情报学报　2018 年 2 月　第 37 卷　第 2 期
Journal of the China Society for Scientific and Technical Information, Feb. 2018, 37(2): 206-219

DOI: 10.3772/j.issn.1000-0135.2018.02.009

融合百度指数的流感预测机理与实证研究

王若佳 [1,2]

（1. 北京大学信息管理系，北京　100871；2. 北京大学海洋研究院，北京　100871）

摘　要　本文通过挖掘网络搜索数据与我国流感疫情的在内在机理，利用关键词的时序特征实现了较为精准的提前预测。研究首先从信息行为、信息搜寻行为等理论概念出发，对百度指数与流感病例数据之间的逻辑关系进行探讨，建立理论框架；然后以理论框架为基础，用范围选词法对百度搜索词进行初步筛选，并利用互相关分析选出具有先行性质的关键词，用于构建预测模型；最后，对比融合百度指数的三种预测模型，评估其预测效果。互相关分析结果大致符合本文提出的逻辑框架，可提前十周预测流感疫情的关键词内容和流感疫苗相关；提前一周的关键词多涉及流感的症状表现；而同步类关键词多为常用搜索词或治疗方法。模型对比结果显示，多元线性回归模型、支持向量机模型和神经网络模型都能有效地进行流感预测，无论提前十周还是提前一周，支持向量机的效果最好。

关键词　流感预测；百度指数；神经网络；支持向量机；多元线性回归

Mechanism and Empirical Research on Forecasting Influenza Epidemic Fused with Baidu Index

Wang Ruojia[1,2]

(1. *Department of Information Management, Peking University, Beijing*　100871;
2. *Institute of Ocean Research, Peking University, Beijing*　100871)

Abstract: This study explores the internal mechanism and possibility of forecasting an influenza epidemic based on both search queries and actual influenza data. First, the logical relationship is explored between online information searches and conventional surveillance data based on the concepts of information behaviors, information seeking behaviors, and so on. Then, the range selection method and cross-correlation analysis are used to select keywords according to the theoretical framework. Finally, three models are established and compared. The results show that (i) the empirical research proves the logical rationality of the theoretical framework: the keywords that could reflect flu trends ten weeks in advance are related to influenza vaccines; those a week in advance are related to influenza symptoms; and most of the simultaneous keywords are frequent terms related to influenza; (ii) all three models can predict influenza effectively, and support vector machine yields the most accurate forecasting result.

Key words: influenza forecasting; Baidu Index; neural network; support vector machine; multiple linear regression

1　引　言

流行性感冒是由流感病毒引起的急性呼吸道感染，也是一种传染性强、传播速度快的流行病。尽可能早地获取流感的流行趋势有助于控制疾病的传播，进而减少损害。近年来，由于互联网信息的时效性及搜索引擎的普遍性，基于网络搜索数据对流感进行预警的研究逐渐增多。

收稿日期：2017-06-25；**修回日期**：2017-12-16

作者简介：王若佳，女，1992 年生，在读博士生，主要研究方向为网络数据挖掘、数据分析，E-mail: wangruojia@pku.edu.cn。

现有基于网络搜索数据的流感预警研究可大致分为以下两种类型：

（1）实时监测（nowcasting）。通过分析搜索词与流感疫情的相关性，筛选出相关系数较高的搜索词，根据其搜索现状来估计同时段的流感病例数。实时监测研究以谷歌流感趋势（Google Flu Trends，GFT）最为著名。2008 年，Google 发现，搜索流感相关主题的人数与实际患有流感症状的人数之间存在着密切关系[1]，他们依据这种数量关系提出了谷歌流感趋势，受到了广泛关注并被众多学者使用[2-4]。需要注意的是，GFT 预警功能实现的关键在于一系列流感关键词（如温度计、流感症状、感冒、感冒药等），只要用户在 Google 搜索引擎中输入这些关键词，系统就会自动跟踪分析。然而，由于 GFT 并不能覆盖全世界各个地区，因此一些学者使用其他工具（如谷歌趋势）进行了相应研究[5-6]。这些研究虽然在使用工具、统计方法、数据来源等方面有所区别，但具有一个共通的地方，即仅仅关注关键词与真实病情的相关程度或监测效果，反而忽略了预测的本质，如关键词选择的依据、关键词的特征识别等方面。

（2）提前预测（forecasting）。是指基于现有数据对未来还未真实发生的流感疫情进行估计，多使用时间序列分析等方法。“提前预测”研究在传统流感预测中较为常见[7]，但基于网络搜索数据建立预测模型的研究较少，且预测效果一般。

据此，本研究提出了一套有关搜索数据在流感预测中应用的完整框架，包括理论基础与应用模型、网络搜索词的选择方法、关键词时差关系的判定，进而通过融合官方流感数据和搜索数据构建多种流感预测模型，最后对多种模型的预测效果进行比较从而得出最优模型。

2　文献回顾

2.1　网络搜索数据与预测对象的内在机理研究

网络搜索数据可直接或间接反映互联网用户的行为与心理[8]，一些针对社会经济活动的研究尝试剖析了搜索数据与预测对象的内涵关系。例如，宏观经济方面，张崇等[9]从商品市场的供求关系角度分析了网络搜索数据与居民消费价格指数（CPI）之间的关系；旅游管理方面，王炼等[10]基于旅游需求预测和信息搜索理论探讨了网络信息搜索在旅游需求预测中的潜在作用；消费者行为方面，Kulkarni 等[11]认为搜索引擎是消费者想了解新产品时的首选工具，并提出搜索词可揭示消费者对产品的兴趣所在这一假设；社会心理方面，Song 等[12]指出自杀是包括“自杀观念形成—制定自杀计划—尝试自杀—自杀行为”等多个步骤的复杂事件，持有自杀想法的人在这一系列过程中会多次获取相关信息，但由于他们不善于或不愿意与他人沟通，导致网络成为重要信息渠道。最后，在医疗健康方面，尽管也有文献提到很多用户会在去医院寻求医生帮助之前在互联网上搜索相关疾病或症状[13]，但对流感疫情与搜索数据的在内在机理方面挖掘得不够充分，未形成系统的理论框架。

2.2　基于网络搜索数据的疾病预测研究

现有基于网络搜索数据的疾病预测研究中主要使用的方法包括以下三种：

1）关键词的时序变化

网络搜索关键词与相关事物之间除了具有相关关系，还具有时间轴上的对应关系，即关键词搜索量的峰值与事物变化趋势的峰值在时间上的对应。采用时差相关分析法和峰谷对应分析法可识别出关键词的时序特征，其中具有领先特征的关键词可以预测未来。卢洪涛等[14]对 H7N9 禽流感关键词的时序变化特征进行了分析，发现具有领先特征的关键词主要集中于禽流感病毒、防治及疾病名称上，他们指出可以采用这些关键词组进行 H7N9 禽流感爆发初期趋势的预测。

2）状态空间模型（state space model）

状态空间模型以隐含的时间为自变量，分析指定的时间序列数据之间是否存在同时、领先、滞后及回馈等相关关系。杨艳红等[15]使用该模型分析了关键词“乙肝”的谷歌趋势和实际乙肝的发病数据，并预测了 25 周内的乙肝情况，总误差为 8.02%。

3）动态线性模型（dynamic linear model）

动态线性模型由测量方程和状态方程两部分组成，前者可根据时刻 t 的参数向量描述此时对应的因变量 Y，后者可建立时刻 t 的参数向量与下一时刻参数向量之间的联系，从而达到预测的目的。Cao 等[16]利用最基本的动态线性模型比较了单一流感数据和多种流感数据对流感疫情的预测效果，结果表明多种数据的预测效果更加精准。

综上所述，现有基于网络搜索数据的疾病预测研究较少，且多使用线性模型方法。然而网络搜索数据和流感数据之间的相关性会受多种因素影响[17]，

二者之间的相关关系可能呈非线性局势，也就是说线性模型可能无法较好地拟合。此外，虽存在探讨禽流感关键词时序变化的研究，但仅停留在关键词特征的层面，并没有在此基础上做进一步的预测。因此，针对国内外研究现状的不足，本文拟根据关键词时序变化的趋势，融合百度搜索数据及官方流感数据，拟合并比较多种模型对流感的预测效果，为相关搜索引擎和疾病防控中心提供参考。

