Web舆情的长期趋势预测方法

高 辉[1]，王沙沙[2]，傅 彦[1,2]

(1. 电子科技大学计算机科学与工程学院 成都 610054；2. 电子科技大学软件学院 成都 610054)

【摘要】针对传统预测方法无法有效预测Web舆情的长期趋势中拐点的不足，提出一种长期趋势预测方法。该方法首先通过周期分析和层次聚类为每类已发生舆情事件的发展趋势建立类模型库，然后通过对待预测舆情事件已知发展趋势进行自适应变换后，应用最小二乘法从相应的类模型库中选取均方误差和最小的模型来预测该事件的未来发展趋势。实验证明，与传统方法相比该方法在预测舆情事件发展的长期趋势时有较高的关联度，能有效预测长期趋势中的拐点。

关 键 词 分类； 聚类； 周期分析； 长期趋势预测； 关联度分析

中图分类号 TP311.13 **文献标识码** A doi:10.3969/j.issn.1001-0548.2011.03.022

Prediction Model for Long-Term Development Trend of Web Sentiment

GAO Hui[1], WANG Sha-sha[2], and FU Yan[1,2]

(1. School of Computer Science and Engineering, University of Electronic Science and Technology of China Chengdu 610054;
2. School of Software, University of Electronic Science and Technology of China Chengdu 610054)

Abstract In this paper we present a novel approach for long-term prediction of the development trend of Web sentiment. For each class of social events, the class model library of the development trend of Web sentiment is established by cycle analysis and hierarchical clustering. Then the adaptive transform is applied to the already known development trend of a new social event, and the min-sum of MSE from the library is selected to predict the future development trend of web sentiment. Experiments show that, compared with the traditional methods, the approach presented in this paper yields a higher correlation in predicting the long-term development trend of web sentiment, and can predict the turning points of the development trend more effectively.

Key words classification; clustering; cycle analysis; long-term prediction; relevancy analysis

舆情是在一定时期、一定范围内民众对社会现实的主观反映，是群体性的思想、心理、情绪、意见和要求的综合表现[1]。随着互联网的快速发展，网络媒体作为一种新的信息传播形式已经深入人们的日常生活，公众在网络上的言论活跃程度也达到了前所未有的地步。不论是国内还是国际重大事件，都能迅速在网络上传播开来，并引起公众的极大关注和热烈讨论，进而产生巨大的舆论压力，达到任何部门和机构都无法忽视的地步。

网络的特性决定了Web舆情表达快捷、信息多元、方式互动的特点，也从根本上改变了传播者与受传者之间的关系，具备传统媒体无法比拟的优势。一种新的舆情类型——Web舆情逐渐形成，但互联网的虚拟性、隐蔽性、发散性、渗透性、随意性、即时性等特点决定了Web舆情的直接性、突发性和偏差性。文献[2-4]分别根据网络舆情的概念、特点、表达及传播方式，对舆情的变动规律和我国网络舆情的研究与发展现状进行了分析。

Web舆情的产生，不仅打破了传统媒介对社会舆论的相对垄断，改变了传统的舆论形态，而且还迅速显现出其强势。可以说，互联网已成为思想文化信息的集散地和社会舆论的放大器，如果引导不善，负面的Web舆情将会对社会公共安全形成较大威胁。对相关政府部门来说，加强对网络舆情的及时监测和有效引导，提前预测网络舆情的发展趋势，以积极化解网络舆论危机，对维护社会稳定和促进国家发展具有重要的现实意义；加强对网络舆情的监测和引导也是创建和谐社会的应有内涵。

近几年，预测方法被广泛应用于各个领域并且起到了很好的作用。较早期的预测方法主要有自回

收稿日期：2010－01－21； 修回日期：2010－11－18

基金项目：国家高技术研究发展计划(2007AA01Z440)；国家自然科学基金(60973069, 90924011)；四川省应用技术研究与开发项目支撑计划(2008GZ0009)；中国博士后科学基金(20080431273)

作者简介：高 辉(1969－)男，博士，副教授，主要从事方向复杂网络的并行数据挖掘、高效并行算法的设计与程序验证等方面的研究.

