基于百度搜索数据预测全国艾滋病

病发数的实证分析及研究

**周申豪 兰后 胡炘**

**（西南财经大学，经济信息工程学院，2014级金融智能与信息管理实验班）**

**摘 要**

自从1981年艾滋病被首次发现，由于其尚无免疫预防、无有效药物治愈，且病死率极高的特点，艾滋病在全世界的广泛流行已经成为了严重的公众卫生问题和社会问题。自2001年以来，中国已进入艾滋病发病和死亡的高峰期，为了有效防控艾滋病日益增长的发病率，针对艾滋病流行趋势的分析和预测对合理分配卫生资源、制定有效防控策略意义重大。然而，受制于社会群体的复杂性和艾滋传染的特点，当前数据的来源过于有限，质量也存在许多不足，艾滋病的分析和预测工作受到极大限制。

而在当今互联网的时代，搜索引擎由于其隐匿性较强可以很好保护人们的隐私，已经成为人们日益常用的信息获取方式。本文从艾滋病疑似发病者可能会提前通过搜索引擎了解艾滋病的相关信息的这一行为特征作为出发点，分析构建了百度搜索数据与艾滋病发病数的关联机制。然后收集了艾滋病病发数数据，通过文献综述法、百度相关推荐词推荐法、手动查找补充等方式确定了搜索关键词。最终建立了综合搜索指数模型、BP神经网络、支持向量回归模型。

本文构建好预测模型后，利用历史百度搜索数据以及艾滋病发病数据对2015年10月、11月、12月的艾滋病发病数作出预测，发现百度搜索指数模型要优于传统的灰度预测、时间序列预测模型，而在百度搜索指数模型中，支持向量回归模型预测效果最好。本文研究成果对于相关部门提前预估艾滋病的传染疫情是否变化并作出相应的防控措施有一定的帮助。

**关键词：**艾滋病发病数预测 百度指数 回归模型 BP神经网络 支持向量回归

目录

[1 前言 4](#_Toc485217301)

[1.1艾滋病传染现状 4](#_Toc485217302)

[1.2我国艾滋病预测现状 5](#_Toc485217303)

[1.3网络数据使用现状 5](#_Toc485217304)

[2 国内外研究现状 6](#_Toc485217305)

[2.1 艾滋病病发数预测的研究现状 6](#_Toc485217306)

[2.2网络搜索数据应用的研究现状 8](#_Toc485217307)

[3关联机制分析 9](#_Toc485217308)

[4数据采集与预处理 11](#_Toc485217309)

[4.1数据采集 11](#_Toc485217310)

[4.1.1艾滋病发病数 11](#_Toc485217311)

[4.1.2百度搜索指数 11](#_Toc485217312)

[4.2数据预处理 12](#_Toc485217313)

[4.2.1关键词的选取 12](#_Toc485217314)

[4.2.2关键词的清洗 12](#_Toc485217315)

[4.2.3关键词月数据合成 12](#_Toc485217316)

[4.3数据集划分 12](#_Toc485217317)

[5综合搜索指数模型 12](#_Toc485217318)

[5.1构建综合搜索指数 12](#_Toc485217319)

[5.1.1时差相关分析 13](#_Toc485217320)

[5.1.2逐步回归筛选法 13](#_Toc485217321)

[5.1.3合成综合搜索指数 14](#_Toc485217322)

[5.2模型构建 14](#_Toc485217323)

[5.2.1平稳性检验与协整检验 14](#_Toc485217324)

[5.2.2模型建立 15](#_Toc485217325)

[5.3模型求解与检验 16](#_Toc485217326)

[5.3.1模型求解 16](#_Toc485217327)

[5.3.2模型检测 17](#_Toc485217328)

[5.4预测效果分析 18](#_Toc485217329)

[6 BP神经网络模型 18](#_Toc485217330)

[6.1算法简介 18](#_Toc485217331)

[6.2数据集划分 20](#_Toc485217332)

[6.3模型构建 20](#_Toc485217333)

[6.4参数调节 21](#_Toc485217334)

[7 支持向量回归模型 22](#_Toc485217335)

[7.1算法简介 22](#_Toc485217336)

[7.2数据集划分 23](#_Toc485217337)

[7.3参数确定与模型构建 23](#_Toc485217338)

[8总结与建议 24](#_Toc485217339)

[8.1三种模型对比 24](#_Toc485217340)

[8.2总结与建议 25](#_Toc485217341)

[9 任务分工 26](#_Toc485217342)

[参考文献 26](#_Toc485217343)

[附录一 29](#_Toc485217344)

[附录二 30](#_Toc485217345)

# 1 前言

## 1.1艾滋病传染现状

自1981年发现首例艾滋病起，艾滋病在全球快速扩展。目前，对于艾滋病尚无免疫预防疫苗、无有效治愈办法，且病死率极高，艾滋病已经成为了全世界范围内严重的公众卫生问题和社会问题。艾滋病的主要感染人群年龄为15岁至49岁，这将造成社会劳动人口减少，影响社会人口结构，阻碍社会经济的发展；人们对艾滋病病毒携带者和艾滋病病人的歧视，使得艾滋病患者没有公平的人身权利，将导致严重的社会问题；大量艾滋病患者自身无法支付高昂的医疗费用，会给整个社会带来沉重的经济负担。

1985年6月，我国境内发现首例艾滋病病人。根据世界卫生组织在2015年11月30日公布的资料显示，2014年底全球已有3690万艾滋病病毒感染者，新增艾滋病感染者约200万人。我国自1985年报告首例艾滋病病例以来，疫情呈上升趋势，尤其是2000年来，HIV感染人数以每年30%的速度增长。截至2015年10月，全国累计报告现存活艾滋病感染者57.5万例，约占全球的2%。由于我国人口基数大，感染者绝对数大，艾滋病防治形势不容乐观，已逐渐由高危人群向一般人群扩散。我国的艾滋病流行现状有以下6个特点：①艾滋病疫情呈上升趋势，但上升速度有所减缓。尽管历年累计报告数仍在增加，但每年的报告人数增加幅度与2004年之前相比有所减缓。②性传播成为主要传播途径，血液传播的比重显著下降。异性性传播和男男同性性传播已经逐渐成为主要传播途径，且男男性行为人群报告发病呈现逐年增高的态势。③艾滋病流行波及范围广，地区分布差异大。全国低流行与局部地区和特定人群中的高流行并存，疫情上升趋势明显。④自2001年以来，中国已进入艾滋病发病和死亡的高峰期。⑤艾滋病疫情具有从高危人群向一般人群扩散的趋势。⑥导致艾滋病流行的危险因素广泛存在。有利于艾滋病流行的危险因素包括：多性伴、无保护性行为、不安全注射行为、新型毒品泛滥、人群艾滋病知识缺乏、社会歧视较为严重等。

无论国际还是国内，有效预测和预防控制艾滋病的流行是刻不容缓的一项工作。

## 1.2我国艾滋病预测现状

艾滋病流行趋势分析和预测对合理分配卫生资源、制定有效防控策略意义重大。然而，我国目前能够获得的可用于分析和预测的数据质量存在缺陷，数据来源也非常有限，大大限制了分析和预测工作。

