

УДК 004.724.4, 519.246.8

МОДЕЛИРОВАНИЕ СЕТЕВОГО ТРАФИКА И ПРОГНОЗИРОВАНИЕ С ПОМОЩЬЮ МОДЕЛИ ARIMA

Гребенников Александр Витальевич¹, Крюков Юрий Алексеевич²,
Чернягин Денис Викторович³

¹Ассистент;

ГОУ ВПО «Международный Университет природы, общества и человека «Дубна»,
Институт системного анализа и управления;
141980, Московская обл., г. Дубна, ул. Университетская, 19;
e-mail: grebennikovav@mail.ru.

²Кандидат технических наук, доцент Института системного анализа и управления;

ГОУ ВПО «Международный Университет природы, общества и человека «Дубна»,
Институт системного анализа и управления;
141980, Московская обл., г. Дубна, ул. Университетская, 19;
e-mail: kua@uni-dubna.ru.

³Ассистент;

ГОУ ВПО «Международный Университет природы, общества и человека «Дубна»,
Институт системного анализа и управления;
141980, Московская обл., г. Дубна, ул. Университетская, 19;
e-mail: dancher2000@mail.ru.

Прогнозирование сетевого трафика представляет значимый интерес в таких областях как отслеживание перегрузок в сети, контроль потоков данных и сетевое управление. Тщательно подобранная модель трафика способна выявить и предугадать важнейшие характеристики сетевого трафика, такие как кратковременно и долговременно зависимые процессы, самоподобность на больших временных масштабах. В данной статье подбирается модель ARIMA с минимальным числом параметров, имеющая адекватный прогноз. Также представлена процедура оценки параметров модели ARIMA и выбора модели с минимальным числом параметров. Приведены сравнения оценок качества прогноза для полученных моделей.

Ключевые слова: Прогнозирование сетевого трафика, модель трафика, ARIMA.

MODELING NETWORK TRAFFIC AND FORECASTING WITH MODEL ARIMA

Grebennikov Alexandr¹, Krukov Yury², Chernyagin Denis³

¹Assistant;

Dubna International University of Nature, Society, and Man,
Institute of system analysis and management;
141980, Dubna, Moscow reg., Universitetskaya str., 19;
e-mail: grebennikovav@mail.ru.

²Candidate of Science in Engineering, associate professor;

Dubna International University of Nature, Society, and Man,
Institute of system analysis and management;
141980, Dubna, Moscow reg., Universitetskaya str., 19;
e-mail: kua@uni-dubna.ru.

³Assistant;

Dubna International University of Nature, Society, and Man,
Institute of system analysis and management;
141980, Dubna, Moscow reg., Universitetskaya str., 19;
e-mail: grebennikovav@mail.ru.

The predictability of network traffic is a significant interest in many domains such as congestion control, admission control, and network management. An accurate traffic prediction model should have the ability to capture prominent traffic characteristics, such as long-range dependence and self-similarity in the large time scale. In this paper the model ARIMA with minimal amount of parameters has been selected to accurate the traffic prediction. Also the procedure of the estimating of the parameters of the ARIMA model has been exhibited and compared the estimations of accurate the traffic prediction for the getting models.

Keywords: Traffic prediction, traffic modeling, ARIMA.

Введение

Рынки информационных услуг и широкополосного доступа в Интернет стремительно растут, растет и конкуренция среди провайдеров. Одним из главных факторов в конкурентной борьбе для провайдера, на сегодняшний день, является сетевая производительность. С ростом количества пользователей в сети увеличиваются объемы передаваемой информации, что приводит к обострению проблемы пиковых нагрузок. Она осложняется тем, что, как правило, пики нагрузок наблюдаются в короткое время, притом, что в целом загрузка серверов провайдера может не достигать и 70%, и наращивать количество серверов явно нецелесообразно. Не является решением и расширение серверных площадок по причине увеличения эксплуатационных расходов. Между тем, даже кратковременные сбои в работе провайдера отрицательно сказываются на имидже компании и лояльности клиентов, особенно в условиях жесткой конкуренции на этом рынке.

Возникновение проблем, связанных с надежностью сети (перегрузке транзитных узлов, потере информации), как правило, в конечном итоге приводят к ухудшению сетевой производительности. Поэтому одной из важнейших задач, относящихся к управлению сетью, является выполнение действий, направленных на устранение блокировок. Избыток трафика, который ввёл сеть в состояние перегрузки, можно

- заблокировать, т.е. удалить соответствующие пакеты из передачи (как правило, данное действие приводит к повторной передаче заблокированных пакетов, что только усугубляет ситуацию перегрузки);
- доставить адресату с худшими показателями качества, например, за большее время или с большей долей потерянных пакетов.

Последнее из упомянутых действий выглядит предпочтительней, если процесс управления доминирующим трафиком, т.е. трафиком приложений, приведшему к перегрузке линии связи, будет осуществляться в динамическом режиме. Таким образом, решению задачи управления могут способствовать дополнительные сведения, которые можно получить из данных прогноза о пропускной способности линии связи на основе прогнозной модели. Тщательно подобранная модель трафика способна выявить и предугадать важнейшие характеристики сетевого трафика, такие как кратковременно и долговременно зависимые процессы, самоподобность на больших временных масштабах.