归模型(AR)、滑动平均模型(MA)、自回归滑动平均模型(ARMA)、历史平均模型(HA)和Box-Cox法等。随着研究的逐渐深入，又出现了一批更复杂、更精确的预测方法，总体可以分为两类：1) 以现代科学技术和方法为主要研究手段而形成的预测模型，包括非参数回归模型、KARIMA算法、基于小波理论的方法、基于多维分形的方法、谱分析方法、状态空间重构模型和多种与神经网络相结合的预测模型等，这类模型的共同特点是采用模型和方法，不追求严格意义上的数学推导和明确的物理意义，更重视对真实数据的拟合效果；2) 以数理统计和微积分等传统的数学和物理方法为基础的预测模型，包括时间序列模型、卡尔曼滤波模型、参数回归模型、指数平滑模型等。

随着互联网的快速发展，公众在网络上发表言论的活跃程度达到了前所未有的地步，对容易滋生社会舆情的Web舆情事件的发展态势做出及时准确的预测显得越来越重要。准确的长期趋势预测可为相关部门制定相应的应对措施，并为各大主流网站做出正确的舆论引导赢得宝贵时间。但是，目前我国对于网络舆情的预测还处于探索阶段[5-6]，主要是将现有成熟的时间序列预测和人工智能技术应用于Web舆情的趋势分析[7]。时间序列短期趋势预测方法在网络舆情中的应用效果不错，但是该方法很难做出长期趋势预测，尤其是对拐点的预测，并且预测时需要假设所选预测模型满足某一函数分布，比如多项式回归中多项式最高次数的选择等。

针对传统预测方法无法有效预测Web舆情长期趋势拐点的不足，本文提出一种长期趋势预测方法。该方法首先通过周期分析和层次聚类为每类已发生舆情事件的发展趋势建立类模型库，然后通过对预测舆情事件已知发展趋势进行自适应变换后，应用最小二乘法从相应事件类别的类模型库中选取均方误差和最小的模型预测事件未来的发展趋势。

1　预测模型

模型预测允许预测人员对预测条件做一定程度的假设，本文提出的事件长期趋势预测模型是基于历史会重演的假设。研究发现，不仅同一类事件的发展趋势有较高的相似性，而且同一事件的发展会经历不同的周期。为了进一步提高模型的拟合精度和预测效果，本文首先对事件进行分类和切取周期处理，然后为每类事件按周期建立类模型并形成类模型库，再从中挑取与待预测事件均方误差和最小的类模型进行长期预测。

模型的建立需要数据的支撑，为了获取历史事件的时间序列，首先使用网络爬虫从网络上获取数据，并将数据存储到数据库中；通过使用基于向量空间的LP聚类算法[8]对数据库中描述同一个事件的数据进行自动标记，形成事件集；根据舆情的特点，通过分类方法[9-10]将事件分为刑事案件、恐怖袭击、经济安全、自然灾难、公共卫生事件和社会安全事件等事件类别；根据预测的需要，可获取数据库中某事件类别包含的所有事件对应的时间序列。时间序列的过去值会影响将来值，影响的大小及影响的方式可由时间序列中的趋势、周期及非平稳等特征来刻画，因此可采用一个事件的时间序列进行预测。

时间序列的获取可以根据实验条件选取。本文所处理的时间序列值来源于Google trends所统计的数据。所谓Google trends数据并不是原始的搜索量，而是在过去的一段时间里，相对于在Google上执行的总搜索量即某个字词被搜索了多少次，经过标准化并以0～100的缩放结果值表示。Google trends所有的数据都从2004年1月4日开始，建模的具体流程如图1所示。