3 理论框架与模型原理

本节首先从理论上对网络搜索数据与流感病例数据之间的逻辑关系进行探讨，建立理论框架模型；然后具体介绍关键词的选择依据与过程；最后结合相关文献依次对多元线性回归、BP 神经网络及支持向量机这三种模型的原理进行说明。

3.1 理论基础与逻辑框架

有学者认为，所谓健康信息需求，是指个体出现自我感觉身体不适或曾有高危行为导致其对健康状况表示怀疑或不确定时，主动寻求相关健康知识的状态[18-19]。这种健康信息需求会驱使人们产生一系列健康信息搜寻行为。在搜寻的过程中，人们可以使用各种方法来发现并使用某种信息资源（如报纸、图书、人际沟通、网络检索等），其中强调和计算机系统进行交互的健康信息检索行为是搜寻行为的一部分[20]。由于互联网的迅速普及和私密、便捷等优点，网络正逐渐成为人们获取健康信息的高效渠道，因此健康信息检索行为也映射到了互联网中，体现为对网络搜索或其他服务的使用。搜索引擎作为最重要且最常用的网络搜寻工具，其数据具有很强的代表性，可以在一定程度上提高流感预警的准确性，所以本文主要聚焦于网络搜索数据的挖掘与提炼。

基于以上理论基础，笔者融合个体健康情况、健康信息需求和网络健康信息搜寻行为三方面来解释搜索数据与流感病例数之间的关联性，关联框架如图 1 所示。

对于流感这一具体疾病来说，个体的患病过程大体可以分为三个阶段：流感预防阶段、出现症状阶段和确诊阶段。流感预防阶段指个体由于身边患流感人数增加或流感高发季节将至而提前采取相关预防措施，此时其信息需求倾向于如何防止自己患上流感，相对应的常用搜索关键词有“流感预防”“如何预防流感”“流感疫苗”等。这一阶段是患病过程的起始点，领先于确诊的时间周期最长。若预防成功，则信息搜寻行为停止；若预防失败，则进入出现症状阶段。顾名思义，该阶段是指个体出现了类似流感感染的临床症状，此时的信息需求多为确认该症状是否为流感或其他疾病，在搜索数据上会表现出以症状为主的搜索词，如“发烧”“咳嗽”“感冒”等，这一阶段领先于确诊的时间周期相对较短。通过获取相关知识或咨询医生，流感患者将会进入确诊阶段，此时的关注点主要在流感的治疗方法和药物上，相关词汇通常为“流感治疗”或“感康”“快克”“奥司他韦”等针对流感的药物名称。确诊阶段可直接表现为

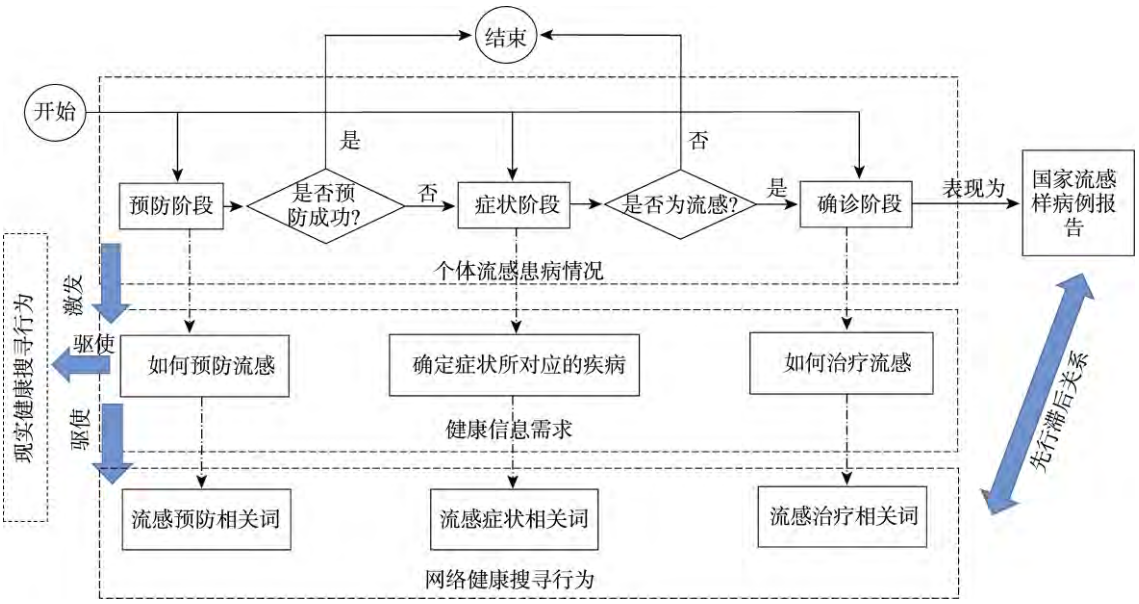

图 1 网络搜索数据与国家流感样病例报告的关联框架

国家流感样病例数的增加，但由于哨点医院对病例的收集和处理需要经过一个复杂的流程，因此该阶段具有较为有限的领先性，同时也可能出现一些关键词同步或滞后于流感样病例报告的情况，其中具有滞后关系的搜索数据并不具有监测作用。

3.2　关键词选择原理

关键词选择是整个研究的第一步，其具体步骤又可分为关键词初选和互相关分析（图 2）。其中，互相关分析的目的一方面在于判断初选词是否与流感病例数相关，另一方面则有助于判断相关词与流感样病例报告的先行滞后关系。

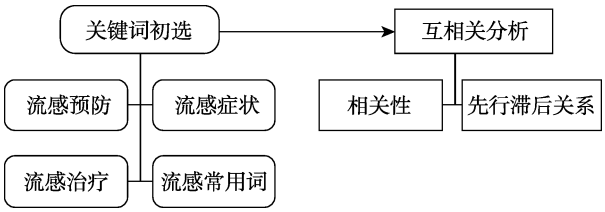

图 2　关键词选择的具体步骤

3.2.1　关键词初选

关键词的初选方法目前主要有三种：技术选词、范围选词以及直接选词。技术选词指尽可能地将所有常用搜索词列入研究范围，然后利用高性能的计算机系统以及自主编译的软件程序依次对这些搜索词进行筛选。使用该方法的代表文献为 Ginsberg 等[1]发表的“Detecting influenza epidemics using search engine query data”，他们对 Google 搜索引擎数据中 5000 万常用搜索词的周平均数作了统计，最后筛选出 45 个相关词。虽然技术选词具有精度高、客观性强等优点，但由于时间、资金、数据来源和技术设备等多方面限制，利用该方法选取关键词的研究并不多。

相对于技术选词，范围选词的方法在很大程度上降低了研究人员对信息技术的依赖性。因为关键词在进入统计学相关性筛选步骤之前，研究人员首先需要根据待研究对象确定一个大致的关键词选择范围，该范围的划定可大大减少关键词数量，以确保适当的工作量。李秀婷等[21]在探讨网络搜索数据对流感快速预警方面所起的作用时，就将搜索词样本分为流感症状、治疗及预防和一般搜索词三个构成范围。这种方法虽然没有技术选词的高精准性，但可以最大限度地减少核心关键词的遗漏；不足之处在于确定范围的标准通常根据研究者的经验，经验主义带来的主观性降低了结果的准确度。

第三种方法操作最为简洁，但同时也最为主观。直接选词法指研究人员根据主观经验直接确定一个或几个最为相关的词汇，例如，杨艳红等[17]在拟合乙型肝炎预测模型时直接使用“乙肝”的搜索量建模，李锐等[22]在探讨禽流感病毒 H5N1 的暴发监测时选取了“H5N1”“发烧”“流感”和“咳嗽”作为关键词。这种方法虽然容易实现，但可能错过一些重要词汇，导致较为严重的偏差。

基于对以上方法的分析，本文倾向于采取范围选词法进行关键词搜集。首先，笔者根据搜索数据与流感病例数的关联框架图，将关键词分为流感预防、流感症状和流感治疗三部分。此外，以“流感”“流行性感冒”“甲型流感”等疾病名称为代表的关键词贯穿于整个患病过程中，因此将该类型关键词划分为流感常用词。综上所述，笔者将研究范围缩小到流感预防、流感症状、流感治疗以及流感常用词四个维度。除了通过先验知识和流感基本发病信息搜集关键词外，笔者还总结了前人相关研究中的关键词，并运用了百度搜索引擎的关键词推荐功能以及站长之家的长尾关键词挖掘工具进行词汇补充，使得关键词的初选工作更为完善。