На сегодняшний день существует множество работ, посвященных разработке прогнозных моделей для сетей данных. Наиболее популярной моделью для прогнозирования являются модели авторегрессии и интегрированного скользящего среднего (ARIMA). Это важный класс параметрических моделей, позволяющий описывать нестационарные ряды.

Целью данной статьи является выявление ARIMA-модели с минимально необходимым порядком параметров, адекватно отражающую поведение сетевого трафика, на основе которой можно совершать достоверные краткосрочные прогнозы.

Модель ARIMA

Характерная запись модели ARIMA(p, d, q) имеет следующий вид:

$$(\Delta^d X_t) = \sum_{i=1}^p \varphi_i (\Delta^d X_{t-i}) + \varepsilon_t + \sum_{j=1}^q \theta_j (\Delta^d \varepsilon_{t-j}), \quad \varepsilon_t \sim N(0, \sigma_\varepsilon^2). \quad (1)$$

Также можно использовать более краткую запись:

$$\varphi(B)(1-B)^d X_t = \theta(B)\varepsilon_t, \quad (2)$$

где $\varphi(\cdot)$, $\theta(\cdot)$ полиномы степени p и q , а B лаговый оператор ($B^j X_t = X_{t-j}$, $B^j \varepsilon_{t-j}$, $j = 0, \pm 1, \dots$), d – порядок взятия последовательной разности ($\Delta X_t = X_{t+1} - X_t = (1-B)X_t$, $\Delta^2 X_t = \Delta X_{t+1} - \Delta X_t = (1-B)^2 X_t, \dots$)

Алгоритм построения модели ARIMA(p, d, q)

Впервые систематический подход к построению модели ARIMA был изложен Боксом и Дженкинсом в 1976 г. Методология построения ARIMA-модели для исследуемого временного ряда включает следующие основные этапы [1]:

- идентификацию пробной модели;
- оценивание параметров модели и диагностическую проверку адекватности модели;
- использование модели для прогнозирования.



Рис. 1. Укрупненная структурная схема подбора модели ARIMA

Таким образом, сначала (в блоке 1 – 3) необходимо получить стационарный ряд. На этом этапе рекомендуется проводить анализ автокорреляционной функции (АКФ) и частной автокорреляционной функции (ЧАКФ). Быстрое затухание значений АКФ – простой тест на стационарность. Также на этом этапе используются статистические тесты на наличие единичного корня (расширенный тест Дики-Фуллера или ADF-тест) [1].

Если в соответствии со статистикой Дики-Фуллера или оценок АКФ ряд является нестационарным, то для перехода к стационарному ряду традиционно применяют оператор взятия последовательных разностей, тем самым определяется значение параметра d (порядка разности). Таким образом, значение одного параметра в модели ARIMA(p, \underline{d}, q) уже известно.

В блоке 4 после получения стационарного ряда исследуется характер поведения выборочных АКФ и ЧАКФ и выдвигаются гипотезы о значениях параметров p (порядок авторегрессии) и q (порядок скользящего среднего). На входе блока 4 может формироваться базовый набор, включающий одну, две или даже большее число моделей, другими словами, портфель моделей.

В блоке 5 после осуществления идентификации модели необходимо оценить их параметры. Для этих целей используется метод максимального правдоподобия (ММП).

В блоке 6 для проверки каждой пробной модели на адекватность анализируется ее ряд остатков. У адекватной модели ряд остатков должен быть похож на белый шум, т.е. их выборочные АКФ не должны отличаться от нуля. Также для проверки гипотезы о том, что наблюдаемые данные являются реализацией «белого шума», используется Q -статистика.

Q -статистика Льюинга-Бокса определяется как

$$Q^* = n(n+2) \sum_{k=1}^m \frac{r_k^2}{n-k}, \quad (3)$$

где n – объем выборки, m – максимальное количество лагов, r_k – коэффициенты автокорреляционной функции.

Если в результате проверки несколько моделей оказываются адекватны исходным данным, то при окончательном выборе следует учесть два фактора:

- повышение точности (качество подгонки модели);
- уменьшение числа параметров модели.

Воедино эти требования сведены в информационные критерии Акайка и Шварца. В данной статье выбран информационный критерий Шварца, в котором усилено требование уменьшения количества параметров модели:

$$SBIK = \ln \left(\frac{\sum_{t=1}^n e_t^2}{n} \right) + \frac{(p+q) \ln(n)}{n}. \quad (4)$$

С помощью модели в блоке 7 можно строить точный и интервальный прогноз на L шагов вперед. Для прогнозирования была выбрана рекурсивная модель, т.е. начальное наблюдение фиксировано, а наблюдение из контрольной выборки добавляется по одному к рабочей. При этом прогнозный горизонт все время остается одинаковым.

Для оценки точности прогноза используется ряд стандартных показателей.

Средняя абсолютная процентная ошибка ($MAPE$):

$$MAPE = \frac{100\%}{L} \sum_{t=1}^L \left| \frac{X_t - \hat{X}_t}{X_t} \right|, \quad (5)$$

где X_t – реальное значение, \hat{X}_t – прогнозное значение, L – интервал прогноза. Если **$MAPE < 10\%$** , то прогноз сделан с высокой точностью, **$10\% < MAPE < 20\%$** – прогноз хороший, **$20\% < MAPE < 50\%$** – прогноз удовлетворительный, **$MAPE > 50\%$** – прогноз плохой.

