Поиск и оценка аномалий сетевого трафика на основе циклического анализа

И.М. Ажмухамедов, А.Н. Марьенков
ФГБОУ ВПО «Астраханский государственный технический университет»,
г. Астрахань

Сетевые технологии стали неотъемлемой частью жизнедеятельности современного общества. При этом для эффективной работы сетей большое значение имеет надежность передачи данных по каналам связи.

Одной из главных причин, влияющих на эффективность работы вычислительной сети (ВС) являются аномалии трафика. Аномалии в трафике ВС могут быть вызваны неисправностью сетевого оборудования, случайными или преднамеренными действиями со стороны легитимных пользователей, неверной работой приложений, действиями злоумышленников и т.д.

Таким образом, для надежной передачи данных в ВС могут быть приняты меры по своевременному выявлению аномалии, поиску ее источника или источников и принятию мер по ее устранению (оповещение о неисправности, фильтрация аномального трафика и т.п.). Следовательно, для обеспечения надежной передачи данных в ВС большое значение приобретает разработка новых методов обнаружения аномалий и меры по ее устранению.

На сегодняшний день одними из наиболее распространенных средств, используемых для выявления аномалий, являются средства обнаружения атак (СОА). Данные средства идентифицируют подозрительную (аномальную) активность, направленную на вычислительные или сетевые ресурсы и реагируют на нее.

Однако ни одно из существующих средств обнаружения атак не способно полностью выявлять аномальную активность в трафике ВС. По статистике около 80% нарушений совершаются внутренними нарушителями, т.е. сотрудниками организации. Используемые средства обнаружения атак малоэффективны при выявлении негативных воздействий со стороны внутренних злоумышленников. В целом можно выделить следующие недостатки средств обнаружения атак на ВС:

- высокая стоимость коммерческих систем обнаружения атак;
- большое количество ложных срабатываний, а также высокий процент пропуска реальных атак на вычислительные сети;
- слабые возможности для обнаружения новых и видоизмененных атак;
- проблемы при определении источника нарушения и целей атакующего в случае антропогенной угрозы;
- невозможность определения некоторых нарушений на начальных этапах;
- большие требования к вычислительным ресурсам систем, работающих в режиме реального времени;
- высокая квалификация экспертов по выявлению атак, необходимая при внедрении СОА.

В связи с существующими недостатками современных СОА возникает необходимость разработки новых методов обнаружения аномального сетевого трафика позволяющих выявлять и оценивать величину аномалии, а также принимать решения о необходимости ее устранения.

В работах [1, 2] описана общая схема управления трафиком ВС на основе выявления аномалий. В данной схеме могут быть выделены следующие функциональные блоки:

- извлечение информации о сетевых пакетах;
- построение прогноза;
- поиск и оценка аномалии;
- реагирование на аномалию;
- заполнение и редактирование базы правил (БП).

На первом этапе из трафика извлекается вся необходимая для прогнозирования информация. Поскольку цель прогнозирования на основе имеющихся данных о загрузки сети получить значение объема трафика на определенный период времени в будущее, из заголовка IP-пакета необходимо выделять информацию об общей длине пакета, а также сохранять дату и время получения пакета. При фильтрации может быть использована информация об адресе источника и адресе назначения IP-пакета. Таким образом, для прогнозирования трафика из IP-пакета извлекается следующая информация:

- объем IP-пакета;
- IP-адрес источника;
- IP-адрес назначения;
- дата получения IP-пакета;
- время получения IP-пакета.

На основе собранной статистики производится прогнозирование сетевого трафика. Построим математическую модель прогнозирования трафика на базе циклического анализа временных рядов. В основу алгоритма прогнозирования сетевого трафика могут быть положены идеи, изложенные в [3].

Рассмотрим основные шаги, необходимые для проведения циклического анализа.

Отбор данных. На первом этапе необходимо определить форму и количество данных, на которых будет производиться прогнозирование. Циклический анализ сильно зависит от однородности данных. Используемые данные должны иметь однородную структуру, иначе неоднородность данных при анализе, скорее всего, изменит структуру циклов. Таким образом резкое изменение в работе ВС (например подключении большого количество хостов или изменение в расписании), которые могут изменить форму циклов, должны учитываться при поиске и оценке аномалии.

