ARIMA – модель прогнозирования значений трафика

Ю.А. Крюков, Д.В. Чернягин

Аннотация. 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 предсказать важнейшие характеристики сетевого трафика: кратковременно и долговременно зависимые процессы, самоподобность на больших временных масштабах. В данной статье представлена модель ARIMA с минимальным числом параметров, имеющая адекватный прогноз. Предложена процедура оценки параметров модели ARIMA и выбора модели с минимальным числом параметров. Приведены сравнения оценок качества прогноза для полученных моделей.

Ключевые слова: Сетевой трафик, прогнозирование сетевого трафика, модель трафика, ARIMA.

Введение

Производительность компьютерной сети является наиболее значимым параметром как для провайдера услуг связи, так и для пользователей корпоративных сетей. Большинство крупных операторов связи ежегодно констатируют нелинейное увеличение объемов передаваемых данных в сравнении с ростом абонентской базы. Доминирующую роль в формировании роста объемов трафика играют современные файлообменные протоколы, формирующие дополнительный градиент пульсаций мгновенной пропускной способности линий связи. Неравномерность распределения объемов передаваемого трафика во времени остается главной проблемой в сетях с коммутацией пакетов.

Наряду с этим наблюдается рост количества конвергентных сетей, предполагающих трансляцию трафика времязависимых сервисов, таких как IPTV и VoIP. Интеграция сервисов в рамках одной сети передачи данных имеет множество преимуществ как со стороны пользователей, так и со стороны провайдера. Однако непрерывный рост объемов трафика файлообменных сервисов может приводить к снижению качества трансляции изображения и звука. Несмотря на значительные инвестиции в оборудование и коммуникации со стороны провайдера, повышение общей пропускной способности сети, внедрение процедур QoS (поддержки приоритетов для различных типов трафика) далеко не всегда обеспечивают необходимый запас производительности для качественной работы критичных приложений.

Одной из приоритетных задач сетевых компаний становится поддержка сбалансированной нагрузки на сегменты сети. Возникновение перегрузок в одном из транзитных узлов может приводить к переполнению внутренних буферов на коммутационном оборудовании, потере части пакетов и, в конечном итоге, к снижению производительности сети в целом. Обеспечение мер, направленных на устранение перегрузок на сетевом оборудовании за счет процедур управления трафиком сервисов, не критичных к потерям пакетов, позволит значительно сократить расходы на обслуживание и модернизацию существующих сетевых сегментов.

Процедуры управления сетевым трафиком производят динамическую конфигурацию коммутационного оборудования с помощью инст-

рументария различных фильтров и политик и обеспечивают временное ограничение интенсивности трафика определенных сервисов в течение периодов максимальной нагрузки. Избыток трафика, который может ввести линию связи в состояние перегрузки, можно:

- заблокировать, т.е. удалить соответствующие пакеты из сети передачи (как правило, данное действие приводит к повторной передаче источником потерянных пакетов, что только усугубляет ситуацию перегрузки в последующие временные интервалы);
- временно запретить источнику генерировать пакеты определенного типа (большинство существующих сервисов не предусматривают внешнего динамического управления);
- доставить пакеты 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 будет осуществляться в динамическом режиме и ограничивать поток данных лишь в кратковременные периоды, минимизируя эффект «снежного кома». Учитывая временные интервалы, необходимые для осуществления управляющего воздействия, решения о необходимости такого воздействия должны приниматься с упреждением.

Таким образом, решению задачи управления могут способствовать дополнительные сведения, которые можно получить из данных прогноза о загрузке линии связи на основе прогнозной модели. Тщательно подобранная модель трафика способна учесть 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 прогнозирования являются модели авторегрессии и проинтегрированного скользящего среднего (ARIMA). Это - важный класс параметрических моделей позволяет описывать нестационарные ряды.