当前预测艾滋病病发数主要障碍有以下几个方面：①各类高危人群的规模大小不清楚。无论是吸毒人群、性乱人群、男男同性恋人群，还是既往卖血人群等，目前都还没有掌握人群规模方面较完整的信息，从而为估算这些人群中感染者数目和趋势变化造成障碍。②各类人群的危险行为特点和变化情况不清楚。近几年由于行为监测工作的开展对某些人群的危险行为特点和变化有了一定了解，然而，这些工作还远远不满足分析流行趋势并预测趋势变化的需求。③目前实际HIV感染病例数目不清，报告病例所占实际病例的比例不清楚。由于没有较完整的病例报告资料，目前很难对HIV流行状况有比较精确地描述。④对重点人群在HIV流行中的作用大小不清楚。如靠卖淫养吸的女性吸毒者将HIV从吸毒人群传至性乱人群中的作用，男男同性恋迫于社会压力将病毒从同性性行为人群传至一般人群中的作用。⑤缺乏对干预措施在HIV流行趋势影响方面的了解。比如目前开展的安全套推广、“四免一关怀”等干预措施对不同地区和人群中HIV流行的影响程度需要评估，并进一步估计对我国HIV流行趋势的影响大小。⑥对社会转型过程中一些社会因素对HIV流行可能产生的影响缺乏了解。随着社会经济变化，人们的性观、婚姻观、自我价值认定等方面都发生了很大的变化，而这些变化的趋势都直接或间接地影响着我国HIV流行趋势的变化。⑦对与艾滋病相关其他来源数据的利用不够充分。

疾病的流行趋势分析和预测是一项非常复杂的工作，需要多学科的合作，需要借鉴国际上先进的经验和技术，也需要结合我国的实际情况。

## 1.3网络数据使用现状

搜索引擎作为在互联网上获取信息的最常用工具，是连接信息资源与用户需求的枢纽，其在满足不同信息需求的同时也记录了用户的各种搜索行为，这些搜索行为记录被称为网络搜索数据，如艾滋病疑似发病者将会将自身已经表现出来的症状在通过搜索引擎搜索来进行自我诊断等。网络搜索数据蕴含着数以亿计用户的关注及需求，能够折射出现实社会中一些颇具影响力的事件，且这一数据资源可被视为某些事件（活动）发展趋势与演变规律的直接测度指标，其为相关的学术研究提供了数据支持，同时也使得基于网络搜索数据视角的理论和实证研究具有可行性。

截至2016年6月，我国搜索引擎用户规模达5.93亿，使用率为83.5%，用户规模较2015年底增加2635万，增长率为4.7%；手机搜索用户数达5.24亿，使用率为79.8%，用户规模较2015年底增加4625万，增长率为9.7%。在整体网民、手机网民中，搜索引擎都是第二大互联网应用。网民在互联网上搜索信息并产生丰富的数据，这些数据被搜索引擎记录下来并且具有对用户行为敏感性的特征，学者可以通过对这些数据的探索来发现网民的行为特征，从而产生了新的研究方向：利用网络数据对社会各种活动进行预测。

互联网搜索数据具有及时性的特征，而且由于网络用户分布范围广泛，用户群体量大，因此获取网络搜索数据是一种接近于全样本的“准自然”采样统计，其普通性和客观性比传统数据具有明显优势；它与流行性疾病疫情（尤其是艾滋病等涉及伦理隐私性的疾病）之间存在较强的关联性。

网络搜索数据的使用为本文对艾滋病疫情预测的思考提供了新的切入点。艾滋病疑似患者出于伦理的顾忌，大都直接通过网络搜索引擎进行匿名搜索以获得相关信息，同时百度搜索引擎则实时地将用户查询数据记录下来形成海量的用户搜索数据。因此，本文将使用中国百度搜索引擎搜索数据来预测中国艾滋病疫情，试图探索出一种监测中国艾滋病疫情蔓延情况的新方法。

# 2 国内外研究现状

## 2.1 艾滋病病发数预测的研究现状

有文献分别从艾滋病的疫情估计和预测方法，以及建立和完善社会防空体系等方面展开研究。Perelson等人在Science发表的论文利用传染病数学模型分析了5例实施HVI-1蛋白酶抑制剂的患者数据，得出了病毒清除率、感染细胞的存活时间和病毒的世代时间。UNAIDS估计、模型模拟及预测顾问组创建了利用监测资料估计和预测HIV感染率的软件包EPP(Estimation and Projection Package)，使用监测数据拟合不同地理区域的流行曲线，然后累计起来产生对全国流行率的估计。

在国内，80年代后对于传染病的监测开始逐渐发展起来。对于中国艾滋病疫情预测方面，颜康康等(2015)基于灰色系统理论建立了艾滋病预测的GM(1，1)模型，利用2003到2012年有关艾滋病、淋病、梅毒发病率数据，进行建模拟合并预测变化趋势。郭璐等(2014)利用运用数学模型，依据均方根误差和平均绝对误差绝对值最小原则确立最优模型，对HIV/AIDS病人进行估计与预测，以南京市2006-2013年艾滋病的疫情资料为例，结合指数平滑法进行模型拟合和评价，探讨模型的可行性，为南京市艾滋病监测和防控工作提供科学依据。陈婕卿等人(2013)收集345名感染男男性接触者的例总淋巴细胞计数、血红蛋白、血小板计数、白细胞计数和红细胞计数，分别建立基于logistic回归、线性判别分析、决策树、反向传播多层感知机和径向基函数神经网络的预测模型，并对预测效果进行折交叉验证。利用敏感度、特异度、阳性预测值、阴性预测值和受试者操作特征(ROC)曲线下面积对种预测模型的性能进行评价。张晓玲(2013)使用BP神经网络建立了艾滋病的疫情预测模型，为大理州的艾滋病监测和预警提供了一种新的方法，对大理州艾滋病预防工作部门的工作具有一定的理论意义和实践价值。王强等(2012)建立我国艾滋病流行预测模型，设计不同行为改变方案，预测艾滋病流行趋势。李巧巧等人运用EPP模型预测分析我国MSM人群艾滋病流行趋势，依据2005-2012年全国MSM人群综合监测资料和2012年全国MSM人群估计规模数，运用EPP模型对全国MSM人群艾滋病疫情进行估计和预测。

也有不少文献运用演化博弈理论建立传染病流行的动力学模型，可以从社会系统工程角度对艾滋病的中长期发展趋势进行预测，并探讨政府采取各种疫情防控措施的作用。

而目前应用中国百度搜索数据预测中国艾滋病发病数的研究成果还很少见。

## 2.2网络搜索数据应用的研究现状

随着信息技术的快速发展，以及互联网的普及，大数据时代的来临，搜索引擎可以为用户提供更加全面的数据，并且逐渐成为用户网上搜索数据及信息最方便快捷的工具，用户使用搜索引擎获取信息的同时也将自己的搜索痕迹和行为留在了搜索引擎中，一条条的搜索记录信息汇聚在一起则包含互联网用户现实生活中的行为意图，更是群体用户的社会性、经济性规律和发展趋势的反映。目前基于网络搜索数据的经济社会类行为预测已成为一个学术热点，目前在国内外取得了一定的研究成果，但是此类文献大多数在2008年之后才出现。

国际上，基于网络搜索数据的预测研究始于医药卫生领域。Eysenbach(2010)、Polgreen(2008)等研究得出搜索指数与流感发病率存在长期稳定关系，利用指数提前预测出流感发病情况及流感的死亡率。Jeremy Ginsberg(2009)发现Google日志中与流感发病率拟合最好的45个关键词搜索量与美国疾病控制和防治中心发布的流感看诊量数据有很强的相关性，由此构建了基于Google搜索数据的监测模型，可以比传统方法提前1-2周预测流感发病率，该文在Nature的发表引发学者广泛关注。Joshua Ritterman(2009)利用蜘蛛程序在UGC(UserGenerateContent)媒体抓取的关键词，利用流感的历史数据和关键词指标组合预测的效果很好。

在宏观经济领域，通过网络搜索与宏观经济的究主要集中在失业率、消费、股市指数、经济周期与经济衰退等方面，结果表明网络搜索数据对经济现象的预测效果较好。在大宗商品领域，学者们对零售业、汽车、汽油价格、房地产市场、旅游业等与网络搜索的相关性进行了研究，结果得出搜索指数对未来的销量和价格具有预测能力。

综上所述，网络搜索数据对社会性、经济性的相关性研究具有一定的科学性和合理性，而且网络搜索数据具有很强的客观性，是用户现实生活中的真是反映和自我写照，同时，网络搜索数据对传统数据作了更好的延伸和扩展。