Поскольку циклический анализ предполагает работу с рядом данных, необходимо сформировать имеющиеся данные по сетевому трафику в виде ряда значений, описывающих изменение объема трафика во времени. Для этого необходимо провести дискретизацию потока трафика. Рассмотрим этот процесс на примере (рис. 1).

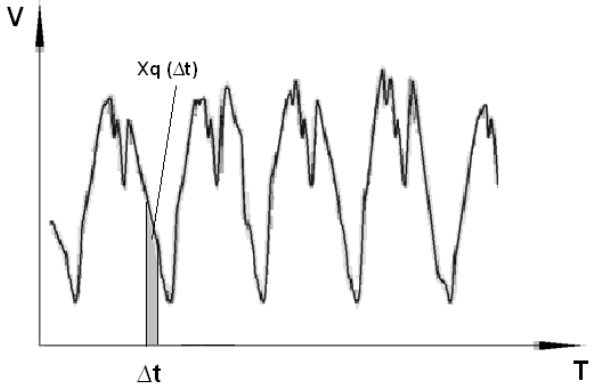


Рис. 1. Дискретизация трафика

На рисунке представлен график, изображающий поток сетевого трафика: на оси абсцисс представлено время t, на оси ординат отложен объем трафика V. Пусть имеется статистика по трафику, собранная за период времени T. Чтобы получить ряд данных, разделим период времени T на Q равных интервалов Δt:

$$Q = \frac{T}{\Delta t}.$$

Далее для каждого интервала Δt складываем объемы сетевых пакетов, попавших в данный интервал времени:

$$X_q(\Delta t) = \sum_{j=1}^{R} V_{\text{пакета}},$$

где: R – количество сетевых пакетов, попавших в интервал Δt, q – номер интервала, q= $1, \dots, Q$, X_q – ряд упорядоченных данных, описывающий изменения объема трафика во времени, с частотой дискретизации Δt.

Сглаживание данных. Определившись с данными, необходимо исключить из трафика случайные колебания. Для этого предусмотрен шаг по сглаживанию данных.

Для устранения случайных колебаний используется метод краткосрочной центрированной скользящей средней ряда данных.

Количество точек для сглаживания данных возьмем равным L. При вычислении скользящей средней по L точкам, из первоначального ряда данных будет выброшено $L-1$ точек: $\frac{L-1}{2}$ – в начале и в конце ряда. Таким образом, длина нового ряда данных $\bar{X}_k$ равна: $N = Q - (L-1), k = 1, \dots, N$:

$$\bar{X}_k = \frac{1}{L}\sum_{j=k}^{(k+L-1)} X_j.$$

Поиск возможных циклов. Устранив случайные колебания, можно приступить к непосредственному поиску циклов. Чтобы определить частотные составляющие рассматриваемого ряда, используем метод спектрального анализа. Математической основой спектрального анализа является преобразование Фурье [4]. Поскольку обрабатываемая статистика сетевого трафика имеет вид цифрового ряда, для определения частотных составляющих подойдет метод дискретного преобразования Фурье. С помощью прямого дискретного преобразования Фурье найдем комплексные амплитуды ряда данных $\bar{X}_k$:

$$Y_n = \sum_{k=1}^{N} \bar{X}_k e^{-\frac{2\pi i}{N}nk},$$

где N – количество элементов ряда данных $\bar{X}_k$ и количество компонентов разложения, i – мнимая единица.

Модуль комплексного числа может быть найден как:

$$|Y_n| = \sqrt{Re^2(Y_n) + Im^2(Y_n)}.$$

На основе комплексных амплитуд Y_k вычисляется спектр мощности:

$$R_n = |Y_n|^2 = Re^2(Y_n) + Im^2(Y_n),$$

Изобразим спектр мощности графически (рис. 2) [3].

