第 35 卷 第 6 期 计 算 机 工 程 2009 年 3 月
Vol.35 No.6 Computer Engineering March 2009
·软件技术与数据库· 文章编号：1000—3428(2009)06—0063—03 文献标识码：A 中图分类号：TP393

网络论坛的自相似性及其模型

曾剑平，张世永
(复旦大学计算机学院，上海 200433)

摘 要：网络论坛是一种主要的互联网应用，人们对它的研究主要集中在话题分析、变化趋势分析等方面。该文研究网络论坛中文章数随时间变化的统计特性，通过方差分析、R/S 分析方法发现自相似性存在于网络论坛中。为描述这种自相似性，提出一种基于时延厚尾分布的产生式模型，对该模型进行理论与仿真分析，验证了该模型具有表达网络论坛自相似性的能力，模型的计算复杂度小。
关键词：网络论坛；自相似性；产生式模型

Self-similarity and Modeling of Web-based Forum

ZENG Jian-ping, ZHANG Shi-yong
(School of Computer, Fudan University, Shanghai 200433)

【Abstract】 Web-based forum is one of important applications on Internet. More and more research has been focused on it, such as topic analysis, trend analysis. This paper concentrates on the statistical character of the posting number in Web-based forum. By means of variance analysis and R/S analysis, finds that self-similarity generally exists in Web-based forum. A generative model based on time-delay levy distribution is proposed to describe the kind of property. Theoretic and simulation analysis are done to verify the effectiveness of the model and show that the computation complexity is low.
【Key words】 Web-based forum; self-similarity; generative model

1 概述

近年来，随着 Web2.0 的提出，可写式互联网得到人们越来越多的关注，网络论坛、博客等正逐步成为互联网应用的热点。网络论坛中的话题热点分析成为目前主要的研究方向，采用的技术主要是文本分析、聚类和分类算法等[1]。另一方面，由于各种商业活动决策的需要，如产品市场调查或广告，从宏观上了解论坛或博客的人气变化，用户参与度的变化正在成为一个新的热点，对论坛中文章数随时间变化规律的研究最近也得到了人们的关注[2-4]。例如，采用隐 Markov 模型(HMM)对话题的生命周期进行建模型，并按照模型进行变化规律的预测[2]。文献[3]结合话题语义和 HMM 模型进行生命周期的分析。

虽然 Markov 类的概率模型能描述时间序列的相关特性，但是它们是一种短时相关模型，对于长相关的过程则无法描述，因此，上述的趋势分析方法的适用性有限。本文针对整个论坛或论坛中独立的频道所表现出来的宏观文章数变化，研究文章数序列的长相关特性。本文的主要创新点是：基于实际的论坛数据采用统计分析方法发现了论坛文章数随时间变化的自相似性，提出一个适合描述网络论坛自相似性的产生式模型，并进行了验证与分析。

2 网络论坛中的自相似性分析

2.1 理论基础

在互联网应用领域中进行自相似性研究比较早的是 Leland, Taqqu 等人，他们在对网络流量进行大量观察分析时，发现了在较宽广的时间粒度上网络数据包的统计特性是相似的[5]。

对于一个时间序列 X_t, t=1,2,…，如果对所有 $m \in N$，以下公式成立，则称序列是严格自相似[6]。如果对所有 $m \to \infty$，该式成立，则称序列是渐进自相似。

$$X_t \overset{d}{=} m^{-H} \sum_{i=(t-1)m+1}^{tm} X_i^{(m)} \tag{1}$$

其中，$\overset{d}{=}$ 表示概率分布上的相等；$X_i^{(m)}$ 称为序列 X_t 的 m-聚集序列，可通过如下式子计算得到：

$$X_i^{(m)} = \frac{1}{m} \sum_{t=(k-1)m+1}^{km} X_t \tag{2}$$

式(1)中的参数 H 反映了时间序列自相关函数的衰减速度，称为 Hurst 指数，描述了序列的自相似度。自相似过程往往表现出长相关，而具有长相关特性的过程存在如下的自相关函数[6]：

$$r(k) \sim k^{-\beta}, \quad 0 < \beta < 1 \tag{3}$$

β 与 H 具有如下的关系：

$$H = 1 - \beta/2 \tag{4}$$

因此，对于自相似过程有 0.5<H<1。H 越大表示自相似度越大。计算 Hurst 值是一种判断自相似性的主要方法。

2.2 论坛文章数的自相似性

通过网络爬虫方式从某高校的 bbs 上下载 2 个频道的每篇文章(包括回帖)，这 2 个版面具有不同的特点：一个是学生专门用来讨论与生活相关的各种话题(频道 1)；另一个是有关国内外的时事政治的话题(频道 2)。所选择的不同的时间段(表 1)构成了分析用的数据集。

