Ye. Tykhonov, V. Zhebka, A. Bondarchuk

USE OF STATISTICAL AND ANALYTICAL METHODS FOR SOLVING PROBLEMS OF «BIG DATA»

The analysis of large data is increasingly becoming a popular practice, which is accepted by many organizations to generate valuable information from large volumes of data. Many data represent opportunities, and some challenges for statisticians. After all, using the data we deal with many problems, for example, expensive equipment, many non-structured data, which prevents us from quickly finding valuable information, speed, as we work with billions of gigabytes. From the point of view of data processing in regression analysis, the key operations are the calculation of the current total difference and the adjustment of parameter values. If the first operation is parallelized in an obvious way, then the second is more complicated. In the most general case, when adjusting weights, a well-known mathematical fact is used: the function of several parameters increases in the direction of the gradient and decreases in the direction opposite to the gradient. In turn, the calculation of the gradient consists in the calculation of the partial derivatives of the function for each of the parameters, which is reduced to discrete differentiation based on the calculation of weighted sums. As a result, the adjustment of parameter values is also reduced to summation, which can be parallelized. The problem of Big Data clustering is that the existing algorithms imply the possibility of directly referring to any information entity in the source data. In turn, the source data can be distributed across different servers, and it is not guaranteed that each cluster is stored strictly on one server. If the distribution of data across servers is made transparent to the clustering algorithm, he believes that the data is located in some distributed virtual memory, then this will inevitably lead to copying large amounts from one server to another. This article will discuss statistical and analytical methods to combat «bad» data.

Keywords: large data (big data); analytics; analysis of big data; sampling; statistical methods.

УДК 621.398.96

Ю. В. МЕЛЬНИК, канд. техн. наук, ст. наук. співробітник;
С. В. ПАНАДІЙ;
В. І. КОРСУН, аспірант,
Державний університет телекомунікацій, Київ

ЗАБЕЗПЕЧЕННЯ ЦІЛІСНОСТІ ІНФОРМАЦІЇ В МУЛЬТИСЕРВІСНИХ МЕРЕЖАХ ЗВ'ЯЗКУ

Розроблено метод забезпечення цілісності та доступності інформації на базі технологій мережного рівня мультисервісних мереж зв'язку. Запропоновано правило прийняття рішення щодо оцінювання цілісності інформації.

Ключові слова: імовірність модифікації; доступність інформації; імовірність цілісності інформації.

Вступ

Одним зі шляхів забезпечення комплексного захисту інформації без зниження QoS є використання ресурсів мультисервісної мережі зв'язку (ММЗ). Користувачеві при цьому достатньо визначитись із профілем захисту — кількісними оцінками конфіденційності, цілісності та доступності інформації. Система керування згідно з результатами моніторингу вільних ресурсів ММЗ реалізує не тільки з'єднання, яке підтримує QoS для даного додатка, а й заявлений користувачем профіль вимог. Втілення в життя зазначеного підходу цілком можливе за рахунок протоколів маршрутизації і сигналізації.

Основна частина

Здійснення паралельного передавання та обробки інформації в точці прийому є одним з ефективних методів, що забезпечують надійність обчислювальних і телекомунікаційних систем [1–3]. Застосуємо такий підхід для забезпечення цілісності інформації з підтриманням показників QoS високошвидкісних додатків мереж зв'язку, що функціонують у реальному масштабі часу.

Нехай у мережі між вузлом-джерелом (ВД) і вузлом-отримувачем (ВО) передається повідомлення, що являє собою потік бітів $M = \{M_1, M_2\}$ із відповідними апріорними ймовірностями $P(M_1)$ і $P(M_2)$.

Повідомлення передається від ВД до ВО по n паралельних з'єднаннях через m транзитних вузлів (ТВ) у кожному з'єднанні (рис. 1).

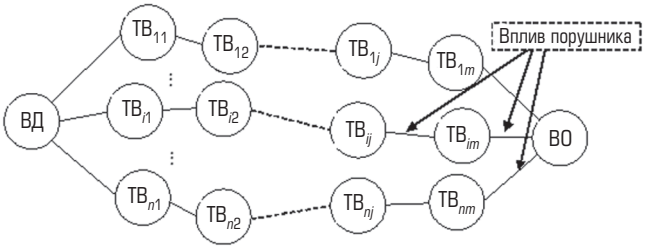

Рис. 1. Організація паралельних з'єднань

Нехай $P_M^{(i)}$ — імовірність модифікації повідомлення внаслідок атаки порушника у відповідному i-му з'єднанні $\left(i = \overline{1,n}\right)$. У цьому разі цілісність інформації досягається за рахунок прийняття рішення у ВО за n прийнятими символами. У результаті значення M^* на виході вирішувального

пристрою (ВП) відповідатиме переданому значенню M_1 або M_2.

