мобильные платежи — устройство действует как платежная карточка; электронная доска — мобильный телефон используется для чтения RFID меток с уличных досок для объявлений, чтобы на ходу получать информацию. Возможны и другие приложения NFC в недалеком будущем: удостоверение личности; карты путешественника; мобильная торговля; электронные деньги; электронная покупка билетов (авиабилеты, билеты на концерт и др.); электронные ключи — ключи от машины, ключи от дома/офиса, ключи от гостиничного номера и т. д. Существуют варианты криптографии, используемые в метках. Области применения меток рассмотрены на конкретных примерах. В частности, показано использование брелка с NFC меткой «Prestigio PKR1». Изучена проблема защиты информации и предложены способы ее решения. Предложено расширить сферы применения технологии NFC внедрением ее в различные телекоммуникационные системы.

Ключевые слова: технология NFC; метка NFC; технология RFID; QR код; модуляция; протокол; интерфейс; криптография; мобильное устройство NFC; технология Wi-Fi.

V. M. Chorna, O. M. Tkalenko, O. V. Polonevich, O. V. Senkov

FEATURES OF INFORMATION PROTECTION IN NFC

At present, the implementation of wireless technologies in various applications is being observed. They replace advanced technologies and make communication between devices more convenient and easy for the user. Near Filed Communication (NFC) is evolving with technologies such as Wi-Fi, Wi-MAX. This technology is designed to transmit information over short distances. NFC technology is used on mobile devices. It is a logical continuation of RFID technologies (Radio Frequency Identification). NFC supports RFID standards ISO 14443 / mifare, Feli Ca as well as ISO / IEC 18092. Devices can operate in both active and passive modes. The passive mode operates according to the same principles as the non-contact RFID card. This mode increases the autonomy of the portable device and allows you to use the NFC technology even when the power is off. NFC can be used in all cases where contactless cards are used, and card standard compliance allows the use of existing infrastructure. For example, mobile purchase of tickets in public transport is an extension of the existing non-contact infrastructure; mobile payments — the device acts as a payment card; Electronic Board — A mobile phone is used to read RFID tags, from outdoor bulletin boards to get information on the go. Also, other NFC applications in the future may include: Identity; traveler's card; mobile trade; electronic money; electronic purchase of tickets (air tickets, concert tickets, and others); electronic keys — car keys, home / office keys, hotel room keys, etc. There are variants of cryptography that are used in the labels. The scope of the labels is examined on specific examples using the NFC Prestigio PKR1 keychain. The problem of protection of information is studied and ways of its solution are proposed. It is proposed to expand the scope of application of NFC, introducing it into various telecommunication systems.

Keywords: NFC technology; NFC tag; RFID technology; QR code; modulation; protocol; interface; cryptography; NFC mobile device; Wi-Fi technology.

УДК 004.62

Є. С. ТИХОНОВ, аспірант;
В. В. ЖЕБКА, канд. техн. наук;
А. П. БОНДАРЧУК, доктор техн. наук, доцент,
Державний університет телекомунікацій, Київ

ВИКОРИСТАННЯ СТАТИСТИЧНИХ І АНАЛІТИЧНИХ МЕТОДІВ ДЛЯ РОЗВ'ЯЗАННЯ ПРОБЛЕМ «ВЕЛИКИХ ДАНИХ»

Аналіз великих даних дедалі частіше стає популярною практикою, яку впроваджують численні організації, маючи на меті створення цінної інформації з величезних обсягів даних. Великі дані пропонують принципово нові можливості, а також виклики для статистиків. Адже використовуючи дані, ми стикаємося з багатьма супутніми проблемами, такими як висока вартість обладнання, неструктурованість масивів даних, що заважає негайно знаходити потрібну інформацію, блискавична швидкодія — опрацювання мільярдів гігабайт. У статті розглянуто статистичні та аналітичні методи боротьби з «поганими» даними.

Ключові слова: великі дані (big data); аналітика; аналіз великих даних; дискретизація; статистичні методи.