2 个频道各包含一段学生放假的时间，在这段时间内用户访问论坛的次数相对较少。对这 4 个时间段对应的数据集进行统计，得到每分钟内的文章数序列 X，检验该序列是否

作者简介：曾剑平(1973－)，男，博士，主研方向：信息安全；张世永，教授、博士生导师
收稿日期：2008-08-10 **E-mail：**zeng_jian_ping@hotmail.com

具有自相似性，并计算其自相似度量值。

表 1　分析用的实际论坛数据集

数据集	版面	时间段	文章数
1	频道 1	2006-012-18~2007-01-19	40 794
2	频道 1	2007-01-21~2007-02-14	22 116
3	频道 2	2006-08-22~2006-09-30	28 740
4	频道 2	2006-11-02~2006-12-14	20 259

判断一个序列 X 是否具有自相似性，可以通过方差分析和 R/S 统计 2 种方式[6]，假设时间序列 X_t, t=1,2,⋯,N，表示在 t 时刻论坛中的文章数，t 的单位取秒。

(1)方差分析

选择一个正整数 m=10 s, 100 s, 1 000 s, 8 000 s，把文章数序列 X 分割成大小等于 m 的 N/m 个子序列，按照式(2)计算 m-聚集序列 $X_i^{(m)}$，并计算该序列的方差 var($X_i^{(m)}$)。在 lg-lg 坐标图上，随着 m 的增大，lg(var)近似按照线性递减的方式变化，直线的斜率即为式(3)中的指数(-β)。

(2)R/S 统计

选择 X 的子序列 $Y=\{X_1, X_2,\cdots,X_n\}$，分别计算 Y 的部分和 $Y^{(m)}=\sum_{i=1}^{n}X_i$ 及 Y 的样本方差 $S(n)$。计算 R/S 统计量：

$$RS(n)=\frac{1}{\sqrt{S(n)}}[\max_{1\le t\le n}(Y(t)-\frac{t}{n}Y(n))-\min_{1\le t\le n}(Y(t)-\frac{t}{n}Y(n))]$$

改变 n 的值，从而得到 $RS(n)$与 n 的关系，在 lg-lg 坐标上，这种关系也近似于线性关系，直线的斜率为 H 值。

图 1、图 2 分别是网络论坛数据集 3 的方差分析和 R/S 统计分析的结果，通过采用最小均方误差拟合可以得到相应的拟合直线的斜率分别为-0.292, 0.843，因此，它们对应的 H 值分别为 H=0.854, 0.843，可见 2 种方法计算得到的 H 值很接近。对以上 4 个数据集的估计结果(H 值)如表 2 所示。

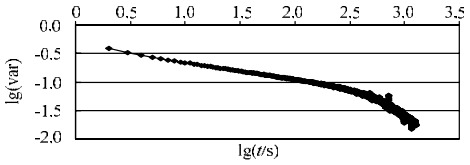

图 1　数据集 3 的方差分析结果

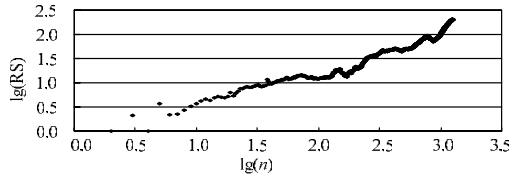

图 2　数据集 3 的 *RS* 分析结果

表 2　各个数据集的 *H* 值

数据集	方差法	*R/S* 分析
1	0.891	0.875
2	0.889	0.870
3	0.854	0.843
4	0.903	0.894

对这些数据集的测试结果可见 H>0.5，因此，这些文章数序列具有自相似性。

3　自相似模型——GMWF

著名复杂系统专家 Barabási 于 2005 年在 Nature 上公布的最新的研究中，利用接收和发送 E-mail 的日志数据，发现了人的行为通常存在一个突发现象，而这种突发现象通常是自相似性的一个重要表现。他提出了一种工作序列的优先级模型，认为与用户相关的事件时间间隔存在厚尾分布，这种分布导致了突发行为的出现。