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

Модель ARIMA

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

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

Можно использовать и более краткую запись:

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

где $\varphi(\bullet)$, $\theta(\bullet)$ полиномы степени p и q, B - лаговый оператор $\left(B^j X_t=X_{t-j}, B^j\varepsilon_{t-j}, j=0,\pm1,...\right)$, d порядок взятия последовательной разности $\left(\Delta X_t=X_{t-1}-X_t=(1-B)X_t, \Delta^2 X_t=\right.$
$\left.=\Delta^2 X_{t+1}-\Delta X_t=(1-B)^2 X_t,...\right)$.

Алгоритм построения модели ARIMA (p, d, q)

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

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

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. На этом этапе используются также 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

Рис. 1. Укрупненная структурная схема подбора модели ARIMA

ряду традиционно применяют оператор взятия последовательных разностей, тем самым определяется значение параметра d (порядка разности). Таким образом, значение одного параметра в модели ARIMA(p,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. Для проверки гипотезы о том, что наблюдаемые данные являются реализацией «белого шума», используется также Q-статистика.

Q-статистика Льюинга-Бокса определяется как

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

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:

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

С помощью модели в блоке 7 можно строить точный и интервальный прогноз на L шагов вперед. Для прогнозирования была выбрана рекурсивная модель, в которой 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):

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

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):

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

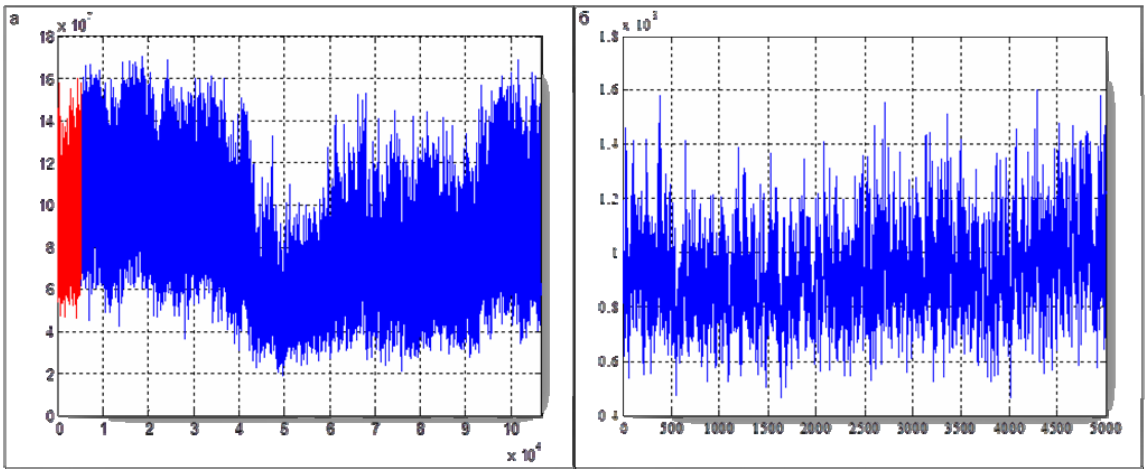

Рис. 2 - Временной ряд, соответствующий реальному трафику
а) – исследуемый временной ряд; б) – тренировочный участок

Методика эксперимента

В соответствии 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 длинной, равной 5000 значениям;
- на данном тренировочном участке оценивались параметры упреждающей модели по формуле 6;
- 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 наиболее часто встречающаяся модель;
- 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 полученная абсолютная ошибка прогноза вплоть до e_{t+1};
- тренировочный участок сдвигался на один шаг вперед в предположении, что к наступившему моменту времени уже стало известно действительное значение только что спрогнозированного отсчета t+1, и т.д.

С помощью приведенного 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.

В соответствии с формулой (6) ряд (Рис.2,б) проверяется на стационарность с помощью статистики Дики-Фуллера. Согласно значениям критерия ADF-теста ряд является нестационарным, поскольку нулевая гипотеза о наличии единичного корня подтверждается на 5% уровне значимости, а фактическое значение t-критерия Стьюдента (-0,0404) находится правее табличного (-1,9416), 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