3.2.2　互相关分析

互相关分析（cross correlation）是利用互相关系数 r 来估计整个时间序列中两个序列之间相关程度的一种标准方法。其原理在于将被选指标相对于基准指标前后移动若干个时间单位，然后对移动后的序列和基准指标序列求相关系数，最大的相关系数所对应的移动时间就是该指标领先或延迟的时间段。具体计算公式如下：

假设基准指标 $Y(i)$，被选指标 $X(i)$，其中 $i=1,2,3,\cdots,n$，则移动 d 时的互相关系数 r 为

$$r(d)=\frac{\sum_{i}^{n}\left[(x_{(i+d)}-\overline{x})(y_{(i)}-\overline{y})\right]}{\sqrt{\sum_{i}^{n}(x_{(i+d)}-\overline{x})^2}\sqrt{\sum_{i}^{n}(y_{(i)}-\overline{y})^2}}$$

$$d=0,\pm1,\pm2,\ldots,\pm D$$

其中，d 为延迟数，取不同值时具有不同含义。$d=0$ 时表示不移动，代表两个序列同步；d 取负值时表示被选指标向前移动，代表被选指标相对于基准指标先行；相反，d 取正值时表示被选指标向后移动，即被选指标相对于基准指标滞后。D 表示最大延迟数。

在本研究中，基准指标 Y 为国家流感中心发布的流感检测阳性数，被选指标 X 为初选词的百度搜

索指数。对于某一初选词，计算并比较延迟数 d 取不同值时的相关系数 r，并取 r 的最大值。其中，r 的最大值可反映出初选词的百度指数曲线与流感阳性数曲线的最大相似程度，当二者的最大相似程度达到一定阈值，才能认为该初选词可以在一定程度上反映流感疫情。以相关系数 r 为指标对初选词进行过滤后，可得到与流感数据具有较强相关性的关键词。此时，需继续判断关键词在时间上的性质类别。对于某一关键词，其 r 值最大时所对应的 d 值即为该关键词的百度指数 X 与我国流感阳性数 Y 之间的时间关系；若 d=0，则表示 X 和 Y 在时间上同步，说明该词可实时对我国流感疫情做出反馈；若 d 取负值，则表示 X 在时间上领先于 Y，说明该词可提前对我国流感疫情做出预测；本研究将前者称为同步关键词，后者称为先行关键词，其中只有先行关键词才具有预测价值。

3.3 算法原理

通过关键词的选择，仅能得到网络搜索数据与流感疫情具有相关性，但该数据对流感样病例数的具体解释能力如何，需要选择一个合适的算法，用来近似地表达变量间的变化关系。本节分别介绍了多元线性回归、BP 神经网络和支持向量回归三种算法的原理和特点，并探讨了其在流感领域的应用情况。

3.3.1 多元线性回归

多元线性回归模型的一般形式可表示为

$$y=\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_kx_k+\varepsilon$$

其中，$\beta_0,\beta_1,\beta_2,\cdots,\beta_k$ 是模型的参数，ε 为误差项。它表明 y 是 $x_1,x_2,\cdots,x_k$ 的线性函数（$\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_kx_k$ 部分）加上误差项 ε。误差项反映了除 $x_1,x_2,\cdots,x_k$ 与 y 的线性关系之外的随机因素对 y 的影响，是不能由 $x_1,x_2,\ldots,x_k$ 与 y 之间的线性关系所解释的变异性。

在流感监测领域，Culotta[23]构建了多元线性回归模型

$$\text{logit}(P)=\beta_1\text{logit}(Q(\{w_1\},D))+\cdots+\beta_k\text{logit}(Q(\{w_k\},D))+\beta_{k+1}+\varepsilon$$

其中，P 为患流感症状的人占所有人口的比例，$Q(W,D)$为所有文档集合 D 中包含关键词集合 W 的百分比，关键词集合 W 包含从 w_1 到 w_k 的 k 个关键词。

本研究中将关键词的领先周期纳入到多元线性回归模型当中，表示为

$$y_t=\beta_0+\beta_1x_{1(t-a1)}+\beta_2x_{2(t-a2)}+\cdots+\beta_kx_{k(t-ak)}+\varepsilon$$

其中，y_t 为 t 时间段内流感阳性数占流感样病例总数的百分比；x_1 至 x_k 分别代表 k 个关键词中每一个关键词的百度搜索指数；a_1 至 a_k 为关键词 x_1 至 x_k 相对于流感样病例报告发布时间的先行周期。

3.3.2 BP 神经网络

多元线性回归中，多重共线性可能会造成模型不稳定，且在向后剔除变量的过程中，与因变量之间相关性较小的自变量将会逐渐移去，从而减少原数据包含的信息。人工神经网络模型（artificial neural networks, ANN）由多个神经元组成，神经元之间通过可调的连接权值连接，其优点在于强大的非线性趋近性和泛化能力。BP 神经网络是最常见的神经网络模型，其基本思想如图 3 所示。

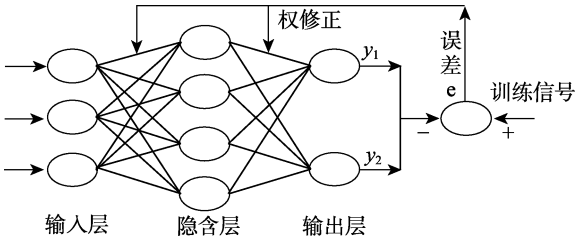

图 3 BP 神经网络模型的运行流程

输入变量从输入层进入，经过隐含层处理后，由输出层输出，如果输出层的输出值与真实输出变量相差太多，则进入误差反传播阶段，以某种形式将误差反传至隐含层的各个单元。各单元调整权值后重新输出，周而复始地进行，直到网络误差减小到可以接受的范围内，训练过程结束。

已有一些学者尝试将神经网络算法应用于流感预测中。Xu 等[24]利用不同的神经网络模型模拟流感数据与搜索引擎查询数据之间的关系，并采用交叉验证法选取最优神经网络进行预测。Xu 等[25]比较了传统线性模型和前馈神经网络模型（FNN）对流感样病例的预测效果，发现 FNN 能够更好地预测流感爆发高峰。

3.3.3 支持向量机

由于 BP 神经网络的泛化能力较低，且模型结构难以确定，导致该算法在实际中的应用受到一定限制。支持向量机基于结构风险最小化原则，具有严格的理论和数学基础，泛化能力较强，因此笔者将该模型纳入流感预测研究中。

由于本研究中待预测的流感阳性数所占比例为数值型变量，因此主要对支持向量机中的支持向量回归机做简要介绍。简单来说，支持向量回归机是通过升维，在高维空间中构造线性决策函数来实现线性回归。此外，为适应训练样本集的非线性，支

持向量回归算法采用核函数使原来的线性算法“非线性化”。

支持向量机在流感预测领域中的应用并不多见。韩国学者 Woo 等[26]基于社会媒体数据和 Web 查询扩展建立了支持向量回归模型，证明使用搜索查询提高韩国流感疫情监测具有可行性。我国学者卢汉体等[27]从流感相关疾病病例数、各类气象因素以及流感病原阳性率中选取与流感样病例数相关的因素，对比了支持向量回归模型和统计预测模型对流感疫情的预测效果，发现前者更加精准。

4　实证分析

4.1　数据来源

本研究的数据来源主要包括两部分，一部分为我国对流感疫情监测的官方数据，另一部分为同时间段内的搜索引擎数据。

4.1.1　中国国家流感中心流感周报

流感样病例（influenza-like illness，ILI）是国内外对流行性感冒病情监测的一个主要指标，它是指哨点医院所有门诊病例中发热（体温≥38℃）并伴有咳嗽或咽痛症状之一者。我国的流感样病例数据由分布在全国 31 个省（自治区、直辖市）的 185 家国家级流感样病例监测哨点医院和 155 家非国家级流感样病例监测哨点医院提供，其中南方省份每家哨点医院每周采集 5~15 份流感样病例标本；北方省份每家哨点医院在 4—9 月每月采集 5~15 份流感样病例标本，10 月至次年 3 月每周采集 10~15 份流感样病例标本[28]。这些数据按周上报给国家流感中心，再由该机构汇总、统计并发布流感周报，通常较实际疫情延后 7~10 天。