Отношение сигнала к шуму (SER):

$$SER = 10 \lg \left(\frac{\sum_{t=1}^L X_t^2}{\sum_{t=1}^L (X_t - \hat{X}_t)^2} \right). \quad (6)$$

Методика эксперимента

В соответствие с методикой, описанной в [3], были получены дампы трафика, соответствующие временному промежутку с 17 до 00 часов нескольких суток. Далее дампы приводились в эквидистантный вид с временем агрегации 10 мкс. Эксперимент по выявлению минимального необходимого числа параметров модели для адекватного прогнозирования был поставлен по следующему алгоритму:

- в исходном дискретном временном ряде X_t , соответствующем трафику, выделялся так называемый тренировочный участок с фиксированной длиной, равной 5000 значениям;

- на данном тренировочном участке оценивались параметры упреждающей модели по алгоритму из пункта 2;
- полученные результаты о количестве параметров упреждающей модели на каждом тренировочном участке сводились в таблицу, отражающую частоту появления модели с определенным количеством параметров на всем рассматриваемом временном ряде X_t ;
- из полученной таблицы выбиралась модель наиболее часто встречающаяся;
- на основе полученной модели на случайно выбранном тренировочном участке исходного ряда оценивались параметры прогностической модели по алгоритму из пункта 2, причем формировался прогноз \hat{X}_{t+1} (на 10 шагов вперед) $t+10$ значения ряда X_t , следующего за концом тренировочного участка;
- фиксировалась получившаяся абсолютная ошибка прогноза вплоть до e_{t+10} ;
- сдвигался тренировочный участок на один шаг вперед, считая, что к наступившему моменту времени уже стало известно действительное значение только что спрогнозированного отсчета $t+1$ и т.д.

С помощью вышеуказанного алгоритма были проведены исследования по выявлению модели ARIMA с минимальным числом параметров, позволяющей сделать прогноз. Исследования проводились на основе программы Matlab 7.9. На Рис.2,а представлен IP-трафик с выделенной тренировочной областью, область между отчетами 1 и 5000.

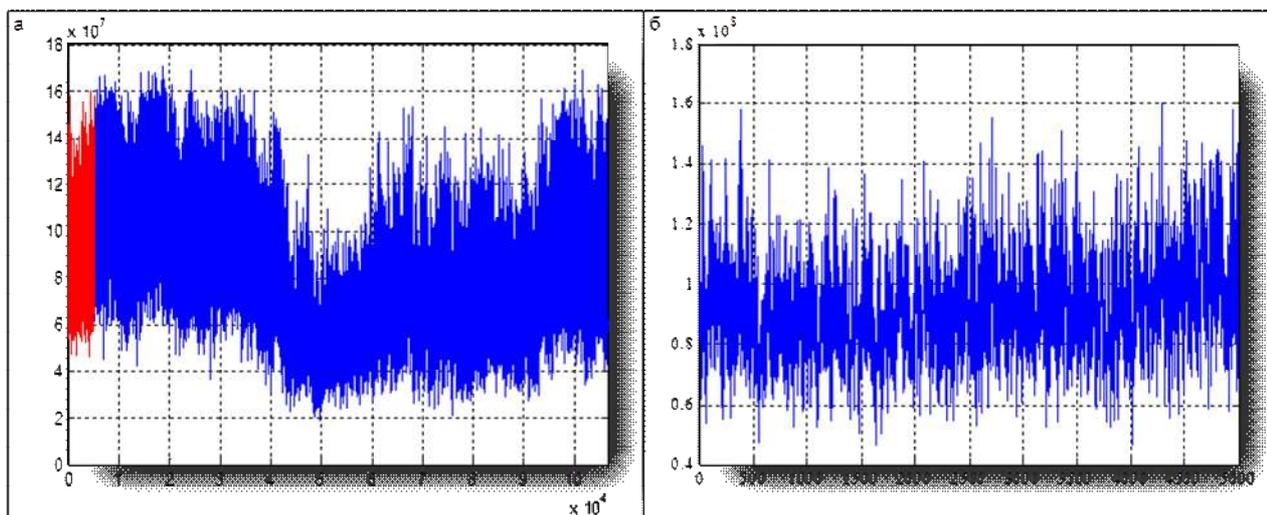


Рис. 2. Временной ряд, соответствующий реальному трафику:

а) исследуемый временной ряд; б) тренировочный участок

В соответствие с алгоритмом из пункта 2 ряд (Рис. 2,б) проверяется на стационарность с помощью статистики Дики-Фуллера. Согласно значениям критерия ADF-теста ряд является нестационарным, поскольку нулевая гипотеза о наличие единичного корня подтверждается на 5% уровне значимости, а фактическое значение t-критерия Стьюдента (-0,0404) находится правее табличного (-0,9416), при этом количество лагов, которое нужно включать в модель при применении ADF-теста, равнялось 60.

Таким образом, ряд подвергается взятию последовательной разности, и тест повторяется. Можно увидеть, что после трансформации среднее значение ряда эквивалентно нулю (Рис. 3).

