第44卷第6期 南京航空航天大学学报 Vol. 44 No. 6
2012年12月 Journal of Nanjing University of Aeronautics & Astronautics Dec. 2012

多时间序列关联规则分析的论坛舆情趋势预测

钱爱玲[1] 瞿彬彬[2] 卢炎生[2] 陈攀攀[1] 陈国栋[1]
(1.台州学院数学与信息工程学院,台州,317000; 2.华中科技大学计算机科学与技术学院,武汉,430074)

摘要:为了预测论坛舆情及其动态演变趋势,基于多时间序列的关联分析,集中分析了论坛中3个量的时间序列之间的关联规则:活跃者之间的关系强度的时间序列、坚定支持者人数的时间序列以及坚定支持者成员的变化频度的时间序列。然后给出了一种新的基于多时间序列关联分析的论坛舆情预测算法(Forum sentiment trend prediction based on multi time series association rule analysis,TPMTSA),并在真实数据集和拟合数据集上进行了大量的实验。结果表明:TPMTSA算法具有有效性和较高的运行效率。研究结果可用于论坛舆情预警监控。

关键词:论坛舆情;趋势预测;时间序列;关联分析
中图分类号:TP311.13 **文献标识码**:A **文章编号**:1005-2615(2012)06-0904-07

Forum Sentiment Trend Prediction Based on Multi Time Series Association Rule Analysis

Qian Ailing[1], *Qu Binbin*[2], *Lu Yansheng*[2], *Chen Panpan*[1], *Chen Guodong*[1]
(1. School of Mathematics and Information Engineering, Taizhou University, Taizhou, 317000, China;
2. School of Computer Science & Technology, Huazhong University of Science and Technology, Wuhan, 430074, China)

Abstract: In order to predict the evolving trend of forum sentiment, based on the association analysis of multi time series, the association rules of three-quantity time series over forum sentiment are anlyzed, namely, the strength of relationship between actors, the number of pillars, and the changing frequency of pillars. Then a novel prediction algorithm, forum sentiment trend prediction based on multi time series association rule analysis (TPMTSA), is proposed. Extensive experiments over real and synthetic datasets are conducted. Results show the effectiveness and the efficiency of TPMTSA. The research results can be used to monitor the forum opinion.

Key words: forum sentiment; trend predicton; time series; association analysis

近年来,各国政府越来越重视论坛舆情。论坛上的焦点事件和敏感话题,往往迅速吸引着成百上千万的网民,形成庞大的论坛网络,其节点之间的联系状态复杂,持续动态演变,使得以前在规模较小的网络上有效的方法和算法,到了论坛网络中,往往因为计算复杂度太大而不能有效,甚至成为NP难[1-3]。论坛舆情的分析和预测,面临着一系列新的挑战,吸引了从社会科学到自然科学各个学科领域的研究[1-5]。

从数据挖掘的角度看,挖掘分析论坛网络中的一些关键要素的时间序列的演化规律和关联分析,可以预测论坛网络的演变趋势[1,3]。从复杂性科学看,论坛网络是一种复杂的社会网络,通过对社区结构与网络中活动者之间的关系强度的相关分析,可以跟踪预测网络的动力行为特征和演化趋势[1-5]。社区检测是各个领域关于社会网络的研究热点,其中基于谱的方法吸引了众多研究者[2,6],并提出了一些著名算法,如Betweenness算法[7-8],External Optimization方法[9-10], Greedy算法[11-12], Lanczos方法[13], Power方法[14]等。在这

基金项目:国家自然科学基金(61100060)资助项目;教育部人文社会科学研究青年基金(11YJCZH008)资助项目。
收稿日期:2012-03-09;**修订日期**:2012-09-19
通讯作者:钱爱玲,女,博士,副教授,1967年10月生,E-mail:alinghust@126.com。

