УДК 004.45(045)

[1]Минаев Ю.Н. – д-р техн. наук,
[1]Гузий Н.Н. – канд. техн. наук,
[2]Филимонова О.Ю – канд. техн. наук

ИДЕНТИФИКАЦИЯ АНОМАЛЬНЫХ СОСТОЯНИЙ ТРАФИКА КОМПЬЮТЕРНЫХ СЕТЕЙ НА ОСНОВЕ ПАРАДИГМЫ МНОГОМЕРНЫХ СЕТЕЙ

[1]Национальный авиационный университет
[2]Киевский национальный университет строительства и архитектуры

Рассмотрены вопросы представления трафика компьютерной сети инвариантами тензора 4-го ранга. Трафик рассматривается как поток тензоров 2-го ранга многомерной сети. Показана эффективность применения предложенного подхода для идентификации аномальных состояний компьютерной сети

Введение

Компьютерные сети (КС) относятся к классу сетей с коммутацией пакетов, общей технологией которых является семиуровневая эталонная модель взаимодействия открытых систем. Вместе с тем, анализ трафика выполняется в большинстве случаев на уровне обобщенного трафика без учета его компонент[1, 2].

Одним из актуальных направлений исследования *NGN*, *FGN* является моделирование на основе многомерных сетей. Использование многомерной структуры сети создает комплекс принципиально новых задач, в частности выявления аномалий трафика[3,4].

Для идентификации аномального трафика компоненты трафика представляются отдельным измерением многомерного массива (тензора). В силу этого требуется разработка принципиально новых алгоритмов идентификации аномальных состояний трафика. Процессы, связанные с передачей информации в многомерной интегрированной сети достаточно сложны, одним из способов анализа таких сетей являются многомерные тензоры и их декомпозиции, в частности, сечения тензоров.

Цель исследования

Создание виртуальных многомерно-матричных моделей сетевых структур в качестве средства для исследования трафика КС.

Современное состояние исследований.

В работах Г.Крона [5] показано, что представление объекта исследования (измерения) в виде тензора есть более адекватным. Тензорная модель, которая рассматривается как матричная проекция, позволяет анализировать объект в разных системах координат.

Анализ трафика широко использует понятие «нормального» и «нестандартного» (необычного) трафика. Ожидаемый (нормальный) трафик зто трафик, который передается чаще всего. Известно, что пакетный трафик – это дискретный положительный процесс с сингулярной внутренней структурой. Наиболее полно он может быть представлен на уровне иерархических структур.

Проблема выявления аномалий в роботе КС на основе анализа трафика относится к т.н. трудно формализуемым проблемам, решение которых современная наука видит в использовании интеллектуальных технологий [6]. Параметры трафика авторами в общем виде определены следующим образом: $x. = \{ x_i \}$, $i = 1,9$: x1 – Protocol ID протокол, связанный с событием (TCP=0, UDP=1, ICMP=2, unknown=3); x2 – номер порта источника; x3 – номер порта хоста назначения; x4 – IP-адрес источника; x5 – IP-адрес приемника; x6 – ICMP Type тип ICMP-пакета (Echo Request or Null); x7 – ICMP Code кодовое поле из ICMP-пакета

(None or Null); x8 – Raw Data Length длина данных в пакете; x9 – Raw Data порция данных в пакете.