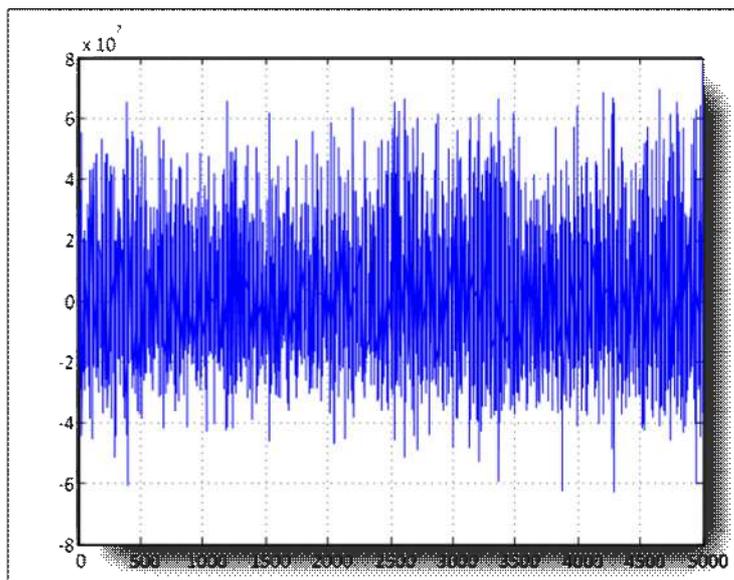


Рис. 3. Временной ряд после взятия последовательной разности

ADF-тест тоже пройден и, поскольку, операция последовательной разности была применена один раз, то значение коэффициента $d = 1$. Получаем модель следующего вида: $ARIMA(p, 1, q)$. Поскольку ряд является стационарным, производим оценку порядка параметров p и q модели $ARMA(p, q)$, которая, в свою очередь, представляет из себя модель $AR(p)$ и $MA(q)$. Для этого воспользуемся ЧАКФ и АКФ, соответственно.

Порядок модели $AR(p)$ выбирается из ЧАКФ и соответствует последнему ненулевому коэффициенту ЧАКФ (Рис. 4, а). Соответствующая процедура применяется для нахождения порядка модели $MA(q)$, только используется АКФ. Таким образом, получив, порядок всех коэффициентов, мы имеем следующую модель $ARIMA(29, 1, 6)$.

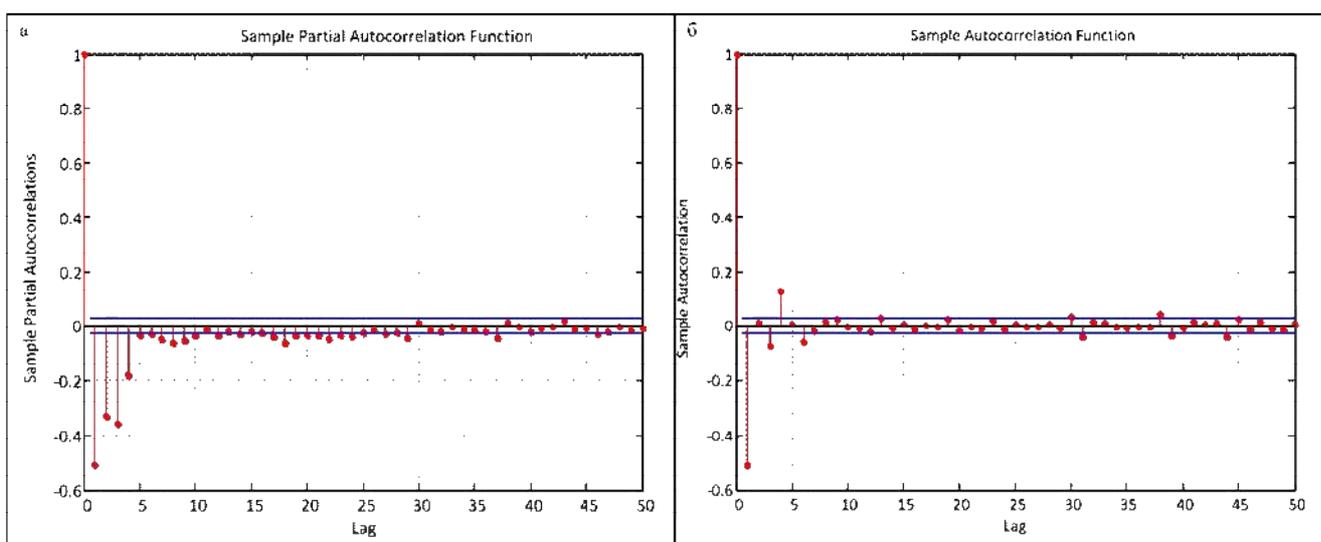


Рис. 4. Функции: а) ЧАКФ; б) АКФ

Каждая полученная модель проходит проверку на адекватность посредством анализа ряда остатков этой модели. Как было сказано выше (блок 6), модель адекватна описываемому процессу, если ряд остатков представляет собой случайную компоненту, соответствующую белому шуму, т.е. АКФ остатков не должна существенно отличаться от нуля.

При проверке значимости коэффициентов АКФ используются два подхода:

- проверка значимости каждого коэффициента автокорреляции отдельно;
- проверка значимости множества коэффициентов автокорреляции как группы.

В соответствии с первым подходом распределение коэффициентов автокорреляции должно приближаться к нормальному распределению с нулевым математическим ожиданием и дисперсией $1/n$.

Поэтому, если выборочный коэффициент автокорреляции выходит за интервал $\pm t_{\alpha}/\sqrt{n}$, то нулевая гипотеза о равенстве этого коэффициента нулю отвергается (Рис. 5, б).