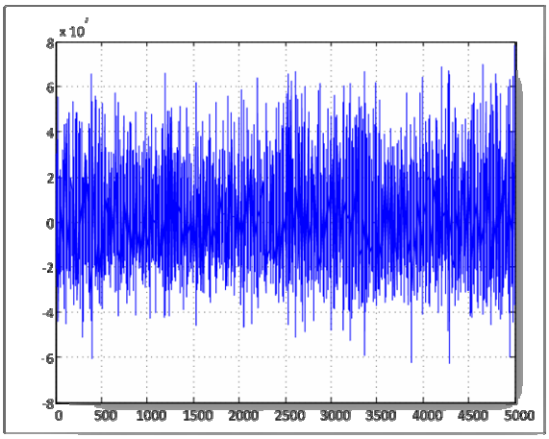

Рис. 3. Временной ряд после взятия последовательной разности

среднее значение ряда эквивалентно нулю (Рис.3). ADF-тест тоже пройден, и поскольку операция последовательной разности была применена один раз, то значение коэффициента $d = 1$. Получаем модель 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 собой 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), при этом используется АКФ (Рис.4,б). Таким образом, получив порядок всех коэффициентов, мы имеем следующую модель ARIMA(29,1,6).

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 $\pm t_\gamma / \sqrt{n}$, то нулевая гипотеза о равенстве этого коэффициента нулю отвергается (Рис.5,б).

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.

