УДК 004.4’242

СТАТИСТИЧЕСКИЙ АНАЛИЗ ТЕЛЕКОММУНИКАЦИОННЫХ СЕТЕЙ НА ОСНОВЕ ИССЛЕДОВАНИЯ ИНФОРМАЦИОННЫХ ПОТОКОВ, ПРЕДСТАВЛЕННЫХ В ВИДЕ ВРЕМЕННЫХ РЯДОВ

© 2006 А. К. Скуратов

Государственный научно-исследовательский институт
информационных технологий и телекоммуникаций, г. Москва

В целях статистического анализа телекоммуникационных сетей рассматриваются марковская модель системы, моделирование сетевого трафика фрактальным броуновским движением, моделирование временных рядов. Приведены модели временных рядов, разработанные для статистической системы мониторинга телекоммуникационных сетей.

Учитывая объективно сложившуюся неоднородность как телекоммуникационных сетей, сетевых информационных ресурсов, так и аудитории, которой данная информация адресована, необходимо создание и надежное функционирование достаточно большого набора инфокоммуникационных сервисов, обеспечивающих эффективную работу пользователя с разнородной информацией в гетерогенной телекоммуникационной сети. Практика использования и эксплуатации гетерогенных телекоммуникационных сетей, связанная с недостаточной их прозрачностью, сложностью, организационными ограничениями и спецификой, определяет необходимость использования статистических методов их анализа и мониторинга на основе открытой потоковой информации, которую можно легко получить, используя доступные методы и средства.

В результате обработки статистической информации о функционировании телекоммуникационной сети можно определить нормальный профиль сети (этап анализа). Выявление и предсказание отклонений от нормального профиля сети (этап мониторинга) проводится системным администратором с целью определения возникновения нештатной ситуации и принятия соответствующего решения об изменении конфигурации сети.

Таким образом, является актуальной разработка методов сбора первичной статистической информации о функционировании телекоммуникационной сети, обработки первичной информации с использованием выбранных статистических методов анализа и выработка рекомендаций по реконфигурации сети.

Работы выполняются при поддержке Российского фонда фундаментальных исследований, проект 05-07-90360.

Исследование и анализ информационных потоков, циркулирующих в телекоммуникационных сетях с целью выбора математической модели

С целью выбора наиболее адекватной модели для анализа и мониторинга телекоммуникационных сетей рассмотрим наиболее распространенные модели системы.

Марковская модель системы. В качестве исходной информации для построения марковской модели рассматриваются так называемые события, например, все действия пользователя, связанные с безопасностью: локальная авторизация, запросы на удаленный доступ и т. п.

Пусть событие – это одно из возможных случайных значений состояния системы $\vartheta_1, \vartheta_2, ..., \vartheta_k, ..., \vartheta_K$. Тогда система описывается дискретным во времени случайным процессом с множеством значений $\vartheta_1, \vartheta_2, ..., \vartheta_k, ..., \vartheta_K$, каждое из которых является определенным событием, фиксируемым операционной системой. Интервалы между различными событиями определяются отдельными действиями пользователя, вызвавшими то или иное событие, и, следовательно, могут быть неодинаковыми. Однако это не имеет существенного значения для построения марковской

модели, так как в ней важна последовательность действий, а не интервал между ними. Тогда $\theta_n = \theta(t_n)$ - случайная величина, характеризующая состояние системы через n шагов, а $\theta_0 = \theta(t_0)$ - случайное начальное состояние системы.

Полное вероятностное описание поведения рассматриваемой системы задается совместными конечномерными вероятностями $P(\theta_0, \theta_1, ..., \theta_n)$ при всех n.