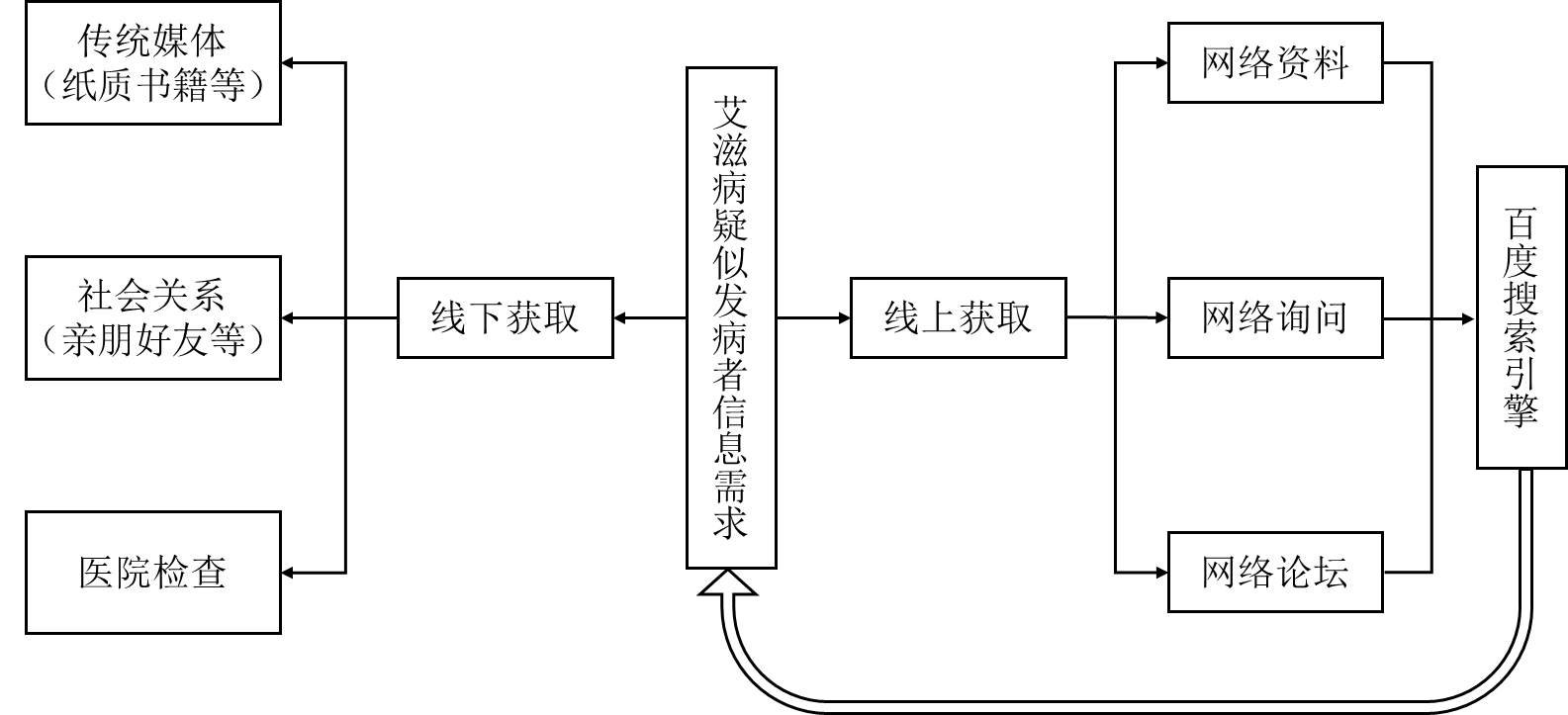
# 3关联机制分析

艾滋病疑似发病者怀疑自己可能感染后，会通过各种信息渠道获取关于艾滋病病状的相关信息以检验自己是否发病或者发病情况是否严重。现实中可供选择的信息渠道非常多，可以分为两类，第一类是下线方式，疑似发病者可以向身边亲朋好友咨询，可以去医院检查，可以外出购买或查看相关纸质书籍资料等等；第二类是在线方式，艾滋病疑似发病者可以借助互联网络提供的虚拟网络平台通过搜索引擎输入关于艾滋病有关的词语等进入网上专家咨询、网上论坛、网络书籍资料等界面进行信息在线收集。

随着互联网络在中国的快速发展与普及，我国能够使用电脑以及互联网的人数在不断增加，网络搜索已经成为很多人获取信息的重要途径之一，人们可以在网络上获取非常丰富的信息。搜索引擎为人们提供了更符合私密需求的、更开放快捷的信息获取渠道。艾滋病疑似患者出于伦理的顾忌，大都不会向身边的亲朋咨询或者直接去医院检查，而是直接通过网络搜索引擎进行匿名搜索获得有关信息，进而进行自我诊断。搜索引擎下搜索行为的匿名性恰好可以满足艾滋病疑似患者的隐私顾忌，同时搜索引擎下信息的丰富性又可以为艾滋病疑似患者提供充足的信息进行自我患病诊断。

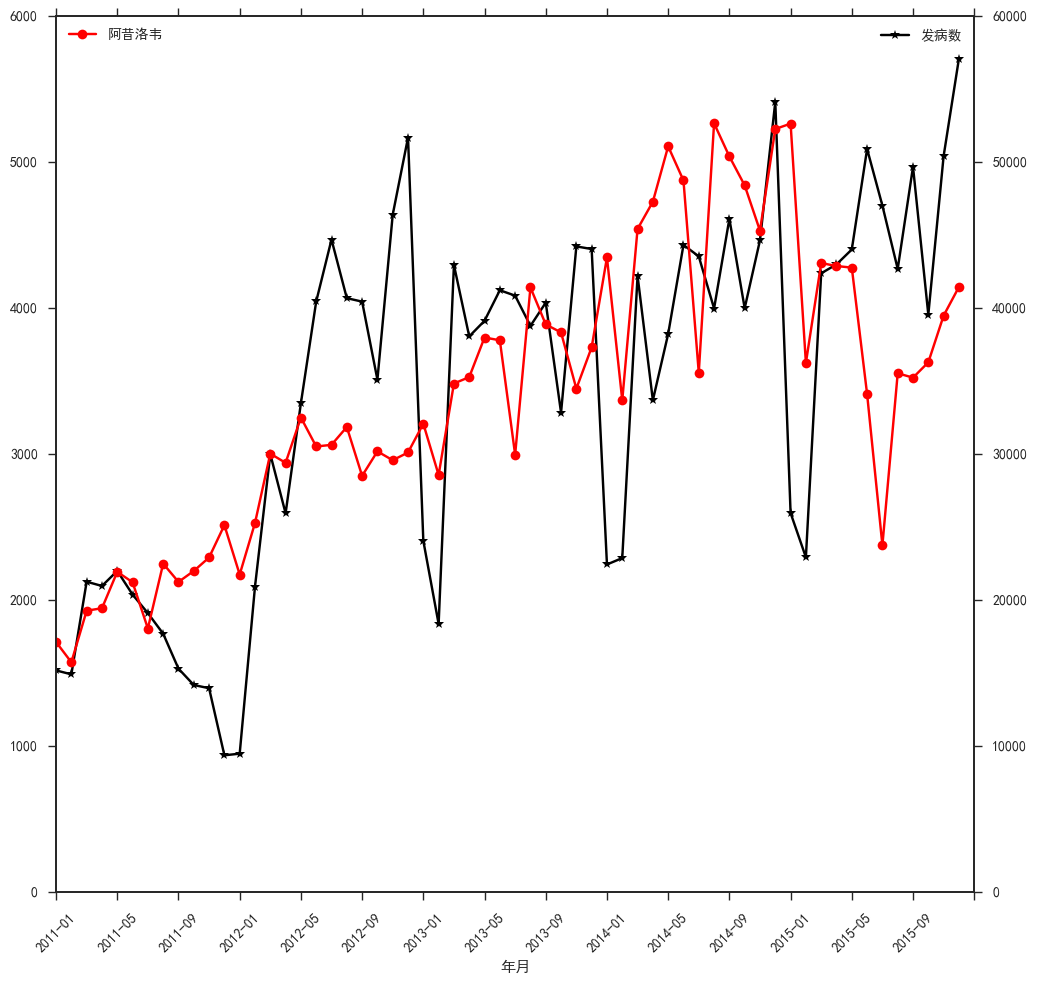
在中国，百度已经成为搜索的主要引擎。百度搜索指数以每日用户在引擎中输入关键词的搜索量为基础，将搜索量相对较大者处理成具有统一指标的日绝对数据，基本上可以在当天就获得当天的搜索量数据。这些关键词及其搜索量可以显示搜索者在搜索这一显性行为下所蕴涵的隐性意图，可以反映搜索者在网络这一虚拟空间中所想要寻求到的现实世界中的行为帮助和趋势指导。这种网络搜索数据的合集被称作用户的“意图数据库”。