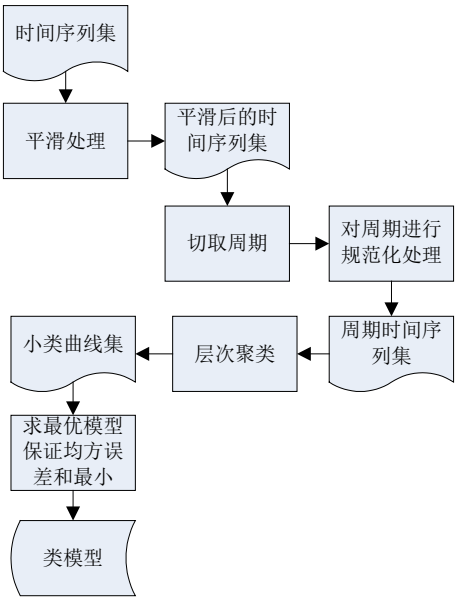

图1　建模流程图

1.1　时间序列的预处理

经过事件分类处理后，可得到事件类别集合 $C=\{C_1,C_2,\cdots,C_n\}$。在建立事件类别C_k的类模型库时，需要对事件类别C_k中包含的每一个事件i所对应的时间序列 $X_i^{(k)}=\{X_{i1}^{(k)},X_{i2}^{(k)},\cdots,X_{it}^{(k)}\}$ 进行曲线拟合。曲线拟合问题是在诸多试验和工程实际中广泛

应用的数据处理方法，拟合的目的主要是根据已有的时间序列寻找曲线的特征，从而求解曲线的相关参数。该问题的解决通常要根据线性的特点选取一定的数学模型，以待求的线性参数作为未知参数，采用最小二乘法进行处理。普通的最小二乘法以观测值残差平方和极小为准则，忽略了自变量的误差，拟合结果使拟合曲线沿一个方向与实际曲线最佳逼近。而采用的正交最小二乘法[11]以正交距离的残差平方和极小为准则，顾及了因变量和自变量的误差，拟合的结果从整体上保持最佳。

然而时间序列的生成可能会受到噪声数据的影响，为减少噪声数据的影响，本文使用一维中值滤波法对获取的原始时间序列 $X_i^{(k)}=\{X_{i_1}^{(k)},X_{i_2}^{(k)},\cdots,X_{i_t}^{(k)}\}$ 所对应的曲线进行平滑处理，时间序列：

$$x_{ij}^{(k)}=\text{median}(X_{i(j-\lfloor r/2\rfloor)}^{(k)}:X_{i(j+\lfloor r/2\rfloor)}^{(k)}) \tag{1}$$

式中，k为噪声数据所属的事件类别，i为该数据所属的事件；j为该事件的第j个数据；r为平滑窗口的大小；median表示取时间序列 $X_{i(j-\lfloor r/2\rfloor)}^{(k)}\sim X_{i(j+\lfloor r/2\rfloor)}^{(k)}$ 的中值；$\lfloor r/2\rfloor$表示取$r/2$的下整数。

现有的预测方法在进行短期预测时都取得了一定的成效，但均无法预测长期趋势中的拐点，影响长期预测的效果。研究发现，一个事件的发展可能会经历几个周期的循环，因此，为了更好地预测长期趋势中的拐点，以进行较准确的长期趋势预测，在对曲线进行平滑处理后，应对每条曲线进行切取周期处理，以获取事件发展的不同周期。本文提出的切取周期的方法为：

1) 遍历原始曲线，保留明显的关键转折点，用直线把这些关键转折点连接起来形成折线图。选择关键转折点的具体做法是：将曲线开始和结尾的点选为关键转折点，然后从第一个关键转折点开始，尝试用直线连接它和它后面的每一个点，直到中间有点与该直线的距离超过给定范围值d，该超出给定范围值的点就被认为是一个新的关键转折点。再从该新的关键转折点开始，重复上面的过程，直到曲线的最后一个点。

2) 采用遍历折线图上各关键转折点的方法寻找每个周期T的开始和结束的位置，以避免无关起伏的干扰。周期开始的判断标准为：从第一个关键转折点开始，当折线图中相邻两个关键转折点构成的线段的斜率超过人为给定的阈值(如本文实验中取为3，可以根据具体的实验数据进行调整)时，就判定周期开始。周期结束的判断标准为：周期开始后，满足下列两个条件之一，就判断周期结束。① 趋势的起伏在一个给定的范围值d内，即在给定范围值d内选择关键转折点，并且该关键转折点距周期开始的时间跨度至少为min$T$，曲线的当前高度不超过周期开始时的2倍；② 周期的长度已经超过给定的最大时间跨度max$T$。