Одним из способов выявления аномальных состояний трафика является идентификация при помощи стандартных персептронов. Для заданных состояний КС -аномального - $X^{(a)} = \{x_i^{(a)}\}$, i=1, 9 и нормального $X^{(n)} = \{x_i^{(n)}\}$, i=1, 9 сформированы множества $\{{}^R X^{(a)}\}_{j=1}^J$ и $\{{}^R X^{(n)}\}_{j=1}^J$, J= 1000÷10000, у которых каж-дая компонента трафика лежит в интервале ± 15 % от заданных, т.е. $x_i^{(a)} - 0.15\, x_i^{(a)} \le {}^R x_i^{(a)} \le x_i^{(a)} + 0.15\, x_i^{(a)}$ $(\forall i)$, ${}^R x_i^{(a)} \in \{{}^R X^{(a)}\}_{j=1}^J$ и соответственно $x_i^{(n)} - 0.15\, x_i^{(n)} \le {}^R x_i^{(n)} \le x_i^{(n)} + 0.15\, x_i^{(n)}$ $(\forall i)$, ${}^R x_i^{(n)} \in \{{}^R X^{(n)}\}_{j=1}^J$. Нейронная сеть (рис.1), обученная на одном из наборов, например, наиболее часто встречающимся- $\{{}^R X^{(n)}\}_{j=1}^J \to 1$, $\{{}^R X^{(a)}\}_{j=1}^J \to 0$, позволяет практически однозначно идентифицировать состояние.

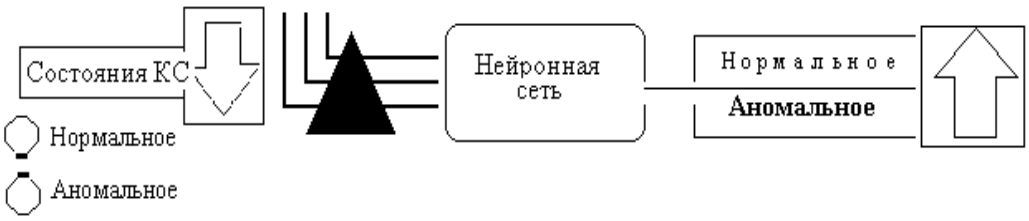


Рис. 1. Идентификация состояний КС при помощи НС

Недостаток данного подхода состоит в том, что предварительное определение базовых состояний $X^{(a)}$ и $X^{(n)}$ является экспертной оценкой и по своей сложности адекватно исходной задаче.

Современные приложения (трафик *Internet*, телекоммуникационные записи и др.) генерируют большие массивы данных. Тензоры (мультимерные массивы) обеспечивают естественное представление для таких данных. Тензорные декомпозиции стали инструментальными средствами для обобщения и анализа временных рядов [6–8] .

На основе тензорных декомпозиций можно обнаруживать общности групп объектов, предсказывать потерю связей, ранжировать объекты и многие другие задачи [9]. На основе обобщения низкоранговой матричной аппроксимации. *Rank-1* тензор определяется как внешнее произведение r векторов *x(1), . . . , x(r-1), x(r)* заданных через r-мерный массив у которого *(i1 , . . . , ir)* -тый вход есть *xi1(1), x i2(2), . . . , x ir(r)*; это обозначено как $x(1) \otimes x(2) \otimes \cdots x(r)$. Для любого r-мерного массива A существует аппроксимация через сумму малого числа *rank*- 1 тензоров.

Основная часть исследования

Многомерный трафик КС представляется многомерным массивом с достаточно большим количеством компонент, для анализа которого целесообразно использовать аппарат тензорного анализа и мультилинейной алгебры. Уникальное свойство тензоров состоит в том, что индексы могут иметь произвольную топологию [10]. Тензор-декомпозиции – модели тензора в смысле наименьших квадратов: если T- тензор, то $\tilde{T}$ -модель тензора (в общем случае меньшей размерности), полученная, если минимизировать $\delta = |T - \tilde{T}|2$ или точнее

$$\delta = \sum_{i_1, i_2, \ldots} \left| T(i_1, i_2, \ldots) - \tilde{T}(i_1, i_2, \ldots) \right|^2 .$$

Тензор может храниться в т.н. отложенной форме как сумма rank-1 тензоров [11]. На рис.2 показан многомерный тензор как 3-мерный массив данных размера *I1 ×I2 ×I3*.

Определение 1. Пусть *A* - тензор размерности $I1 \times I2 \times \cdots \times IN$. Порядок *A* есть *N*, *n*-ая размерность (способ или

путь) A есть размер *In*. Нотация двоеточия использована, чтобы обозначить полную область данного индекса (*i*-тая колонка матрицы дается *A(i, :)* и j-тый столбец – *A(:, j))*.