方面的研究中，Newman 于 2006 年提出的矩阵模块化方法及其Q 函数[2]，是社会网络社区结构分析和划分方法中最为著名的方法。尤其是Newman 的Q 函数方法不需要事先知道网络中的社区个数，而其他方法都需要，在实际的动态社会网络中，这是不容易事先知道的。由于计算的复杂性，早期的社区结构研究都是对静态数据集的，而 Newman 的Q 函数方法对动态网络一样有效。随后，众多的研究者跟踪研究了使用该Q 函数分析社区的关系强度，表明了在实际社会网络中Q 函数用于分析和划分社区结构的有效性和高效率[14-16]。2009 年，文献[14]中使用该Q 函数的模块化方法，对美国众院和上院 200 多年中的 109 届选举进行研究发现，一个政党在其模块化关系强度居中时特别不稳定，而在模块化关系强度小或者强的时候，很稳定，很少能改变两院大多数人的固有意见。2007 年，Palla 等人[15]在 Nature 杂志上著文他们的研究发现：在社会网络中，对于小社区，如果其成员不断动态变换，则其稳定性较小；如果其成员不经常变换，则较稳定，这与人们通常的想法相一致。另一方面，他们的研究还发现，对于较大的社区，如果其成员经常动态变换，其稳定性和持续性较强；如果其成员很少变换，则往往易于分裂，这与人们通常的想法相反，也更吸引研究者们。总体上，目前已有的对于论坛网络的研究，在研究内容和研究方法上，大多是对舆情演进中各个因素的单方面的研究。尽管时间序列关联分析是数据挖掘的一个经典研究主题，但是多时间序列趋势之间的关联分析仍然还有很大的研究空间。

本文的研究目标集中在预测论坛舆情的演变趋势：趋向危机，还是趋向平缓。主要受 Newman[2,9-11]、Waugh 等人[14]和 Palla 等人[15]的研究结果的启发，基于数据挖掘中多时间序列之间的关联分析，在文献[17]提出的论坛舆情趋势预测方法的研究基础上，进一步给出了基于多时间序列关联分析的论坛舆情趋势预测算法(Forum sentiment trend prediction based on multi time series association rule analysis，TPMTSA)。其中，首次对多时间序列趋势之间的关联规则给出定义。实验验证算法的有效性和效率。

1　符号和定义

1.1　有关符号及其时间序列

将论坛社会网络记为图 $G=<V, E>$，其中，G 为无向图，V 为顶点集合，E 为图中连接边的集合。用 v_i 和 v_j 表示图中的任意两个顶点，也即论坛上的两个参与者，如一个顶点 v_i 发贴，另一个顶点 v_j 跟帖，则该两个顶点建立了连接边 e_i，用 $e_i=(v_i, v_j)$ 表示。

表 1 列出了常用的符号、意义描述及其所形成的时间序列，更为具体的描述参见文献[17]。

表 1　符号、意义描述和对应的时间序列

在时间点 t 时的符号	意义	时间序列
T	时间序列的时间段长度，比如一个小时、一天、两天或一周。	
$\boldsymbol{G}_t$	在时间点 t 的图。	$G^S=\{G_1, G_2, \cdots, G_t, \cdots, G_n\}$
s_t	论坛上参与者人数。	$s^S=\{s_1, s_2, \cdots, s_t, \cdots, s_n\}$
n_t	过滤掉 s_t 中的只观看而不发评论意见的潜伏中立者后，从而得到的净参与者人数。	$n^S=\{n_1, n_2, \cdots, n_t, \cdots, n_n\}$
Q_t	在时间点 t 的社区关系强度值 Q_t。	$Q^S=\{Q_1, Q_2, \cdots, Q_t, \cdots, Q_n\}$
Q_{t+1}	Q_t 的预测值。	$S^S=\{S_1, S_2, \cdots, S_t, \cdots, S_n\}$
p_t	某个话题的赞同和反对的坚定支持者人数在时间点 t 时的值。	$p^S=\{p_1, p_2, \cdots, p_t, \cdots, p_n\}$
f_t	从时间点 $t-1$ 到 t 坚定支持者的变更频度，简称为在时间点 t 的变更频度。	$f^S=\{f_1, f_2, \cdots, f_t, \cdots, f_n\}$