预处理的最后一步工作就是对切取的周期曲线的时间长度进行规范化处理。根据建立类模型库需确保度量一致性的原则，将所有周期曲线的时间长度统一规范化为max$T$。因此需要对周期曲线进行插值处理，具体的插值方法为：假设某周期曲线对应的时间序列为 $x=\{x_1,x_2,\cdots\}$，长度为 $\text{len}(x)$，时间序列 x 经过插值后，时间长度规范化为max$T$的时间序列 $y=\{y_1,\cdots,y_i,\cdots,y_{\max T}\}$，则有：

$$y_i=x_l+(x_u-x_l)(q_i-l) \tag{2}$$

式中，$1\leqslant i\leqslant \max T$；$q_i=\dfrac{i\times\text{len}(x)}{\max T}$；$l=\lfloor q_i\rfloor$；$u=\lceil q_i\rceil$。

1.2 类模型库的建立

对某事件类别包含的所有事件进行预处理后，可获得规范化的周期曲线，再使用层次聚类算法将规范化的周期曲线进行聚类。确定数据集的聚类数目是聚类分析中一项基础性的难题，文献[12]提出了一种基于层次聚类思想的计算方法，不需要对数据集进行反复的聚类，其主要步骤为：

1) 首先扫描数据集获得聚类特征统计值。

2) 然后自底向上地生成不同层次的数据集划分，增量地构建一条关于不同层次划分的聚类质量曲线，该曲线极值点所对应的划分用于估计最佳的聚类数目。

将周期曲线聚类后得到的各个聚类簇视为小类，对于每一小类的类模型，应用最小二乘法求出与该小类包含的所有周期曲线均方误差和最小的类模型。具体方法为：设某小类包含的周期曲线集为 $\{y_1,y_2,\cdots,y_n\}$，每个周期曲线 y_i 对应的时间序列为 $\{y_{i1},y_{i2},\cdots,y_{im}\}$，其中 $1\leqslant i\leqslant n$。定义所求的该小类的类模型为：

$$\hat{y}_{ij}=a_0+a_1y_{ij}^1+a_2y_{ij}^2+\cdots+a_ky_{ij}^k \tag{3}$$

式中，$\hat{y}_{ij}$ 为预测值；y_{ij} 为实际观察值；a_0、a_1、…、a_k为类模型中待求的系数；k的取值范围可根据具体情况从[3,20]中选取。该小类n个周期曲线对应的时间序列的实际观察值与类模型中给出的预测值的均方误差总和为：

$$e=\sum_{i=1}^{n}\sum_{j=1}^{m}(\hat{y}_{ij}-y_{ij})^2 \tag{4}$$

将式(3)代入式(4)后，式(4)可视为关于a_0、a_1、…、a_k的多元函数，根据多元函数求极值的方法，分别对a_0、a_1、…、a_k求一阶偏导，并令其等于零得到非齐次线性方程组：

$$\begin{cases}\sum_{i=1}^{n}\sum_{j=1}^{m}(a_0+a_1y_{ij}^1+a_2y_{ij}^2+\cdots+a_ky_{ij}^k-y_{ij})y_{ij}^0=0\\ \sum_{i=1}^{n}\sum_{j=1}^{m}(a_0+a_1y_{ij}^1+a_2y_{ij}^2+\cdots+a_ky_{ij}^k-y_{ij})y_{ij}^1=0\\ \vdots\ \ \vdots\\ \sum_{i=1}^{n}\sum_{j=1}^{m}(a_0+a_1y_{ij}^1+a_2y_{ij}^2+\cdots+a_ky_{ij}^k-y_{ij})y_{ij}^k=0\end{cases} \tag{5}$$