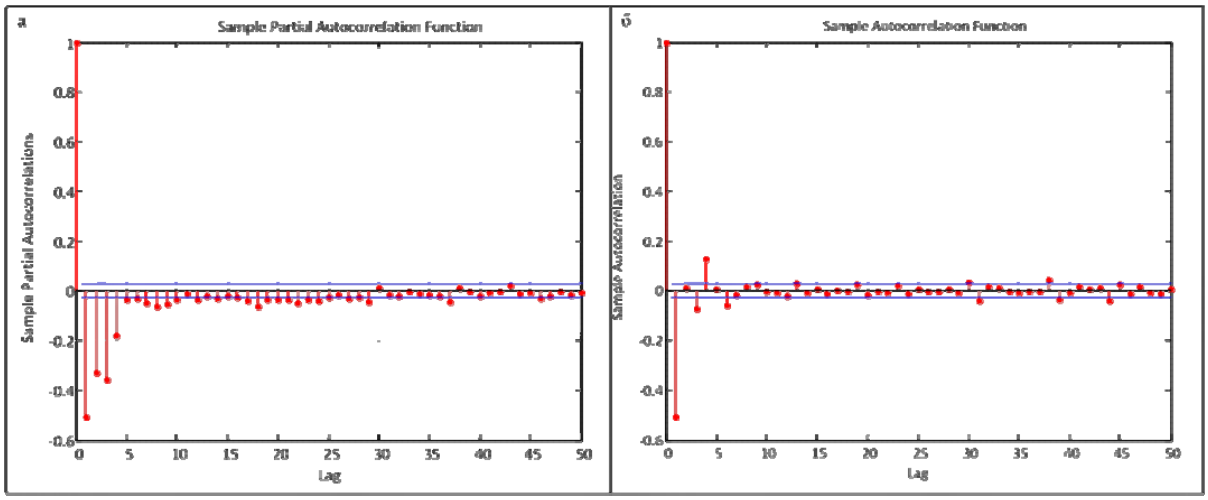

Рис.4. Функции: а) – ЧАКФ; б) - АКФ

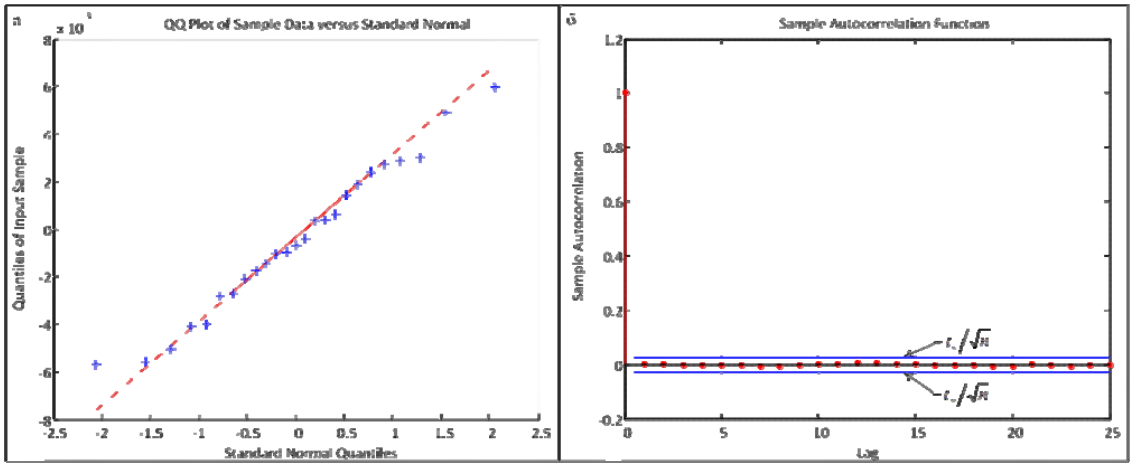

Рис. 3. Графики: а) – коэффициенты автокорреляции остатков модели ARIMA(29,1,6) на нормальной вероятностной бумаге; б) – функция автокорреляции остатков модели ARIMA(29,1,6).

Второй подход опирается на Q-статистику Льюинга-Бокса (3). При нулевой гипотезе об отсутствии автокорреляции статистика Q имеет χ^2-распределение. Если же $Q > \chi^2_{кр}$, 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 $\chi^2_{кр} = 37.6525$, что подтверждает нулевую гипотезу о независимости коэффициентов автокорреляции вплоть до 25 лагов.

В результате проведенных расчетов 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, 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, выбранные модели позволяют делать достоверные прогнозные оценки поскольку расчет прогноза осуществляется на малых масштабах времени, для экономии времени расчета предпочтительна та модель, порядок коэффициентов которой минимален.

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
	11	32	2	4	9	9	2	0	0	2	0	0	0
	12	22	2	4	1	3	1	1	1	0	0	0	0

Для оценки точности прогноза использовались показатели 5 и 6. Нужно отметить, что

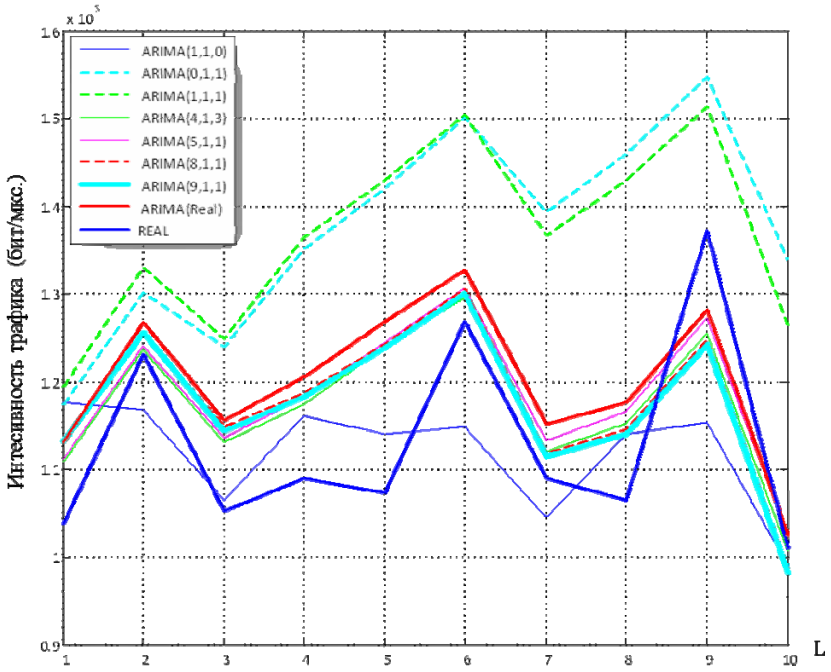

Рис. 6. Прогноз значений трафика для одного из случайных участков временного ряда

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 соответствии 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 перечисленных 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.

Анализ полученных данных

Далее на Рис.7–Рис.10 приведены результаты усредненных оценок, характеризующие качество прогнозных оценок MAPE и SER.