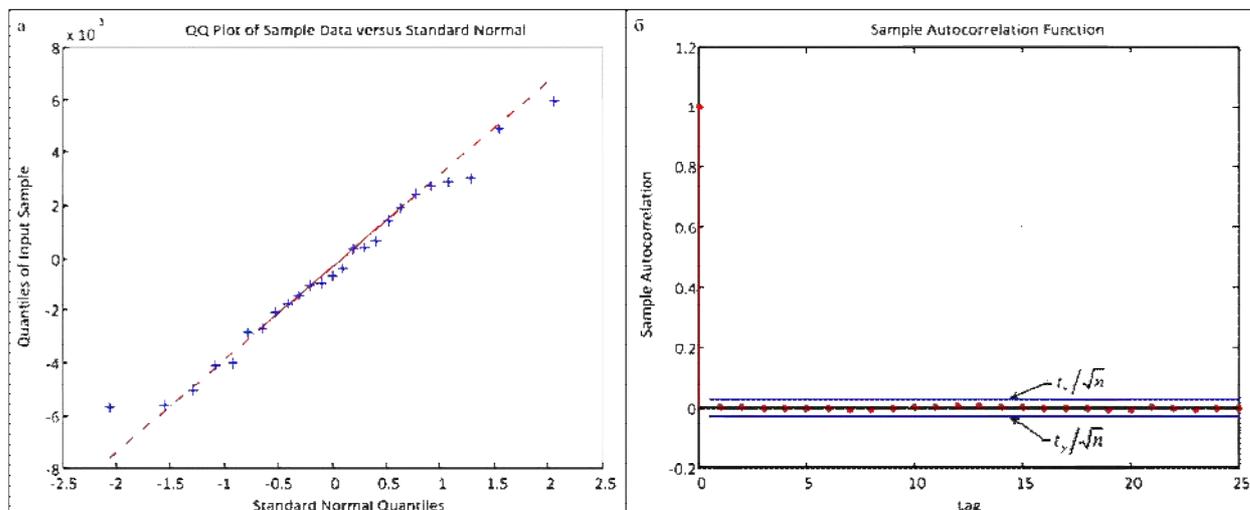


Рис. 5. Графики: а) коэффициенты автокорреляции остатков модели ARIMA(29,1,6) на нормальной вероятностной бумаге; б) функция автокорреляции остатков модели ARIMA(29,1,6)

На Рис. 5, а показаны коэффициенты автокорреляции на нормальной вероятностной бумаге для полученной модели ARIMA(29,1,6). На основе этого графика можно судить о нормальности распределения коэффициентов.

Второй подход опирается на Q -статистику Льюинга-Бокса (3). При нулевой гипотезе об отсутствии автокорреляции статистика Q имеет χ^2 -распределение. Если же $Q > \chi_{кр}^2$, то группа первых коэффициентов автокорреляции значимы, т.е. не все коэффициенты равны нулю. В нашем случае для полученной модели ARIMA(29,1,6) $Q = 1,2309$, а $\chi_{кр}^2 = 37,6525$, что подтверждает нулевую гипотезу о независимости коэффициентов автокорреляции вплоть до 25 лагов.

Проделав соответствующие расчеты, был получен портфель моделей, в который входило больше 100 моделей кандидатов. Затем по информационному критерию SBIC выбиралась наиболее адекватная модель.

Таким образом, для данного тренировочного участка была получена адекватная модель следующего порядка ARIMA(3,1,7). Затем тренировочный участок сдвигался на 2500 значений и процедура, описанная в алгоритме на Рис. 1, повторялась.

Всего было получено 2010 ARIMA-моделей. Коэффициенты p и q полученных моделей были сведены в таблицу 1, которая отражает частоту появления модели на всем рассматриваемом промежутке с соответствующими параметрами, при этом у всех моделей коэффициент $d = 1$ и в таблице не учитывался. В таблице выделены зоны наиболее часто встречающихся моделей на участке антропогенного воздействия. Модели, соответствующие этим зонам, имеют следующие порядки

ARIMA(4,1,3), ARIMA(5,1,1), ARIMA(8,1,1), ARIMA(9,1,1). Отсюда можно предположить, что с помощью выбранных моделей можно делать достоверные прогнозные оценки. Важно отметить, что в связи с тем, что расчет прогноза осуществляется на малых масштабах времени, для экономии времени расчета предпочтительна та модель, порядок коэффициентов которой минимален.

Таблица 1. Частота появления модели ARIMA с соответствующими параметрами p и q

		q											
		1	2	3	4	5	6	7	8	9	10	11	12
p	1	0	24	9	23	23	29	47	30	12	13	9	6
	2	0	8	3	7	16	13	14	14	4	3	8	7
	3	1	3	15	17	48	84	21	7	9	3	6	4
	4	44	16	103	57	43	22	11	5	4	0	1	3
	5	99	25	52	45	36	15	4	2	2	1	1	1
	6	78	6	11	17	24	10	1	2	0	0	0	0
	7	86	9	8	9	10	5	2	0	2	0	0	0
	8	117	13	7	6	4	4	2	0	2	0	0	0
	9	207	10	13	6	6	0	1	3	0	0	0	0
	10	76	1	7	3	3	2	0	0	10	0	0	0
	1	32	2	4	9	9	2	0	0	2	0	0	0
	2	22	2	4	1	3	1	1	1	0	0	0	0

Как было сказано выше, для оценки точности прогноза использовались показатели 5 и 6. Нужно отметить, что в прогнозном сценарии использовалось несколько направлений для проверки точности прогноза полученных моделей. В первом случае сравнивался прогноз на один шаг вперед, во втором – на 2 шага вперед, в третьем – на 3 шага вперед и в четвертом – на 10 шагов вперед.