解该非齐次线性方程组可以求出所有驻点$(a_0,a_1,\cdots,a_k)$，并与边界值上的最大值和最小值进行比较，最小值所对应的驻点即为所求类模型式(4)中的各个系数，从而可建立该小类的类模型。采用同样的方法，可建立其他小类的类模型，从而建立该事件类别的类模型库。

1.3　长期趋势预测

当新的舆情事件发生时，首先确定该事件所属的事件类别，并获取该事件已发生的时间序列；将该时间序列进行自适应缩放变换后，逐一与其所属事件类别对应的类模型库中的类模型进行匹配，选取类模型库中与待预测事件已知时间序列均方误差和最小的类模型作为待预测事件的长期预测模型，从而实现对新舆情事件的长期预测，具体流程如图2所示。

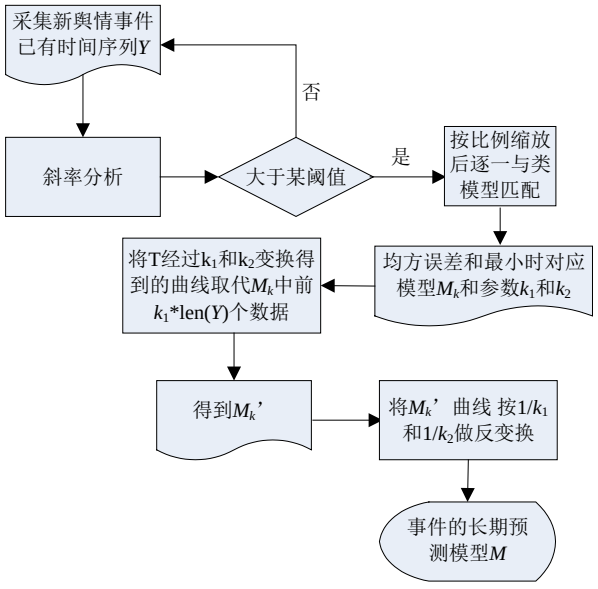

图2　趋势预测流程

为了提高周期性长期预测的准确率和有效性，当识别到新的舆情事件发生时，按一定的时间间隔采集其现有的时间序列Y，并对时间序列Y对应的曲线的斜率进行分析，如果斜率大于或等于某阈值，说明该事件已经开始被广泛关注，开始将该事件已有的时间序列Y与其所属事件类别的类模型库里的类模型进行匹配，设时间序列Y的长度为$\mathrm{len}(Y)$。具体方法为：

1) 对时间序列Y对应曲线的横坐标和纵坐标分别按照比例k_1和k_2进行缩放变换。为了寻找合适的缩放比例k_1和k_2，采用双重循环进行遍历查找，设$1\leqslant k_1\leqslant 100$，$1\leqslant k_2\leqslant 100$，循环遍历的步长为0.1。每一次循环，对时间序列$Y$对应曲线中的横坐标和纵坐标进行缩放调整，经过缩放变换以后曲线的横坐标x_i和纵坐标y_i分别为：

$$x_i=\frac{i}{k_1} \tag{6}$$

$$y_i=(Y_l+(Y_u-Y_l)(x_i-l))k_2 \tag{7}$$

式中，$l=\lfloor x_i\rfloor$；$u=\lceil x_i\rceil$；$1\leqslant i\leqslant\lceil \mathrm{len}(Y)\times k_1\rceil$，$\mathrm{len}(Y)$为时间序列$Y$的长度。将缩放变换后产生的时间序列$Y'=\{y_1,y_2,\cdots y_i,\cdots\}$逐一与类模型库$M$中每个类模型的时间序列的前$\lceil \mathrm{len}(Y)\times k_1\rceil$个数据进行比较，求出均方误差和，并记录循环中均方误差和取最小值时对应的类模型M_k和缩放比例k_1和k_2。

2) 对于所记录的均方误差和最小时对应的类模型M_k和缩放比例k_1和k_2，将时间序列T经过k_1和k_2缩放变换后得到的时间序列Y'替代类模型M_k对应的时间序列中前$\lceil \mathrm{len}(T)*k_1\rceil$个数据得到时间序列$M_k'$。