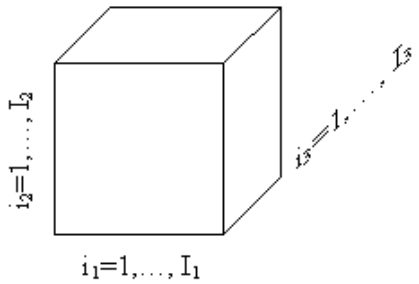


Рис. 2. Представление 3-х мерного массива

Рассмотрим тензор 3-го-порядка (рис.2). В этом случае определение единственного индекса связано с вырезкой (*slice*), которая является матрицей в специфической ориентации. Так *A(i, :, :)* формируют *i*-тую горизонтальную вырезку, *A(:, j, :)* - *j*-тая боковая вырезка и *A(:, :, k)* *k*-тая фронтальная вырезка.

Определение 2. Обобщенная матрицизация. Пусть *A* есть $I_1 \times I_2 \times \cdots \times I_N$ - тензор *r* и предположим, что необходимо перестроить его в матрицу размером $J_1 \times J_2$. Число входов в матрицу должно быть таким же, что и число входов в тензор $\prod_{n=1}^{N} I_n = J_1 \cdot J_2$. Данные J_1 и J_2 удовлетворяют следующему свойству: отображение может быть задано любым количеством измерений такой длины, что имеем отображение π один в один, так что

$$\pi:\{1,\ldots,I_1\}\times\{1,\ldots,I_2\}\times\ldots\times\{1,\ldots,I_N\}\rightarrow$$
$$\rightarrow\{1,\ldots,J_1\}\times\{1,\ldots,J_N\}.$$

Класс *tensor_as_matrix (tensor toolbox Mat-Lab)* поддерживает конверсию тензора в матрицу. Рассмотрим *n*-мерную матрицизацию. Обычно тензор матрицизирован так, что все его волокна связаны с отдельными измерениями, выровненными как колонки результирующей матрицы. Это специальный случай обобщенной матрицизации, где только одна размерность отображается в строке, так что *K* = 1 и {*r*1} = ={*n*}. Результирующая матрица типично обозначена как *A(n)*. Колонки могут быть упорядочены по любому направлению.

Одно из преимуществ матрицизации состоит в том, что тензор, хранимый в матричной форме, может быть манипулирован как матрицы, сокращая *n*-мерное умножение. Если $B = A \times n M$, тогда $B(n) = MA(n)$. Последовательность n-мерных произведений, которая записана в матричной формулировке, может быть представлена как Кронекерова последовательность произведений, включающих *U* матрицы. Последовательность n-мерных произведений имеет вид

$$B_{(n)} = U^{(n)} A_{(n)} (U^{(c_{n-1})} \otimes U^{(c_{n-2})} \otimes \cdots \otimes U^{(c_1)}$$

Можно также создать два дополнительных класса для поддержки представления тензоров в декомпо-зированной форме в виде суммы *rank*-1 тензоров. *rank*-1 тензор – это тензор, который может быть записан как внешнее произведение векторов, т. е

$$A = \lambda u(1) \circ u(2) \circ \ldots u(N),$$

где λ - скаляр и каждый *u(n)* – *In*-вектор для *n = 1 . . . N*. Символ °обозначает внешнее произведение; в случае *(i1 i2 . . . iN)* входа *A* представляется в виде

$$A(i_1, i_2, ..., i_N) = \lambda u_{i_1}^{(2)} u_{i_2}^{(2)} \cdots u_{i_N}^{(N)},$$

где *ui* обозначает *i*-ый вход вектора *u*. Существует несколько декомпозиций тензора. Декомпозиция *SVD* (показана на рис.3), применяемая к тензору, дает

$$A = \sum \times_1 U \times_2 V \times_3 W,$$

$$A = \sum_{i=1}^{r_1} \sum_{j=1}^{r_2} \sum_{k=1}^{r_3} \sigma_{ijk} u_1 \circ v_j \circ w_k, \quad \sum \text{ - }$$

корневой (сердцевинный) тензор, *U, V, W* – компоненты или факторы, корневой тензор – диагональный, колонки в компонентах ортонормальны.