Умовні ймовірності прийняття рішення на користь M_1 або M_2 визначаються відповідно як [4]

$$P\left(M_1/\left(x_i;i=\overline{1,n}\right)\right)=\frac{P(M_1)\left\{\prod\limits_{i;x_i=M_1}\left(1-P_M^{(i)}\right)\cdot\prod\limits_{i;x_i=M_2}P_M^{(i)}\right\}}{P\left(x_i;i=\overline{1,n}\right)},$$

$$P\left(M_2/\left(x_i;i=\overline{1,n}\right)\right)=\frac{P(M_2)\left\{\prod\limits_{i;x_i=M_2}\left(1-P_M^{(i)}\right)\cdot\prod\limits_{i;x_i=M_1}P_M^{(i)}\right\}}{P\left(x_i;i=\overline{1,n}\right)}.$$

Візьмемо відношення цих виразів. Якщо результат виявиться більшим за одиницю, то приймаємо рішення на користь M_1, в іншому випадку — на користь M_2. Після логарифмування відношення і виконання деяких перетворень дістанемо вираз

$$\ln\frac{P\left\{M_1/\left(x_i;i=\overline{1,n}\right)\right\}}{P\left\{M_2/\left(x_i;i=\overline{1,n}\right)\right\}}=a_0+\sum_{i=1}^{n}x_i\cdot a_i, \qquad (1)$$

де

$$a_0=\ln\frac{P(M_1)}{P(M_2)};\ a_i=\ln\frac{1-P_M^{(i)}}{P_M^{(i)}}.$$

Отже, справджується таке правило прийняття рішення [2]:

$$a_0+\sum_{i=1}^{n}x_i\cdot a_i\begin{cases}>0\Rightarrow M^*=M_1,\\<0\Rightarrow M^*=M_2.\end{cases} \qquad (2)$$

Функціональну схему ВП наведено на рис. 2.

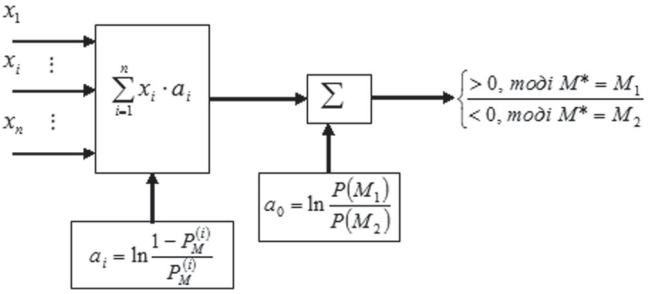

Рис. 2. Функціональна схема вирішувального пристрою

Для оцінювання ймовірності цілісності інформації в мережі впровадимо такі обмеження:

- імовірності модифікації $M = \{M_1, M_2\}$ по всіх з'єднаннях між ВД і ВО рівні між собою, тобто $P_M = P_M^{(i)};\ i=\overline{1,n}$, і незалежні (див. рис. 1);
- кількість n паралельних з'єднань між ВД і ВО непарна, причому $n \geq 3$.

Тоді ймовірність цілісності інформації (див. рис. 2) визначається виразом [2]:

$$P_{\text{цВП}}=1-\sum_{i=0}^{(m-1)\cdot 2}C_n^{(n+1+2i)/2}\cdot\left(1-P_M\right)^{(n-1-2i)/2}\cdot P_M^{(n+1+2i)/2}, \qquad (3)$$

де $C_n^{(n+1+2i)/2}$ — кількість сполучень із n по $(n+1+2i)/2$.

Результати оцінювання цілісності інформації, розраховані за формулою (3), наведено на рис. 3.