Из Рис.7 и Рис.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).

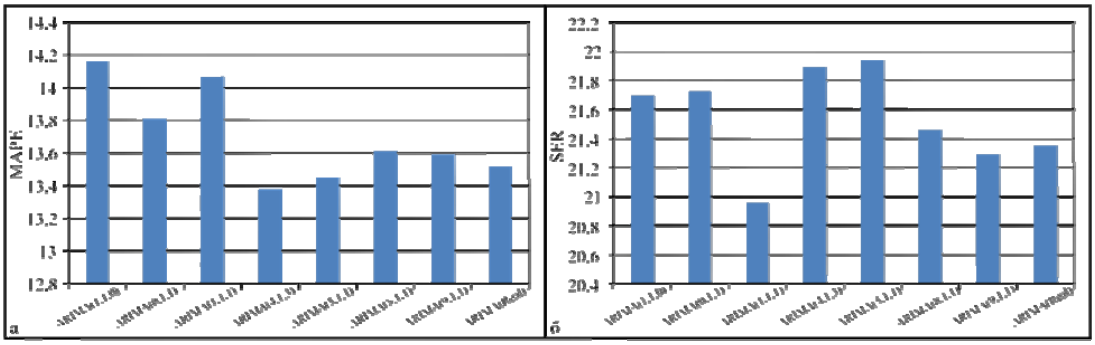

Рис. 7. Среднее значение коэффициентов MAPE и SER для прогноза на 1 шаг вперед

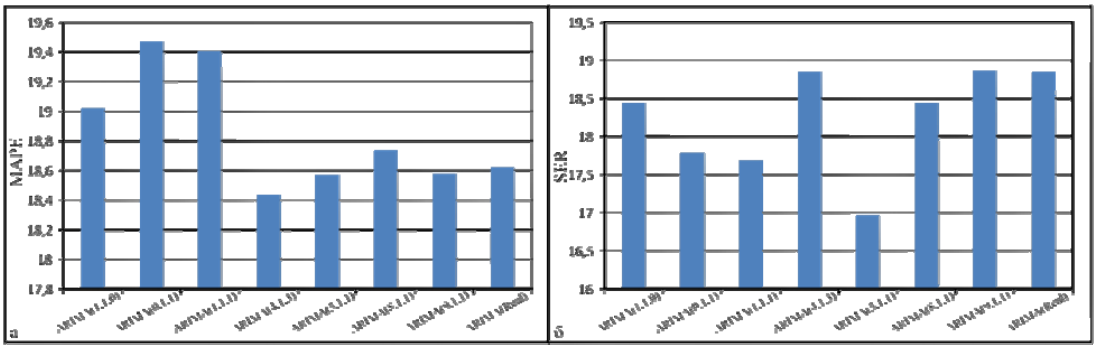

Рис. 8. Среднее значение коэффициентов MAPE и SER для прогноза на 2 шага вперед

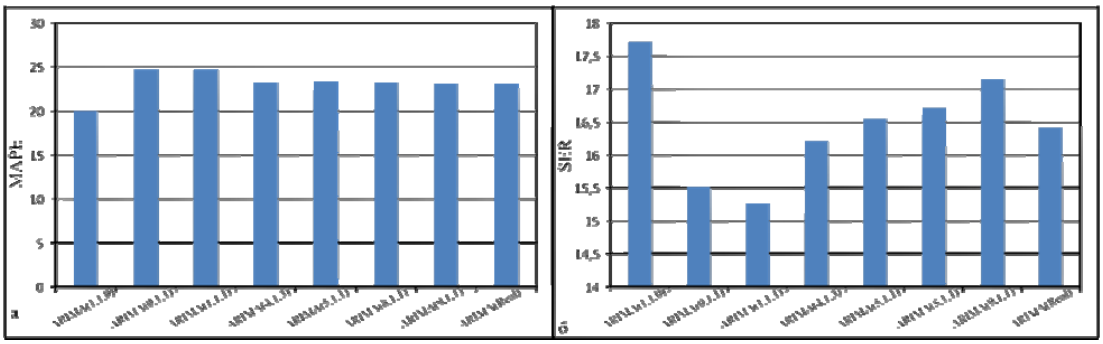

Рис. 9. Среднее значение коэффициентов MAPE и SER для прогноза на 3 шага вперед

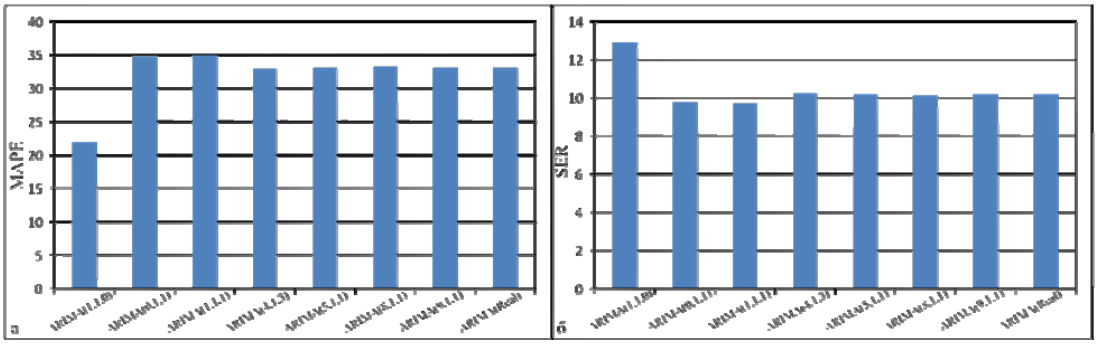

Рис. 10. Среднее значение коэффициентов MAPE и SER для прогноза на 10 шагов вперед

Заключение

Построение систем on-line мониторинга с возможностью динамического управления выделенной пропускной способностью линии связи для определенных видов сервиса — важная перспективная задача. Широкое распространение сервисов распределенного хранения данных P2P, отсутствие «узкого горла» как элемента системы, ограничивающего общий объем транслирующихся данных, позволяют рассматривать дальнейшее развитие сетей P2P как фактор, вносящий дисбаланс в существующую сетевую инфраструктуру. Внедрение систем динамического управления сетевым трафиком способно сбалансировать нагрузку, возникающую в различные временные периоды, и исключить возникновение перегрузок в сегментах сети без существенного влияния на качество работы распределенных файлообменных сервисов.

