Б. Я. Советов, Т. М. Татарникова, В. В. Цехановский
Санкт-Петербургский государственный электротехнический университет
«ЛЭТИ» им. В.И. Ульянова (Ленина), Санкт-Петербург

АВТОРЕГРЕССИОННЫЕ МОДЕЛИ ПРОГНОЗИРОВАНИЯ СЕТЕВОГО ТРАФИКА

Актуальность прогнозирования сетевого трафика обусловлена требованием хранения большого количества данных в течение продолжительного времени. В связи с этим возникает необходимость прогнозирования объемов трафика с целью принятия необходимых мер по защите и сохранению данных. В работе обосновывается применение моделей авторегрессионного класса для построения прогнозных оценок. Приведены результаты, позволяющие обосновать перспективные требования к объемам памяти узлового оборудования инфокоммуникационной сети.

Введение. Развитие инфокоммуникационных технологий с каждым десятилетием преодолевает новую планку ограничений пропускной способности и приводит к увеличению количества пользователей сети Интернет [1].

Простота в использовании и доступность гаджетов привносят новые медиа ресурсы, вследствие чего, интернет трафик становиться разнородным [2]. В начале XXI века обнаружены новые особенности сетевого трафика, такие как самоподобность и «пакетирование», которые влияют на нагрузку сетевых узлов.

Основным свойством самоподобия является накопление памяти – сильная зависимость от предыдущих значений. Интернет-трафик при сглаживании имеет определенную структуру с трендом, на которую стохастически влияют редкие «всплески» пакетов – пакетирование. Такого рода всплески влияют на математические моменты временной последовательности как в локальных масштабах времени, так и на больших размахах. Пакетирование может являться причиной потери данных. В связи с этим возникает необходимость прогнозирования явления пакетирования с целью принятия необходимых мер по защите и сохранению данных.

Актуальность проблемы обусловлена законодательным требованием хранения большого количества данных в течение продолжительного времени.

Представленные графики прогноза роста трафика опубликованы на официальном сайте Cisco с обновлениями от 7 февраля 2017 г. [3]. Данные имеют отношение к глобальному трафику и датируются 2016 г.

Согласно исследованию Cisco, в 2016 г. объем мобильного трафика вырос на 63 % по сравнению с 2015 г. К концу 2016 г он достиг 7,2 ЭБ в месяц при объеме в 4,4 ЭБ к концу 2015 г. По доле в общем потоке IP-трафика и по общему приросту интернет-трафика по-прежнему будет доминировать видео: 80 % всего интернет-трафика к 2021 г., тогда как в 2016 г. этот показатель составлял 67 %. К 2021 г. в мире будет около 1,9 млрд пользователей интернет-видео не считая тех, кто пользуется исключительно мобильной связью, тогда как в 2016 г. таких было 1,4 млрд. К 2021 г. через мировой интернет в месяц будет передаваться 3 трлн минут видео. Установлено, что наибольшее статистическое самоподобие имеет канал с передачей видео. Ввиду этого будет возрастать востребование к прогнозу такого канала передачи информации.

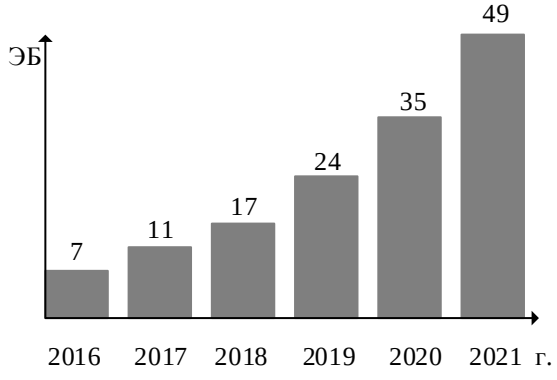

Рис. 1. Прогноз роста трафика от Cisco

Обзор существующих решений. Модели прогнозирования разделяют на два вида: статистические и структурные. Для статистических моделей функциональная зависимость задается аналитически. К таким моделям относятся: регрессионные, авторегрессионные и экспоненциального сглаживания. В свою очередь структурные модели основаны на зависимости структур. К ним относятся: нейросетевые модели, модели на базе цепей Маркова и деревья принятия решений. В таблице ниже приведены достоинства и недостатки выше перечисленных методов.