此外，在通过网络搜索成功获取艾滋病有关信息后，当时的搜索行为主体又可能会将自己的体会、经验等信息以匿名方式分享在网络上，这又会成为后期艾滋病疑似发病者搜索的信息源，如此往复形成信息传导链环。

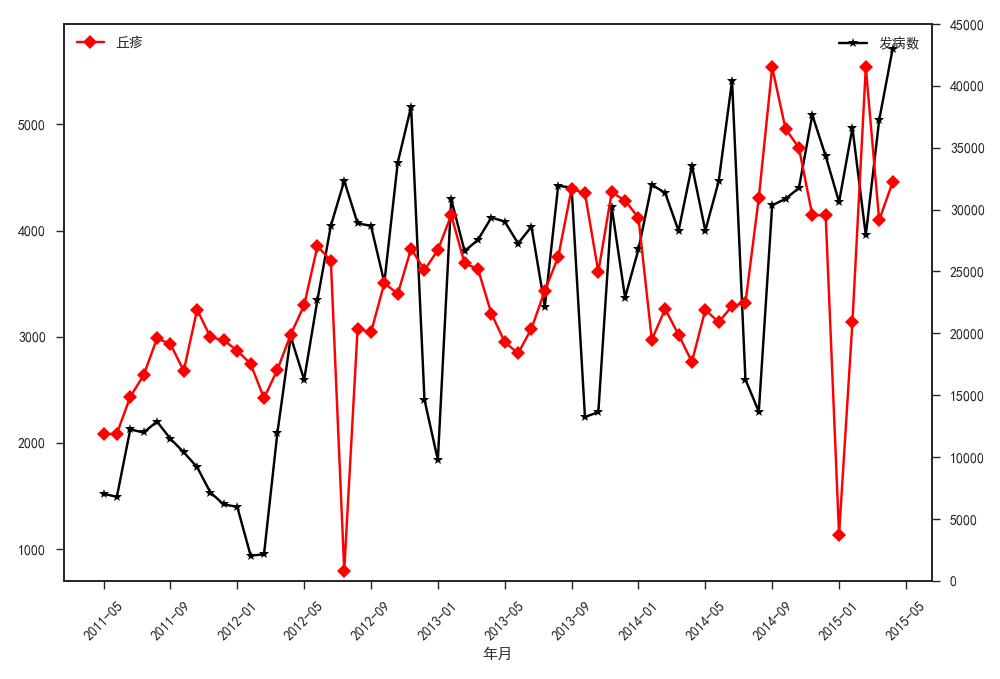


**图3-1 艾滋病疑似发病者信息获取流程**

与艾滋病有关的搜索量数据不一定都是艾滋病疑似发病者搜索产生的，但从统计学角度上百度中与艾滋病有关的部分关键词搜索量与艾滋病发病情况之间存在一定的相关性。以与艾滋病有关的关键词中的“阿昔韦洛”“丘疹”为例，其搜索量的变化趋势与艾滋病发病数的变化趋势确实具有较强的相关性。



**图3-2 “阿昔洛韦”搜索量与艾滋病病发数变化曲线图**

****

**图3-3 “丘疹”搜索量与艾滋病病发数变化曲线图**

总体说来，互联网搜索数据具有及时性的特征，而且由于网络用户分布范围广泛，用户群体量大，因此获取网络搜索数据是一种接近于全样本的“准自然”采样统计，其普通性和客观性比传统数据具有明显优势；它与流行性疾病疫情（尤其是艾滋病等涉及伦理隐私性的疾病）之间存在较强的关联性。

因此，本文提出假设，百度中与艾滋病有关的部分关键词搜索量对艾滋病发病数具有一定的预测能力。

# 4数据采集与预处理

## 4.1数据采集

### 4.1.1艾滋病发病数

本文采用公共卫生科学数据中心发布的艾滋病发病数月度数据，共可获得自2011年1月至2015年12月共60个月的月度数据。艾滋病病发数数据在本文模型中作为被解释变量。

### 4.1.2百度搜索指数

百度指数是以百度网民搜索的关键词的数据为基础的数据分享平台，包括每日整体搜索指数和周平均整体搜索指数。根据搜索数据来源的不同，百度指数又可分为PC搜索指数和移动搜索指数。前者积累了2006年6月至今的数据，而后者仅包括2011年1月至今的数据。

本文采用自2011年1月1日至2015年12月31日官网提供的百度指数的每日整体搜索指数，即包含PC搜索指数和移动搜索指数的数据。关键词的百度指数在本文中将用于构建解释变量。

## 4.2数据预处理

### 4.2.1关键词的选取

本文运用文献综述法、百度相关关键词推荐并结合手工查找补充的办法选取与艾滋病相关的关键词，最终可获得百度指数的关键词有68个，如HIV阳性、艾滋病论坛、艾滋病症状、食欲不振等等。

### 4.2.2关键词的清洗

本文的关键词清洗主要指对含义相近的关键词进行筛选、剔除。例如，对于“艾滋病症状”和“艾滋病有哪些症状”这两个意思相近的关键词，删除搜索量相对较低的关键词而留下搜索量较高的。

### 4.2.3关键词月数据合成

由于艾滋病病发数是月数据，而关键词的搜索指数是日数据，因此需要对关键词的数据进行合成。基于关键词每天搜索量是相互独立的假设，本文将日数据进行简单加和得到月数据。

## 4.3数据集划分

根据比例0.05对2011年1月至2015年12月共60个月进行划分，将后5%作为验证集，其余月份作为训练集。

# 5综合搜索指数模型

## 5.1构建综合搜索指数

根据国内外基于网络搜索数据所作的经济社会相关性研究文献，可以得出搜索指数的合成主要涉及关键词时差相关分析、关键词筛选和合成综合搜索指数三个过程。

### 5.1.1时差相关分析

时差相关分析法，是利用时差相关系数来确定整个时间序列内两个或更多个序列之间的平均关系的一种方法。在本文中，我们需要计算被选指标与基准指标之间的先行、一致或滞后关系，以及它们的相关系数大小。

假设基准指标（发病数）为，被选指标（关键词）为，时差相关系数为。

 （5-1）

式（4-1）中，，

这里时代表同步，取负值时代表先行，相反取正值时表示滞后，被称为时差数或延迟数。表示关键词和发病数指标取齐后的数据个数。在对指标进行计算时，得到不同时差数下的时差相关系数，在这些值中，选取绝对值的最大值，得到对应时差数，若，则剔除该关键词，否则保留。

至此，我们筛选出了部分关键词，得到了每个关键词的最大时差相关系数与其对应的时差数（先行期）。

### 5.1.2逐步回归筛选法

首先，依据弱相关的标准0.3，将最大时差相关系数小于0.3 的关键词剔除。

其次，由于这些关键词之间存在信息重叠问题，只需要将主要的关键词筛选出即可反映所有关键词的信息。本文采用逐步筛选策略(stepwise)，设定入选的关键词应满足其系数在0.05 的显著性水平下不为0。根据输出结果，筛选出26个关键词，先行期数以及最大时差相关系数分别如下表所示。详细数据见附录。