Вступ

Великі дані являють собою ті чи інші відомості в масовому (стосовно обсягу, інтенсивності та складності) масштабі, опрацювання яких перевершує можливості стандартного програмного забезпечення у плані управління та аналізу. Пошукові системи в інтернеті (наприклад, Google і YouTube) та інструменти соціальної мережі (скажімо, Facebook і Twitter) генерують щодня мільярди даних про суспільну активність. Ці дані включають у себе текстовий контент — структурований, напівструктурований або неструктурований, мультимедійний контент (відео, аудіо) на безлічі платформ. Щодня людство виробляє близько 2,5 квантилійних байт даних (2,5 мільярда гігабайт), переважно (до 90%) неструктурованих [1].

Нині це настільки ж актуально, як нанотехнології та квантові обчислення. По суті, Big Data є артефактом людського індивіда, а також колективним інтелектом, який генерується і поширюється переважно через технологічне середовище, де практично все може бути документально зафіксовано, виміряно й сфотографовано в цифровій формі, а отже, за допомогою відповідного процесу перетворено на дані. Величезний обсяг і надвисока швидкість їх обробки можуть призвести до накопичування шумів, фальшивої кореляції та побічної однорідності, що породжує проблеми стосовно обчислювальної техніко-економічної та алгоритмічної стабільності.

Для високоякісної реалізації аспекта великих даних потрібні нове статистичне мислення та спеціальні методи. Щодо поглибленого аналізу даних існують два обчислювальні бар'єри:

♦ по-перше, дані можуть бути занадто великі для зберігання в пам'яті комп'ютера;

♦ по-друге, виконання завдання комп'ютером може тривати настільки довго, що очікувати результатів не доводиться.

Методи аналізу великих даних можна поділити на три групи:

1) описова аналітика;

2) прогнозна аналітика;

3) пресцентивна аналітика.

Основна частина

Як зазначалося в [2], загальні проблеми великих даних можуть бути згруповані на основі життєвого циклу даних за трьома головними категоріями: дані, процеси та проблеми управління (див. рисунок).

• Проблеми з даними пов'язані з характеристиками самих даних (наприклад, обсяг даних, різноманітність, швидкість, правдивість, непостійність, якість, відкриття та догматизм).

• Випробування процесу пов'язані з низкою методів, що з'ясовують, як зафіксувати дані, як інтегрувати дані, як перетворювати дані, як вибрати правильну модель для аналізу та як подати результати.

• Завдання управління включають у себе конфіденційність, безпеку, власне управління та етичні аспекти.

Статистичні методи дозволяють визначити рівняння зв'язку вхідних і вихідних параметрів, проаналізувати параметри технологічного процесу, побудувати математичну модель процесу, або, іншими словами, установити взаємну залежність між різними факторами і технологічними результатами процесу.

Статистичне дослідження охоплює такі питання:

• визначення законів розподілу параметрів процесу, аби з'ясувати можливості застосування тих чи інших статистичних методів обробки результатів;

• визначення тісноти та форми зв'язку між окремими параметрами процесу;

• отримання статистичної моделі процесу у вигляді регресійного рівняння та оцінювання його адекватності;

• визначення динамічних характеристик процесу.

Методологію стосовно великих даних можна поділити на три категорії:

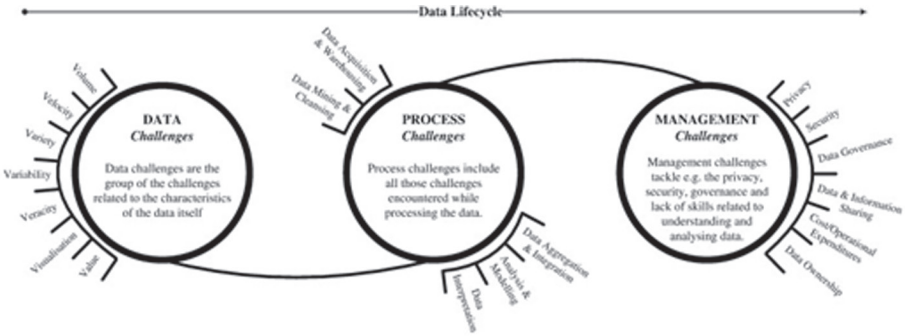
1) передискретизація;

2) поділ та підпорядкування;

3) он-лайн оновлення.