Таблица

Модели и методы прогнозирования

Модели и методы	Достоинства	Недостатки
Регрессионные	Простота в моделировании и проектировании	Сложность нахождения оптимальных коэффициентов и функциональной зависимости
Авторегрессионные	Простота в моделировании и анализа	Нельзя моделировать нелинейные процессы
Экспоненциальное сглаживание	Простота моделирования	Узкая применимость моделей
Нейросетевые	Большое разнообразие архитектур; нелинейность	Сложность в выборе архитектур; размер обучающей выборки; большие временные и ресурсоемкие затраты на обучение
Цепи Маркова	Единообразие проектирования	Узкая применимость моделей; нельзя моделировать долгосрочные процессы
Классификационно-регрессионные деревья	Простота обучения модели; возможность масштабирования	Сложность построения алгоритма дерева

В работе [4] анализируются статистические и структурные модели прогнозирования, где делается вывод о применимости статистических моделей для краткосрочного прогноза и структурных – для среднесрочного и долгосрочного прогноза. В рамках рассматриваемой задачи прогнозирования трафика, с целью выделить достаточный ресурс для хранения трафика в узловом оборудовании на краткосрочном периоде времени, авторегрессионная модель имеет ряд преимуществ перед структурными (обучаемыми) моделями. В условиях малого количества данных и отсутствия времени на обучение, данная модель способна прогнозировать с меньшей ошибкой на коротких промежутках времени.

Прогнозирование реальных данных. Для получения среднесрочных и долгосрочных прогнозов необходимы длинные записи трасс трафика [5]. Это, казалось невыполнимое условие, было выполнено благодаря открытой публикации данных интернет-трафика исследовательской группы Японии MAWI проекта WIDE. В тренировочном наборе задействованы 24-часовые трассы за 05.09.2018 г. и 04.09.2019 г., 48-часовые трассы за 01.09-11.2007 г., 72-часовые трассы за 18-20.03.2008 г., 96-часовые трассы за 30.03.2009 г. и 01-04.02.2009 г.

К полученным данным реального сетевого трафика будут применяться модели из авторегрессионного класса: ARIMA – интегрированная модель авторегрессии и скользящего среднего, а также SARIMA – модификация ARIMA с выделением сезонности. Для определения оптимальных параметров моделей, был проведен полный перебор по возможным значениям. Выявлены наилучшие значения параметров для модели ARIMA (p, d, q):

p – порядок отставания = 3;
d – степень разности = 1;
q – порядок скользящей средней = 2.

Для модели SARIMA (p, d, q), (P, D, Q), m и управление детерминированным трендом:

p – порядок авторегрессии тренда = 2;
d – порядок изменения тренда = 1;
q – тренд скользящей средней = 2;
P – сезонный порядок авторегрессии = 1;
D – порядок сезонных разниц = 0;
Q – сезонный порядок скользящих средних = 2;
m – количество временных шагов за один сезонный период = 0.

Для обеих моделей выбрана одна дата для качественной оценки сравнения значений предсказания. Сравнение моделей будет проводиться на данных за (двое суток) 09-10.04.2019 г.

Разделение данных на тренировочные и тестовые приводятся к соотношению 3 к 1. Исходные данные нормализованы, то есть вся числовая последовательность поделена на максимальные значения из ряда. Основным критерием отклонения от исходных данных выбран – MSE (Mean Squared Error) – среднеквадратическая ошибка:

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2, \qquad (1)$$

где Y_i – значение оригинала, $\hat{Y}_i$ – значение предсказания.