1.2　多时间序列关联规则定义

定义 1　某两个时间序列 TS^t 和 TS^s 在趋势上存在关联规则，当且仅当 TS^t 和 TS^s 在趋势上的关联满足最小置信度阈值和最小支持度阈值。其中，两个时间序列 TS^t 和 TS^s 在趋势上的关联，是指在某个时间点以后时间序列的数据点取值符合某种定义，该定义可以是如下几种事务中的某一种：增长趋势；下降趋势；在某个时间点以后数据点值均大于某个值；在某个时间点以后数据点值均小于某个值；在某个时间点以后数据点值均等于某个值；在某个时间点以后数据点值均在某个值附近波动；在某个时间点以后数据点值均满足其他某种定义。其置信度和支持度定义为

$$\text{置信度}(TS^t \Rightarrow TS^s)=\frac{\text{包含 } TS^t \text{ 和 } TS^s \text{ 的事务数}}{\text{包含 } TS^t \text{ 的事务数}} \quad (1)$$

$$\text{支持度}(TS^t \Rightarrow TS^s) = \frac{\text{包含 } TS^t \text{ 和 } TS^s \text{ 的事务数}}{\text{总事务数}} \tag{2}$$

2 TPMTSA 算法

基本思想：很多文献研究发现，社会网络不同于一般的网络，社会网络既不是规则网络，也不是随机网络，而是具有很强的社会统计特征的网络[1,2,8]，因而揭示社会网络演变规律的方法往往基于观察和实验。在文献[17]提出的论坛舆情趋势预测方法研究的基础上，本文进一步从3个时间序列 Q^S、p^S 和 f^S 趋势之间关联的角度，给出了基于多时间序列关联分析的论坛舆情预测算法TPMTSA，如算法1所示。然后通过实验验证关联规则的置信度和支持度，从而验证算法的有效性。

算法1 TPMTSA 算法

Input：论坛网络图 G_t 以及其时间序列 G^S，平滑常数 λ，时间段长度 T，净参与者人数阈值 N，特征值的个数 κ

Output：危机等级：High，Middle，Primar 或者 Potential

```
for 每一个时间点 t do
  统计净参与者人数 n_t
  if n_t >= N then
    报警：危机等级为 Potential
    启动子过程 Q-calculate 和 Q-predict
    for i=1 to κ do
      计算出图 G_t 的 u_i 和 β_j
      选择 s 使之尽可能与 u_i 并行
    end for
    由式(3)计算出 Q_t，并放进时间序列 Q^S 中
    由式(4)计算出 Q_{t+1}
    if QL<=Q_{t+1}<=QU then
                    // 如果Q_{t+1}值小于QL，没有危机，不跟踪
      报警：危机等级为 Primary
      提示应用特别关注 Q_{t+1}值的演进情况
      启动子过程 p-count 和 f-count
    else
      if QU <=Q_{t+1}then      // Q_{t+1}值大到超过了阀值QU
        if (p_{t+1}<P and f_{t+1}<F) or (p_{t+1}>=P and f_{t+1} >= F) then
                    // p_{t+1}和 f_{t+1}正关联
          报警：危机等级为High
          else                // p_{t+1}和 f_{t+1}负关联
          报警：危机等级为Middle
        end if
      end if
    end if
  end if
end for
```

TPMTSA 分为4步，对应于4级可能的舆情趋势，其中3个时间序列 Q^S、p^S 和 f^S 的计算方法简述如下，更为具体的计算方法和4个计算子过程参见文献[17]。

(1)从跟踪时间序列 n^s 开始，当其在某个时间点的值 n_t 达到一个给定的阈值 N 时，表明论坛舆论有发展为敏感的潜在可能，需要加以关注，舆情评级为潜在危机(Potential tension)。N 往往是一个很大的量值，可以由用户设定。

(2)同时启动子过程 Q-calculate 和 Q-predict。给定下界阈值 QL、上界阈值 QU，如果 $Q_{t+1}<QL$，社区关系强度低，意味着论坛上参与者的观点多种多样，关系松懈，这时往往没有危机，不会形成舆情。