3) 将M_k'的横坐标和纵坐标分别按$1/k_1$和$1/k_2$进行缩放变换(反变换)得到长期预测曲线M，当采集到该舆情事件的新数据时，重复上面的步骤即可得到新的长期预测曲线。

2　实验部分

本部分预测实验针对属于公共卫生类的猪流感事件，选取的测试数据为从Google趋势上获取的“猪流感”在2009年3月～2009年7月期间的Google trends时间序列。

2.1　实验效果

对公共卫生类事件的预测需要以该类事件的类模型库为基础，为了构建类模型库，需要对相似性较高的该类事件的曲线进行聚类并为各小类建立类模型。对已有公共卫生类其他事件聚类和建模的效果图如图3所示，其中用虚线分开的4个区域分别表示该类事件按层次聚类方法所聚的4个小类，每个区

域中位于上方的图为该小类曲线聚类的结果，位于下方的图为该小类按前面介绍的多元函数求极值的方法所建的类模型。

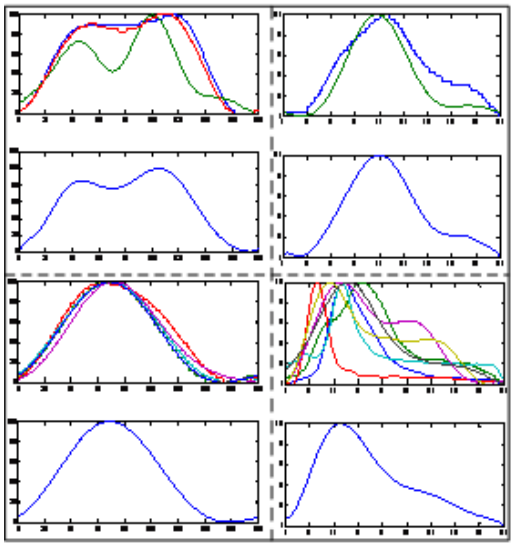

图3 聚类和建模效果图

为验证本文所提长期趋势预测方法的有效性，取从Google趋势上获取的猪流感时间序列前10天的数据作为训练数据，10天以后的数据作为测试数据，进行长期预测，具体过程为：

1) 对采集到的猪流感前10天的数据进行自适应缩放变化。鉴于Google trend数据最大标准化值为100，设定横坐标缩放比例k_1和纵坐标缩放比例k_2的取值区间均为[1,100]，步长为0.1。当k_1=1.5，k_2=1时，从公共卫生事件类模型库中选取的第4小类模型(即图4中的实曲线)与进行缩放后的猪流感数据的均方误差和最小。

2) 选定类模型后，对该类模型的横坐标和纵坐标分别按$1/k_1$和$1/k_2$进行缩放变换，并将缩放后的前10天的数据替换为猪流感事件给定的前10天的数据，得到猪流感事件的长期趋势预测曲线，即图4中的虚曲线，图4中的实曲线表示从Google趋势上获取的猪流感的实际数据。

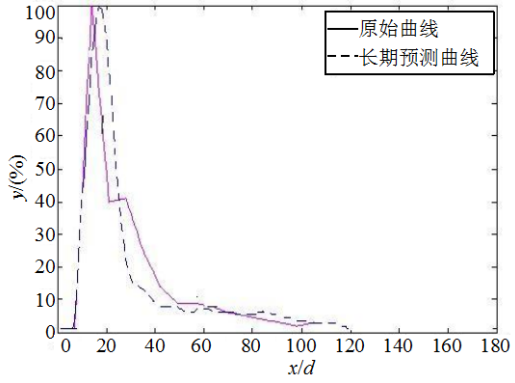

图4 预测效果图

从图4的预测效果来看，本文方法能在事件发生初期较好地预测事件长期发展趋势的拐点。

2.2 对比分析

建立预测模型后，必须检验模型预测的有效性。现有检验方法中的关联度检验法被广泛用于衡量模型预测的精度。因此，本文采用关联度分析检验预测模型的精度，并将本文提出的预测方法与几种传统的预测方法进行对比。关联度检验法主要是比较实际时间序列和各预测时间序列中实际值与各预测值的相对大小，找出差别的最大值和最小值，进而求得实际数据与各预测数据之间的关联度。