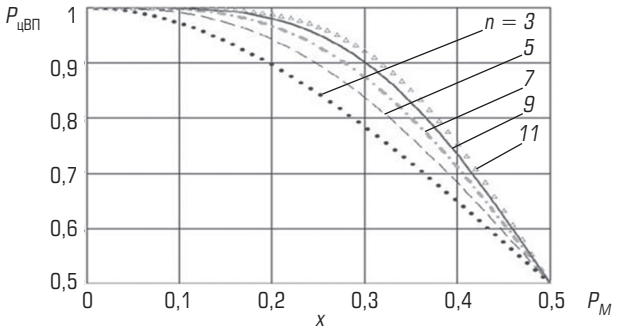

Рис. 3. Результати теоретичного розрахунку $P_{\text{цВП}} = f(P)$ для різних значень n

Для уникнення фінансових та організаційних труднощів при перевірці функціонування ВП (див. рис. 2), що реалізує алгоритм (2), на діючій мережі зв'язку у разі підтвердження теоретичних результатів оцінювання ймовірності цілісності інформації на виході ВП доцільно скористатися методом статистичного моделювання [4].

Вихідні дані алгоритму моделювання такі:

- $P(M_1)$ і $P(M_2)$ — апріорні ймовірності появи $M=\{M_1, M_2\}$ на виході ВД за умови $P(M_1)+P(M_2)=1$;
- n — кількість з'єднань між ВД і ВО;
- $P_M = P_M^{(i)};\ i=\overline{1,n}$ — імовірності модифікації бітового потоку $M = \{M_1, M_2\}$ на виході ВД з n з'єднаннями між ВД і ВО;
- N_0 — кількість переданих значень $M = \{M_1, M_2\}$ між ВД і ВО (кількість незалежних випробувань при статистичному моделюванні).

Здійснюємо N_0 випробувань, кожне з яких складається з п'яти етапів.

На ***першому етапі*** формується випадковий бітовий потік $M = \{M_1, M_2\}$ за правилом:

$$M=\begin{cases}+1, \text{ якщо } z_k\leq P(M_1);\\-1, \text{ якщо } z_k>P(M_1),\end{cases}$$

де z_k — випадкове число, яке генерує датчик випадкових чисел із рівномірним законом розподілу $0\leq z_k\leq 1;\ k=\overline{1,N_0}$.

На ***другому етапі*** модифікуються значення $M = \{M_1, M_2\}$ у кожному з n встановлених паралельних з'єднань між ВД і ВО:

$$x_i=\begin{cases}\text{якщо } z_k\leq P_M, \text{ то модіфікація є, } x_i=M\cdot(-1);\\\text{якщо } z_k>P_M, \text{ то модіфікації нема, } x_i=M.\end{cases}$$

На ***третьому етапі*** розраховуються коефіцієнти $a_0=\ln\frac{P(M_1)}{P(M_2)};\ a_i=\ln\frac{1-P_M^{(i)}}{P_M^{(i)}}$ і приймається рішення за правилом (2).

На ***четвертому етапі*** перевіряється правильність прийняття рішення за правилом (2) і виконується підрахунок $N+$, тобто кількості правильно прийнятих значень $M = \{M_1, M_2\}$ з N_0 переданих.

На ***п'ятому етапі*** визначається оцінка ймовірності цілісності інформації на виході вирішувального пристрою:

$$P_{\text{цВП}}=\frac{N+}{N_0}.$$

Метод статистичного моделювання є наближеним. Похибка результату обчислення має статистичну природу. Кількісний взаємозв'язок між абсолютною похибкою і кількістю випробувань N_0 визначається як [5]:

$$\Delta_a = N_0^{-0,5}\,\sigma\, t_\beta\,, \qquad (4)$$

де Δ_a — абсолютне значення похибки (половина довірчого інтервалу); σ — середньоквадратичне відхилення від $P_{цВП}$; β — достовірність отримуваної оцінки; t_β — таблична функція, обернена до нормальної при аргументі $(1+\beta)^{-1}$.

Визначимо, скільки необхідно провести випробувань, щоб забезпечити задану абсолютну або відносну похибку обчислення. Із (4) легко дістати шукане значення:

$$N_0 = t_\beta^2\,\sigma^2\Delta_a^{-2}. \qquad (5)$$

Результати оцінювання цілісності інформації на виході ВП (див. рис. 2) методом статистичного моделювання (програмну реалізацію виконано в середовищі MatLab) подано на рис. 4.