本研究采用的流感疫情数据来自中国国家流感中心网站发布的流感样病例周报（http://www.cnic.org.cn/chn/down/?typeid=20），样本时间段为 2012 年第 52 周（2012 年 12 月 29 日始）至 2015 年第 14 周（2015 年 4 月 5 日止）。

4.1.2　百度指数

网络数据来源于百度搜索引擎的百度指数（http://index.baidu.com/）。百度指数是以百度海量网民行为数据为基础的数据分享平台，其搜索指数以网民在百度的搜索量为数据基础，以搜索词为统计对象，通过科学的方法计算出每个搜索词在百度搜索引擎中被搜索频次的加权和。

4.2　关键词选择

4.2.1　关键词初选

本文根据流感的发病阶段和常用流感搜索词，选定了“流感疫苗”“感冒”“流感治疗”“流感药物”“H1N1”等较为原始的搜索词，然后通过对其他相关研究所选关键词的总结以及搜索引擎的关键词推荐，扩充每种关键词类型的词语数量。

需要注意的是，在百度指数中，个别关键词由于搜索频次过低导致搜索指数显示为 0，如“气短”“副流感病毒”；有些词因为不常见所以并没有被百度收录，从而不予显示其搜索量，如“偏肺病毒”“冠状病毒”。这些关键词都属于无效词汇，笔者将其剔除后，对剩余有效关键词进行归纳，形成初始词表，具体如表 1 所示。

4.2.2　互相关分析结果

本研究利用 SPSS 的 Cross Correlation 功能计算出初始词表中每个词和官方数据的互相关系数。对于每个关键词，需要在计算所得结果的 30 个相关系数中，找出与流感疫情官方数据的相关系数绝对值最大值，该最大值需满足大于或等于 0.5 的条件，以保证数据之间的相关性。在确定具有相关性之后，可继续判断该关键词在时间上的性质类别（先行、同步还是滞后）。例如，对于“流感疫苗”关键词，其互相关分析结果如表 2 所示。表中相关系数绝对值最大的为 0.610，相对应的先行时间为 11 周，因此被划分为先行指标。

同理，对于关键词“发烧”，表 3 中相关系数绝对值最大的为 0.680，相对应的先行滞后时间为 0 周，因此确定为同步指标。滞后类指标无论是用于监测还是预测都缺乏时效性，因此被剔除。

根据相关性和先行滞后性的考察，最后得到 22 个关键词，其中有 13 个属于先行类，9 个为同步类，具体关键词编号和先行时间如表 4 所示。

由表 4 可知，先行十周左右的关键词编号为 k1，k10 和 k12，所对应的内容为“流感疫苗”“流感疫苗副作用”和“流感疫苗有必要打吗”，都属于流感预防阶段；先行一周左右的关键词内容有关于流感症状的，如编号为 k4 的“感冒”、编号为 k21 的“高烧”，也涉及具体的药物名称，如编号为 k39 的“泰诺”、编号为 k51 的“康泰克”；同步类关键词主要为常用搜索词或治疗词，如编号为 k70 的“甲流是什么”、编号为 k49 的“甲流治疗”等。总之，搜索关键词和流感疫情的时差关系与本文给出的关联框架大致相同，在一定程

表 1 关键词初选结果及编号

类型	关键词(编号)	扩展词(编号)	类型	关键词(编号)	扩展词(编号)	类型	关键词(编号)	扩展词(编号)
预防阶段	流感疫苗(k1)	流感疫苗副作用(k10) 禽流感疫苗(k11) 流感疫苗有必要打吗(k12)	治疗阶段	打喷嚏(k27) 头痛(k28) 寒颤(k29)	头疼(k40)	常用词	阿莫西林(k57) 金刚烷胺(k58) 连花清瘟胶囊(k59)	
	流感预防(k2)	禽流感预防(k13)		呼吸道感染(k30)	上呼吸道感染(k41)		流行性感冒(k60)	
	预防流感(k3)	如何预防流感(k14) 怎样预防流感(k15) 预防禽流感(k16)		肺炎(k31) 支气管炎(k32)	肺炎症状(k42) 气管炎(k43)		流感(k61)	猪流感(k68) 流感病毒(k69)
症状阶段	感冒(k4)	病毒性感冒(k17) 肠胃感冒(k18) 感冒症状(k19)		流感症状(k33)	猪流感症状(k44) 甲流症状(k45) 甲型流感症状(k46) 甲流的症状(k47) 禽流感症状(k48)		甲流(k62)	甲流是什么(k70) 甲流病毒(k71)
	发烧(k5)	发热(k20) 高烧(k21)		流感治疗(k34)	甲流治疗(k49)		甲型流感(k63)	甲型流感病毒(k72)
	体温(k6)			流感药物(k35)	流感吃什么药(k50)		禽流感(k64)	禽流感最新消息(k73) 禽流感病毒(k74) 禽流感传播途径(k75)
	鼻炎(k7)	鼻塞(k22) 流鼻涕(k23) 鼻窦炎(k24)		感冒药(k36)	康泰克(k51) 感康(k52) 快克(k53)		H1N1(k65)	H1N1 流感(k76) 甲型 H1N1 流感(k77)
	咽喉炎(k8)	咽炎(k25) 咽喉痛(k26)		特敏福(k37)	达菲(k54)		H7N9(k66)	H7N9 型禽流感(k78) H7N9 禽流感最新消息(k79)
	咳嗽(k9)			扎那米韦(k38)	奥司他韦(k55)		H5N1(k67)	
				泰诺(k39)	白加黑(k56)			

表 2 “流感疫苗”关键词的互相关分析结果

先行（滞后）周数	先行值与官方数据的相关系数	滞后值与官方数据的相关系数
0	–0.008	–0.008
1	0.046	–0.083
2	0.074	–0.148
3	0.107	–0.175
4	0.175	–0.178
5	0.262	–0.173
6	0.354	–0.162
7	0.445	–0.153
8	0.520	–0.145
9	0.583	–0.137
10	0.609	–0.128
11	**0.610**	–0.130
12	0.585	–0.145
13	0.536	–0.179
14	0.476	–0.214
15	0.395	–0.247

表 3 “发烧”关键词的互相关分析结果

先行（滞后）周数	先行值与官方数据的相关系数	滞后值与官方数据的相关系数
0	**0.680**	**0.680**
1	0.662	0.595
2	0.599	0.458
3	0.506	0.321
4	0.406	0.193
5	0.312	0.078

表 4 “先行”关键词与“同步”关键词

类型	关键词	相关系数	先行时间	类型	关键词	相关系数	先行时间
先行	k39	0.63	–1	先行	k12	0.572	–10
先行	k51	0.623	–1	先行	k1	0.61	–11
先行	k34	0.613	–1	同步	k70	0.764	0
先行	k52	0.588	–1	同步	k5	0.68	0
先行	k4	0.583	–1	同步	k46	0.613	0
先行	k60	0.572	–1	同步	k50	0.607	0
先行	k21	0.556	–1	同步	k49	0.601	0
先行	k56	0.555	–1	同步	k30	0.6	0
先行	k22	0.502	–1	同步	k9	0.58	0
先行	k18	0.549	–2	同步	k63	0.555	0
先行	k10	0.61	–10	同步	k43	0.5	0

度上印证了理论基础的可行性。此外，由于只有先行类词语可用来拟合流感预测模型，因此笔者仅选择了 13 个先行类关键词进入下一步的模型建立与比较。

4.3　流感预测模型的建立

由前文对关键词的分析结果可知，根据先行时间的不同，先行类关键词可分为“先行十周”和“先行一周”，因此在建立预测模型时也可按先行时间的不同分别拟合。图 4 为本研究拟建立的模型汇总图。