**表5-1 关键词信息**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 关键词 | HIV阳性 | 艾滋病症状 | 带状疱疹 | **…** |
| 先行期 | -4 | 0 | 0 | **…** |
| 相关系数 | 0.52 | 0.9 | 0.626 | **…** |

### 5.1.3合成综合搜索指数

将筛选后的关键词按先行期期数与艾滋病病发数对齐，从2011年6月起，用于预测发病数的关键词搜索指数数据完整，因此以2011年6月作为第一个月，2011年12月作为最后一个月，共55个月。

记26个关键词中第m个关键词第i个月的搜索指数为，综合搜索指数Q为：

 （5-2）

经过person相关性检验，综合搜索指数Q与发病数Y存在相关系数为0.662的强相关关系。

## 5.2模型构建

此时，我们已得到用于构建模型的解释变量综合搜索指数Q与被解释变量发病数Y，其具体数值见附录。

### 5.2.1平稳性检验与协整检验

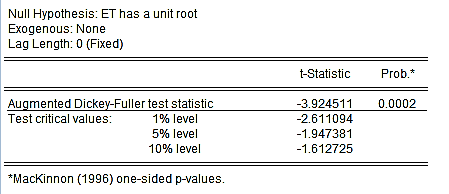
考虑到本文使用数据为时间序列数据，需要检验其平稳性，并使用EG两步法考察它们之间是否存在协整关系。

根据协整关系的检验方法，首先需确定Y与Q序列是否为非平稳序列，即考察其单整阶数。检验结果表明，Y与Q序列都是一阶单整的，即Y~I(1)，Q~I(1)。

**表5-2 变量平稳性检验**

|  |  |  |  |
| --- | --- | --- | --- |
| **变量** | **t-Stat** | **Prob.** | **检验结果** |
| Y | -2.1441 | 0.2289 | 非平稳 |
|  | -4.7243 | 0.0003 | 平稳 |
| Q | -0.4033 | 0.9003 | 非平稳 |
|  | -4.6947 | 0.0004 | 平稳 |

在上述检验的基础上，为了分析Y与Q之间是否存在协整关系，先作两变量之间的回归，然后验证回归残差的平稳性，估计结果如图5-1所示：

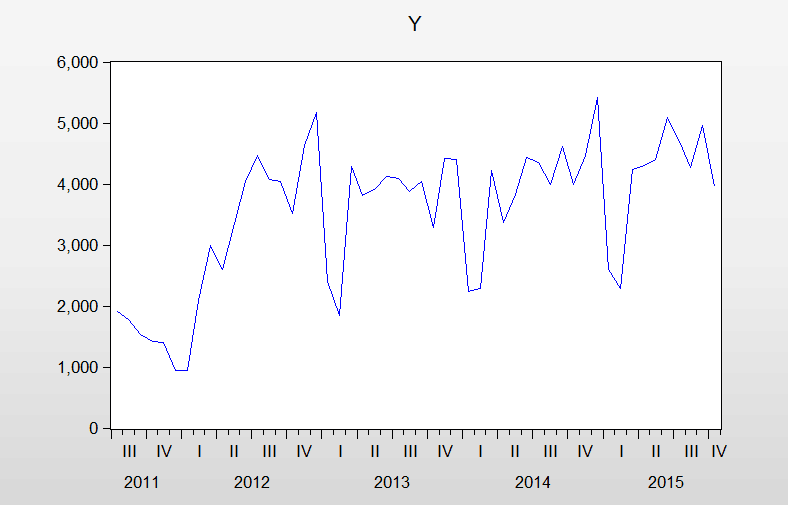


**图5-1 残差序列平稳性检验估计结果**

在5%的显著性水平下，t检验统计量为-3.9245，小于相应临界值，从而拒绝，表明残差序列不存在单位根，是平稳序列，说明Y与Q之间存在协整关系。

### 5.2.2模型建立

因为艾滋病发病人数Y和综合搜索指数Q均为平稳序列，可以进行回归。本文选择艾滋病发病数作为因变量，由关联机制分析可知综合搜索指数Q与发病数有一定的先行滞后关系，而从艾滋病的传染特性可知，前一期的发病数对本期发病数有一定影响作用。同时，在对作趋势曲线图时可以发现，发病数Y呈现出以一年为周期的循环变动特点，如下图所示。



**图5-2 艾滋病发病数月度趋势图**

为了更好的显示百度指数对于艾滋病发病数的预测效果，我们必须剔除季节因素和不规则因素，根据虚拟变量个数设置规则，设置12-1=11个代表月份的虚拟变量，当表示第i个月时，，表示其他月份时，。

综上所述，我们建立如下所示的艾滋病病发数与综合搜索指数之间的模型：





## 5.3模型求解与检验

### 5.3.1模型求解

经过多元回归分析求解得模型结果如下表所示。

**表5-3 模型结构**

|  |  |  |
| --- | --- | --- |
| 解释变量对应的系数 | 模型 | |
| **系数** | **P-值** |
| *β0* | 588.485 | .134 |
| *β1* | 0.001 | .109 |
| *β2* | 0.722 | .000 |
| *β3* | -1953.154 | .000 |
| *β4* | -466.210 | .220 |
| *β5* | 1213.648 | .002 |
| *β6* | -562.3750 | .123 |
| *β7* | 88.769 | .804 |
| *β8* | 344.114 | .339 |
| *β9* | -147.554 | .666 |
| *β10* | -447.522 | .192 |
| *β11* | -752.877 | .031 |
| *β12* | 390.255 | .282 |
| *β13* | 77.777 | .828 |

故艾滋病病发数与综合搜索指数之间的模型得：



### 5.3.2模型检测

模型检测结果如下表所示

**表5-4 模型检测**

|  |  |
| --- | --- |
| *R2* | .850 |
| *Adjusted R square* | .798 |
| *回归方程F检验概率P-值* | .000 |
| *D.W.* | 1.969 |

由表所示结果可知，可决系数R2为0.850，修正可决系数R2为0.798，表明该模型对样本的拟合程度较好。回归方程F检验概率P-值为0.000，说明自变量的回归系数不同时为零，模型中的D.W.为1.969，可认为残差之间是相互独立的。

总而言之，该模型是比较理想的，回归方程中综合搜索指数Q 的系数为正且显著，说明与艾滋病有关的百度搜索数据与艾滋病发病数之间存在显著正相关关系，与艾滋病有关的网络搜索量的增加意味着艾滋病发病数的增加。因此，通过对相关关键词搜索量的研究可以预测艾滋病发病数的情况，提前了解艾滋病疫情的蔓延程度。

## 5.4预测效果分析

为进一步考察二者之间是否具有预测关系，对变量Y与Q做Granger因果关系检验。检验结果表明二者存在Granger因果关系，说明Q可以显著地Granger引起Y，Q在Y的预测中有显著的解释。

**表5-5 Granger因果关系检验结果**

|  |  |  |  |
| --- | --- | --- | --- |
| **原假设** | **观察值** | **F 统计量** | **P 值** |
| Q does not Granger Cause Y  Y does not Granger Cause Q | 50 | 7.02312  4.52735 | 0.0022  0.0162 |

结合回归方程中Q对Y的系数显著，充分证实了网络搜索指数确实对艾滋病发病数具有一定的预测能力。根据以上模型对2015年11月和12月艾滋病病发数进行预测，可以看到，模型预测结果良好。

**表5-6 预测结果**

|  |  |  |
| --- | --- | --- |
|  | 2015年11月 | 2015年12月 |
| 预测值 | 4930.805 | 5364.617 |
| 实际值 | 5040 | 5707 |
| 误差 | 2.167% | 5.999% |