В [4] показано, что адекватный прогноз значений сетевого трафика можно осуществить с помощью моделей: ARIMA(0,1,1) на 2 шага вперед, ARIMA(1,1,0) – на 6 шагов вперед, ARIMA(1,1,1) на 50 шагов вперед. Поэтому целесообразно включить в эксперимент по прогнозированию данные модели. Следует отметить, что для чистоты эксперимента случайным образом выбирались тренировочные участки, производился прогноз в соответствие с алгоритмом на рис. 1, а затем для полученной модели и для моделей ARIMA(0,1,1), ARIMA(1,1,0), ARIMA(1,1,1), ARIMA(4,1,3), ARIMA(5,1,1), ARIMA(8,1,1), ARIMA(9,1,1), высчитывались значения MAPE и SER по формулам 5 и 6, соответственно, для прогноза на 1, 2, 3 и 10 шагов вперед. Всего было выбрано 230 случайных участков временного ряда, на каждом были получены значения MAPE и SER для вышеупомянутых моделей. Затем вычислялось среднее значение данных показателей для разных интервалов прогноза.

На рис. 6 показан пример прогноза значений трафика на 10 шагов вперед для одного из участка временного ряда. Модель ARIMA(Rial) соответствует выбранной модели на данном участке по критерию BIC, а REAL – это фактические значения данных трафика. Из рисунка видно, что абсолютная ошибка, т.е. разница между фактическим значением и прогнозным значением, минимальна для модели ARIMA(4,1,3) до двух шагов.

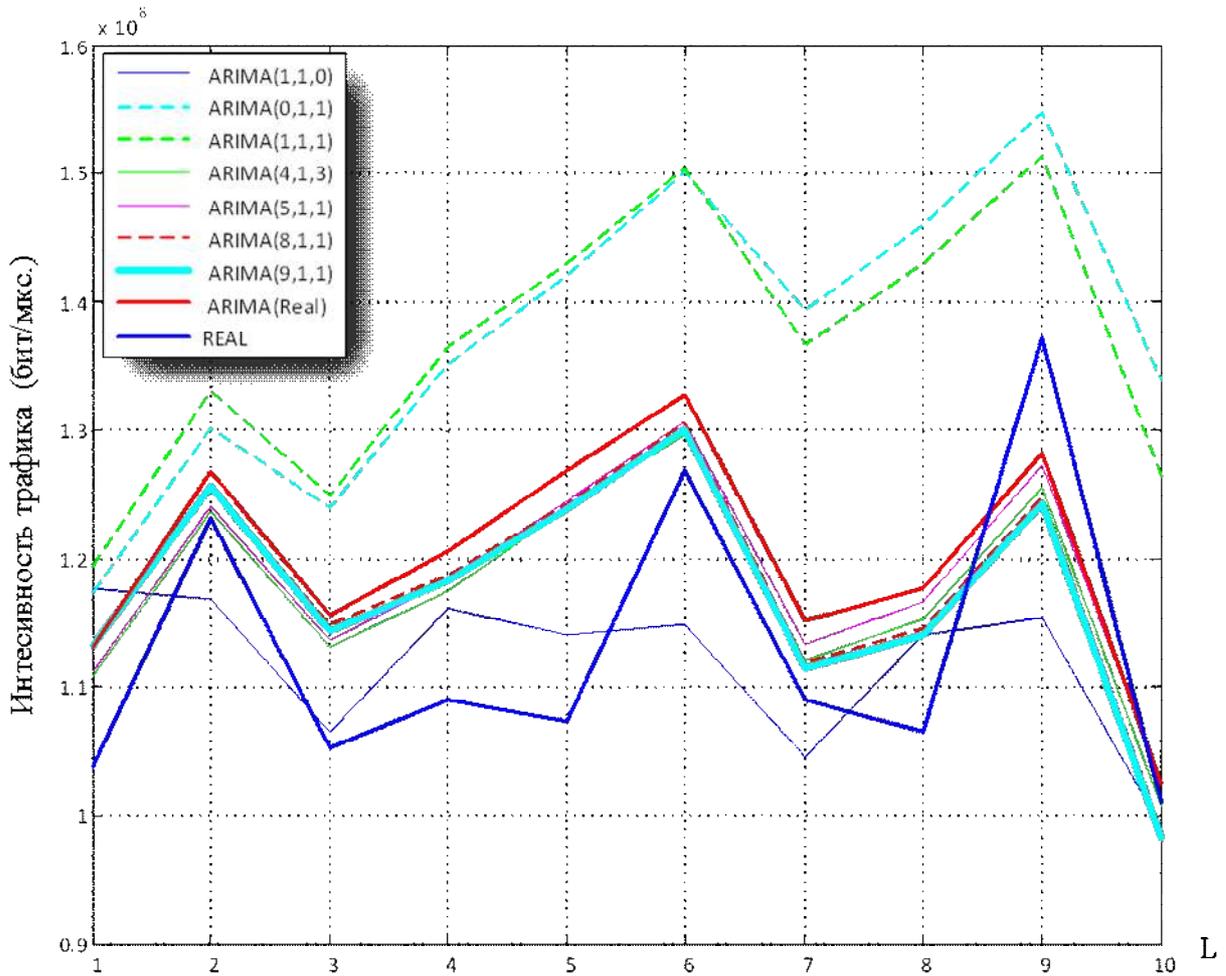


Рис. 6. Прогноз значений трафика для одного из случайных участков временного ряда

Анализ полученных данных

Далее приведены результаты усреднения оценок, характеризующие качество прогнозных оценок MAPE и SER.