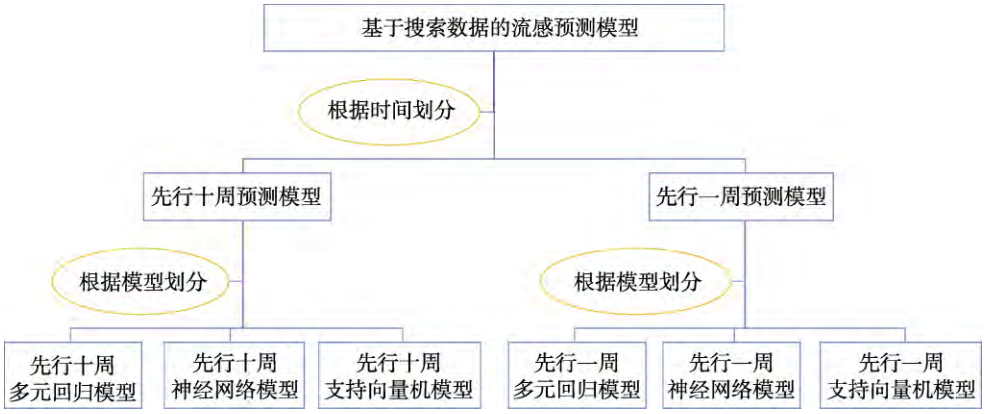

图 4　模型汇总图

4.3.1　“先行十周”预测模型

本研究选取 2012 年第 52 周至 2014 年第 52 周的数据用于参数率定及模型训练。需要注意的是，对于“先行十周”预测模型来说，输入数据和输出数据之间需有 10 周的时间差，因此输入数据的时间范围为 2012 年第 52 周至 2014 年第 42 周，输出数据范围为 2013 年第 9 周至 2014 年第 52 周。

4.3.1.1　“先行十周”多元线性回归模型

多元线性回归模型在基于搜索数据的流感疫情预测中最为常见。本节以 SPSS 为建模工具，建立“先行十周”的多元线性回归模型。

1）变量选择

对于可提前十周预测的多元回归模型，其变量选择如下：被解释变量为流感阳性数占流感样病例总数的百分比 y，解释变量依次为“流感疫苗”的百度指数（x_1）、“流感疫苗副作用”的百度指数（x_2）、“流感疫苗有必要打吗”的百度指数（x_3）以及提前十周时可获得的流感检测阳性数（x_4）。

2）模型构建与检验

选取“向后”剔除法，剔除对于因变量无显著影响（$P>0.1$）且和其他自变量有密切关系的变量。根据 SPSS 输出结果，得到多元线性回归模型：

$$y_t=0.008x_{1(t-10)}+0.047x_{2(t-10)}-0.015x_{3(t-10)}-0.234\,x_{4(t-10)}+3.421 \quad (1)$$

方程总体线性显著性检验结果显示，F 统计量的概率值为 0.00，说明自变量和因变量间存在显著线性关系；变量显著性检验结果显示，x_1 至 x_4 变量的 P 值分别为 0.071、0.000、0.067 和 0.002，均小于 0.1，因此并没有被剔除；拟合优度检验结果显示，R^2为 0.626，说明自变量可以解释因变量 62.6%变化，拟合程度一般。

4.3.1.2　“先行十周”BP 神经网络模型

R 语言作为一种免费的统计分析软件，可加载大量第三方功能包，方便快捷地实现数据挖掘功能。笔者使用 R 中的 neuralnet 包建模计算，该包中的 neuralnet 函数可以构建神经网络模型，plot 函数可绘制神经网络示意图，compute 函数可计算新观测值的预测值。

1）读取数据

以“流感疫苗”“流感疫苗副作用”“流感疫苗有必要打吗”这三个关键词的百度指数以及提前十周时可获得的流感检测阳性数为输入变量，以流感阳性数占流感样病例总数的百分比为输出变量。

2）模型构建与检验

首先对训练样本做最大最小规范化，然后根据以下设置进行 BP 神经网络训练：隐层 hidden 为 7，迭代停止条件 threshold 为 0.01，损失函数 err.fac 为 see（误差平方），linear.output 取值为 FALSE，表示输出节点的激活函数为非线性函数，学习率 learningrate 为 0.001，算法 algorithm 为 backprop，即传统 BP 反向传播网络。得到神经网络拓扑图（图 5）。

拟合优度结果显示 R^2=0.871，效果良好。

4.3.1.3　“先行十周”支持向量机模型

笔者使用 R 语言中的 e1071 包建立流感预测的支持向量机模型。e1071 包可实现支持向量机，朴素贝叶斯、模糊聚类、装袋聚类等多种算法，其中，svm

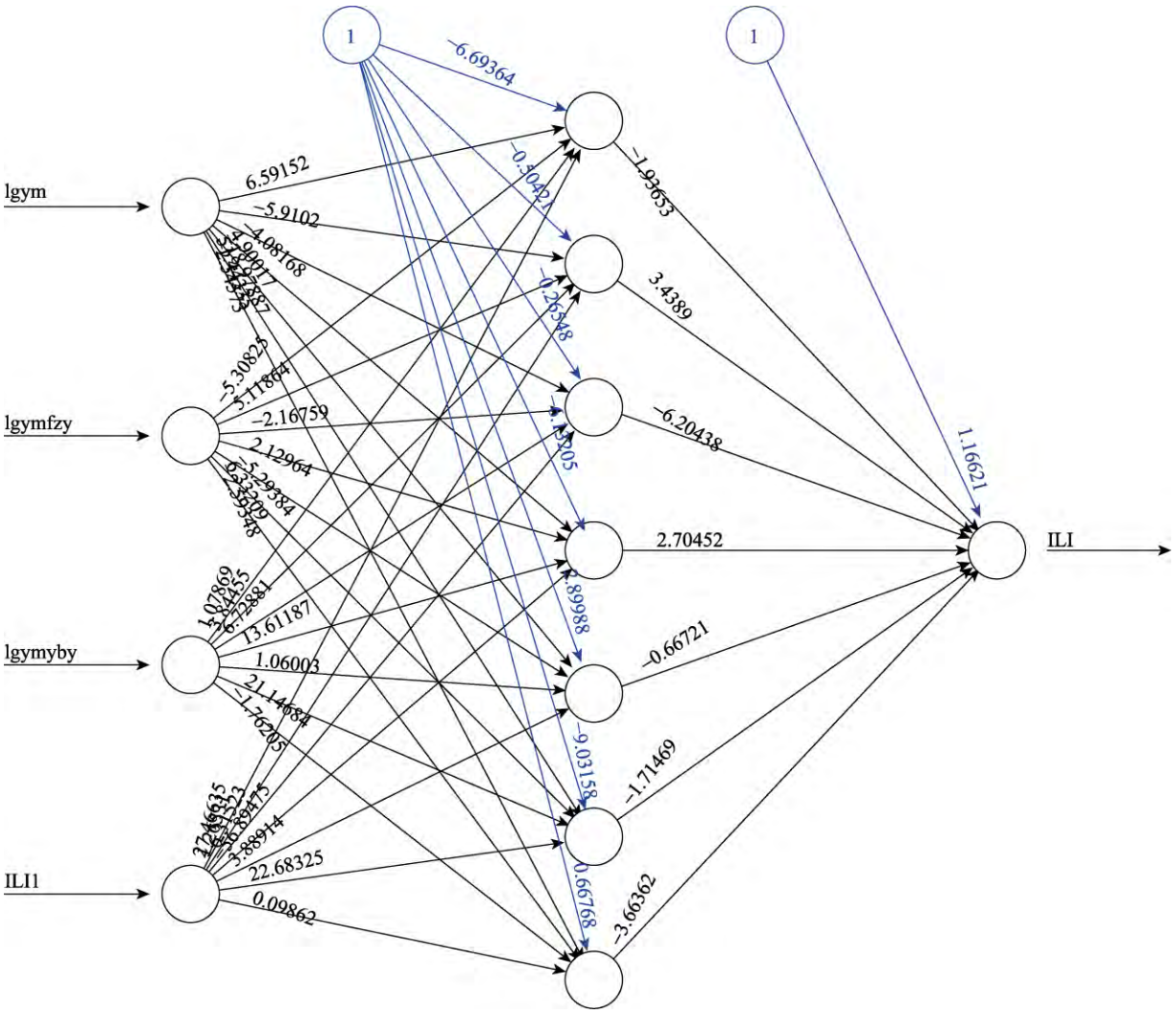

图 5 “先行十周”神经网络拓扑图

函数可实现支持向量回归。

1）读取数据

输入变量及输出变量与建立神经网络模型的数据相同。

2）模型构建与检验

对已最大最小规范化后的训练数据建立支持向量回归模型，参数设置中 type 值为“eps-regression”，表示以 eps 不敏感损失函数为基础的支持向量回归。通过计算，可得 R^2=0.657，小于神经网络的拟合优度（0.871），但和多元线性回归（0.626）相比有所提高。

