III МЕЖДУНАРОДНАЯ КОНФЕРЕНЦИЯ
«УСТОЙЧИВОСТЬ И ПРОЦЕССЫ УПРАВЛЕНИЯ»

СПбГУ 5-9 октября 2015

ОПЕРАЦИОННАЯ И ИНФОРМАЦИОННАЯ ЧУВСТВИТЕЛЬНОСТЬ: НОВЫЕ АКТУАЛЬНЫЕ ОЦЕНКИ КАЧЕСТВА КОМПЬЮТЕРНЫХ АЛГОРИТМОВ

М. В. Ульянов

д-р. техн. наук, проф., в.н.с.

Институт проблем управления им В.А. Трапезникова РАН, г. Москва,

ВМК МГУ

muljanov@mail.ru

АКТУАЛЬНОСТЬ

Методики выбора алгоритма, рационального для данной области применения актуальны при разработке алгоритмического обеспечения программных средств.

Ряд областей (задачи реального времени, распределённые вычисления) предъявляют жёсткие требования к алгоритмам по временной устойчивости. Для многопроцессорных систем это равносильно требованию балансировки загрузки процессоров. Актуальным является построение такой оценки качества алгоритма, которая отражала бы эти требования.

Очевидно, что качество временных прогнозов по среднему во многом определяется тем, насколько велико влияние различных входов фиксированной длины на трудоёмкость, т. е. насколько велика информационная чувствительность исследуемого алгоритма.

В связи с этим, в аспекте проблематики прогнозирования временной эффективности программных реализаций алгоритмов и оценки их стабильности по времени, представляет интерес задача введения количественной оценки качества алгоритма, отражающей его операционную стабильность.

ТЕРМИНОЛОГИЯ И ОБОЗНАЧЕНИЯ

D — вход алгоритма A: конечное множество слов фиксированной длины в бинарном алфавите, задающее конкретную решаемую задачу;

$n = \lambda(D)$ — длина входа алгоритма: $D \underset{\lambda}{\to} N$, целочисленная функция, в общем случае определяемая как функция от мощности множества D: $\lambda(D) = \lambda(|D|) = n$;

$f_A(D)$ — трудоёмкость алгоритма A на входе D — целочисленная функция, значение которой есть число базовых операций (в принятой модели вычислений), заданных алгоритмом A на входе D;

$D_n = \{D \mid \lambda(D) = n\}$ — множество всех входов алгоритма A, имеющих длину n;

$f_A^{\wedge}(n)$ — трудоёмкость алгоритма в худшем случае на всех допустимых входах длины n, т. е. максимум $f_A(D)$ на множестве D_n;

$f_A^{\vee}(n)$ — трудоёмкость алгоритма в лучшем случае, минимум $f_A(D)$ на множестве D_n.

При этом для всех классов алгоритмов всегда выполнено: $f_A^{\vee}(n) \le f_A(D \in D_n) \le f_A^{\wedge}(n)$.

ИНФОРМАЦИОННАЯ ЧУВСТВИТЕЛЬНОСТЬ АЛГОРИТМОВ

Понятие «информационная чувствительность» отражает тот факт, что на разных входах D, имеющих фиксированную длину n, алгоритм задаёт различное число базовых операций принятой модели вычислений. Разброс значений трудоёмкости алгоритма на различных входах фиксированной длины означает его ненулевую информационную чувствительность. В общем случае причиной такой вариации является влияние значений элементов входа и других содержательных особенностей на число базовых операций.

Теоретические границы такого разброса на входах длины n задаются значениями функций $f_A^{\wedge}(n)$ и $f_A^{\vee}(n)$. Размах варьирования:

$$R(n) = f_A^{\wedge}(n) - f_A^{\vee}(n),$$

Рассматривая трудоёмкость алгоритма на фиксированной длине входа как дискретную ограниченную случайную величину, можно либо теоретически обосновать задаваемый алгоритмом закон распределения, либо на основе экспериментальных данных вычислить статистические оценки наблюдаемого распределения и предложить некоторую аппроксимацию, удовлетворяющую одному из критериев согласия.

КВАНТИЛЬНАЯ ОЦЕНКА ИНФОРМАЦИОННОЙ ЧУВСТВИТЕЛЬНОСТИ (I)

Идея рассмотрения трудоёмкости алгоритма при фиксированной длине входа как ограниченной случайной величины, аппроксимируемой некоторой известной функцией плотности распределения вероятностей, привела к введению понятия доверительной трудоёмкости на основе вычисления γ-квантиля аппроксимирующего закона распределения. На основе этого подхода была предложена и квантильная оценка информационной чувствительности алгоритмов. Основная идея состоит в определении длины сегмента нормированных значений трудоёмкости, по которому интеграл от функции плотности равен заданной вероятности (надёжности) γ.

$$x_f = \frac{f - f_A^{\vee}(n)}{f_A^{\wedge}(n) - f_A^{\vee}(n)} = \frac{f - f_A^{\vee}(n)}{R(n)}.$$

Содержательно такая количественная оценка с обозначением $\delta_{IQ}(\gamma)$ есть доля теоретического сегмента варьирования трудоёмкости, в которой с заданной вероятностью γ будут наблюдаться значения трудоёмкости алгоритма на произвольных входах фиксированной длины.

КВАНТИЛЬНАЯ ОЦЕНКА ИНФОРМАЦИОННОЙ ЧУВСТВИТЕЛЬНОСТИ (II)

Очевидно, что для практического применения и теоретического исследования алгоритма необходимо рассматривать $\delta_{IQ}(\gamma)$ не только как функцию вероятности γ, но и как функцию длины входа n, т. е. $\delta_{IQ} = \delta_{IQ}(\gamma, n)$.