当 $QL \leqslant Q_{t+1} < QU$ 时，社区关系强度居中，论坛上参与者的观点存在差异，论坛呈弱稳定性，这种情况较容易进行有目标的疏导。舆情评级为初级危机(Primary tense)。

(3)随着论坛的发展演绎变化，社区强度 Q_{t+1} 可能达到 QU。当 $Q_{t+1} \geqslant QU$ 时，表明论坛上的观点高度一致，且稳定，需要更加关注，同时启动两个子过程 p-count 和 f-count，往往存在上界阈值 P 和 F。如果 $p_{t+1}<P$ 且 $f_{t+1}>F$，或者 $p_{t+1} \geqslant P$ 且 $f_{t+1}<F$，表明社区的主要支持者人数少且变更频度大，或者表明主要支持者人数较多且变更频度小，这时如果要改变论坛上的舆论观点的局面，往往难度大。舆情评级为中级危机(Middle tense)。

(4)当 $Q_{t+1} \geqslant QU$ 时，如果 $p_{t+1}<P$ 且 $f_{t+1} \leqslant F$，主要支持者人数少且变更频度小，表明论坛中可能有某个组织，组织的成员再通过论坛有目的地发布流言；如果 $p_{t+1} \geqslant P$ 且 $f_{t+1} \geqslant F$，表明主要支持者人数众多且变更频度大，这表明论坛上众多的人误解意见高度一致，意味着论坛上的舆论观点或者可能是合理的，或者可能众多的人有误解。舆情评级为高度危机(High tense)。

社区关系强度时间序列数据的计算方法

$$Q_t = \sum_i^{\kappa} \frac{1}{4m} a_i^2 \beta_j \tag{3}$$

式中：κ 为由应用给定的常数，比如可以是1,2,3等；β_i 为图 $\boldsymbol{G}_t$ 的邻接矩阵的特征值，且 $\beta_1 \geqslant \beta_2 \geqslant \cdots$

$\geqslant \beta_n$；$a_i = \boldsymbol{u}_i^{\mathrm{T}} s$，$s = \sum a_i \boldsymbol{u}_i$，$\boldsymbol{u}_i$ 为特征向量。将 Q_t 加入到时间序列 Q^s 中。

跟踪 Q^s，使用加权移动平均（Exponentially weighted moving average，EWMA）方法，对于任意时间片 T，将对 Q_t 的平滑值作为其预测值 Q_{t+1}

$$Q_{t+1} = \lambda \times \sigma_{t-1} + (1-\lambda) Q_t \tag{4}$$

式中：λ 为平滑常数；σ_{t-1} 为时间片 $t-1$ 时的估计值。根据式(3)，最大特征值 β_1 及其系数 a_1 对 Q_{t+1} 的影响最大。选择 s_1，使得其尽可能平行于 $\boldsymbol{u}_1$，以使得相应的顶点作为主要支持者尽可能影响 Q_{t+1} 的值，并统计 $a_i=1$ 的元素个数，即主要支持者人数，记为 p_t，则

$$p_t = \sum_i^{\kappa} (a_i \text{ 为系数为 1 的元素个数的计数和}) \tag{5}$$

用 p^s 和 s_1，计算从时间点 $t+1$ 到 $t+2$ 的元素的变动个数，该个数即为主要支持者的变更频度，记为 f_{t+2}，则

$$f_{t+2} = \frac{|s_{1,t+2} - s_{1,t+1}|}{s_{1,t+1}} \tag{6}$$

式中：$s_{1,t+1}$ 为 s_1 在时间片 $t+1$ 的值。

3　复杂度分析

从算法 1 可见，计算的消耗主要在第 6～9 行，其余行均是线性的。第 7 行最耗时，该行的目的是计算出图 G_t 对应的邻接矩阵的前 κ 主特征向量，采用经典的 Power 方法或者 Lanczos 方法计算需要的时间复杂度为 $O(\kappa n^2)$，这决定了 TPMTSA 算法需要的时间复杂度为 $O(\kappa n^2)$，简化为 $O(n^2)$。其中，$n=n_t$，n_t 为矩阵的规模，即矩阵的顶点数，也即论坛上的主要支持者人数。