设实际时间序列为 $X^{(0)}=\{X_1^{(0)},X_2^{(0)},\cdots,X_n^{(0)}\}$，采用预测方法$i$ $(i\geqslant 1)$进行预测得到的时间序列为 $X^{(i)}=\{X_1^{(i)},X_2^{(i)},\cdots,X_n^{(i)}\}$，则预测方法$i$的关联度定义为：

$$\eta^{(i)}=\frac{1}{n}\sum_{j=1}^{n}\frac{\min\limits_{i\geqslant 1}\min\limits_{1\leqslant k\leqslant n}\left|X_k^{(i)}-X_k^{(0)}\right|+\rho\max\limits_{i\geqslant 1}\max\limits_{1\leqslant k\leqslant n}\left|X_k^{(i)}-X_k^{(0)}\right|}{\left|X_j^{(i)}-X_j^{(0)}\right|+\rho\max\limits_{i\geqslant 1}\max\limits_{1\leqslant k\leqslant n}\left|X_k^{(i)}-X_k^{(0)}\right|} \tag{8}$$

式中，$1\leqslant j\leqslant n$；$\left|X_k^{(i)}-X_k^{(0)}\right|$为实际时间序列 $X^{(0)}$ 和预测方法i对应的预测时间序列 $X^{(i)}$ 中第k个数据 $X_k^{(i)}$ 和 $X_k^{(0)}$ 的绝对误差；$\min\limits_{i\geqslant 1}\min\limits_{1\leqslant k\leqslant n}\left|X_k^{(i)}-X_k^{(0)}\right|$为两级最小差；$\max\limits_{i\geqslant 1}\max\limits_{1\leqslant k\leqslant n}\left|X_k^{(i)}-X_k^{(0)}\right|$为两级最大差；$\rho$为分辨率，$0<\rho<1$，一般取$\rho$=0.5。

通常关联度的经验阈值取r_0=0.6，若预测方法i的关联度 $\eta^{(i)}\geqslant r_0$，则表示其预测结果较好。对单位不统一，初值不同的序列，在计算相关系数前应该先进行初始化，即将该序列所有数据分别除以第一个数据。

表1 关联度对比

事件名	类模型	多项式回归归	自回归	灰色理论
猪流感	0.796 2	0.534 8	0.524 1	0.469 7
孙伟铭	0.802 1	0.610 2	0.543 8	0.667 1
开胸验肺	0.813 2	0.587 4	0.627 3	0.486 9
针刺事件	0.845 7	0.348 4	0.298 3	0.424 7
新疆暴乱	0.715 9	0.623 1	0.546 7	0.358 7
胡斌飙车	0.694 6	0.523 9	0.437 1	0.278 0
国庆	0.729 1	0.428 9	0.498 7	0.354 1
上海杀警察	0.671 9	0.459 7	0.593 7	0.427 4
邓玉娇案件	0.824 2	0.439 1	0.357 1	0.347 5
通钢事件	0.631 7	0.637 6	0.283 0	0.278 4
平均关联度	0.752 5	0.519 3	0.471 0	0.410 3

选取10个不同事件的前10天的数据，分别采用式(8)对本文所提出的预测方法与传统预测方法(包括多项式回归模型、自回归模型和灰色理论模型)进行关联度对比分析，关联度对比结果如表1所示。

从表1所示的实验结果来看，本文所提出的预测方法的关联度均大于经验阈值0.6，说明该方法预测的结果是有效的。并且，本文所提出的预测方法的平均关联度比其他3种传统预测方法的最大平均关联度高出40%，因此使用本文提出的预测方法比较适用于事件的长期趋势预测。

3 总 结

互联网的迅猛发展使得社会舆情有了新的载体，网络舆情对社会政治、经济和文化的平稳发展产生了较大的影响，因此有必要利用现有的人工智能和数据挖掘技术实现对舆情的分析和预测，为进一步治理和正面引导网络舆情的发展奠定基础。