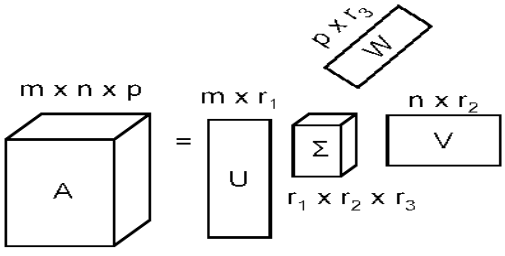
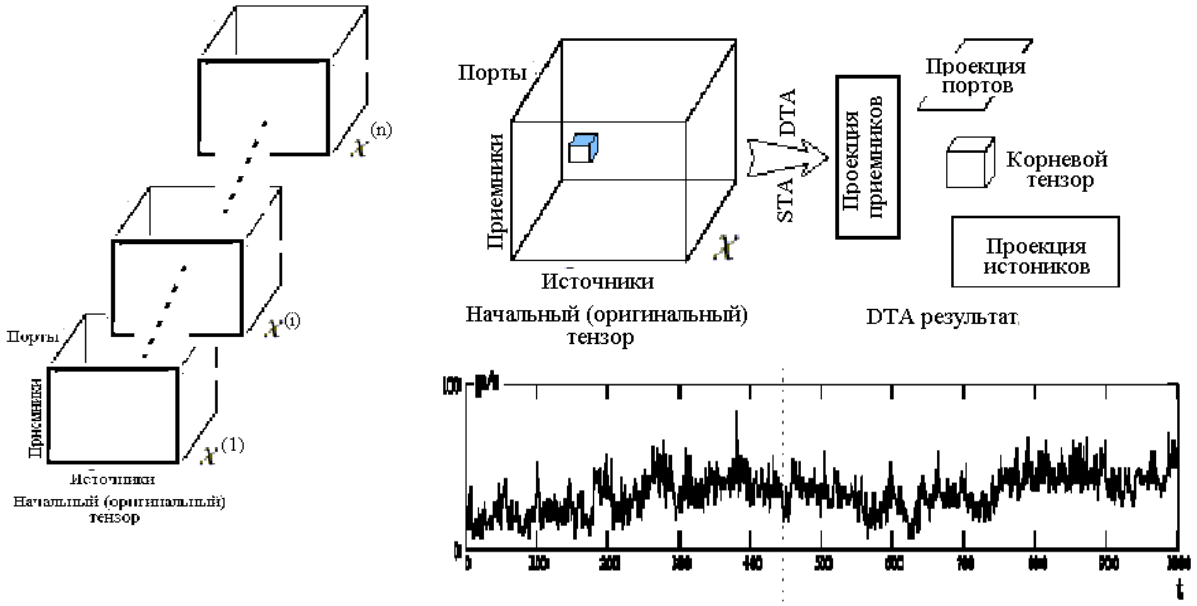


Рис.3. Декомпозиция *SVD*

Определение аномалий трафика рассмотрено в [11]. Трафик сети рассматривается на уровне существования тензоров порядка M= 3 с направлениями "IP-источник ", "IP-приемник", "порт" .

На рис. 4 представлен сетевой поток данных как тензор 3-го ранга, разложенный на корневой тензор и 3 матрицы.



а) б)

Рис.4 Тензор 3-го порядка сетевого потока данных: а) сетевой поток, б) разложение начального тензора и кривая количества пакетов трафика:, пак/с.