根据相关文献，在传统的灰度预测、时间序列预测模型中，预测误差常在15%-25%左右，而我们的百度综合搜索指数预测模型的误差远远低于该数值，百度综合搜索指数预测模型在预测精确度上要优于传统预测模型。

# 6 BP神经网络模型

## 6.1算法简介

BP神经网络(Back-Propagation Neural Network)的基本思想是利用最小均方差(LMS)学习算法，在网络的学习过程中使用梯度下降技术，利用误差逆向传播来修正初始权数，从而实现实际输出与期望输出的均方差最小化。图6-1给出了最简单的带有一个隐藏层的BP神经网络拓扑结构。其中：表示输入层单元； 表示隐藏层单元；表示输出层单元：从输入层单元i到隐藏层单元j的连接权为；从隐藏层单元j到输出层单元k的连接权为；表示所有的连接权。上标表示不同的输入模式；P表示输入模式数()；和分别对应于适当选定的隐藏层和输出层激活函数。对给定的输入模式；隐藏层单元j的输入为：

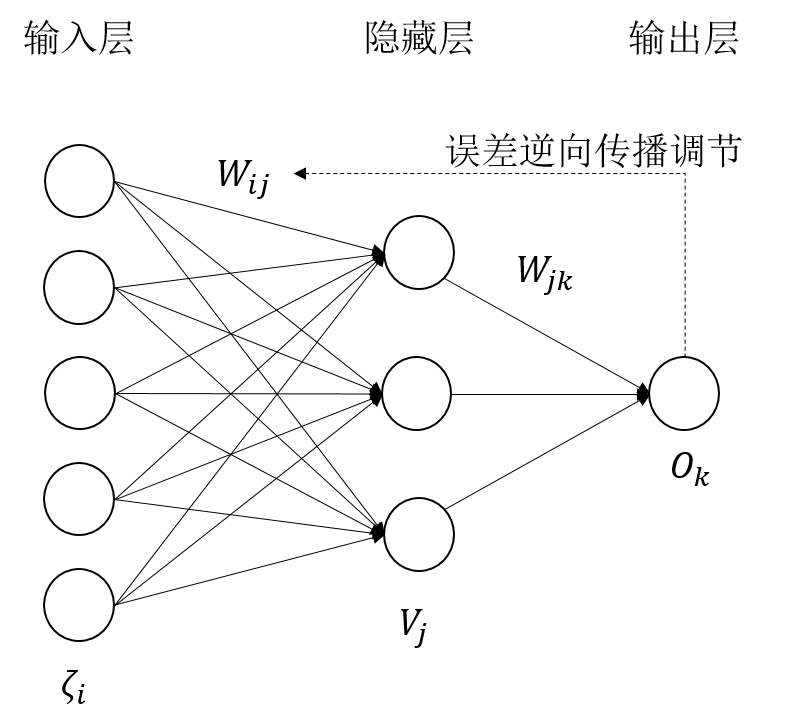
(1)

输出为：

最终输出层单元的输出为：

对于输出单元i，总误差函数为：

所谓的权学习，即对每一个当前的，取适当增量从而进一步得到，使得总误差函数随着迭代次数的增多而减少，从而最终达到最小值。



**图6-1 三层的BP神经网络模型**

## 6.2数据集划分

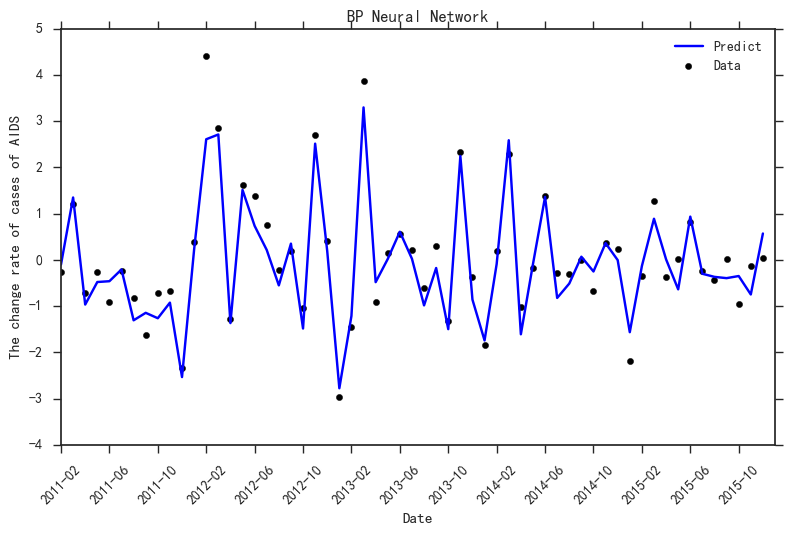
由于原始数据各个分量数值的数量级有很大差异，对于某个输入层节点，如果过大，则相应的更新量过大；这样会使得自身节点影响过大而其他分量丧失调控能力，所以有必要对数据进行标准化处理。而由于变量为时间序列，通过尝试，为了得到更好的预测结果，本文选择对其进行取变化率的数据转化(Preis,2013)。由于本文数据集样本小，仅有60个月的月度数据，于是按5%的比例将数据集划分为训练集与测试集。

## 6.3模型构建

本文选择导入基于Theano和Tensorflow的高层神经网络处理包Keras。使用该处理包可以直接进行建模编译拟合预测，可以方便快捷的进行原型设计。



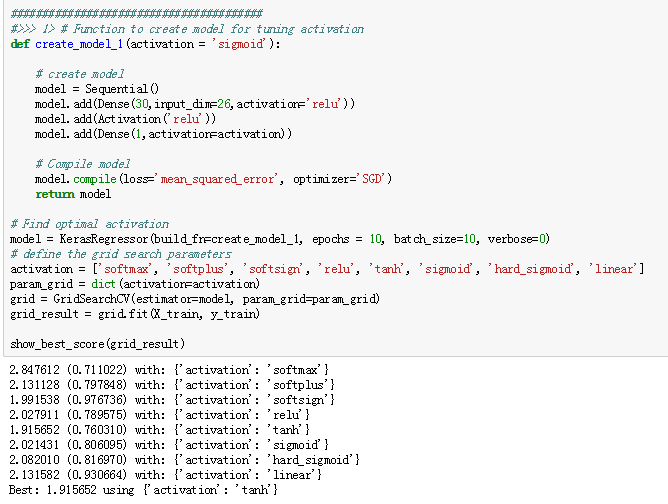
**图6-2 BP神经网络模型(最终参数版)**



**图6-3 BP神经网络模型预测效果**

## 6.4参数调节

本文导入GridSearchCV，直接采取超参数自动检索进行调参。它的主要思想在于构建模型时，将固定的参数设定成可变参数范围，然后程序运行后，自动遍历可能情况找出最佳参数。我们将度量方法确定为MSE均方误差，即均方误差越小，参数越好。但值得注意的是，最后模型匹配度最好，预测均方误差最小的模型并非集合所有参数最优解。因而对于最优参数的探索还需结合经验进一步试探。



**图6-4 调参及结果(以调节激活函数为例)**

# 7 支持向量回归模型

## 7.1算法简介

支持向量回归的基本思想是通过一个非线性映射将数据映射到高维特征空间 ,并在这个空间进行线性回归，能够较好地解决小样本、非线性、高纬数和局部极小点等实际问题，被人称为替代人工神经网络的较好方法。

对于给定 k 个样本数据，其值表示为：，式中 的n维向量， 为相对应的输出变量，回归算法的基本思想是通过一个非线性映射，将数据集映射到高维特征空间H，并在这个空间进行线性回归。具体的函数形式可表示为：

，

由于支持向量机理论只考虑高维特征空间的点积运算，而不是直接使用函数，从而巧妙地解决了因映射函数的未知而无法显式表达的问题。称满足Mercer条件的对称函数为核函数，常用的核函数有：（1）多项式核函数(Poly)、（2）线性核函数(Linear)、（3）高斯径向基函数核函数(RBF)、（4）正切核函数(Sigmoid)