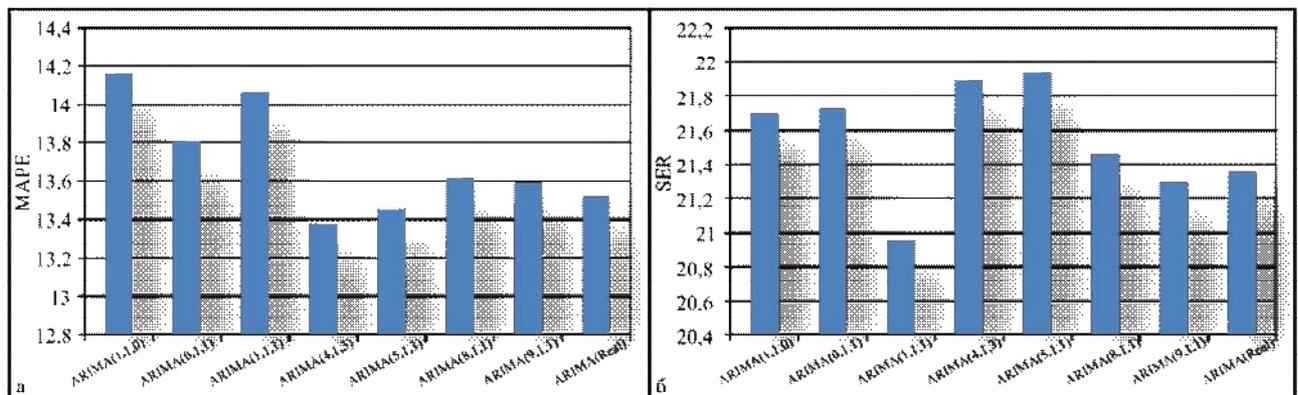


Рис. 7. Среднее значение коэффициентов MAPE и SER для прогноза на 1 шаг вперед

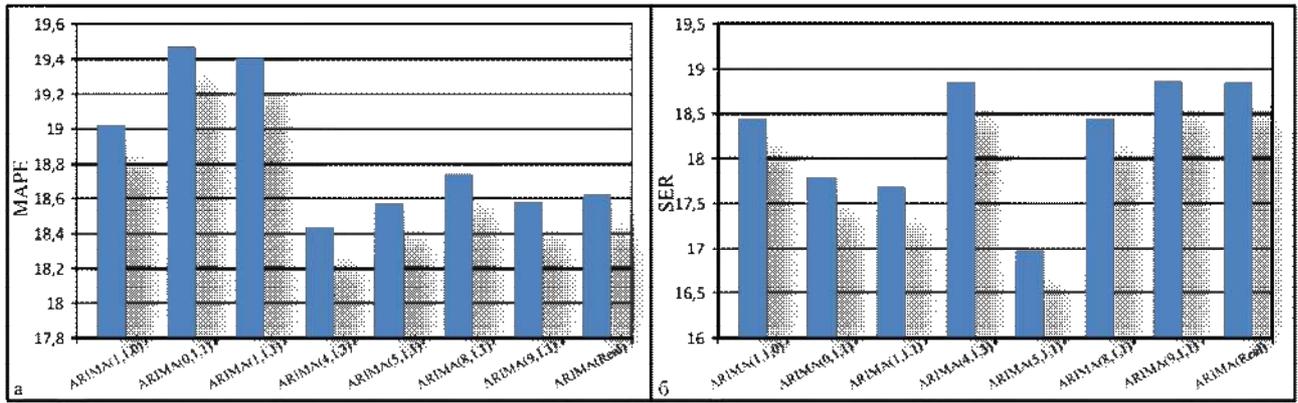


Рис. 8. Среднее значение коэффициентов MAPE и SER для прогноза на 2 шага вперед

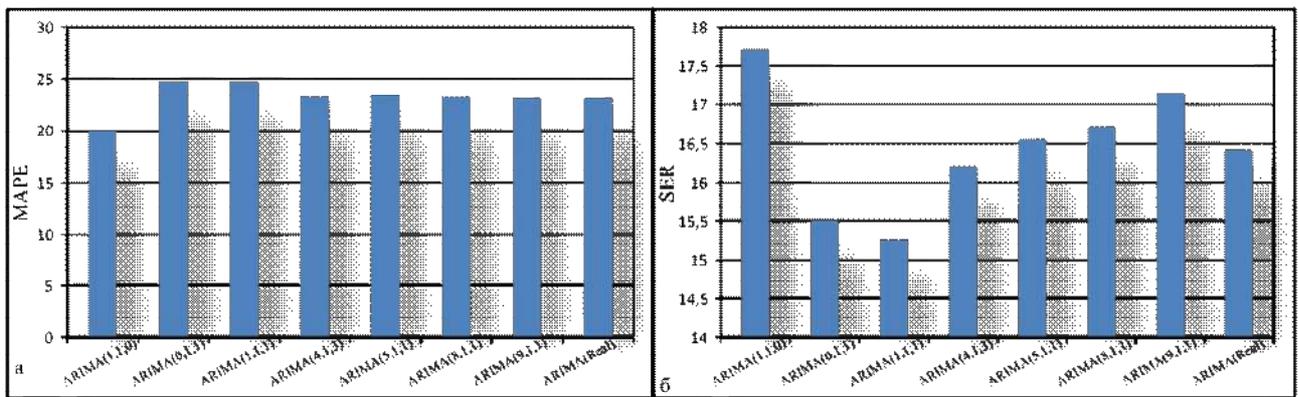


Рис. 9. Среднее значение коэффициентов MAPE и SER для прогноза на 3 шага вперед

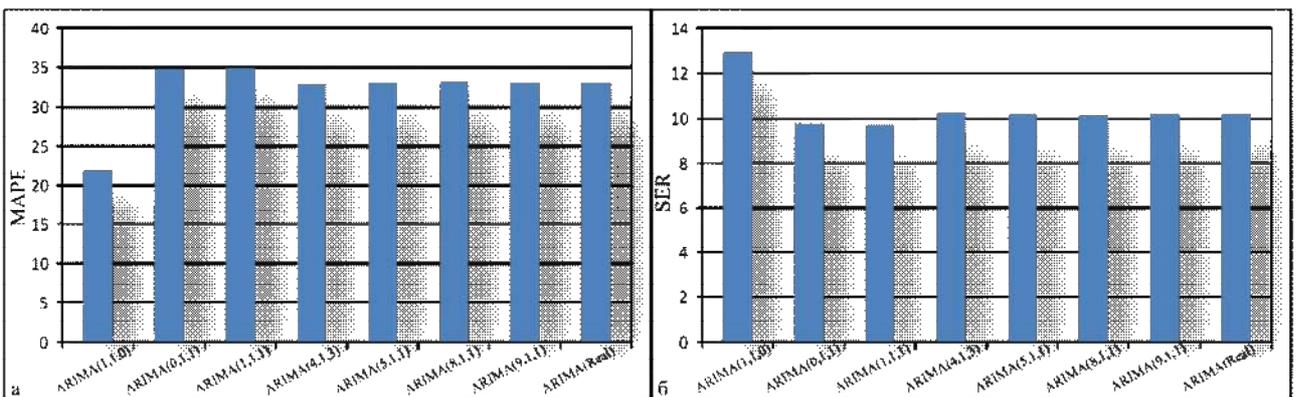


Рис. 10. Среднее значение коэффициентов MAPE и SER для прогноза на 10 шагов вперед

Из рисунков 7 и 8 видно, что оценка MAPE модели ARIMA(3,1,4) минимальна и находится между 10% и 20%, что соответствует хорошему прогнозу. Поэтому данная модель лучше всего подходит для прогнозирования значений трафика на 2 шага вперед. Далее можно заметить, что первенство перенимает модель ARIMA(1,1,0).

Заключение

По мнению современных экспертов, нагрузка на современные сети передачи данных ежегодно увеличивается на 30 – 40%. Современные сети – сети с широким набором всевозможных сервисов и услуг. Часть из них является сервисами времязависимого трафика (сервисы IP-телефонии, потокового видео, различных видов диспетчеризации оборудования), другая часть, например, файлообменные сети FTP-серверов не критичны к текущей пропускной способности. В условиях постоянно растущего объема передаваемого сетевого трафика на одном из сегментов сети может образоваться так называемого «узкое горлышко». Возникновение такого рода ситуаций приводит к увеличению времени ожидания пользователей, увеличению нагрузок на линии передачи данных, а так же к лавинообразному увеличению служебного