Для примера сетевой поток как тензор 3-го порядка, имеющий для рассматриваемого периода времени три измерения: источник, приемник и порт, которые могут быть рассмотрены как 3*D* куб данных (рис.2). Вход *(i, j, k)* в этом тензоре имеет количество пакетов от соответствующего источника (*i*) в приемник *j* через порт *k*, в течение данного периода времени. Динамический аспект исходит из того, что новые тензоры прибывают непрерывно во времени.

В тестовой сети система определяет наличие аномалии, анализируя образцы трафика на уровне множеств – источники – приемники - порты соответственно. Для 1-го уровня (тензор уровень) моделируем аномальность тензора $\aleph$ через реконструкцию ошибки:

$$e_i = \| \aleph_i - \Psi_i \prod_{i=1}^{M} \times_i U_l^T \|_F =$$

$$= \| \aleph_i - \aleph_i \prod_{i=1}^{M} \times_i U_l U_l^T \|_F^2$$

Для 2-го уровня (уровень направления – отдельного измерения), реконструкция ошибки l-го направления включает только одну проекцию матрицы *Ul* для данного тензора $\aleph$:

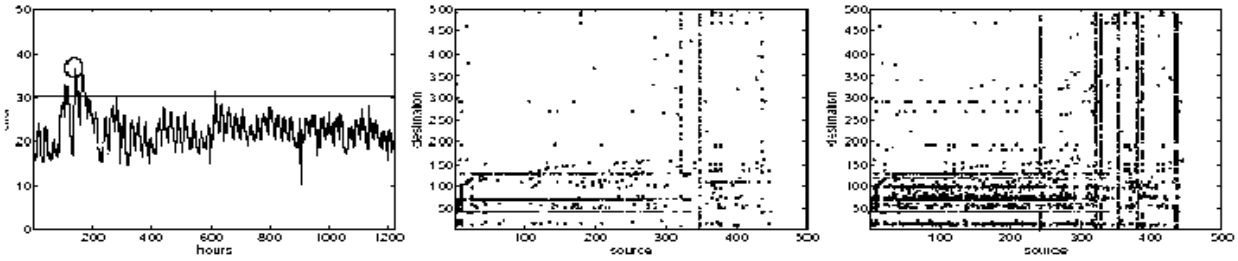
$$e_d = \left| \aleph - \aleph \times_l U_l U_l^T \right|_F^2 .$$

Для 3-го уровня (уровень размерности), ошибка размерности *d* на *l*-ом направлении есть только реконструкция слайса размерности *d* тензора вдоль *l*-го направления. Формально уровень ошибки определяется из выражения

$$e_{T+1} \geq \text{mean}\left(e_i|_{i=1}^{T+1}\right) + \alpha \cdot \text{std}\left(e_i|_{i=1}^{T+1}\right)$$

Условие *et ≥ eT+1* есть признаком аномалии.

Рассмотрим идентификацию реальной аномалии. Главные образцы связей сети должны в быть подобными друг другу на определенных временных интервалах. На рис. 5(а) показана относительная реконструкция ошибки во времени, определення на основе *DTA*.



(a) Ошибка во времени (b) Нормальный трафик (c) Аномальный трафик

Рис.5 Определение аномалии трафика по аппроксимационной точности

Точки аномалии расположены выше строки – порога среднеквадратичного отклонения процента ошибки. Некоторые взрывные ошибки происходят в интервале времени 140-й и 160-й час (рис.5a). Рис. 5b и 5c демонстрируют нормальный и аномальный трафик.

Идентификация аномального состояния КС на основе структурированного тензор-трафика использует специальные математические методы нелинейной и тензорной аппроксимации.

Пусть A – двухуровневая матрица с размерами уровней $n1$ и $n2$. Приблизим ее суммой кронекеровых произведений вида $A = \sum_{k=1}^{r} A_1^k \otimes A_k^2$, где размеры A_1^k, A_k^2 соответственно $n1 \times n1$ и $n2 \times n2$. Пусть $A=Ar$ и r- наименьшее возможное число кронекеровых произведений, сумма которых есть A. Тогда r называется тензорным рангом A. Наилучшие приближения, минимизирующие $||A-Ar||F$, могут быть получены алгоритмом *SVD*, примененным к матрицам, полученных специальным «перемешиванием» элементов.