4.3.2 “先行一周”预测模型

与“先行十周”预测模型类似，“先行一周”预测模型的训练集中，输入数据时间范围为 2012 年第 52 周至 2014 年第 51 周，输出数据范围为 2013 年第 1 周至 2014 年第 52 周。

4.3.2.1 “先行一周”多元线性回归模型

1）变量选择

被解释变量为流感阳性数占流感样病例总数的百分比 y，解释变量依次为“泰诺”“康泰克”“流感治疗”“感康”“感冒”“流行性感冒”“高烧”“白加黑”“鼻塞”“肠胃感冒”“流感疫苗副作用”“流感疫苗有必要打吗”和“流感疫苗”的百度指数（$x_1, x_2, \cdots, x_{13}$），以及提前一周时可获得的流感检测阳性数（x_{14}）。

2）模型构建与检验

选取“向后”剔除法，剔除对于因变量无显著影响（$P > 0.1$）且和其他自变量有密切关系的变量。根据 SPSS 输出结果，得到多元线性回归模型：

$$y_t = 0.008\, x_{1(t-1)} - 0.005 x_{2(t-1)} + 0.005\, x_{10(t-1)} - 0.003\, x_{12(t-1)} + 0.846\, x_{14(t-1)} - 8.003 \quad (2)$$

其中，自变量 $x_3, x_4, \cdots, x_9$ 及 x_{11}、x_{13} 被剔除，从已排除变量的 T 检验结果也可看出以上自变量的 P 值分别为 0.517、0.768、0.887、0.910、0.829、0.827、9.830、

0.425 和 0.453，均大于 0.1，确实不能引入回归模型中；其他自变量的 P 值分别为 0.024（x_1）、0.030（x_2）、0.011（x_{10}）、0.062（x_{12}）、0.000（x_{14}）。拟合优度检验结果表明，在剔除自变量的过程中，R^2由 0.962 降至 0.960，变化较小。