трафика. Эти факторы негативно сказываются на общей работоспособности сети. В качестве решения таких проблем особого внимания заслуживают сервисы распределенного хранения данных P2P. Повсеместное распространение этой технологии приведет к исчезновению проблемы «узкого горла», как элемента системы, ограничивающего общий объем транслируемых данных. Развитие сетей P2P может вносить дисбаланс в существующую сетевую инфраструктуру. Построение систем on-line-мониторинга сетей P2P с возможностью динамического управления выделенной пропускной способностью для определенных видов сервиса – важная перспективная задача.

В данной статье рассмотрен алгоритм прогнозной части такой системы, которая позволит своевременно предупреждать о значительных выбросах сетевого трафика. И как важный компонент такой части, найдена модель ARIMA с минимальным числом параметров, позволяющая совершать достоверные краткосрочные прогнозы. Также показано, что прогноз более 2 шагов вперед хорошо осуществляется с помощью модели ARIMA(1,1,0).

Список литературы

1. Дуброва, Т. А. Статистические методы прогнозирования. – М.: ЮНИТИ-ДАНА, 2003. – С. 206.
2. Канторович Г.Г. Анализ временных рядов // Экономический журнал ВШЭ. – №3. – 2002.
3. Крюков, Ю. А. Исследование самоподобия трафика высокоскоростного канала передачи пакетных данных // sanse.ru: сайт электронного журнала. [Электронный ресурс]. – Режим доступа: <http://sanse.ru/download/33> (дата обращения 09.10.2010).
4. Rutka G. Network Traffic Prediction using ARIMA and Neural Networks Models // Electronics And Electrical Engineering. – №4. – 2008.