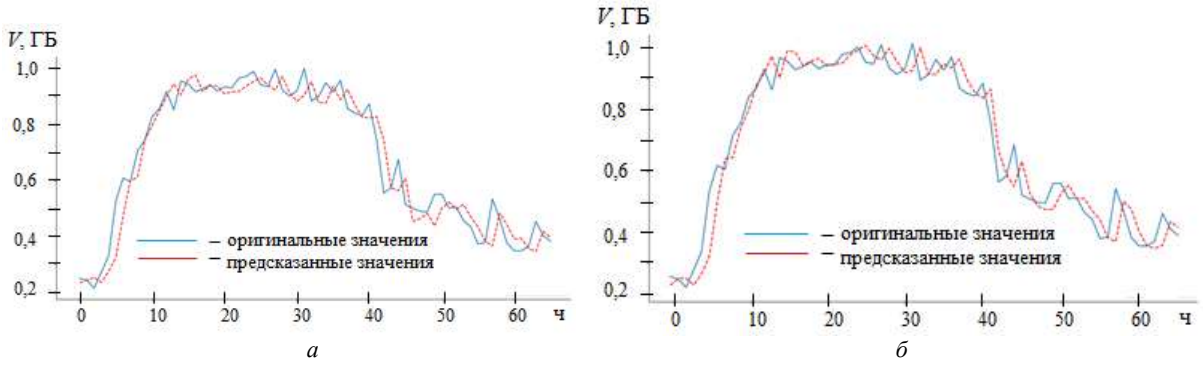

Рис. 2. Сравнение объемов оригинального трафика с предсказанными по модели:
а – ARIMA, MSE=0,004; *б* – SARIMA, MSE=0,006

Из представленных выше результатов предсказания, можно сделать вывод, что модели повторяют геометрию с небольшим отклонением. Предсказания для модели SARIMA можно улучшить путем увеличения исходных данных с визуальной сезонностью.

Модели авторегрессионного класса вполне подходят для интеграции в сетевые устройства как аналитическая программа, работающая на агрегационных данных, что позволит в реальном времени предсказывать поведение объема трафика на коротком промежутке времени.

Заключение. Рост объемов разнородного трафика в инфокоммуникационных сетях актуализирует вопросы обеспечения качества предоставляемых услуг связи, что в свою очередь требует обращения к моделям прогнозирования. Результаты, полученные в работе, позволяют обосновать перспективные требования к объемам памяти узлового оборудования инфокоммуникационной сети.

ЛИТЕРАТУРА

1. Индикаторы информационного общества: 2014. Статистический сборник. М.: НИИ «ВШЭ», 2015. 320 с.
2. **Кузовкова Т.А.** Оценка роли инфокоммуникаций в национальной экономике и выявление закономерностей ее развития. // *Системы управления, связи и безопасности*. 2015 № 4. С. 26-68.
3. Электрон. текстовые дан. www.cisco.com/URL:https: //www.cisco.com/c/ru_ru/ about/press/press-releases/ 2017/06-09b.html (дата обращения: 26.04.2020).
4. **Советов Б.Я., Татарникова Т.М., Пойманова Е.Д.** Организация многоуровневого хранения данных // *Информационно-управляющие системы*, 2019. № 2. С. 68–75. doi:10.31799/1684-8853-2019-2-68-75.
5. **Lorido-Botran Tania, Miguel-Alonso Jose, Lozano Jose A.** A review of auto-scaling techniques for elastic applications in cloud environments // *Journal of grid computing*. 2014. Vol. 12. № 4. P. 559–592.

B.Ya. Sovetov, T.M. Tatarnikova, V.V. Cehanovsky, (Saint Petersburg Electrotechnical University “LETI”, St. Petersburg)

Autoregression Network Traffic Prediction Models

The relevance of forecasting network traffic is due to the requirement to store a large amount of data for a long time. In this regard, it becomes necessary to predict traffic volumes in order to take the necessary measures to protect and preserve data. The paper substantiates the use of autoregressive class models for constructing predictive estimates. The results are given, which make it possible to substantiate the prospective requirements for the memory volumes of the nodal equipment of the infocommunication network.