Алгоритми ***передискретизації*** широко застосовуються при обробці звукових сигналів, радіосигналів та зображень. У разі передискретизації відліки сигналу, які відповідають одній частоті дискретизації, обчислюються на основі відомих відліків цього самого сигналу, що відповідають іншій частоті дискретизації. При цьому вважається, що обидві частоти дискретизації відповідають умовам теореми Котельникова [2]. Ідеальна передискретизація еквівалентна відновленню неперервного сигналу за його відліками з подальшою дискретизацією його на новій частоті.

Розглядають і ***субдискретизацію***, передусім як метод посилення впливу [3]. Автори запропонували використати засоби масової інформації для полегшення наукових відкриттів щодо великих даних із використанням обмежених обчислювальних ресурсів. Згідно з методом впливу береться мала частина даних із певними вагами з повної вибірки, а далі виконуються передбачувані обчислення для повної вибірки. При цьому мала добірка використовується як сурогат. Ключ до успіху



Концептуальна класифікація проблем БД

методів посилення впливу полягає в побудові ваг і неоднорідних варіантів вибірки заради того, аби впливові точки даних були відібрані з великою вірогідністю.

Ці точки слугують для досягнення можливих результатів обчислень навіть тоді, коли доступні прості аналітичні розрахунки.

Зазначені точки дозволяють візуалізувати дані, коли неможливо візуалізувати повну вибірку.

Вони зазвичай використовують нерівні ймовірності вибірки для підпрограми даних.

Цей підхід цілком унікальний у наданні повсюдного доступу для отримання інформації з великих даних без використання високопродуктивних обчислень.

Середнє значення логічної вірогідності [4] запропоновано як підхід до стохастичного наближення з використанням передискретизації при залученні великих геостатистичних даних. Метод використовує середні значення Монте-Карло, розраховані із субдискретизацією, для наближеної кількості необхідних значень у разі повних даних. Унаслідок мінімізації дивергенції Кулбака–Лейблера згадані значення наближаються до розбіжності між середніми, розрахованими із субдискретизацією. Це призводить до максимального середнього за методом оцінювання правдоподібності. Розв'язок для рівняння середньої оцінки отримують із процедури стохастичного наближення, де на кожній ітерації поточна оцінка оновлюється на основі з розміром m, сформованим з повних даних. Оскільки m набагато менше за n, метод масштабується для великих даних. Лян [5] установив узгодженість та асимптотичну норму отриманої оцінки в разі м'яких умов. У симуляційному дослідженні швидкість конвергенції методу майже не залежить від n — вибіркового розміру повних даних.

Алгоритм поділу та підпорядкування, як правило, складається з трьох кроків:

1) поділ великого набору даних на K блоків;

2) обробка кожного блока окремо (або паралельно);

3) агрегування рішення кожного блока з метою сформувати остаточне рішення для повних даних.

Он-лайн оновлення потоку даних. У деяких програмах дані потрапляють у потоки або великі фрагменти, і послідовно оновлений аналіз залишається без зберігання даних.

Мотивованість із погляду баєсівського висновку [6] розширює можливості роботи в кількох важливих напрямках.

По-перше, автори вводять оцінку дисперсії параметрів регресії в лінійній моделі та оцінюють параметри рівнянь. Ці оцінки дозволяють користувачам робити висновки про параметри істинної регресії, виходячи з попередньо розроблених точкових оцінок розподілу згаданих параметрів регресії.

По-друге, автори розробляють ітеративні алгоритми оцінювання та подають статистичні висновки для лінійних моделей. Наводять оцінки рівнянь, які оновлюються при отриманні нових даних. Отже, хоча параметр розділення та перемоги добре піддається обробці паралельних даних для кожної підмножини, підхід до оновлення он-лайн для потоків даних, по суті, послідовний за своїм характером. Відповідні алгоритми було розроблено таким чином, що вони є обчислювально-ефективними та мінімальними в зберіганні, оскільки не припускають доступу до історичних даних чи їх зберігання.

