Extracting patterns and relations from the world wide web // International Workshop on The World Wide Web and Databases. Valencia: Springer, 1998: 172–183

[20] Kleene S C. Representation of events in nerve nets and finite automata. Santa Monica: RAND Corporation, 1951