基于这个研究结果，假设网络论坛上的用户在浏览发表文章的过程中也存在服从厚尾分布的时间延时。

3.1　模型描述

假设在 $t1$, $t2$ 时刻网络论坛上各出现了一个新文章，令时间延时 $\Delta t=t2-t1$，则 Δt 服从下面的 Pareto 分布：

$$p(x=\Delta t)=\alpha x^{-a-1},\ \alpha>0 \tag{5}$$

该分布是一种最简单的厚尾分布。

网络论坛自相似性的产生式模型是基于实际的用户发行为，假设每秒最多只有一个用户发贴，则该产生式模型(GMWF)描述如下：

输入：参数 α，模拟时间长度 T

Step 1　在时刻 $t0$ 出现一个新文章，即 $x(t0)=1$

Step 2　按照均匀分布产生一个[0, 1]的随机变量值 Ψ

Step 3　按照下面的公式产生一个延时长度

$$\Delta t=\mathrm{e}^{-\frac{\lg(\psi)}{\alpha}} \tag{6}$$

Step 4　在 $t0$+1 到 $t0+\Delta t$ 时间范围内产生连续空输出，即 $x(t)$=0, $t=t0$+1, $t0$+2,⋯, $t0+\Delta t$

Step 5　在 $t0+\Delta t$ +1 时刻产生一个输出，即 $x(t0+\Delta t+1)$=1

Step 6　$T \leftarrow T-(\Delta t+1)$

如果 T>0，则：$t0 \leftarrow t0+\Delta t$+1，转 Step 2 执行

Step 7　结束

输出：文章数序列 X

Step 2, Step3 是为了产生一个服从 Pareto 分布的伪随机变量的值。

3.2　模型分析

模型的复杂性分析：该模型在时间上没有复杂的操作，但是对空间有一定要求，即所需要的空间长度为 T。在实际应用中，如果以秒作为时间单位，那么研究 m 个月的连续文章数序列时，T 的长度是 m×30×24×3 600。为了保存这样的序列，需要花费的空间比较大。为此，采用一种简单的压缩序列表示方式，即如果只有一个空输出，则用一个 0 表示，如果有多个空输出，则记录该空输出的个数。

4　实验结果与分析

4.1　实验方法

用 Java 语言实现了 GMWF 模型，并在 WindowsXP 系统上做了测试。实验中设置 T=30×24×3 600，即模拟产生一个月的论坛文章数序列。

同时为了与现有的相关模型的准确性做比较，选择 HMM 模型作为论坛文章数序列的表示，文献[3]采用固定隐状态数(等于 4, 6)的模型，这个模型的输出状态为 0 和 1，分别表示有文章、没有文章 2 种情景。而 HMM 模型的转移矩阵、输出分布和初始分布则采用随机方式产生。

基于 HMM 模型的文章数序列产生过程如下：首先根据初始概率分布选择一个隐藏状态，根据隐藏状态对应的输出分布选择一个 0 或 1 的输出，在下一个时刻，根据转移矩阵选择下一个可能的隐状态，再产生一个输出。重复这个过程，直到产生的序列长度为 T。

对 GMWF 和 HMM 产生的序列采用方差分析和 R/S 统计分析方法，计算其 Hurst 指数。

4.2　实验结果

实验中改变 α 的值，产生 T 长度的序列。对于每个 α，各运行 5 次，计算通过方差分析得到的平均 Hurst 值，如表 3 所示。

表 3　GMWF 产生的序列的 *H* 值

α	*H* 值
1.9	0.638
1.7	0.704
1.5	0.751
1.3	0.808
1.1	0.866
0.9	0.870
0.7	0.891
0.5	0.910

同样，对 HMM 模型产生的序列进行了统计分析，在不同模型参数下的测试结果如表 4 所示。

表 4　HMM 模型产生的序列的 *H* 值

模型	*H* 值
随机模型 1 (隐状态数=4)	0.498
随机模型 2 (隐状态数=4)	0.497
随机模型 3 (隐状态数=4)	0.499
随机模型 4 (隐状态数=6)	0.500
随机模型 5 (隐状态数=6)	0.501
随机模型 6 (隐状态数=6)	0.499

从实验结果可以看出，GMWF 能很好地产生自相似序列，能描述 H>0.5 的自相似性，并且当 α<1.3 时，其 H 值与实际观察到的序列的 H 值很接近。而 HMM 模型所产生的序列的 H 值大都小于或等于 0.5，不具备自相似性。实际上由于 HMM 模型产生的序列的自相关函数是指数衰减的，不能描述长相关特性。改变参数值 GMWF 模型能获得具有不同 H 值的序列，说明了该模型具有描述多种不同类型网络论坛文章数序列的能力。

从模型的计算复杂度分析，取 α 为 1.5，产生的模拟序列分别为 15 天、30 天、45 天、60 天、75 天时的计算复杂度，如图 3 所示，可见计算复杂度与模拟的天数呈线性关系。