4.3.2.2　“先行一周”BP 神经网络模型

1）读取数据

以“先行一周”关键词的百度指数和提前一周时可获得的流感检测阳性数为输入变量，以流感阳性数占流感样病例总数的百分比为输出变量。

2）模型构建与检验

经多次实验，隐层中神经元数量为 5 效果最好，可得神经网络拓扑图（图 6）。从拟合优度来看，R^2=0.955，与“先行一周”多元线性回归模型效果相差不大。

4.3.2.3　“先行一周”支持向量机模型

1）读取数据

输入变量及输出变量与建立“先行一周”神经网络模型的数据相同。

2）模型构建与检验

以规范化后的训练数据建立支持向量回归模型，拟合优度检验结果显示 R^2=0.932，拟合效果较好。

4.4　流感预测模型的效果评估

本研究选取 2015 年第 1 周到第 13 周的数据作为测试样本，分别比较了“先行十周”的 3 个模型以及“先行一周”3 个模型的预测能力。

4.4.1　“先行十周”模型预测能力比较

从预测的整体趋势拟合效果看，图 7 展示的是

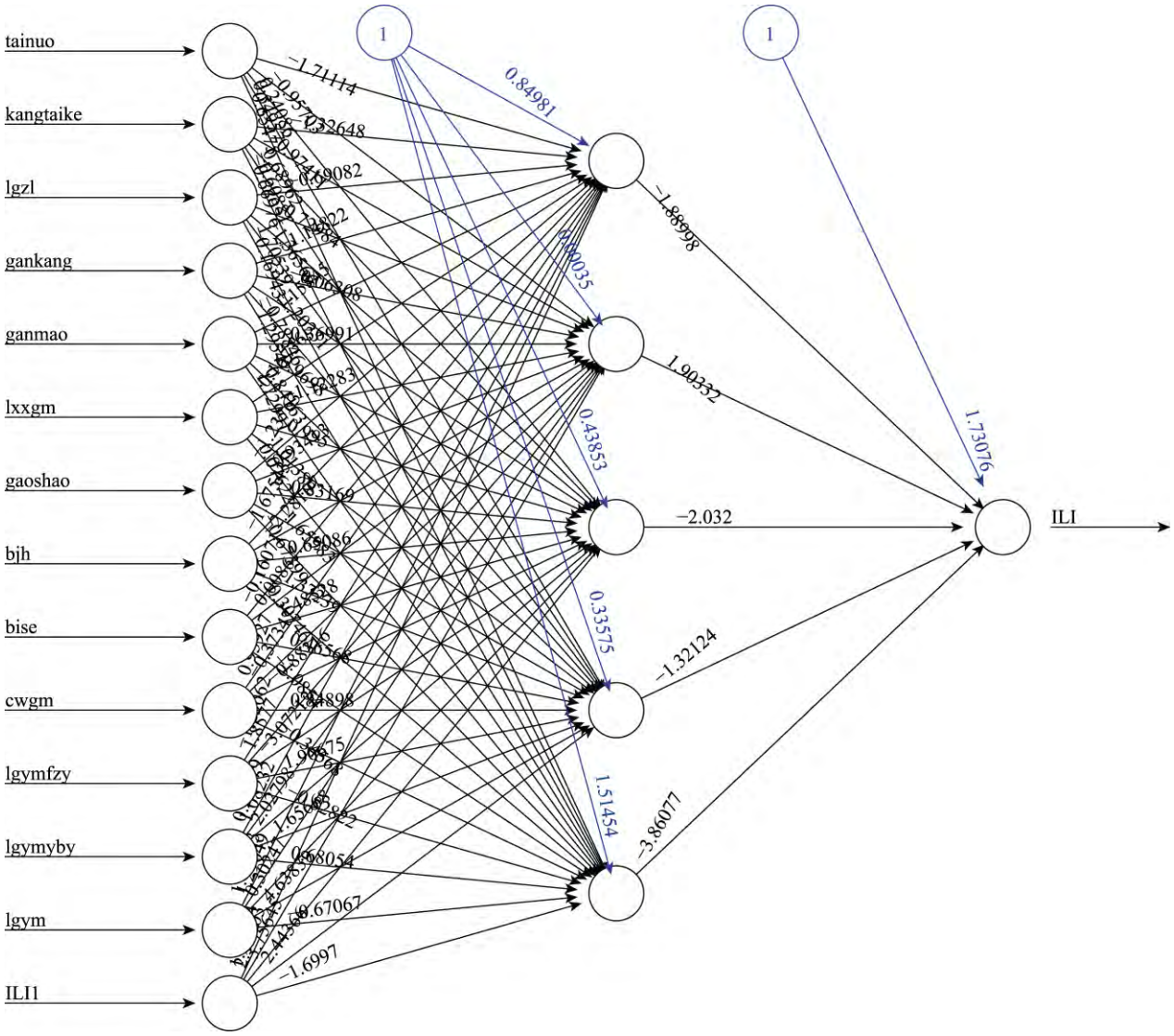

图 6　“先行一周”神经网络拓扑图

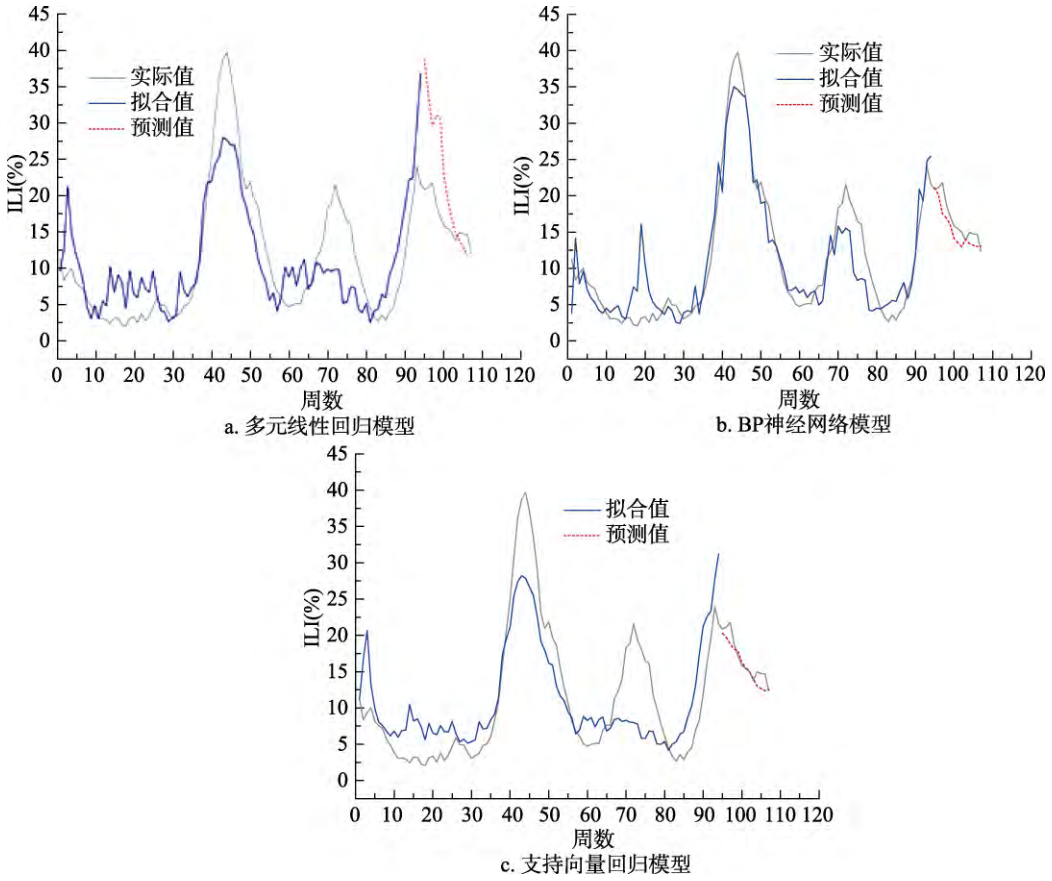

图 7 “先行十周”的流感预测模型的可视化效果图

分别使用三个模型进行预测的可视化效果图，其中浅色实线为流感阳性数占流感样病例总数的百分比，深色实线代表模型拟合值，深色虚线代表模型预测值。三个模型中，从拟合效果看，BP 神经网络模型的拟合效果最好，拟合优度为 0.871；支持向量回归模型和多元线性回归模型的拟合效果相差不大，分别为 0.657 和 0.626；从预测效果看，BP 神经网络和支持向量回归模型的输出更接近实际情况，便于推广。

具体到预测精度上，从表 5 中可以得到，支持向量机模型的平均相对误差(MRE)最小，仅为 0.06；其次是 BP 神经网络模型（0.09）；而多元线性回归模型的误差最大，为 0.35。

4.4.2 “先行一周”模型预测能力比较

图 8 是三个“先行一周”模型仿真值与预测值的可视化效果图，可以看到每个模型的仿真拟合效果都不错。预测能力和“先行十周”模型类似，同样是 BP 神经网络和支持向量回归模型更贴近现实，通用性较强。

具体到预测精度上，从表 6 可以得到，三个模型的平均相对误差（MRE）相差不大，由小到大分别为支持向量机（0.05）、BP 神经网络（0.07）和多元线性回归（0.08）。

表 5 “先行十周”模型预测值与实际值的相对误差

周数	“先行十周”预测模型的相对误差		
	多元线性回归	BP 神经网络	支持向量机
第 1 周	0.87	0.01	0.03
第 2 周	0.60	0.05	0.07
第 3 周	0.35	0.20	0.14
第 4 周	0.65	0.10	0.04
第 5 周	0.79	0.07	0.04
第 6 周	0.43	0.11	0.03
第 7 周	0.27	0.13	0.01
第 8 周	0.13	0.14	0.00
第 9 周	0.08	0.02	0.01
第 10 周	0.08	0.10	0.13
第 11 周	0.12	0.11	0.14
第 12 周	0.20	0.11	0.16
第 13 周	0.01	0.06	0.01
平均相对误差（MRE）	0.35	0.09	0.06

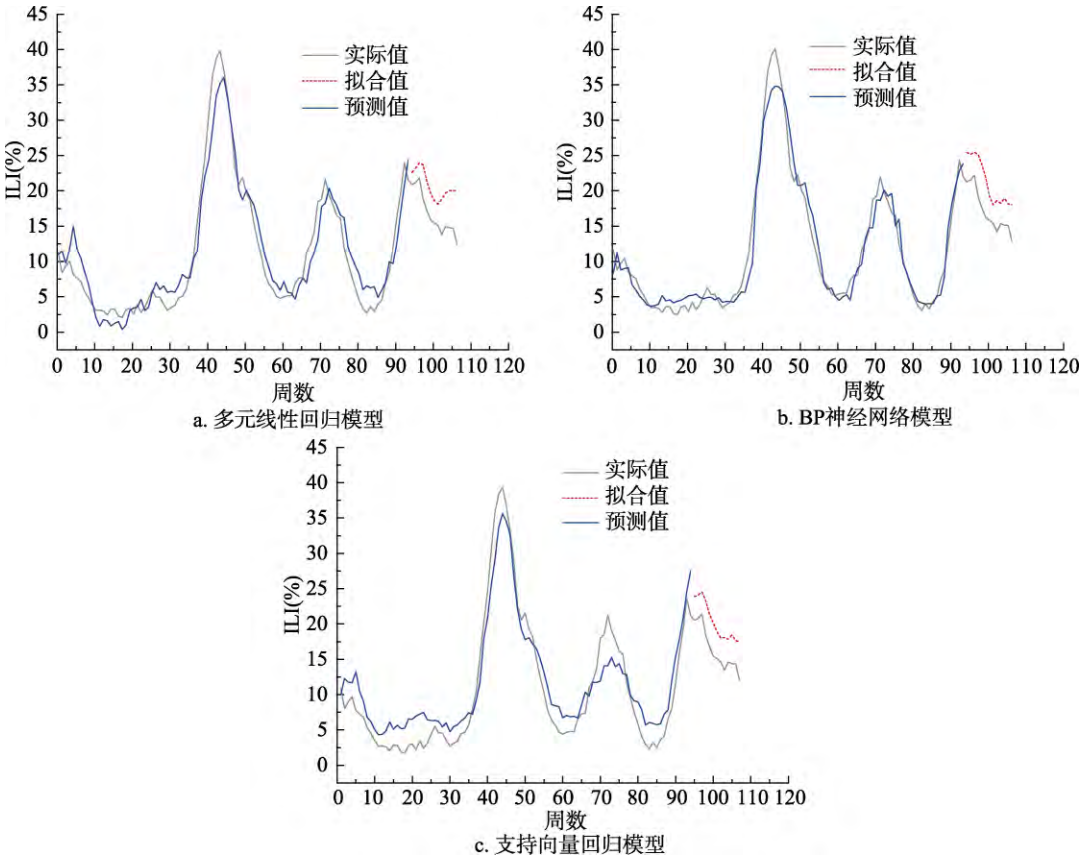

图 8　“先行一周”的流感预测模型的可视化效果图

表 6　“先行一周”模型预测值与实际值的相对误差

周数	“先行一周”预测模型的相对误差		
	多元线性回归	BP 神经网络	支持向量机
第 1 周	0.13	0.02	0.05
第 2 周	0.12	0.04	0.06
第 3 周	0.11	0.06	0.06
第 4 周	0.02	0.08	0.02
第 5 周	0.02	0.09	0.00
第 6 周	0.04	0.07	0.01
第 7 周	0.09	0.07	0.04
第 8 周	0.09	0.13	0.08
第 9 周	0.04	0.01	0.01
第 10 周	0.01	0.11	0.09
第 11 周	0.05	0.05	0.03
第 12 周	0.06	0.10	0.08
第 13 周	0.24	0.06	0.08
平均相对误差（MRE）	0.08	0.07	0.05

5　讨论与结论

近年来，通过搜索引擎和社交网络等用户生产内容平台监控并预测流行性疾病成为一个热门研究领域。无论是基于推特、维基百科、谷歌趋势还是雅虎搜索数据[17,29]，这些研究一方面往往仅关注关键词与流感疫情的相关关系，忽略对其预测本质的探讨；另一方面，使用方法通常是较为常见的线性模型。据此，本研究首先从理论上对网络搜索数据与流感样病例数据之间的逻辑关系进行探讨，融合个体健康情况、健康信息需求和健康信息搜寻行为三方面建立了理论框架模型；然后通过实证研究确定了和我国流感疫情既具有较强的相关关系又具有先行同步性的搜索关键词；最后选取三种代表性算法，比较了它们对流感的预测效果。