По-третє, автори розглядають проблему можливих порушень рангу при роботі з блоками даних і властивостями єдності комбінованих та сукупних оцінок при використанні псевдообернеої матриці. Автори подають також методи оцінювання придатності в налаштуваннях лінійної моделі, оскільки стандартна діагностика на залишковій основі не може бути виконана із сукупними даними без доступу до історичних даних. Замість цього автори пропонують випробування відхилення, що спираються на прогнозовані залишки, котрі базуються на прогностичних значеннях, обчислюваних із сукупної оцінки коефіцієнтів регресії, досягнутих у попередній точці нагромадження. Окрім того, автори вводять нову он-лайн оновлену оцінку коефіцієнтів регресії та відповідну оцінку стандартної помилки в параметрі оціночного рівняння, яке використовує інформацію з попередніх даних.

Існують виклики щодо даних. Це група завдань, пов'язаних із характеристиками самих даних. Різні дослідники мають особливе розуміння характеристик даних.

Наприклад, пропонуються такі групи характеристик даних: 3Vs [обсяг, швидкість та різноманітність]; 4Vs [обсяг, швидкість, різноманітність та мінливість] і 6Vs [обсяг, швидкість, різноманітність, правдивість, мінливість та вартість]. Це можна описати докладно.

• ***Обсяг*** (надвеликі набори даних, вимірювані в тера-, пета-, зеттабайтах, або навіть більші).

Такий масштаб і такий обсяг даних — великий виклик самі по собі. Вони відбивають неоднорідність, повсюдність і динамічний характер різних ресурсів та пристроїв генерування даних. До того ж і кількість самих даних про навколишній світ, що їх визначають, отримують, обробляють, інтегрують і т. ін., неосяжна (наприклад, дані про Всесвіт, бізнес-дані, медичні дані, дані всіляких спостережень). Отже, ідеться про колосальне збільшення великомасштабних даних (Facebook щоденно генерує понад 500 терабайт даних,

а Walmart щогодини збирає більш ніж 2,5 петабайт даних про трансакції своїх клієнтів).

Зрештою висуваються нові виклики щодо технології видобування даних, а це вимагає нових підходів до розв'язання проблеми великих даних [7].

• ***Різноманітність*** (наприклад, кілька форматів даних зі структурованим і неструктурованим текстом / зображенням / мультимедійним вмістом / аудіо / відео / датчиком даних / шуму). Проблеми, пов'язані з різноманітністю даних, також є складним завданням. Величезний обсяг даних не є послідовним і не відповідає конкретному шаблону або формату. Його зафіксовано в різноманітних формах та джерелах. Наприклад, повідомлення (текст, електронна пошта, твіти, блоги) — контент, створений користувачем, трансакційні дані (веб-журнали, бізнес-трансакції), наукові, зокрема експериментальні, дані, веб-дані (зображення, розміщені в соціальних мережах) тощо.

Ці різні форми та нерідко сумнівна якість даних чітко вказують на те, що неоднорідність є природною властивістю Big Data, і це великий виклик для розуміння та управління такими даними.

• ***Правдивість*** — характеристика у великих наборах даних недосяжна (анонімність, неточність, непослідовність). Це стосується не просто якості даних, а передусім їх тлумачення, оскільки практично всі зібрані дані мають невід'ємні відмінності (справжні чи позірні). IBM придумала цю характеристику даних, аби відбити недостовірність, калейдоскопічність, притаманну багатьом джерелам структурованих і неструктурованих даних.

• ***Швидкість*** (лавиноподібне надходження даних із неоднорідною структурою). Виклик, що його адресує нам швидкість, полягає в необхідності керування практично некерованими масивами неоднорідних даних, що призводить до появи нових даних або оновлення тих, що існували досі.

Висновки

♦ Присутність неструктурованих даних у загальному їх наборі створює істотні завади при пошуку цінної інформації.