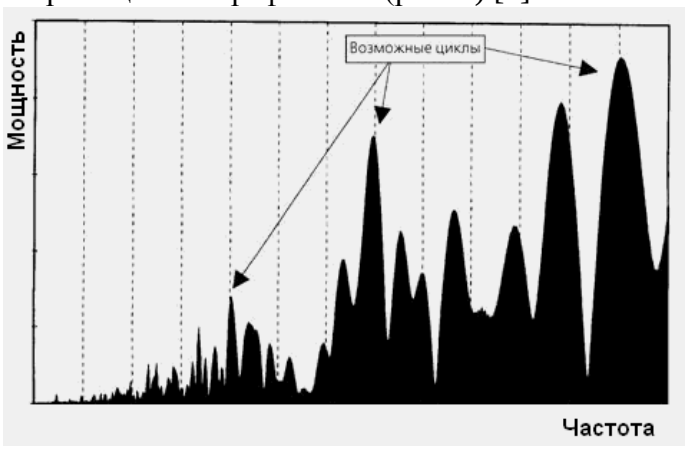


Рис. 2. Спектр мощности данных

На рисунке видно, что высокие значения скапливаются около некоторых частот. Пики в областях скопления высоких значений показывают возможные циклы. Значением частоты цикла будет являться индекс n, при котором наблюдается высокое значение спектра мощности R_n.

Определив возможные циклы и их частоты, рассчитаем обычную (вещественную) амплитуду A и фазу φ. Пусть найдено b возможных циклов, частоты которых составляют множество S, т.е. каждое значение частоты, при которой наблюдается пик в области скопления высоких значений спектра является элементом множества S.

Тогда амплитуды и фазы найденных циклов могут быть вычислены по формулам:

$$A_h = \frac{|S_h|}{N} = \frac{1}{N}\sqrt{Re^2(S_h) + Im^2(S_h)},$$

$$\varphi_h = Arg\ (S_h) = arctg\left(\frac{Im(S_h)}{Re(S_h)}\right),$$

где $h = 1, \dots, b$, $Arg\ (S_h)$ – функция мнимого числа: угол мнимого числа (в радианах), соответствующий (S_h).

Функция, описывающая цикл, выглядит как:

$$f_h(t) = A_h \cos(S_h t + \varphi_h).$$

Однако, как уже было сказано, высокое значение спектра мощности лишь предполагает наличие цикла. Поэтому следующим шагом является подтверждение найденных циклов. Для этого необходимо проверить определенное количество критериев.

Удаление трендовых компонентов в трафике. Качество проверки циклов на статистическую надежность сильно зависит от существования направленности в данных. Поэтому перед проверкой необходимо провести удаление тренда из данных. Для этого можно применить метод отклонения от скользящего среднего. В данном случае, скользящая средняя будет отражать силы роста в данных, следовательно, ее вычитание из данных удалит и трендовую составляющую. Таким образом, чтобы удалить тренд в данных необходимо для каждой найденной частоты рассчитать скользящую среднюю для ряда данных $\bar{X}_k$ с количеством точек сглаживания $L = S_h$:

$$\bar{X}`_k = \frac{1}{L}\sum_{j=k}^{(k+L-1)} \bar{X}_j,$$

где полученный ряд данных будет короче исходного на $L - 1$ точек: $N` = N - (L - 1)$, $k = 1, \dots, N`$.

Далее вычитаем из исходно ряда данных $\bar{X}_k$ полученную скользящую среднею $\bar{X}`_k$:

$$\bar{X}``_k = \bar{X}_k - \bar{X}`_k.$$

Удалив силы роста в данных, можно приступать к проверке найденных циклов на статистическую значимость.

Проверка циклов с точки зрения статистической значимости. Для оценки циклов обычно используют тесты F-коэффициент и хи-квадрат, поэтому они же будут использованы для проверки циклов в сетевом трафике.

Отметим, что результаты теста зависят от количества повторений цикла в данных. Чем таких повторений больше, тем более статистически значим данный цикл.

Комбинирование и проецирование циклов в будущее. Прогнозирование трафика происходит на этапе комбинирования и проецирования циклов. Для этого циклы объединяются и на основе полученного результата можно спрогнозировать их поведение в будущее. Для проецирования циклы математически комбинируются в одну общую кривую.

Допустим, что тесты прошло D циклов. Подтвердившиеся циклы проецируются в общую кривую, описывающую периодичность в ряде данных:

$$\bar{V}(t) = \sum_{j=1}^{D} f_h(t).$$

Данная функция описывает периодичность в трафике, найденную на основе данных за период времени T. Полученная функция может быть экстраполирована в будущее и позволяет получить прогнозируемое значение трафика на период времени $\bar{T}$ в будущее:

$$V_{\text{прогноз}}(t`) = \sum_{j=1}^{D} f_h(t`),$$

где $t` \in (T, \bar{T})$ (рис. 3).