本文提出了一种Web舆情长期趋势预测方法。该方法首先通过周期分析和层次聚类为每类已发生舆情事件的发展趋势建立类模型库。当待预测舆情事件发生时，首先确定其所属事件类别并获取已有的时间序列，将其进行自适应缩放变换后，应用最小二乘法从其所属事件类别的类模型库中选取均方误差和最小的模型预测该事件未来的发展趋势。实验结果显示，本文提出的预测方法比传统方法更适合对舆情事件的长期趋势预测，可以弥补现有预测技术无法预测情事件发展趋势拐点的缺陷，更好地帮助政府和监管部门采取及时有效的措施，提高网络舆情监管的功效。

参 考 文 献

[1] 王来华. 舆情研究概论——理论、方法和现实热点[M]. 天津: 天津社会科学院出版社, 2007.
WANG LaiHua. Public opinion study—theory, method and reality hotspot[M]. Tianjin: Tianjin Social Sciences Press, 2007.

[2] 刘毅. 略谈网络舆情的概念、特点、表达与传播[J]. 理论界, 2007, (1): 11-12.
LIU Yi. Talk about web sentiment’s concept, characteristic, expression and communication[J]. Theory Horizon, 2007, (1): 11-12.

[3] 彭丹, 许波, 宋仙磊. 基于网络评论的网络舆情研究[J]. 现代情报, 2009, 29(12): 4-7.
PENG Dan, XU Bo and SONG XianLei. Web sentiment study based on web comment [J]. Modern Information, 2009, 29 (12): 4-7.

[4] 曾润喜. 我国网络舆情研究与发展现状分析[J]. 图书馆学研究, 2009, 19(1): 24-37.
ZENG Run-xi. Study our country’s web sentiment and analyze its state-of-the-art[J]. Study of Library Science, 2009, 19(1): 24-37.

[5] 腾达. 基于趋势分析的网络舆情监控系统(TANCMS)的研究与实现[D]. 长沙: 国防科学技术大学, 2008.
TENG Da. Research and realization of TANCMS based on trend analysis[D]. Changsha: National University of Defense Technology, 2008.

[6] 程辉. 基于时间序列的网络舆情预测模型[J]. 网际网路技术学刊, 2008, 9(5): 16-17.
CHENG Hui. The prediction model of public opinion based on trend analysis[J]. Journal of Internet Technology, 2008, 9(5): 16-17.

[7] 张珏. 网络舆情预测模型与平台的研究[D]. 北京: 北京交通大学, 2009.
ZHANG Jue. Research on forecasting models and platform of online public opinion[D]. Beijing: Beijing Jiaotong University, 2009.

[8] RAMAGE D, HEYMANN P, CHRISTOPHER D. Manning. Clustering the tagged web[C]//WSDM ’09 Proceedings of the Second ACM International Conference on Web Search and Data Mining. New York, USA: ACM, 2009: 54-63.

[9] COUTO T, ZIVIANI N, CALADO P, et al. Classifying documents with link-based bibliometric measures[J]. Information Retrieval, 2009, 13(4): 355-363.

[10] WANG Pu, HU Jian, ZENG Hua-jun, et al. Using wikipedia knowledge to improve text classification[J]. Knowledge and Information Systems, 2008, 19(3): 265-281.

[11] TERRADO M, BARCELO D, TAULER R. Quality assessment of the multivariate curve resolution alternating least squares method for the investigation of environmental pollution patterns in surface water[J]. Environ Sci Techno, 2009, 43(14): 5321-5326.

[12] 陈黎飞, 姜青山, 王声瑞. 基于层次划分的最佳聚类数确定方法[J]. 软件学报, 2008, 19(1): 24-37.
CHEN Li-fei, JIANG Qing-shan, WANG Sheng-rui, A hierachical method for determining the number of cluster[J]. Journal of Software, 2008, 19(1): 24-37.

编 辑 蒋 晓