3 个知名的社区分析算法的时间复杂度：Betweenness 为 $O(n^3)$[7-8]，External Optimization 为 $O(n^2\log^2 n)$[9-10]，Greedy 为 $O(n\log^2 n)$[11-12]。其中，Greedy 的时间复杂度比 TPMTSA 的小，但是 Greedy 采用的是近似方法，所以 TPMTSA 的精确度较好于 Greedy。由此可见，TPMTSA 时间复杂度较好，将在第 4.4 节实验对此进行验证。

在计算空间要求方面，TPMTSA 需要一个大小为 $n_t \times n_t$ 的矩阵空间和一个大小为 n_t 的一维向量，目前一般普通计算机均能够提供这个空间要求，因此在此不再进一步讨论。

4　实验设计与结果分析

实验环境为：PC 主机 1.60 GHz Intel(R) Core (TM) 2 CPU，内存 2 GB，操作系统 Linux Ubuntu9.10，采用 Matlab7.0 计算矩阵 $\boldsymbol{G}_t$ 的特征向量，以此计算社区强度 Q_t。

4.1　数据集

数据集：采用了2010 年底发生在中国的3 个重大事件在论坛上的真实数据集和一个拟合数据集。

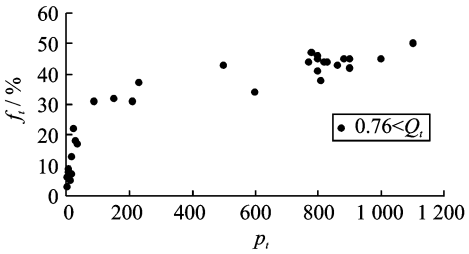
3 个真实数据集的来源：第一个是“没有拆迁就没有新中国”的谬论，第二个是“我爸爸是李刚”的狂言，第三个是“钱云会案”是谋杀还是交通事故的疑问。该 3 个事件，引起了社会各方面的广泛关注，各网络论坛上关注者人数一时间到达数百万。本文分别跟踪了 3 个事件的话题在猫扑和天涯论坛上的演进情况，收集了每个事件的话题从发生当天到随后几周的数据。“我爸爸是李刚”发生在 10 月 16 日，到一周后的 10 月 23 日晚上 12 点（24 日 0 点），猫扑上关于这个话题有 3 349 个线程和27 155 个帖子。钱云会案发生在12 月 25 日，到29 日，温州警方召开新闻发布会仍然确定是交通事故，而不是谋杀，到29 日晚上12 点（30 日0 点），天涯上关于这个话题的帖子数达到了 2 551 377，点击更是超过了100 000 000 次。“没有拆迁就没有新中国”的谬论发生在10 月12 日，第二天天涯上帖子数达到了2 615。随后，人民日报及时于10 月15 日发表了强烈的评论文章，此后每天的跟贴数迅速减少。这 3 个案例事件都有明显相反的两个对立面的评论：赞同的和不赞同的，绝大多数是不赞同的，这也是一般网络论坛上舆论的典型特点。因此，把论坛网络分为两个社区进行分析是合理的。

拟合数据集：由于真实数据集难免不完整和不精确，又参考网上所收集的数据集拟合了 800 个数据，每个拟合数据包括 1～50 K 个节点。

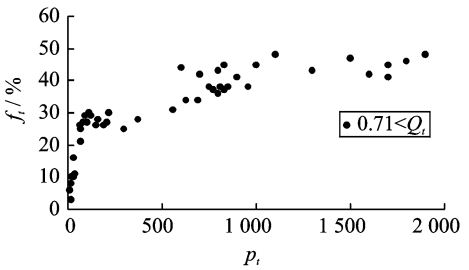
4.2　时间序列 Q^s、p^s 和 f^s 的趋势关联分析

分别对 $Q_t>0.65$ 和 $0.35<Q_t<0.65$ 两种情况，进行对比实验，实验结果如图1(a～d)所示。关于社区强度 Q_t 值，由图 1 可见，3 个真实案例的 Q_t 值都比较高，李刚案为0.76，钱云会案为0.71，强拆案为 0.68，这与实际情况相一致，即在这 3 个案例中，论坛上绝大多数帖子的意见是批评的、不赞同的。对于拟合数据集，将实验分为 $0.65< Q_t$ 和 $0.35< Q_t<0.65$ 两类情况分别进行。