Проблема трилинейной аппроксимации тензора формулируется следующим образом. Пусть дан трехмерный масив (тензор) $A=aijk$ размером $n1 \times n2 \times n3$. Требуется найти такие матрицы $U=ui\alpha$ $V=vja$, $W=wka$ размеров $n1 \times r$, $n2 \times r$, $n3 \times r$, которые минимизируют функционал

$$\left\| a_{ijk} - \sum_{\alpha=1}^{r} u_{i\alpha} v_{j\alpha} w_{k\alpha} \right\|,$$

где $\left\| \varphi_{ijk} \right\| = \left(\sum_{ijk} \varphi_{ijk}^2 \right)^{1/2}$.

Одно из решений задачи состоит в следующем. Пусть даны некоторые приближения U, V, W к решению U^*, V^*, W^* проблемы. Тогда по V и W находится новое значение U из решения задачи наименьших квадратов

$$U = \underset{U}{\operatorname{argmin}} \left(\left\| a_{ijk} - \sum_{a=1}^{r} u_{i\alpha} v_{j\alpha} w_{k\alpha} \right\|^2 \right).$$

Задача распадается на $n1$ независимых подзадач: $i=1,2,\ldots n$.

Представление трафика КС тензором позволяет агрегировать все параметры трафика и унифицировать его анализ. Авторами предложен принцип структурирования трафика, представленного в виде тензора четных рангов [7, 13], что позволило создать обобщенную модель трафика. Создание данного тензора реализовано путем тензорного произведения строки (вектора) безразмерных (приведенных) параметров трафика [13] -X={xj} на колонку (вектор) функций псевдопринадлежности M={μj},j=1,2,…,9; T=(X={xj})•(M= ={μj}T), где • - знак тензорного произведения, T=[tij],i,j=1,9; T - символ транспонирования; x=[x1, x2, x3, x4, x5, x6, x7, x8, x9], $^{(1)}x$ =[1 0 0 0 1 0 0 0 1]T, $^{TRAFIC}T = x \otimes^{(1)} x$

$$^{TRAFIC}T = \begin{pmatrix} x1 & x2 & x3 & x4 & x5 & x6 & x7 & x8 & x9 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ x1 & x2 & x3 & x4 & x5 & x6 & x7 & x8 & x9 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ x1 & x2 & x3 & x4 & x5 & x6 & x7 & x8 & x9 \end{pmatrix}$$

В святи с тем, что трафик содержит 3^2 параметров, получаем тензор 4-го ранга.

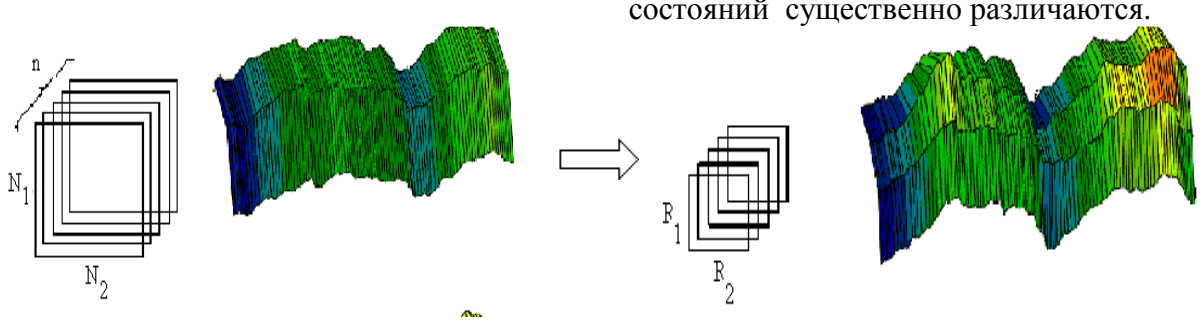
Объект, который определяется совокупностью коэффициентов *aijk…m* полилинейной формы $\varphi = \varphi(x, y, z, ..., w)$, является ортогональным тензором. Совокупность коэффициентов *aij* билинейной формы $\varphi=\varphi(x, y)$, образующая матрицу $A=(aij)$, представляет собой тензор валентности 2. Следовательно, компоненты тензора *bk,…, m* валентности *p* – 2 определяются по формуле *bk…m= a11k…m+ a22k…m+ +a33k…m* [12]. Операция получения тензора *bk…m* из тензора *aijk…m* называется свертыванием тензора *aijk…m* по индексам *i* и *j*. В результате свертывания тензоров валентностей p и q получается тензор валентности *p + q -2*. Свертывание тензоров можно производить по любому количеству *r* таких пар. Валентность результирующего тензора на *2r* единиц меньше суммы валентностей исходных тензоров.