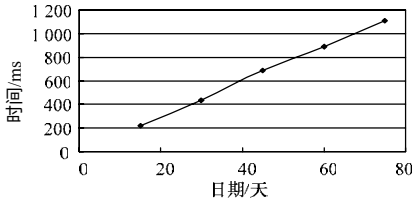

图 3　GMWF 模型的计算复杂度

为了反映模型所需要的空间，当 GMWF 采用压缩序列表示时，选择不同指数和时间长度(15 天、30 天、60 天)，计算所需要的空间需求如图 4 所示。可见当 α 增加到一定值时，模型所需要的空间增加得很缓慢。

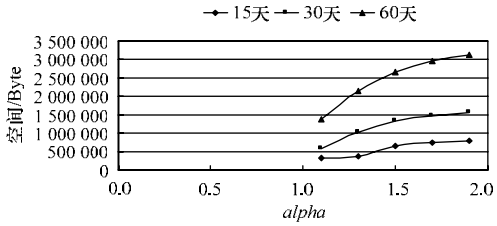

图 4　GMWF 的空间复杂度

5　结束语

网络论坛分析已越来越受到人们的关注，本文研究论坛中的文章数随时间变化的统计特性，采用方差分析和 R/S 分析发现了网络论坛中存在自相似性，并提出一个产生式模型，与目前常用的基于 Markov 或 HMM 模型的表示方法相比，它能有效地揭示论坛文章数随时间变化的自相似性。这种自相似性对于提高论坛文章数趋势预测的准确性以及分析用户行为变化具有很好的指导意义。后续的工作将利用这种自相似性研究论坛文章数变化的预测算法。

参考文献

[1] Ye Huimin, Cheng Wei, Dai Guanzhong. Design and Implementation of On-line Hot Topic Discovery Model[J]. Wuhan University Journal of Natural Sciences, 2006, 11(1): 21-26.

[2] Zeng Jianping, Zhang Shiyong. Predictive Model for Internet Public Opinion[C]//Proc. of International Conference on Fuzzy System and Knowledge Discovery. [S. 1.]: IEEE Computer Press, 2007: 7-11.

[3] Mei Qiaozhu, Xu Ling, Wondra M, et al. Topic Sentiment Mixture: Modeling Facets and Opinions in Weblogs[C]//Proc. of International Conference on World Wide Web. New York, USA: ACM Press, 2007: 171-180.

[4] Bansal N, Koudas N. BlogScope: Spatio-temporal Analysis of the Blogosphere[C]//Proc. of International Conference on World Wide Web. New York, USA: ACM Press, 2007: 1269-1270.

[5] Leland W E, Taqqu M S, Willinger W, et al. On the Self-similar Nature of Ethernet Traffic[J]. IEEE/ACM Transaction on Networking, 1994, 2(1): 2-15.

[6] Crovella M C, Bestavros A. Self-similarity in World Wide Web Traffic: Evidence and Possible Causes[J]. IEEE/ACM Transactions on Networking, 1997, 5(6): 835-846.

编辑　索书志

(上接第 62 页)

由以上 2 个实验可以看出，AIFCM 算法可以很好地处理增量数据，给出一个不错的聚类结果。

7　结束语

目前已经提出了许多聚类算法及其变种，但增量式聚类算法的研究较少。当数据集因更新而发生变化时，数据挖掘的结果也应该进行相应的更新。由于数据量大，在更新后的数据集上重新执行聚类算法以更新挖掘结果显然效率比较低下。本文提出的自适应模糊 C-均值的增量式聚类算法考虑了数据分布的几何结构，能够处理新增数据，产生新类，对新增数据不需要重新聚类，使类内具有较高的凝聚度、类间具有较大的差异性，并且结果不受孤立点的干扰，对于类的分裂，可以自动确定聚类数目和聚类中心，减少了 FCM 聚类效果受主观因素的影响，提高了聚类效率。此外，本算法还能较好地处理噪声数据，提高算法的实用性。

参考文献

[1] 吴　琪, 高　滢, 王晓涛. 一种基于距离的增量聚类算法[J]. 解放军理工大学学报, 2005, 6(6): 537-540.

[2] 王洪春, 彭　宏. 基于模糊C-均值的增量式聚类算法[J]. 微电子学与计算机, 2007, 24(6): 156-161.

[3] Tan Pangning, Steinbach M. 数据挖掘导论[M]. 范　明, 译. 北京: 人民邮电出版社, 2006.

[4] 楼顺天. 基于MATLAB的系统分析与设计——模糊系统[M]. 西安: 西安电子科技大学出版社, 2001.

编辑　张　帆