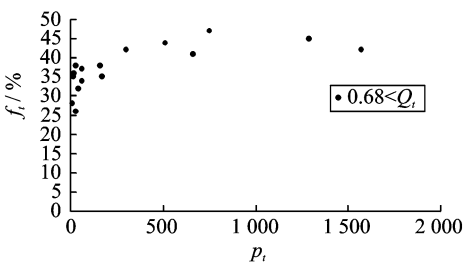
李刚案如图 1(a)所示，实验共取了 30 个时间区段数据作为时间序列的点值。从事件发生至随后 15 d 的每一天为一个时间区段，再之后的 15 周的



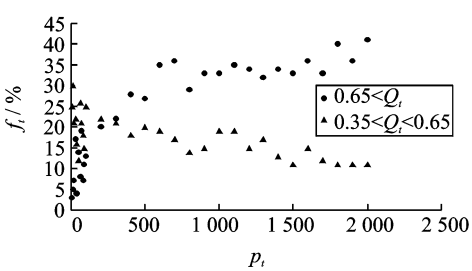
(a) 李刚案



(b) 钱云会案



(c) 强拆案



(d) 拟合数据

图1 Q_t,p_t 和 f_t 3个量的关联关系

每一周为一个时间区段。2011年1月30日,即后面15周中的倒数前一周,法院及时作出了判决,判决之后的一周内,论坛上评论帖子逐渐减少,直至很少再有帖子,舆论平息。

钱云会案如图1(b)所示,实验共取了46个时间区段数据作为时间序列的点值。从事件发生当天至随后46 d的每一天为一个时间区段。2011年2月1日,即实验所计算的46 d中的倒数第八天,警方再次及时举行了听证会,并判决认定为交通事故,而不是谋杀,听证会之后的一周内,论坛上评论帖子逐渐减少,舆情转为缓和。

强拆案如图1(c)所示,实验共取了16个时间区段的数据作为时间序列的点值。从事件发生当天至随后的共14 d的每一天为一个时间区段,之后的2周的每一周为一个时间区段。2010年10月15日,即事件发生后的第3天,人民日报及时发表了强烈的批评文章,论坛上的帖子数很快渐少,舆情缓和。

3个时间序列 Q^S、p^S 和 f^S 的关联方面,3个案例的实验结果相似:图1(a~c)3个图中,右上角的一些点,表明对于较大的 p_t 值有较大的 f_t 值,这是各个事件发生后的前几个时间区段的值的情况;图中左下角的一些点,表明对于较小的 p_t 值有较小的 f_t 值,这是实验中测得的各个事件发生后最后几个时间区段上值的情况;同时,3个事件均有较大的 Q_t 值,均在0.65以上。图1(c)与图1(a,b)有一个不同点,即在事件发生后的第三天,人民日报及时发表了强烈的批评文章,舆情迅速趋向缓和,p_t 迅速较少,而 f_t 没有明显减少,表明尽管还有少数支持者在发表批评评论,但是支持者变更频度比较大,可能没有多少固执者,或者可能没有专门的组织成员在论坛上发贴制造舆论。

这表明,事件一开始发生时,大量的网民被激愤,批评帖子一时间接跌而发,舆论意见是高度一致的,论坛舆情趋向于高级危机。随后,如果有关部门高度重视,并且处理公正,则舆情很快缓和。