Для упрощения предполагается, что система описывается моделью простой цепи Маркова, и тогда вероятности $P(\theta_0, \theta_1, ..., \theta_n)$ определяются известным выражением

$$P(\theta_0, \theta_1, ..., \theta_n) = P_0(\theta_0) \prod_{\mu=1}^{n} \pi_\mu(\theta_\mu \mid \theta_{\mu-1}).$$

Далее делается допущение, что вероятности одношаговых переходов $\pi_\mu(\theta_\mu \mid \theta_{\mu-1})$ не зависят от времени, т. е. $\theta(t)$ рассматривается как простая стационарная цепь Маркова. Предположение о стационарности цепи Маркова вносит еще большие упрощения в модель, а также в вычислительный алгоритм, делая тем самым использование подобной модели удобным на начальных этапах исследования системы. Естественно, что в этом случае ставится вопрос об адекватности модели.

Моделирование сетевого трафика фрактальным броуновским движением [1]. При построении таких моделей сетевого трафика постулируется или доказывается фрактальность происходящих в сетях процессов на базе исследования свойства самоподобия.

В основе экспериментальной проверки фрактальных свойств трафика сети лежат методы, позволяющие по выборочным значениям числа событий на интервалах заданной длительности сформировать и оценить некоторые статистики, которые можно затем использовать для проверки гипотезы о протяженной зависимости трафика.

К числу процессов, аппроксимируемых фрактальным броуновским движением, можно отнести RTT-задержку (round-trip time задержка) [2, 3].

Для стационарного процесса RTT-задержку (обозначим ее T_i) можно записать в виде

$$T_i = T_{1i} + T_{2i} + T_{пр},$$

где i =1, 2, 3... – номера задержек (циклов); T_{1i}, T_{2i} – интервалы, соответствующие времени пересылки пакета от источника к приемнику и обратно, $T_{пр}$ - время обработки информации в приемнике. Для известного маршрута движения пакета величина задержки равна

$$T_i = T_0 + \Delta T_i,$$

где T_0 - постоянная составляющая при отсутствии очередей; ΔT_i – случайная составляющая, связанная с задержками в сети. Пусть $\Delta T_{ср}$ – среднее значение приращения RTT-задержки. Модель фрактального броуновского процесса для момента t_n имеет вид:

$$B_H(t_n) = \sum_{i=1}^{n} \left[T_i - (T_0 + \Delta T_{ñð}) \right].$$

Это позволяет записать выражение для корреляционной функции процесса и перейти к ее исследованию.

Моделирование временных рядов. Моделирование различных составляющих, характеризующих работу сети, таких, как объем трафика, количество потерянных пакетов и др. [4], в виде временных рядов имеет ряд очевидных преимуществ по сравнению с вышеописанными способами. При построении модели временных рядов используется экспериментальная информация (полученная в реально функционирующей сети), требуется меньше допущений и, следовательно, более адекватно отражается реальный объект, т. е. телекоммуникационная сеть. Математическая модель описывает поток информации в зависимости от момента t. При статистическом анализе временных потоков информации необходимо осуществить выделение тренда, выделение периодических составляющих - колебаний относительно тренда с некоторой регулярностью, анализ случайного компонента.

Математическое описание обычно включает в себя одну из подобных составляющих или сумму нескольких из них.

Для такого показателя работы сети, как загрузка каналов, в [5] предложена модель, включающая три составляющие:

$$Y(t) = f(t)+g(T)+\varepsilon(t),$$

где $f(t)$ - тренд, медленно меняющаяся во времени функция, описывающая изменения среднесуточных (средненедельных) загрузок за интервалы времени большие, чем суточная периодичность; $g(T)$ - периодическая составляющая, которая может быть описана конечным рядом Фурье, построенным по экспериментальным данным величин загрузок телекоммуникационного канала; $\varepsilon(t)$ - случайная последовательность, относительно которой делается предположение о равенстве нулю ее математического ожидания $M[\varepsilon(t)] = 0$.

Предлагаются следующие методы исследования данной модели. Моделирование тренда может проводиться с помощью хорошо разработанных методов регрессионного анализа. Для построения ряда Фурье следует применять методы анализа периодограмм и спектрального анализа случайных процессов. Свойства и характеристики случайной последовательности $\varepsilon(t)$ изучаются с помощью классических методов математической статистики и методов анализа случайных последовательностей.