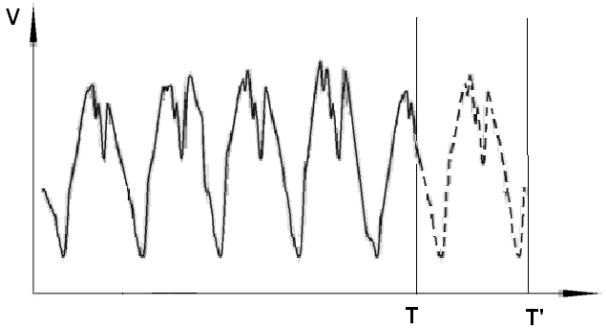


Рис. 3. Прогнозирование трафика

Определив математическую модель прогнозирования сетевого трафика, рассмотрим систему поддержки принятия решений (СППР) поиска и оценки аномалии рис. 4.

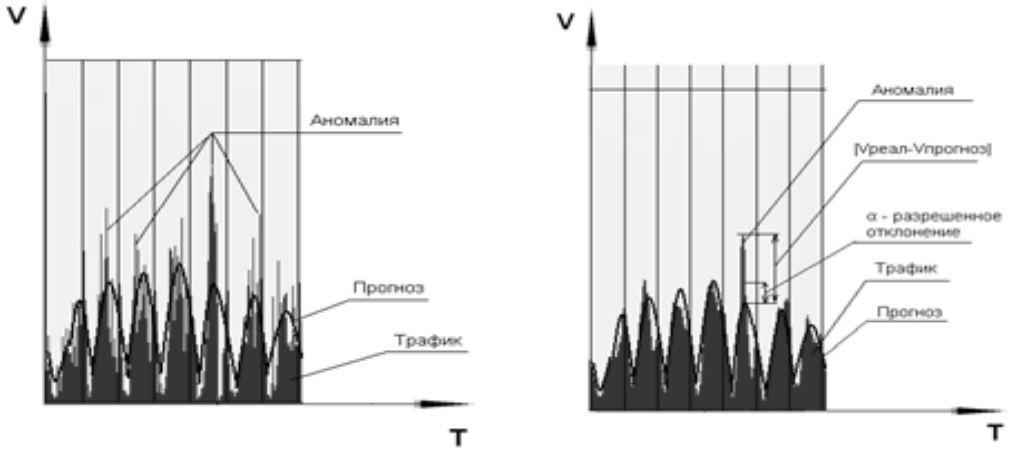
Рис. 4. Поиск и оценка величины аномалии

Для определения объема трафика, поступающего в реальном времени, используется извлекаемая информация о сетевых пакетах. На основе получаемых данных о трафике и текущем прогнозе производится описка аномалии. В случае если аномалия была найдена, результат поиска аномалии передается на блок поиска источников аномалии. Определение источников аномалии осуществляется на основе результата поиска аномалии и информации о сетевых пакетах, поступающих в реальном времени. Далее производится оценка величины аномалии. В оценке используется полученная информация о нарушителях, информация о пакетах, а также лицо принимающее решение (ЛПР) и эксперт. Информация о величине аномалии обобщается и передается для дальнейшего использования.

Поиск аномалии происходит на основе сравнения трафика поступающего в реальном времени с прогнозируемым значением. Для этого в единицу времени t сравниваются два значения $V_{реал}$ – величина объема текущего трафика и $V_{прогноз}$ – прогнозируемое значение объема. Аномальным будет считать отклонение объема, превышающее или равное заданной величине α:

$$|V_{реал} - V_{прогноз}| \geq \alpha.$$

Отметим, что прогноз представлен в виде относительно гладкой кривой. В свою очередь, трафик состоит из кратковременных случайных флуктуаций. Если сравнивать эти два ряда данных, возможны ложные определения аномалий (рис. 5, а). Чтобы этого избежать, реальный трафик сглаживается методом скользящей средней, при этом окно сглаживания смещается по мере поступления нового трафика (рис. 5, б).



а). Трафик до сглаживания б). Трафик после сглаживания

Рис. 5. Аномалия трафика

Если аномалия была обнаружена, происходит поиск источников аномалии. Поиск источников определяется на основе информации, извлекаемой из текущего трафика.

Оценка величины аномалии происходит на основе продукционной базы правил. Ее начальным заполнением занимается эксперт. В дальнейшем ЛПР может корректировать БП исходя из результатов фильтрации трафика.