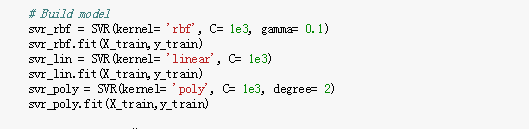
本文选取前三种核函数进行模型拟合预测并比较。

## 7.2数据集划分

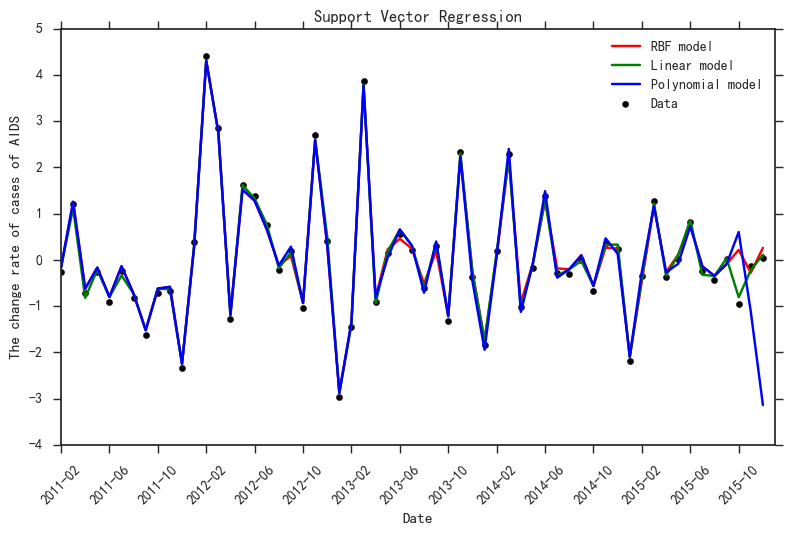
由于同为机器学习模型，本部分与BP神经网络部分处理方式一致，在这里不再赘述。

## 7.3参数确定与模型构建

本文导入python机器学习包sklearn下的SVR函数。而SVR的参数确定与BP神经网络不同，因其模型参数的适用性较强，通常根据经验值确定，而不用多次调节直至最佳。



**图7-1 支持向量回归模型(最终参数版)**



**图7-2 支持向量回归模型预测结果**

# 8总结与建议

## 8.1三种模型对比

（1）相较于传统预测模型，基于百度指数的预测模型增加了人的心理和行为的影响；

（2）综合搜索指数模型充分考虑了先行变量、周期性变化以及季节因素等的影响；

（3）SVM支持向量机和BP神经网络都较适用于样本少且维度高的数据。

①BP神经网络和普通的人工神经网络的不同点在于BP Back-Propogation 逆向传播进行学习。通过优化函数反向调节最开始随机设定的权重。而调节权重的过程事实上就是梯度下降最小化误差值的过程。

②SVM支持向量机可通过非线性映射，将非线性可分问题转化为线性可分问题；同时，考虑到数据样本数小，维度高，试采用SVM并选取径向机核(RBF)、线性核(Linear)、多项式核(Poly)三种核函数进行预测，其中，线性核函数的拟合效果最优。

根据不同的预测精度对比指标，可将模型分为以下两类：

**表8-1 模型对比**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 对比指标 | 传统模型 | | 机器学习 | |
| 灰度预测、时间序列模型等 | 综合搜索指数模型 | BP神经网络 | 支持向量回归模型 |
| 预测误差 | 15-20% | 4.083% | \ | \ |
| 均方根误差 | \ | \ | 0.26 | 0.1148 |

## 8.2总结与建议

本文基于百度搜索关键词的搜索数据对艾滋病发病数构建预测模型，首先分析了百度搜索关键词与艾滋病发病数之间的关联机制，进而采用定量的实证分析方法分析关键词搜索量，建立与艾滋病发病数之间的三种模型，最终求解出三种模型，并发现支持向量回归模型结果最好。本文试图探索出一种监测中国艾滋病疫情蔓延情况的新方法，力图为疾病预防控制中心(CDC)等相关机构提供及时性艾滋病疫情数据参考，提高艾滋病发病数的同步度，降低其滞后程度，为国家防治艾滋病相关部门提供更加及时的信息，以便其及时采取有效措施来监控艾滋病的蔓延。

由于网络搜索数据的更新性非常强，传播性非常强大，在不同的阶段其表现特征可能不同，因此在利用网络搜索数据建立的预测模型中，我们只能运用其来预测未来一定时间范围内的艾滋病发病数，而无法将之运用到后面很长时间。此预测模型可以在后期随时调整，以便更好地预测。由于这种新的预测方法尚处在探索阶段，还有很多难点和问题需要进一步思考研究。如何更好地筛选并合成综合搜索指数，如何更好地调整机器学习模型中的参数，如何去除噪声点等等，都将是我们后续进一步研究的内容。

最后，我们提出以下建议：

（1）本文所做的预测模型是综合利用历史发病数据和网络搜索数据，对下月的艾滋病发病数作出较为准确的预测，若是数量出现了明显的变化，有关部门可以针对百度指数地区分布图找到相对应的疫情变化区，对症下药式地作出相对应的措施以达到有效的防控效果。

（2）家庭和社区要为艾滋病病人及感染者营造一个友善、理解、健康的生活和工作环境，鼓励他们采取积极的生活态度、改变高危行为、配合治疗，有利于提高病人及感染者的生命质量、延长生命，也有利于艾滋病的预防与控制工作和维护社会安定。

（3）建立政府领导、多部门合作和全社会共同参与的艾滋病预防与控制体系，形成有利于艾滋病防治的社会环境是控制艾滋病流行的重要成功经验。我国预防控制艾滋病的策略应该是预防为主、宣传教育为主、动员全社会参与、实行综合治理。

# 9 任务分工

周申豪(41411007)：收集整理使用网络数据进行预测领域的文献，确定选题；收集整理艾滋病病发数等数据；阅读整理灰度预测、时间序列等传统预测模型，用R编程还原文献实现过程；建立四种综合搜索指数模型，编程求解四类模型并选取最优模型，对模型进行检验；使用python编程、作图；本文档撰写；PPT制作展示。

胡炘(41411064)：搜集有关艾滋病预测模型领域的文献；综合搜索指数模型的求解、检验以及该部分的文档书写；模型对比的文档书写；协助Python代码实现数据截屏后的图像识别；PPT制作展示。

兰后(41423018)：用python代码实现模拟登陆对百度指数进行截屏、实现数据爬虫录屏并上传至github；收集博客和文献资料构建BP神经网络模型和SVR支持向量回归模型并调节参数最后用python代码实现；数据可视化；机器学习部分的文档撰写；PPT制作展示。

# 参考文献

[1]Perelson A S, Neumann A U, Markowitz M, et al. HIV-1 dynamics in vivo: virion clearance rate, infected cell life-span, and viral generation time[J]. Science, 1996, 271(5255): 1582-1586.

[2]Ghys P D, Brown T, Grassly N C, et al. The UNAIDS Estimation and Projection Package: a software package to estimate and project national HIV epidemics[J]. Sexually transmitted infections, 2004, 80(suppl 1): i5-i9.

[3] 颜康康, 林雪君, 鲍红红, 等. 灰色 GM (1, 1) 模型在艾滋病, 淋病, 梅毒发病率预测研究 中的应用[J]. 实用预防医学, 2015, 22(3): 371-374.

[4] 郭璐, 张敏, 朱正平, 等. 指数平滑法在南京市艾滋病疫情预测中的应用[J]. 中国艾滋病性 病, 2014, 12: 012.

[5] 陈婕卿, 朱碧云, 陈卉, 等. 基于数据挖掘的 HIV 感染者 CD4 细胞计数预测模型[J]. 北 京生物医学工程, 2013, 32(005): 479-484.