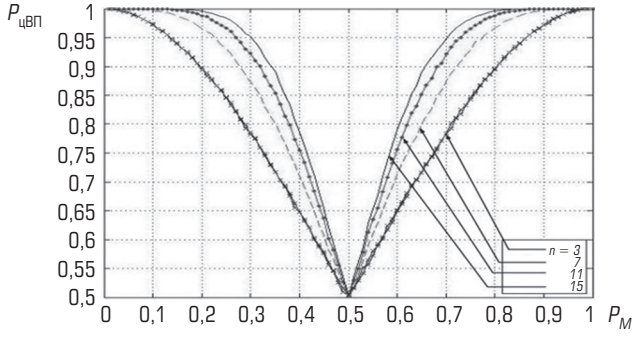

Рис. 4. Результати імітаційного моделювання роботи ВП $P_{ц\,ВП} = f(P_M)$ для різних значень n

Результати імітаційного моделювання підтверджують теоретичні розрахунки за формулою (3).

Резервування каналів зв'язку і дублювання самої інформації є базовими методами забезпечення доступності інформації ТКМ [1; 2]. Це, як правило, реалізується за рахунок організації паралельних з'єднань між ВД і ВО (див. рис. 1). Нехай c_i — вартість i-го з'єднання між ВД і ВО. Тоді загальна вартість організації n паралельних з'єднань буде

$$c_0 = \sum_{i=1}^{n} c_i\,.$$

Якщо вважати дії порушників щодо впливу на кожне з'єднання незалежними подіями, то результуючу ймовірність забезпечення цілісності інформації можна визначити як

$$P_{рез} = 1 - \prod_{i=1}^{n}(1-p_i). \qquad (6)$$

Для забезпечення цілісності інформації за рахунок організації паралельних незалежних з'єднань між ВД і ВО необхідно вибирати ті з'єднання, в яких відношення $\dfrac{\ln(1-p)}{c_i}$ максимальне.

Висновки

Застосування методу інформаційного резервування і резервування елементів інфраструктури дозволяє забезпечити доступність і цілісність інформації в мультисервісних мережах зв'язку з QoS.

Список використаної літератури

*1. **Богатырев В. А.** Надежность двухуровневой отказоустойчивой компьютерной системы при дублировании связей между узлами // Вестн. компьютер. и информ. технологий. 2009. № 1. С. 2–7.*

*2. **Финк Л. М.** Теория передачи дискретных сообщений. 2-е изд., перераб. и доп. Москва: Сов. радио, 1970. 728 с.*

*3. **Хорошевский В. Г.** Архитектура вычислительных систем: учеб. пособие. 2-е изд., перераб. и доп. Москва, 2008. 520 с.*

*4. **Новиков С. Н., Солонская О. И.** Обеспечение целостности в мультисервисных сетях // Докл. ТУСУР. 2009. № 1(19), ч. 2. С. 83–85.*

*5. **Бусленко Н. П.** Моделирование сложных систем. Москва, 1968. 356 с.*

Рецензент: доктор техн. наук, професор **С. В. Козелков,** Державний університет телекомунікацій, Київ.

Ю. В. Мельник, С. В. Панадий, В. И. Корсун

ОБЕСПЕЧЕНИЕ ЦЕЛОСТНОСТИ ИНФОРМАЦИИ НА СЕТЕВОМ УРОВНЕ МУЛЬТИСЕРВИСНЫХ СЕТЕЙ СВЯЗИ

Разработан метод обеспечения целостности и доступности информации на базе технологий сетевого уровня мультисервисных сетей связи. Предложено правило принятия решения по оценке целостности информации.

Ключевые слова: вероятность модификации; доступность информации; вероятность целостности информации.

Yu. V. Melnyk, S. V. Panadii, V. I. Korsun

ENSURING THE INTEGRITY OF INFORMATION AT THE NETWORK LEVEL OF MULTISERVICE COMMUNICATION NETWORKS

In the article the method of ensuring the integrity and availability of information on the basis of technologies of the network level of multiservice communication networks is developed.

The use of the method of parallel transmission and processing of information at the point of reception is shown to ensure the integrity of the information, maintaining the QoS indicators of high-speed applications of communication networks operating in real time.

The decision rule is found on the output of the decoder when modifying the message as a result of the attacker's attack in the appropriate connection.

The functional scheme of the decoding device is presented. Limitations on the effectiveness of information have been introduced.