В данной работе рассмотрен алгоритм прогнозной части системы динамического управления трафиком, которая позволит своевременно предупреждать о последующих значительных выбросах сетевого трафика. Как важный ее компонент, найдена модель ARIMA с минимальным числом параметров, позволяющая совершать достоверные краткосрочные прогнозы. Показано, что прогноз более 2 шагов вперед хорошо осуществляется с помощью модели ARIMA(1,1,0). Использование прогностических моделей с шагом 10 мс хорошо согласуется с временем управляющего воздействия системы на коммутационное оборудование, что наряду с высокой вероятностью прогноза на коротких временных периодах обеспечивает адекватное управление мгновенной пропускной способностью сегмента сети.

Литература

1. Дуброва, Т. А. Статистические методы прогнозирования. [Текст] / Т. А. Дуброва М.: ЮНИТИ-ДАНА, 2003. 206 с.
2. Канторович Г.Г. Анализ временных рядов [Текст] / Г. Г. Канторович. М. : Экономический журнал ВШЭ, №3, 2002.
3. Крюков, Ю. А. Исследование самоподобия трафика высокоскоростного канала передачи пакетных данных / Ю. А. Крюков, Д. В. Чернягин // 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 [Text] / Rutka G. – Electronics And Electrical Engineering, №4, 2008.

Чернягин Денис Викторович. Ассистент ГОУ ВПО «Международный Университет природы, общества и человека «Дубна», Институт системного анализа и управления. E-mail: dancher2000@mail.ru

Крюков Юрий Алексеевич. Доцент ГОУ ВПО «Международный Университет природы, общества и человека «Дубна», Институт системного анализа и управления. Кандидат технических наук. E-mail: kua@uni-dubna.ru