♦ Використовуючи методи статистичного аналізу, можна усунути зайві дані. Але це не означає, що слід позбавлятись від усіх неструктурованих даних. Потрібно насамперед аналізувати, а потім, використовуючи різні методологічні підходи, знаходити для себе справді важливу інформацію.

♦ Великі дані зобов'язані своєю появою та зміцненням позицій у світі бізнесу потужному потоку цифрової інформації. Значною мірою її надлишок і невміння вправно керувати такою лавиною змусили шукати найбільш раціональних підходів. Велика аналітика має бути забезпечена серйозними й зручними інструментами — як програмними, так і безпосередньо аналітичними.

♦ Незважаючи на радикальні зрушення у сфері машинної обробки інформації, сьогодні, як і в перспективі, неможливо обійтися без фахівців, здатних інтенсивно досліджувати дані та формулювати завдання, зрозумілі щодо алгоритму їх виконання. Пошук і усунення помилок у даних — надзвичайно актуальна проблема, яку вирішують справжні професіонали в зазначеній синтетичній галузі наукової практики.

Список використаної літератури

*1. **Добре С., Кхафа Ф.** Інтелектуальні послуги для великих наукових даних // Комп'ютерні системи майбутнього покоління. 2014. Т. 37. С. 267–281.*

*2. **Akerkar R.** Big data computing // CRC Press, Taylor & Francis Group, Florida, USA (2014).*

*3. **Котельников В. А.** О пропускной способности эфира и проволоки в электросвязи // Всесоюз. энергетический комитет: материалы I Всесоюзного съезда по вопросам технической реконструкции дела связи и развития слаботочной промышленности, 1933.*

*4. **Ma P., Sun X.** Leveraging for Big Data Regression // WIREs Computational Statistics. 2014. С. 70–76.*

*5. **A Resampling-Based** Stochastic Approximation Method for Analysis of Large Geostatistical Data / F. Liang, Y. Cheng, Q. Song a. o. // Journal of the American Statistical Association. 2013. Т. 108. С. 325–339.*

*6. **Online** Updating of Statistical Inference in the Big Data Setting / E. D. Schifano, J. Wu, C. Wang a. o. // Technometrics. 2015.*

*7. **Massively** parallel feature selection: an approach based on variance preservation / Z. Zhao, R. Zhang, J. Cox a. o. // Machine Learning. 2013. Т. 92. P. 195–220.*

Рецензент: доктор техн. наук, доцент **В. В. Онищенко**, Державний університет телекомунікацій, Київ.

Е. С. Тихонов, В. В. Жебка, А. П. Бондарчук

ИСПОЛЬЗОВАНИЕ СТАТИСТИЧЕСКИХ И АНАЛИТИЧЕСКИХ МЕТОДОВ ДЛЯ РЕШЕНИЯ ПРОБЛЕМ «БОЛЬШИХ ДАННЫХ»

Анализ больших данных все чаще становится популярной практикой, которую внедряют многие организации с целью создания ценной информации из огромных объемов данных. Большие данные предоставляют принципиально новые возможности, а также вызовы для статистиков. Ведь используя данные, мы сталкиваемся со многими проблемами: дорогостоящее оборудование, масса неструктурированных данных, которые мешают оперативно находить ценную информацию, небывалое быстродействие, измеряемое миллиардами гигабайт в секунду. В статье рассмотрены статистические и аналитические методы борьбы с «плохими» данными.

Ключевые слова: большие данные (big data); аналитика; анализ больших данных; дискретизации; статистические методы.

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).

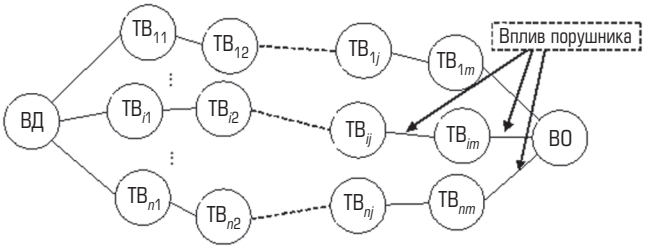


Рис. 1. Організація паралельних з'єднань

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