Рассмотрим структуру БП более подробно. Для оценки величины аномалии используются следующие лингвистические переменные с соответствующими терм-множествами:

- Величина отклонения. V = {низкая, ниже среднего, средняя, выше среднего, высокая}.
- Частота появления аномалии. M = {низкая, ниже среднего, средняя, выше среднего, постоянная}.
- Количество источников аномалий. I = {незначительное, ниже среднего, среднее, выше среднего, большое}.
- Средний объем трафика от одного источника. W = {незначительный, ниже среднего, средний, выше среднего, высокий}.

Выходным параметром является: Величина аномалии E = {незначительная, ниже среднего, средняя, выше среднего, высокая}.

На основе введенных переменных формируется набор правил. Пример правил:

- ЕСЛИ V = {ОТ низкая ДО ниже среднего} и M = {ОТ низкая ДО средняя} и I = {ОТ незначительное ДО среднее}и W = {ОТ незначительный ДО средний} ТО E = {незначительная};

- ЕСЛИ V = {ОТ выше среднего ДО высокая} и M = {ОТ постоянная ДО постоянная } и I = {ОТ выше среднего ДО большое}и W = {ОТ выше среднего ДО выше среднего} ТО E = {высокая};
- …

При формировании набора правил в качестве основы была использована схема с N экспертами, каждый из которых независимо друг от друга, продуцирует набор правил [5].

Суть данного подхода заключается в следующем. Каждый эксперт создает свой набор правил. Сгенерированный набор правил каждым последующим экспертом дополняет базу правил новыми правилами, тем самым увеличивая полноту модели. Поскольку заполнение базы правил экспертами происходит независимо, в каждом последующем наборе правил могут содержаться правила, которые могут повторять уже существующие правила. Также в новом наборе правил могут быть правила, противоречащие правилам из других наборов. Таким образом, появляется проблема проверки базы правил на противоречивость, избыточность и полноту.

Понятие однозначности означает, что каждому сочетанию координат V, M, I, W соответствует только одно значение выходной координаты E. В идеале, правила должны полностью соответствовать понятию однозначности, однако, из-за возможной размытости знаний экспертов и значениях лингвистических переменных, допускается частичная однозначность.

Избыточность подразумевает ситуацию, когда одно правило включает в себя другое правило из общего набора, например:

Правило 1:

ЕСЛИ V = {ОТ низкая ДО ниже среднего} и M = {ОТ низкая ДО средняя} и I = {ОТ незначительное ДО среднее}и W = {ОТ незначительный ДО средний} ТО E = {незначительная}

Правило 2:

ЕСЛИ V = {ОТ низкая ДО ниже среднего} и M = {ОТ низкая ДО средняя} и I = {ОТ незначительное ДО среднее}и W = {ОТ незначительный ДО незначительный} ТО E = {незначительная}

Как видно из примера, в первых трех координатах правила полностью идентичны, однако параметр «средний объем трафика от одного источника» для первого правила измеряется «ОТ незначительный ДО средний», во втором случае это параметр имеет диапазон «ОТ незначительный ДО незначительный» Очевидно, что второе правило входи в первое правило.

Понятие противоречивости. Если два правила имеют на входе одинаковые значения координат V, M, I, W а на выходе значение E различно (нарушение гипотезы однозначности), то данные правила считаются противоречивыми:

Правило 1:

ЕСЛИ V = {ОТ низкая ДО ниже среднего} и M = {ОТ низкая ДО средняя} и I = {ОТ незначительное ДО среднее}и W = {ОТ незначительный ДО незначительный} ТО E = {незначительная}

Правило 2:

ЕСЛИ V = {ОТ низкая ДО ниже среднего} и M = {ОТ низкая ДО средняя} и I = {ОТ незначительное ДО среднее}и W = {ОТ незначительный ДО незначительный} ТО E = {средняя}

Под полнотой понимается отношение доли охвата знаний выходных координат к общему диапазону возможных решений.

Для проверки полноты базы правил необходимо найти отношение количества выходных значений для существующей базы правил к количеству всех возможных значений.