通过以上研究工作，论文的主要发现包括以下几方面：

（1）流感样病例数据和网络搜索数据之间的相关性可用信息行为等理论作为支撑，不同的患病阶段有着不同的健康信息需求，而每种不同需求在搜索行为和数据上又体现为不同的搜索关键词。实证研究中对关键词的时差相关分析结果印证了本文理论部分的分析，可先行十周左右对流感疫情进行预测的关键词内容都和流感疫苗相关，符合大众“如何预防流感”的信息需求；先行一周左右的关键词多涉

及流感的症状表现,对应理论框架中“确认症状”的需求；而同步类关键词多为常用搜索词或治疗方法，具有有限的领先性。

（2）多元线性回归在提前十周预测时误差较大，提前一周预测时准确度明显提高。究其原因，一方面可能因为先行时间为十周的关键词数量较少，另一方面则在于两个多元线性回归模型中起关键作用的自变量——历史数据中的流感检测阳性数。由于历史数据和搜索数据的所含信息具有一定程度的互补性[30]，因此在建立预测模型的过程中，输入变量由关键词的百度指数和流感历史数据共同构成。具体来看，公式(1)中自变量“提前十周流感检测阳性数”（即 x_4）的 P 值为 0.002，公式(2)中“提前一周流感检测阳性数”（即 x_{14}）的 P 值为 0.000，都具有显著统计学意义，说明历史数据中的流感检测阳性数对预测值贡献较大。另外，传统流感预测研究表明，越接近预测时间的流感阳性数与预测值之间相差越小；也就是说，提前一周的 x_{14} 比提前十周的 x_4 包含更多流感疫情信息，因此提前一周预测结果优于提前十周预测结果也在常理之中。

（3）多元线性回归虽然可较好地预测一周后的流感情况，但和其他非线性模型相比还是误差较大，这说明影响流感疫情的因素较多，简单的线性模型并不能很好地表达。对比两种非线性模型，从拟合结果看,BP 神经网络的拟合效果优于支持向量回归，且在提前十周预测时尤为明显。然而，拟合效果好并不一定代表预测精度准，从预测结果看，支持向量回归无论在提前十周还是一周的预测精度都更高，这说明过拟合现象会导致泛化能力变差，与其他领域的研究结果相吻合[31]。

此外，本研究的思路和方法可做进一步应用扩展。文献[19]比较了不同地区的用户搜索词与当地流感数据的相关性，发现地域因素（如当地的人群数量、生活习惯、互联网依赖程度等）对相关性的影响较大。因此在今后的研究中，可在省、市或县的层面进行分析，从而为不同省市的疾病防控中心提供更加有针对性的建议。

参 考 文 献

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

[2] Valdivia A, López-Alcalde J, Vicente M, et al. Monitoring influenza activity in Europe with Google Flu Trends: comparison with the findings of sentinel physician networks - results for 2009-10[J]. Eurosurveillance, 2010, 15(29): 2-7.

[3] Wada K, Ohta H, Aizawa Y. Correlation of “Google Flu Trends” with Sentinel Surveillance Data for Influenza in 2009 in Japan[J]. The Open Public Health Journal, 2011,4: 17-20.

[4] Cook S, Conrad C, Fowlkes A L, et al. Assessing Google Flu Trends performance in the United States during the 2009 influenza virus A (H1N1) pandemic[J]. PLoS ONE, 2011, 6(8): e23610.

[5] Cho S, Sohn C H, Jo M W, et al. Correlation between national influenza surveillance data and google trends in South Korea[J]. PLoS ONE, 2013, 8(12): e81422.

[6] Kang M, Zhong H J, He J F, et al. Using Google Trends for influenza surveillance in South China[J]. PLoS ONE, 2013, 8(1): e55205.

[7] 朱猛, 祖荣强, 霍翔, 等. 时间序列分析在流感疫情预测预警中的应用[J]. 中华预防医学杂志, 2011, 45(12): 1108-1111.

[8] Spink A, Cole C. Human information behavior: Integrating diverse approaches and information use[J]. Journal of the American Society for Information Science and Technology, 2006, 57(1): 25-35.

[9] 张崇, 吕本富, 彭赓, 等. 网络搜索数据与 CPI 的相关性研究[J]. 管理科学学报, 2012, 15(7): 50-59, 70.

[10] 王炼, 贾建民. 基于网络信息搜索的旅游需求预测——来自黄金周的证据[J]. 系统管理学报, 2014, 23(3): 345-350, 358.

[11] Kulkarni G, Kannan P K, Moe W. Using online search data to forecast new product sales[J]. Decision Support Systems, 2012, 52(3): 604-611.

[12] Song T M, Song J, An J Y, et al. Psychological and social factors affecting internet searches on suicide in Korea: A big data analysis of Google search trends[J]. Yonsei Medical Journal, 2014, 55(1): 254-263.

[13] Bardak B, Tan M. Prediction of influenza outbreaks by integrating Wikipedia article access logs and Google flu trend data[C]// Proceedings of the IEEE 15th International Conference on Bioinformatics and Bioengineering, Belgrade, 2015: 1-6.

[14] 卢洪涛, 李纲. 网络搜索关键词时序变化特征研究——以 H7N9 禽流感关键词实验为例[J]. 情报杂志, 2014, 33(11): 175-180.

[15] 杨艳红, 曾庆, 赵寒, 等. 基于谷歌趋势的乙型肝炎预测模型[J]. 上海交通大学学报(医学版), 2013, 33(2): 204-208.

[16] Cao P H, Wang X, Fang S S, et al. Forecasting influenza epidemics from multi-stream surveillance data in a subtropical city of China[J]. PLoS ONE, 2014, 9(3): e92945.

[17] Ortiz J R, Zhou H, Shay D K, et al. Monitoring influenza activity in the United States: A comparison of traditional surveillance systems with Google Flu Trends[J]. PLoS ONE, 2011, 6(4): e18687.

[18] 肖静. 高校教师健康信息行为研究[D]. 南京：南京航空航天大学, 2008.

[19] 张馨遥. 健康信息需求研究的内容与意义[J]. 医学与社会, 2010, 23(1): 51-53.

[20] Wilson T D. Human information behavior[J]. Informing Science: The International Journal of an Emerging Transdiscipline, 2000, 3: 49-56.

[21] 李秀婷, 刘凡, 董纪昌, 等. 基于互联网搜索数据的中国流感监测[J]. 系统工程理论与实践, 2013, 33(12): 3028-3034.

[22] 李锐, 孙利谦, 熊成龙, 等. 基于互联网搜索数据研究全球高致病性禽流感病毒 H5N1 的暴发监测[J]. 中华疾病控制杂志, 2015, 19(8): 773-777.

[23] Culotta A. Towards detecting influenza epidemics by analyzing Twitter messages[C]// Proceedings of the First Workshop on Social Media Analytics. New York: ACM Press, 2010: 115-122.

[24] Xu W, Han Z W, Ma J. A neural netwok based approach to detect influenza epidemics using search engine query data[C]// Proceedings of the International Conference on Machine Learning and Cybernetics, Qingdao, 2010: 1408-1412.

[25] Xu Q N, Gel Y R, Ramirez Ramirez L L, et al. Forecasting influenza in Hong Kong with Google search queries and statistical model fusion[J]. PLoS ONE, 2017, 12(5): e0176690.

[26] Woo H, Cho Y, Shim E, et al. Estimating influenza outbreaks using both search engine query data and social media data in South Korea[J]. Journal of Medical Internet Research, 2016, 18(7): e177.

[27] 卢汉体, 李傅冬, 林君芬, 等. 基于支持向量机的浙江省流感样病例预警模型研究[J]. 浙江大学学报(医学版), 2015, 44(6): 653-658.

[28] 中国国家流感中心. 中国流感监测方案(2010 年版)[EB/OL]. (2016-05-20) [2017-12-17]. http://www.chinaivdc.cn/cnic/fascc/201708/t20170809_149276.htm.

[29] Santillana M, Nguyen A T, Dredze M, et al. Combining search, social media, and traditional data sources to improve influenza surveillance[J]. PLoS Computational Biology, 2015, 11(10): e1004513.

[30] 王若佳, 李培. 基于互联网搜索数据的流感监测模型比较与优化[J]. 图书情报工作, 2016, 60(18): 122-132.

[31] 夏国恩, 金炜东. 基于支持向量机的客户流失预测模型[J]. 系统工程理论与实践, 2008, 28(1): 71-77.

（责任编辑　魏瑞斌）