Модели, построенные на современных частично эвристических методах и предложенные в данной работе для исследования и анализа функционирования телекоммуникационных сетей, будут рассмотрены ниже.

Таким образом, статистические модели телекоммуникационных сетей в виде временных рядов наиболее достоверны, так как основаны на большом числе экспериментальных данных и, следовательно, являются и наиболее информативными для прогноза состояния сети.

Временные ряды и их характеристики для целей статистического мониторинга телекоммуникационных сетей

В случае статистического мониторинга телекоммуникационных сетей при анализе временных рядов наибольший интерес представляет прогнозирование будущих значений ряда. Процедуры предсказания, как правило, основываются на моделировании структуры рядов. Если моделирование осуществляется только с использованием значений самого моделируемого ряда без применения какой-либо дополнительной наблюдаемой переменной (ряда), то говорят об анализе одномерных рядов. Примерами моделей одномерных временных рядов могут служить модель тренда или авторегрессионная модель.

В статистической системе мониторинга телекоммуникационных сетей реализуется процедура автоматического обнаружения подозрительных (т. е. аномально отклоняющихся от тренда) значений. В основе этой процедуры лежит представление о ряде как о сумме тренда и случайной составляющей. Соответственно, выброс – это точка, отстоящая слишком далеко от предполагаемой линии тренда. Для поиска выбросов сначала к ряду применяется процедура медианного сглаживания, состоящая из применения k-точечной скользящей медианы. Затем строится ряд остатков и находится устойчивая оценка его стандартного отклонения (медиана абсолютных отклонений (MAD), деленная на 0,6745 для устранения смещения в случае нормального распределения. В качестве выбросов в исходном ряду рассматриваются точки ряда остатков, превысившие по модулю приблизительно m стандартных отклонений (величина m называется *уровнем детектирования*). Значения параметров k и m задаются пользователем в процессе диалога. Обычно величина m выбирается равной 4. Значения исходного ряда в точках выброса заменяются значениями, полученными при медианном сглаживании.

Под разрывом понимается скачкообразное изменение уровня временного ряда. С данной точки зрения, разрыв – это выброс в ряду значений первых последовательных разностей исходного ряда. Предлагаемая в статистической системе мониторинга телекоммуникационных сетей процедура поиска разрывов (дополнительно к визуальному анализу) устроена следующим образом: сначала к ряду применяется k-точечная скользящая медиана, чтобы отфильтровать возможные выбросы. Затем формируется ряд последовательных разностей сглаженного ряда и получен-

ный ряд обрабатывается с помощью процедуры поиска выбросов.

Модели для временных рядов, разработанные для статистической системы мониторинга телекоммуникационных сетей

Модель авторегрессии. Модель авторегрессии предназначена для описания стационарных временных рядов. Под процессом авторегрессии порядка p (обозначение – АР(p), в английской нотации AR(p)) понимают процесс $X(t)$, удовлетворяющий для некоторой константы c соотношению

$$y_t = \varphi_1 y_{t-1} + ... + \varphi_p y_{t-p} + e_t ,$$

где $y_t = x_t - c$, а e_t – "белый шум" с нулевым средним.

Приведенное уравнение может описывать и нестационарные процессы.

Процесс $X(t)$ стационарен, если все корни полинома $\Phi(z) = 1 - \varphi_1 z - ... - \varphi_p z^p$ лежат вне единичного круга $|z| > 1$.

При слабых дополнительных предположениях стационарный процесс удовлетворяет уравнению авторегрессии бесконечного порядка с убывающими коэффициентами. Поэтому авторегрессионная модель достаточно высокого порядка может хорошо аппроксимировать почти любой стационарный процесс, часто применяется для моделирования остатков в той или иной параметрической модели, например, регрессии или тренда.

Моделью АР(2) хорошо описывается процесс колебаний маятника под действием случайных возмущений.

Для процесса АР(p) теоретические значения частной автокорреляционной функции для лагов, больших p, равны нулю. На основании этого свойства можно выбирать порядок модели авторегрессии для описания выборочных данных. Модель авторегрессии является частным случаем более общей модели АРИСС (ARIMA Бокса – Дженкинса), пригодной и для описания нестационарных рядов.