The calculation formula and results of the evaluation of the integrity of the information are given. It is shown as the confirmation of theoretical results of the estimation of the probability of the integrity of information on the output of the decoder to use the method of statistical simulation. The output data of the simulation algorithm are given.

It is shown that the method of statistical simulation is approximate. For this method, the calculation error is determined.

The number of tests has been determined to provide a given absolute or relative error of calculation. The results of simulation modeling, which confirm theoretical calculations, are presented.

The resulting probability of ensuring the integrity of the information is given, assuming that the influence of offenders on each connection is an independent event.

The rule of decision for estimating the integrity of the information is proposed.

It is shown that the application of the method of information reservation and reservation of infrastructure elements allows to provide the availability and integrity of information in multiservice networks of communication with QoS.

Keywords: probability of modification; availability of information; probability of integrity of information.

УДК 004.055

О. Г. ВАРФОЛОМЕЕВА, канд. техн. наук, ст. науч. сотрудник;
С. Е. МИЛЕНЬКИЙ, С. В. ЛЕЙБОВИЧ, студенты;
Государственный университет телекоммуникаций, Киев

НЕКОТОРЫЕ АСПЕКТЫ ВНЕДРЕНИЯ NGOSS В ДЕЯТЕЛЬНОСТЬ ОПЕРАТОРА ТЕЛЕКОММУНИКАЦИЙ

Информационная компонента бизнес-процесса обеспечивает сотрудников предприятия любой требуемой информацией в режиме реального времени. Эффективность функционирования оператора телекоммуникаций напрямую зависит от того, каким образом он моделировал свои бизнес-процессы. Информационные системы поддержки бизнеса и сопряженных с ним операций являются весьма актуальной и востребованной темой в рамках управления деятельностью оператора и провайдера телекоммуникаций. В статье рассмотрены вопросы, касающиеся эффективного управления деятельностью оператора телекоммуникаций. Предложены модели TNA (технологически нейтральная архитектура) и SOA (сервис-ориентированная архитектура) как механизмы, обеспечивающие взаимодействие между бизнес-процессами и информационной моделью данных. Раскрыты основные преимущества внедрения технологически нейтральной архитектуры в систему управления деятельностью оператора телекоммуникаций.

Сценарии такого внедрения и связанные с ними контракты определяют взаимодействие пользователя информации, содержащейся в SID, и соответствующих процессов eTOM. Все указанные сценарии описываются совокупностью базовых элементов и элементов, специфических для каждого из четырех ракурсов реализации данного сценария: бизнесового, системного, развертывания и внедрения.

Представлен анализ особенностей архитектуры NGOSS и ее основных компонентов. Сформулированы общие подходы к внедрению методологии NGOSS в управление деятельностью оператора телекоммуникаций.

Ключевые слова: аспекты реализации NGOSS; модели eTOM в деятельности оператора связи; системы OSS; телекоммуникационная сеть; система управления; домен; бизнес-модель; информационная система; архитектура; TNA.

Введение

В основу работы эффективного оператора положен инновационный механизм ведения бизнеса, ориентированный на совершенствование бизнес-процесса, т. е. прежде всего на повышение гибкости и скорости реакции бизнеса на те или иные воздействия. Указанный механизм призван также обеспечить сокращение операционных затрат и улучшение качества обслуживания клиентов. Фундамент деятельности эффективного оператора телекоммуникаций составляют три архитектуры (рис. 1).

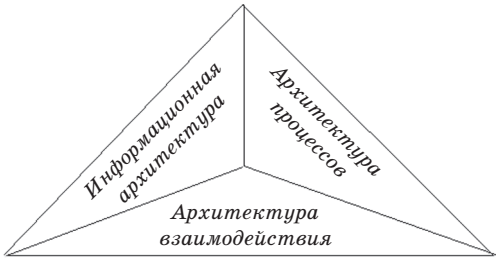

Рис. 1. Базовые архитектуры эффективного оператора

Информационная архитектура предоставляет сотрудникам предприятия информацию в режиме реального времени (on-line). Формируется она при помощи автоматизированных систем управления базами данных и базами знаний, позволяющих обеспечивать удобный интерактивный режим работы с этими базами, а также своевременно и регулярно обновлять хранимую информацию.

Архитектура процессов реализует стратегии и тактики оператора телекоммуникаций для