拟合数据集上的实验结果如图1(d)所示。可以看出:(1)对于 $Q_t>0.65$ 的情况,实验结果与上述3个真实数据集上的基本一致,具体分析讨论参见文献[17];(2)对于 $0.35<Q_t<0.65$ 的情况,测试结果不明显,大致上可见:存在一个粗略阈值 $P=300$ 和 $F=20\%$,对于($p_t<P$)则存在 ($f_t>F$),对于($p_t>P$)则存在($f_t<F$)。即对于较小的 Q_t 值,p_t 和 f_t 大致上呈现负关联,两种情况下,论坛网络社区的关系强度比较弱,社区不稳定,不会趋向危机。

以上实验结果与文献[14,15]的结论基本一致。

4.3 置信度和支持度

从图1(a~d)还可以进一步看出:3个时间序列 Q^S,p^S 和 f^S 的点值呈现着明显的趋势上的关联规则。下面计算其置信度和支持度,验证TPMTSA算法的有效性。在对拟合数据集的实验中,对4组案例的每一种情况分别设计了200次实验,结果如表2所示。

在对每组案例的实验中，先保持 p_t 值在一定的范围内并调节 f_t 值，使得 p_t 和 f_t 取值在该组实验的取值范围内，获得该组实验的 200 个数据，进行 200 次实验。然后，在实验中统计其中 Q_t 值落在预测区域内的次数。(1)保持200< p_t，调节 f_t 使得 20%< f_t，统计得到 Q_t>0.65 的有 187 次，百分比 187/200=93.5%。又保持 p_t<100，调节 f_t 并使得 f_t<15%，统计得到 Q_t>0.65 的有 182 次，百分比 182/200=91%。(2)保持 p_t<300，调节 f_t 使得 25% < f_t，统计得到 0.35<Q_t<0.65 的有 165 次，百分比 165/200=82.5%。又保持 p_t<300，调节 f_t 使得 f_t<20%，统计得到 0.35 < Q_t<0.65 的有 158 次，百分比 158/200=79%。

这表明，p_t 和 f_t 取值的正关联往往和较大的 Q_t 值同时出现，而 p_t 和 f_t 取值的负关联往往和较小的 Q_t 值同时出现，即 3 个时间序列 Q^S，p^S 和 f^S 在趋势上具有关联规则，其置信度即是上述计算的百分比，每组的实验结果均表明置信度比较高。在每组的200 次实验中，直接选择了 p_t 和 f_t 取值符合该组实验条件的 200 个数据作为感兴趣的数据集，而完全排除了不需要的不感兴趣的数据，即每组实验的 200 个数据所组成的数据集中没有不感兴趣的数据，根据式(1～2)，则每组实验的支持度与置信度相等，因此支持度也较高。较高的支持度和置信度验证了关联规则的可信，验证了 TPMTSA 算法的有效性。

表 2　3 个时间序列 Q^S，p^S 和 f^S 的趋势关联规则的置信度和支持度

案例	测试次数	有效次数	置信度/%	支持度/%
Q_t>0.65，200< p_t，20%< f_t	200	187	93.5	93.5
Q_t>0.65，p_t<100，f_t<15%	200	182	91.0	91.0
0.35< Q_t<0.65，p_t<300，25%< f_t	200	165	82.5	82.5
0.35< Q_t<0.65，300< p_t，f_t<20%	200	158	79.0	79.0

4.4　TPMTSA 的运算效率实验

测试TPMTSA 和3 个知名算法Betweenness、External Optimization 和Greedy 在 4 个数据集上运行的CPU 时间的对比，如表3 所示。由表3 可见：TPMTSA 快于Betweenness 和External Optimization，比Greedy 稍慢，然而由于Greedy 是基于近似方法的，所以，总体上，TPMTSA 的运行效率比较高。

表 3　TPMTSA 和 Betweenness、External Optimization、Greedy 运行时间的对比

s

数据集	Betweenness	External Optimization	Greedy	TPMTSA
李刚案	1 219.527	3 225.519	391.291	2 517.615
钱云会案	101.975	150.873	40.192	139.812
拆迁案	0.351	0.950	0.150	0.750
拟合	2.572	3.161	0.823	2.951