Модель скользящего среднего. Модель скользящего среднего СС(q) (в английской нотации MA(q)) описывает стационарные временные ряды и является частным случаем модели Бокса – Дженкинса (АРИСС). Модель записывается в виде

$$x_t = c + e_t - \Theta_1 e_{t-1} - .... - \Theta_q e_{t-q} ,$$

где e_t – "белый шум", c – константа (среднее значение ряда), а Θ_i – коэффициенты модели.

Модель всегда описывает стационарный ряд, но для анализа пригодна лишь такая форма модели, для которой выполняется условие обратимости: все корни полинома

$$\Theta(z) = z^q - \Theta_1 z^{q-1} - .... - \Theta_q$$

лежат внутри единичного круга $|z| < 1$. В этом случае процесс e_t имеет смысл ошибок прогноза на один шаг вперед.

Для процесса СС(q) все значения автокорреляционной функции для лагов, больших q, равны 0. Это свойство является характеристическим.

Важное практическое значение имеют процессы, первая (или более высокая) разность которых стационарна и является процессом СС(q). Подобные процессы устроены как случайные колебания с непостоянным средним уровнем или (для второй разности) непостоянным углом наклона. Для прогнозирования таких процессов часто используется метод экспоненциального сглаживания.

Модель авторегрессии скользящего среднего. Моделями СС(q) и АР(p) за счет выбора их порядков q и p можно удовлетворительно описывать многие реальные процессы. Однако на практике для достижения большей гибкости в подгонке моделей к наблюдаемым временным рядам иногда бывает целесообразным объединить в одной модели и авторегрессию, и скользящее среднее. При этом цель должна состоять в построении моделей наиболее экономных (простых), дающих хорошую аппроксимацию с помощью небольшого числа параметров. Достижению этого помогает рассмотрение *смешанных моделей авторегрессии-скользящего среднего* или *моделей АРСС*(p,q):

$$x_t = \varphi_1 x_{t-1} + ... + \varphi_p x_{t-p} + e_t - \Theta_1 e_{t-1} - ... - \Theta_q e_{t-q}$$

или

$$\varphi(B)x_t = \Theta(B)e_t,$$

где $\Theta(B)$ и $\varphi(B)$ – операторы, определенные, соответственно, для моделей СС(q) и АР(p) и удовлетворяющие сформулированным ранее условиям стационарности и обратимости; e_t – такие же, как и раньше. Подобная модель может оказаться подходящей, например, в том случае, когда наблюдаемый временной ряд является суммой двух или более независимых составляющих, каждая из которых описывается либо моделью АР, либо моделью СС, но которые непосредственно не наблюдаются.

Сезонность. Под сезонностью понимают влияние внешних факторов, действующих циклически с заранее известной периодичностью. Типичными примерами являются эффекты, связанные либо с астрономическими, либо с календарными причинами. Так, в рядах ежемесячных данных часто встречаются сезонные эффекты с периодом 12, в квартальных рядах – с периодом 4. В свою очередь, в информации, собираемой с интервалом 1 ч, могут присутствовать "сезонные эффекты" с периодом 24, а собираемой с интервалом 5 мин – сезонные колебания с периодом 12 (час) и 288 (сутки).

Одна из наиболее простых моделей учета сезонности – модель сезонных эффектов. В аддитивной форме этой модели ряд представляется в виде

$$Y(t) = T(t) + S(t) + e_t,$$

где $T(t)$ – тренд; e_t – ошибка; а $S(t)$ – сезонная составляющая, которая предполагается периодической с периодом L: $S(t) = S(t+L)$.

Фактически функция S определяется своими значениями на периоде длины L, например, $S(1)$, ..., $S(L)$. Для однозначности параметризации модели обычно предполагают, что $S(1) + ... + S(L) = 0$. Значения $S(1)$, ..., $S(L)$ называют индексами сезонности. Поясним их смысл на примере. Пусть $Y(t)$ – ряд суточных данных, а период сезонности – неделя. Соответственно, $L = 7$. Для определенности положим, что момент $k = 1$ соответствует понедельнику. Тогда коэффициент $S(1)$ выражает среднестатистическое отличие понедельников от среднего по всем дням недели. В свою очередь, $S(2)$ – аналогичная характеристика вторников и т. д.