Свертка тензора $^{TRAFIC}T$ валентности 4 дает тензор валентности 2:

$$^{TRAFIC}T = x \otimes^{(1)} x \rightarrow \begin{pmatrix} x_1 & x_2 & x_3 \\ x_4 & x_5 & x_6 \\ x_7 & x_8 & x_9 \end{pmatrix} = (t_{ij}),$$

который в дальнейшем будем называть тензор-трафик (рис.6). Возвращаясь к тензор-трафику, получаем множество матриц 3×3 (рис.7), представленных как $3 \times 3 \times K$ тензор *M (*где *K* количество отсчетов трафика).

Для идентификации аномальных режимов трафика можно использовать инварианты тензор-трафика, значения которых для нормального и аномального состояний существенно различаются.



а) б)

Рис.6. а) реальный тензор-трафик 9×9×n, б) структурированный тензор-трафик 3×3×n

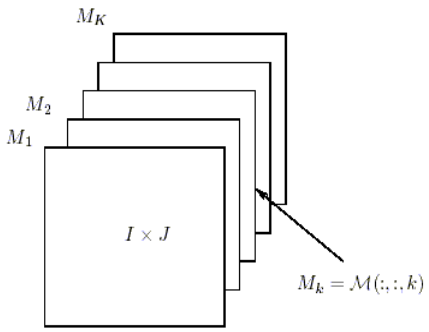


Рис. 7. Матричная модель тензор-трафика

Предположение 1. Аномальное состояние КС определяем по величинам инвариантов тензора и системи правил: если инварианты $I=\{I_1, I_2,\ldots,I_9\}$ или собственные значения матрицы тезор-трафика $\Lambda=\{\lambda_1, \lambda_2,\ldots, \lambda_9\}$ для моментов времени $t^{(1)}$ и $t^{(2)}$ существенно отличаются, т.е. $I^{(1)} \neq\approx I^{(2)}$ или $\Lambda^{(1)} \overset{\neq}{\approx} \Lambda^{(2)}$, где знак $\overset{\neq}{\approx}$ обозначает отсутствие субъективной близости, то имеет место аномальное состояние [13].

Предположение 2. Аномалией может считаться необычное сочетание значений инвариантов тензорного произведения вектора параметров трафика и специального единичного вектора с последующ $I^{(s)}=\{1\ 0\ 0\ 0\ 1\ 0\ 0\ 0\ 1\}$, ⊗ - символ Кронекерова произведения.

Введем понятия последовательности тензоров и тензорного потока [11]. Последовательность тензоров M-го порядка ℵ1 . . . ℵn, где каждый ℵi ∈ RN1× · · ·×NM (1 ≤ i ≤ n), есть тензорная последовательность, если *n* есть фиксированное натуральное число и n есть кардиналом тензорной последовательности. Последовательность тензоров *M*-го

порядка ℵ1 . . . ℵn, где каждый ℵi ∈ RN1× · · ·×NM (1 ≤i ≤ n), названа тензорным потоком, если n есть целое, увеличивающееся во времени.