5　结束语

本文结合了复杂系统科学中社区结构分析的方法和成果，从数据挖掘的角度，采用多时间序列关联规则分析方法，提出了一个新的论坛舆情趋势预测算法 TPMTSA，实验验证了其有效性和效率。该研究结果可提供给有关部门用于舆情监控。在本文的研究基础上，可以进一步研究挖掘发现论坛网络中更多有关量之间的关联规则，从而预测网络各种可能的演进趋势。

参考文献：

[1] Shi Xiaolin, Zhu Jun, Cai Rui, et al. User grouping behavior in online forums[C]//International Conference on Knowledge Discovery and Data Mining. Paris: ACM, 2009: 777-785.

[2] Newman M E J. Finding community structure in networks using the eigenvectors of matrices[J]. Physical Reviem E, 2006,74(3):036104.

[3] Xu Kaiquan, Li Jiexun, Stephen S Y. Sentiment community detection In social networks[C]//Proceedings of the 2011 iConference. Seattle: ACM, 2011:804-805.

[4] Backstrom L, Huttenlocher D, Kleinberg J, et al. Group formation in large social networks: membership, growth, and evolution[C]//Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Philadelphia: ACM, 2006: 44-54.

[5] Aldashev G, Carlletti T. Benefits of diversity, communication costs, and public opinion dynamics[J]. Complexity, 2009, 15(2): 54-63.

[6] 刘永建，朱剑英，夏洪山，等. 飞机复杂系统故障诊断的灰色粗集推理方法[J]. 南京航空航天大学学报，2009,41(2):227-231.
Liu Yongjian, Zhu Jianying, Xia Hongshan, et al.

Grey-rough set reasoning of fault diagnosis in aircraft complex systems[J]. Journal of Nanjing University of Aeronautics & Astronautics, 2009, 41(2): 227-231.

[7] 刘大伟,陶来发,吕琛, 等. 飞机机电系统PHM的综合诊断推理机设计[J]. 南京航空航天大学学报, 2011, 43(S1): 114-118.
Liu Dawei, Tao Laifa, Lv Shen, et al. Design for integrated diagnosis inference engine for PHM of aircraft electromechanical system[J]. Journal of Nanjing University of Aeronautics & Astronautics, 2011, 43(S1): 114-118.

[8] Chi Yun, Song Xiaodan, Zhou Dengyong, et al. On evolutionary spectral clustering[J]. ACM Transactions on Knowlege Discovery from Data, 2009, 3(4): 1-30.

[9] Newman M E J. Analysis of weighted networks[J]. Physical Review E, 2004, 70(5): 056131.

[10] Newman M E J, Girvan M. Finding and evaluating community structure in networks[J]. Physical Review E, 2004, 69(2): 026113.

[11] Newman M E J. Fast algorithm for detecting community structure in networks[J]. Physical Review E, 2004, 69(6): 066133.

[12] Sundar S, Bhagavan B K, Datta A. Computing eigenvalues: Lanczos algorithm with a new recursive partitioning method[J]. Computers & Mathematics with Applications, 1999, 38(5/6): 99-107.

[13] Michel J, Yurii N, Peter R, et al. Generalized power method for sparse principal component analysis [J]. Machine Learning Research, 2010, 11(2): 517-553.

[14] Waugh A S, Pei Liuyi, Fowler J H, et al. Party polarization in congress: A social networks approach [EB/OL]. 2009-07-20 [2012-06-18]. http://jhfowler.ucsd.edu/party_polarization_in_congress.pdf.

[15] Palla G, Barabasi A L, Vicsek T. Quantifying social group evolution[J]. Nature, 2007, 446(7136): 664-667.

[16] Chen S M, Tanuwijaya K. Multivariate fuzzy forecasting based on fuzzy time series and automatic clustering techniques[J]. Expert Systems with Applications, 2011, 38(8): 10594-10605.

[17] Qian Ailing, Qu Binbin, Lu Yansheng, et al. A modularity analysis method for forum situation prediction[J]. Wuhan University Journal of Natural Science, 2011, 16(2): 148-154.