Для рядов, содержащих явно выраженный тренд, часто более естественна мультипликативная форма модели. В этом случае в качестве условия нормировки используется условие $S(1) \times ... \times S(L) = 1$.

Индексы сезонности рассматриваются в статистической системе мониторинга телекоммуникационных сетей как периодические функции с бесконечной областью определения и в таком качестве могут участвовать в любых арифметических операциях над временными рядами. Наличие сезонных эффектов проявляется в виде острых узких пиков в периодограмме на соответствующей частоте (при несимметричной форме сезонной волны – и на кратных частотах). В выборочной автокорреляционной функции также присутствуют выбросы для лагов (запаздываний), кратных периоду сезонности, но эти выбросы могут быть завуалированы присутствием тренда или большой дисперсией случайного компонента.

В статистической системе мониторинга телекоммуникационных сетей для прогнозирования при использовании нескольких временных рядов будем применять либо линейную авторегрессионную модель

$$u_t = a_1 u_{t-1} + q_2 u_{t-2} + ... + b_1 y_{t-1} + \\ + b_2 y_{t-2} + ... c_1 z_{t-1} + c_2 z_{t-2} + ...,$$

либо нейронную сеть с несколькими промежуточными слоями (линейная авторегрессионная модель может рассматриваться как крайний случай нейронной сети без промежуточных слоев). Для оценки коэффициентов авторегрессионной модели и нейронных сетей сначала выбираются ряды $y, z, ...$, которые будут участвовать в предсказании, и формируется матрица данных X со строками вида

$$u_t u_{t-1}, ..., u_{t-k_1}, ..., y_{t-1}, ..., y_{t-k_2}, z_{t-1}, ..., z_{t-k_3}.$$

Таких строк (объектов) в матрице данных будет $n - k + 1$, где $k = max(k_1, ..., k_q)$ и q –

число используемых рядов. Величина лагов k_i, как и состав предсказывающих рядов, специфицируется пользователем. В полученной матрице данных X имеется $\sum_{i=1}^{q} k_i + 1$ переменных. Первая переменная $x = u_t$ является прогнозируемой, а остальные – предсказывающими.

Теперь для оценки коэффициентов выбранной модели могут использоваться все методы регрессионного анализа.

Список литературы

1. Городецкий А. Я., Заболоцкий В. С. Фрактальные процессы в компьютерных сетях. – СПб, Издательство СПбГТУ, 2000.

2. Mandelblot B. B., Van Ness J. W. Fractional Brownian motions, fractional noises and applications// SIAM Reiew, № 10, 1968, p 422-437.

3. Qiong Li, David L.Mills. On the long-range dependence of packet round-trip delays in Internet// Processings of IEEE ICC'98, v. 2, 1998.

4. Айвазян С. А. и др. Прикладная статистика. Классификация и сокращение размерностей/С. А. Айвазян, В. М. Бухштабер, И. С. Енюков, Л. Д. Мешалкин. – М.: Финансы и статистика, 1989.

5. Статистический анализ и мониторинг научно-образовательных интернет-сетей/ И. С. Енюков, И. В. Ретинская, А. К. Скуратов; Под. ред. А. Н. Тихонова. - М.: Финансы и статистика, 2004.

STATISTICAL ANALYSIS OF TELECOMMUNICATION SYSTEM ON THE BASIS OF STUDYING INFORMATION FLOWS PRESENTED AS TIME SERIES

© 2006 A.K. Skuratov

State research institute of information technologies and telecommunications, Moscow

In an effort to analyse telecommunication systems statistically the author considers markov's system model, net traffic simulation by fractal Brownial motion, simulation of time series. Time series models developed for the statistical system of telecommunication system' monitoring are given.