Чтобы определить число существующих выходных значений, необходимо сложить все решения от каждого созданного правила:

$$S_{\text{сущ}} = (\sum \text{Решения правила1} + \sum \text{Решения правила 2} + \cdots + \sum \text{Решения правила } N)$$

Полное множество возможных решений может быть рассчитано перебором всех возможных значений входных координат:

$$S_{\text{общ}} = S_v * S_m * S_i * S_w,$$

где S_v, S_m, S_i, $* S_w$ – количество значений лингвистических переменных. Тогда полное множество возможных решений равно:

$S_{\text{общ}} = 5 * 5 * 5 * 5 = 625$

Тогда полнота базы правил рассчитывается как:

$$\text{П} = \frac{S_{\text{сущ}}}{S_{\text{общ}}} * 100\%$$

Если отношение меньше 100%, производится поиск правил, которые небыли учтены экспертами. На основе полученного результата формируется новый набор правил, который передается для оценки экспертам. Далее снова проводится проверка базы правил на противоречивость, избыточность и полноту до тех пор, пока база не будет полностью сформирована.

После рассмотрения всех индивидуальных наборов правил экспертов, формирование правил заканчивается.

Если в процессе эксплуатации произойдет ситуация, решение которой не будет в базе правил (например, ошибки при обучении), предлагается два варианта действий: блокировка аномального трафика и ожидание действий ЛПР. Получив решение, система формирует на его основе правило и пополняет базу правил. Во втором случае, система выбирает наиболее подходящее правило из базы правил и выполняет действие, исходя из найденного решения. При этом ЛПР может либо добавить новое правило в базу правил или подтвердить выбор СППР и тогда новое правило будет добавлено автоматически.

Таким образом, величина аномалии объема сетевого трафика может быть представлена в виде системы:

$$\text{e} = \begin{cases} \{\{V\}, \{\text{M}\}, \{\text{I}\}, \{\text{W}\}, \{\text{E}\}\}, \left|\text{V}_{\text{реал}} - \text{V}_{\text{прогноз}}\right| \geq \alpha, \\ 0, \left|\text{V}_{\text{реал}} - \text{V}_{\text{прогноз}}\right| < \alpha \end{cases},$$

где $\{V\}, \{\text{M}\}, \{\text{I}\}, \{\text{W}\}$ – лингвистические переменные, применяемые для оценки величины аномалии объема трафика, $\{\text{E}\}$ – множество значений лингвистической переменной, определяющей выходные значения базы правил, $\text{V}_{\text{реал}}$ – объем трафика, поступающего из сети в реальном времени, $\text{V}_{\text{прогноз}}$ – прогнозируемой значение трафика, α – величина, определяющая какое отклонение трафика, поступающего из сети в реальном времени от прогнозируемого значения можно считать аномальным.

Процесс реагирования на аномалию может быть представлен в виде трех основных блоков:

- определение необходимости фильтрации;
- фильтрация пакетов;
- подготовка отчета об аномалии.

Необходимость фильтрации определяется на основании информации об аномалии, а также за счет настроек ЛПР, на основе которых формируются исключения для фильтрации $\{Z\}$ (список заблокированных источников; источники, которые запрещено блокировать и т.п.). Далее происходит непосредственная фильтрация трафика и подготовка отчета об аномалии.

Рассмотрим блок фильтрации (рис. 6).