Тензорный анализ трафика рассматриваем в виде. Дана последовательность тензоров ℵ1 . . . ℵn, где ℵi ∈ RN1× · · ·×NM (1 ≤ i ≤ n), находим ортогональные матрицы Ui ∈ RNi×Ri $|_{i=1}^{M}$, для каждого направления, так что ошибка реконструкции есть минимальной:

$$e = \sum_{t=1}^{n} \left\| \mathcal{X}_t - \mathcal{X}_t \prod_{i=1}^{M} \times_i (\mathbf{U}_i \mathbf{U}_i^T) \right\|_F^2$$

Входные данные 3*D* тензора могут быть рассмотрены как $X = \{X_{ijk}\}_{i=1\,j=1\,k=1}^{n1\;n2\;n3}$, где каждый Xi есть 2D матрица (изображение) размерностью n1× n2.

Выводы

Тензор, представлений в виде мультимерного массива, обеспечивает естественное представление трафика КС.

Многомерные компьютерные сети, один из вариантов которых рассматривает каждую отдельную компоненту трафика в отдельном измерении, дают возможность решения задач анализа трафика.

Предложен метод структурирования многомерного трафика, представленного в виде тензора четных рангов. Показана возможность применения предложенных моделей для выявления аномалий трафика.

Список литературы

1. *Мінаєв Ю.М. , Філімонова О.Ю.* Тензорна модель трафика комп′ютерних систем. Сборник трудов конференции Моделирование-2008, том 2. – С. 461–466

2. *Коновалов Г.В.* Многомерные сети – будущее инфокоммуникационных сетей. «Электросвязь», № 4, 2008. – С.28–34

3. *Соколов Н.П.* Пространственные матрицы и их приложения / М.: Физматгиз, 1960. – 352 с.

4. *Муха В.С.* Анализ многомерных данных / Минск.: Технопринт,2004.– 124 с.

5. *Крон Г.* Тензорный анализ сетей: Пер.с англ./Под ред. Л.Т.Кузина, П.Г.Кузнецова. – М.: Сов. Радио, 1978. – 720 с.

6. *Минаев Ю.Н., Филимонова О.Ю., Гузий Н.Н.* Интеллектуальные технологии в системах идентификации и прогнозирования атак на компьютерные сети Электронное моделирование, Т.27. №6. 2005. – С. 37–52.

7. *Mark Sears, Brett Bader, Tammy Kolda*. Parallel Implementation of Tensor Decompositions for Large Data Analysis. - SIAM AN09 July 8, 2009. – P. 17--25.

8. *Brett W. Bader, Tamara G. Kolda*. MATLAB Tensor Classes for Fast Algorithm Prototyping. - SANDIA REPORT SAND2004-5187 Unlimited Release Printed October, 2004 http://www.ntis.gov/ordering.htm

9. *Kolda, T.G. Jimeng Sun*. Scalable Tensor Decompositions for Multi-aspect Data Mining Data Mining, 2008. ICDM '08. Eighth IEEE International Conference on. P. 363–372

10. *Evrim Acar Daniel M. Dunlavy Tamara G. Kolda* Link Prediction on Evolving Data using Matrix and Tensor Factorizations.http://www.csmr.ca.sandia. gov/~tgkolda/ref/.

11. J*imeng Sun, Dacheng Tao, Christos Faloutsos*. Beyond Streams and Graphs: Dynamic Tensor Analysis. 2006 ACM 1595933395/ 06/0008.- http://www.cs.cmn.edu/~christos/PROJECTS/GRAPH-MINING/.

12. *Акивис М. А., Гольдберг В. В.* Тензорное исчисление: Учеб. пособие. - 3-е изд., перераб. - М.: Физматлит, 2003. – 304 с.

13. *Мінаєв Ю.Н., Філімонова О.Ю.* Інтелектуальні технології прогнозування часових рядів на підставі тензорних інваріантів //Зб. наук. праць " Проблеми інформатизації та управління" , №2(26), 2009. –С.104–112

Подано до редакції 01.10.2010