[6] 张晓玲. 基于 BP 神经网络的大理州艾滋病流行现状分析和疫情预测模型的研究[D]. 云南 大学, 2013.

[7] 王强, 袁建华. 行为改变下的艾滋病流行预测[J]. 数学的实践与认识, 2008, 6: 026.

[8] 刘黎, 彭志行, 施超, 等. EPP 模型在江苏省艾滋病疫情估计, 预测与分析中的应用[J]. 中 华疾病控制杂志, 2010 (6): 479-483.

[9]Eysenbach G. Infodemiology and infoveillance: tracking online health information and cyberbehavior for public health[J]. American journal of preventive medicine, 2011, 40(5): S154-S158.

Chew C, Eysenbach G. Pandemics in the age of Twitter: content analysis of Tweets during the 2009 H1N1 outbreak[J]. PloS one, 2010, 5(11): e14118.

[10]Polgreen P M, Chen Y, Pennock D M, et al. Using internet searches for influenza surveillance[J]. Clinical infectious diseases, 2008, 47(11): 1443-1448.

[11]Ginsberg J, Mohebbi M H, Patel R S, et al. Detecting influenza epidemics using search engine query data[J]. Nature, 2009, 457(7232): 1012-1014.

[12] Ritterman J, Osborne M, Klein E. Using prediction markets and Twitter to predict a swine flu pandemic[C]//1st international workshop on mining social media. 2009, 9.

[13] Choi H, Varian H. Predicting the present with google trends[J]. Economic Record, 2012, 88(s1): 2-9.

[14] Lu X, Zhao X. Differential Effects of Keyword Selection in Search Engine Advertising on Direct and Indirect Sales[J]. Journal of Management Information Systems, 2014, 30(4): 299-326.

[15] Ripberger J T. Capturing curiosity: Using Internet search trends to measure public attentiveness[J]. Policy Studies Journal, 2011, 39(2): 239-259.

[16] LIU Y, LV B, PENG G. Predictive Power of Internet Search Data for Stock Market: A Theoretical Analysis and Empirical Test [J][J]. Economic Management Journal, 2011, 1. [17] Pauwels K, Leeflang P S H, Teerling M L, et al. Does online information drive offline revenues?: only for specific products and consumer segments![J]. Journal of Retailing, 2011, 87(1): 1-17.

[18] Choi H, Varian H. Predicting the present with google trends[J]. Economic Record, 2012, 88(s1): 2-9.

[19] Azar J. Electric cars and oil prices[J]. Available at SSRN 1474023, 2009.

[20] Wu L, Brynjolfsson E. The future of prediction: how Google searches foreshadow housing prices and quantities[J]. ICIS 2009 Proceedings, 2009: 147..

[21] 王博永, 杨欣. 基于网络搜索的房地产政策调控效果研究[J]. 管理评论, 2014, 9: 009. [22] Choi H, Varian H. Predicting the present with google trends[J]. Economic Record, 2012, 88(s1): 2-9.

[23] 张斌儒, 黄先开, 刘树林. 基于网络搜索数据的旅游收入预测——以海南省为例[J]. 经济问题探索, 2015 (8): 154-160.

[24]田翔, & 邓飞其. (2005). 精确在线支持向量回归在股指预测中的应用. 计算机工程, 31(22), 18-20.

[25]吴微, 陈维强, & 刘波. (2001). 用bp神经网络预测股票市场涨跌. 大连理工大学学报, 41(1), 9-15.

[26]Preis, T., Moat, H. S., & Stanley, H. E. (2013). Quantifying trading behavior in financial markets using Google Trends.

[27]Mitchell, T. M. (2009). Mining our reality. Science, 326(5960), 1644-1645.

# 附录一

筛选后的关键词如下表所示：

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **关键词** | HIV阳性 | 艾滋病论坛 | 弓形体病 | 淋病 | 梅毒 | 艾滋病初期症状图片 | 艾滋病检查 | 艾滋病检查费用 | 艾滋病能活多久 |
| **滞后期** | -4 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **MAX\_R** | 0.52 | 0.438 | 0.564 | 0.637 | 0.424 | 0.598 | 0.629 | 0.64 | 0.385 |
| **关键词** | 艾滋病症状 | 艾滋病治疗 | 抽搐 | 带状疱疹 | 盗汗 | 恶性肿瘤 | 呼吸困难 | 肌肉痛 | 咳嗽 |
| **滞后期** | 0 | 0 | 0 | 0 | -5 | -2 | -1 | -1 | -4 |
| **MAX\_R** | 0.9 | 0.512 | 0.591 | 0.626 | 0.546 | 0.621 | 0.522 | 0.385 | 0.506 |
| **关键词** | 食欲不振 | 嗜睡 | 头痛 | 消瘦 | 胸痛 | 血小板减少 | 阿昔洛韦 | 丘疹 |  |
| **滞后期** | 0 | 0 | -1 | 0 | -1 | 0 | 0 | -6 |  |
| **MAX\_R** | 0.603 | 0.747 | 0.544 | 0.58 | 0.599 | 0.61 | 0.61 | 0.393 |  |

# 附录二

解释变量综合搜索指数Q与被解释变量发病数Y数据如下表所示：

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 年月 | 2011.07 | 2011.08 | 2011.09 | 2011.10 | 2011.11 | 2011.12 | 2012.01 | 2012.02 | 2012.03 |
| Y | 1912 | 1771 | 1532 | 1420 | 1398 | 939 | 950 | 2093 | 3000 |
| Q | 338169.8 | 338547.3 | 333357.7 | 343805.5 | 348508.1 | 385050.1 | 348429.5 | 379496 | 381814.4 |
| 年月 | 2012.04 | 2012.05 | 2012.06 | 2012.07 | 2012.08 | 2012.09 | 2012.10 | 2012.11 | 2012.12 |
| Y | 2595 | 3349 | 4045 | 4469 | 4068 | 4044 | 3510 | 4639 | 5167 |
| Q | 399975.6 | 414840.3 | 439054 | 435575.9 | 458877.7 | 457462.3 | 433127.4 | 437980 | 471895 |
| 年月 | 2013.01 | 2013.02 | 2013.03 | 2013.04 | 2013.05 | 2013.06 | 2013.07 | 2013.08 | 2013.09 |
| Y | 2407 | 1838 | 4295 | 3806 | 3913 | 4122 | 4085 | 3878 | 4037 |
| Q | 416057.3 | 422469.8 | 476534.8 | 484876.4 | 490242.4 | 530336.4 | 513620 | 542652.7 | 565699.1 |
| 年月 | 2013.10 | 2013.11 | 2013.12 | 2014.01 | 2014.02 | 2014.03 | 2014.04 | 2014.05 | 2014.06 |
| Y | 3284 | 4422 | 4404 | 2245 | 2289 | 4221 | 3369 | 3823 | 4432 |
| Q | 538881.5 | 499577 | 613003.1 | 477020.1 | 455691 | 565640.3 | 659454.2 | 677160.9 | 698527.5 |
| 年月 | 2014.07 | 2014.08 | 2014.09 | 2014.10 | 2014.11 | 2014.12 | 2015.01 | 2015.02 | 2015.03 |
| Y | 4354 | 3997 | 4611 | 3999 | 4464 | 5411 | 2597 | 2294 | 4238 |
| Q | 681729.9 | 738025.8 | 781234 | 704597.5 | 682130.2 | 760150.6 | 702453.2 | 633649.7 | 730563.9 |
| 年月 | 2015.04 | 2015.05 | 2015.06 | 2015.07 | 2015.08 | 2015.09 | 2015.10 | 2015.11 | 2015.12 |
| Y | 4299 | 4401 | 5089 | 4701 | 4268 | 4967 | 3955 | 5040 | 5707 |
| Q | 783434.2 | 771303.5 | 827275 | 793096.4 | 782442.4 | 796705 | 796125.9 | 1096555 | 1138314 |