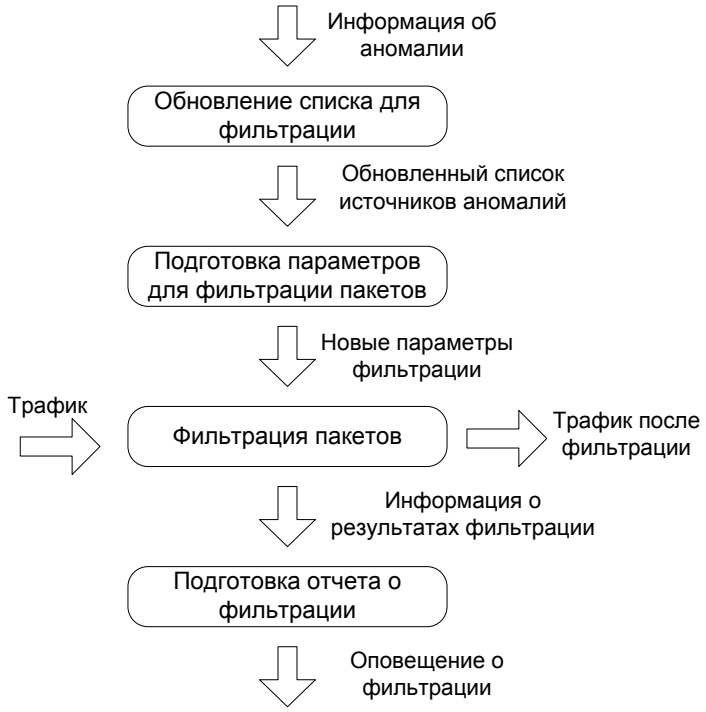


Рис. 6. Фильтрация трафика

Первым действием является обновление списка для фильтрации. В данном блоке добавляются новые источники в список фильтрации, а также удаляются источники, время блокирования которых истекло. Далее подготавливаются параметры для фильтрации, на основе которых в фильтре пакетов формируются правила фильтрации трафика.

Чтобы различать трафик из разных подсетей, необходимо учитывать IP-адрес и маску подсети. Это позволит индивидуально настраивать фильтрацию трафика для каждой подсети. Также должно быть предусмотрено раздельное отслеживание входящего и исходящего трафика.

Поскольку трафик из внешних и внутренних сетей имеет разную информативность и, как следствие, различия при построении модели прогноза, то имеет принципиальное значение разделение трафика из внешней и внутренней сети.

Таким образом, при поиске аномалий объема сетевого трафика, необходимо использовать следующие характеристики:

- Величина отклонения реального трафика от прогнозируемого;
- Размер окна для сглаживания реального трафика;
- IP-адрес подсети и маски;
- Направление трафика (входящий или исходящий);
- Внешняя или внутренняя сеть.

Для определения параметров фильтрации также используется база правил, которая на основе величины аномалии определяет время фильтрации $\{F\}$. В качестве входного параметра используется величина аномалии e, описанная ранее Пример правил:

- ЕСЛИ E = {незначительная} ТО Время блокирования = 1 минута;
- ЕСЛИ E = {высокая} ТО Время блокирования = 120 минут;
- ….

Полученные новые параметры фильтрации используются для настройки фильтра пакетов. Также формируется отчет о проведенной фильтрации.

Общая схема формирования правил фильтрации может быть представлена как:

$$G = \{e, \{Z\}, \{F\}, \{U\}\},$$

где $\{Z\}$ – сетевые адреса источников аномалии, $\{F\}$ – время фильтрации, определяется в зависимости от величины e, $\{U\}$ – список исключений фильтрации.

Таким образом, предложена математическая модель прогнозирования трафика на базе циклического анализа временных рядов, позволяющая определять загрузку сети на основе поиска периодичности в сетевом трафике. Также разработана СППР о наличии аномалии и необходимости ее устранения, позволяющая на основе полученного прогноза выявлять и оценивать величину аномалии, а также генерировать предупреждения о возможных нештатных ситуациях в работе вычислительной сети. Полученные результаты могут быть использованы для поиска неисправностей сетевого оборудования, выявления ошибок в настройке программного обеспечения, выявления случайных и преднамеренных действий со стороны пользователей, а также действия злоумышленников.

Разработанная модель и СППР могут быть использованы в системе управления трафика, позволяя анализировать состояние вычислительной сети с целью выявления аномалий в трафике, предупреждая персонал о необходимости принятия мер, по устранению аномалии (ремонт неисправного оборудования, фильтрация трафика и т.п.)

Литература:

1. Марьенков А.Н., Ажмухамедов И.М. «Повышение безопасности компьютерных систем и сетей на основе анализа сетевого трафика». Инфокоммуникационные технологии. Том 8, №3 / 2010 , стр.106-108
2. Марьенков А.Н., Ажмухамедов И.М., «Обеспечение информационной безопасности компьютерных сетей на основе анализа сетевого трафика», Вестник АГТУ. Серия «Управление, вычислительная техника и информатика» №1 / 2011, стр.141-148
3. Д.Н. Швагер «Технический анализ. Полный курс». М.: издательство Альпина Бизнес Букс, 2007.
4. Медведев С.Ю., «Преобразование Фурье и классический цифровой спектральный анализ», http://www.vibration.ru/preobraz_fur.shtml
5. Проталинский О.М. «Применение методов искусственного интеллекта при автоматизации технологических процессов». Астрахань: Изд-во АГТУ, 2004.