Пусть $\Im = \{f(n,x)\}$, $x \in [0,1]$ есть семейство непрерывных функций плотности распределения, аппроксимирующих нормированные значения трудоёмкости, параметризированное аргументом n — длиной входа алгоритма. Зафиксируем некоторое значение длины входа n, тогда для любой непрерывной, не обращающейся в нуль на интервалах, функции плотности распределения вероятностей $f(n,x)$ из семейства $\Im$ соответствующая интегральная функция распределения вероятностей $F(n,x)$ является монотонно возрастающей, в силу чего имеет обратную функцию. Обозначим через $F^{-1}(n,x)$ функцию обратную к $F(n,x)$, тогда

$$\delta_{IQ}(\gamma, n) = F^{-1}\left(n, \frac{1}{2} + \frac{\gamma}{2}\right) - F^{-1}\left(n, \frac{1}{2} - \frac{\gamma}{2}\right).$$

КВАНТИЛЬНАЯ ОЦЕНКА ИНФОРМАЦИОННОЙ ЧУВСТВИТЕЛЬНОСТИ (III)

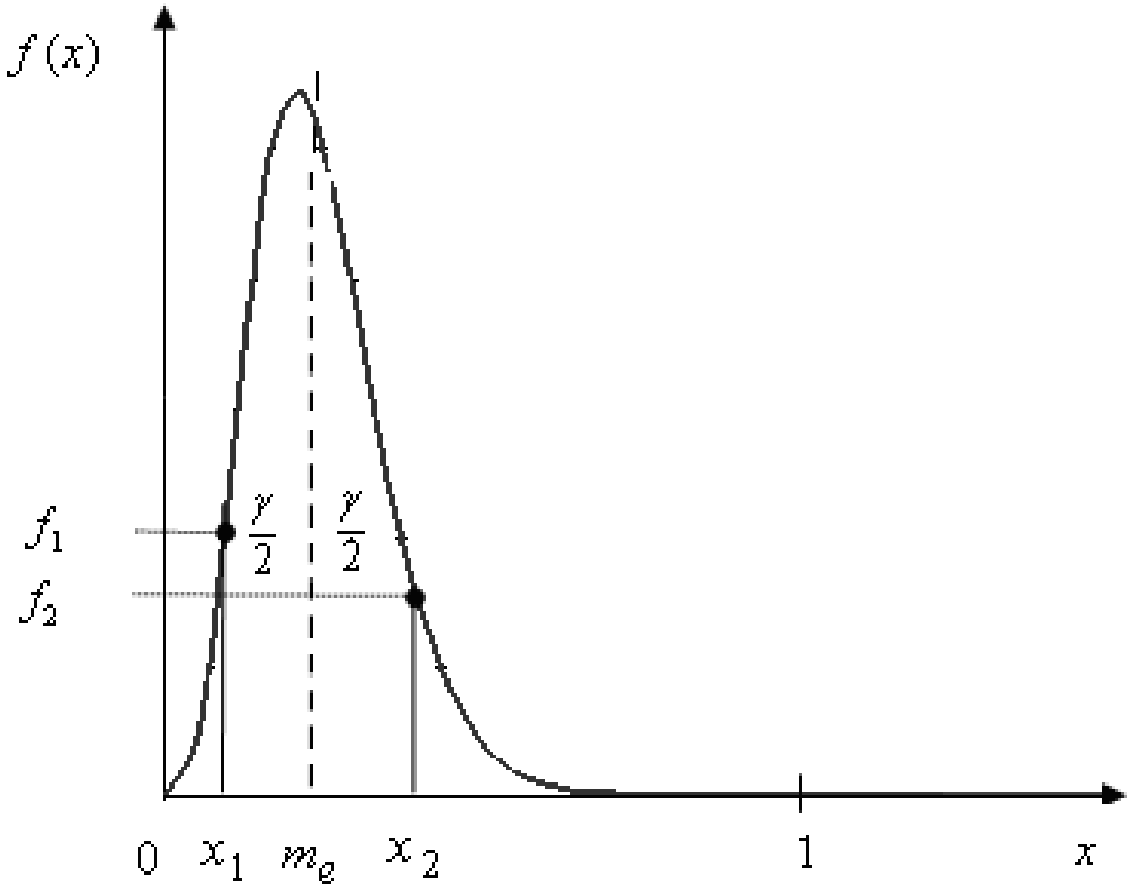

Рис.1. Квантильная количественная мера информационной чувствительности (в обозначениях данного рисунка $\delta_{IQ}(\gamma, n) = x_2 - x_1$).

КОЛИЧЕСТВЕННАЯ ОЦЕНКА ОПЕРАЦИОННОЙ ЧУВСТВИТЕЛЬНОСТИ (I)

Для разработчиков представляет интерес не только собственно длина сегмента, но и его положение, и зависимость этих значений от длины входа. Модифицированная квантильная оценка информационной чувствительности имеет вид:

$$\delta_{IQ}{}^{*}(\gamma, n) = \left(x_{\gamma}^{(1)}(n),\ x_{\gamma}^{(2)}(n),\ x_{\gamma}^{(2)}(n) - x_{\gamma}^{(1)}(n)\right),$$

и задаётся тремя числами: значениями левой и правой границ нормированного сегмента, по которому интегрируется заданная вероятность γ, и длиной этого сегмента.

Более актуальной является информация, представленная не в относительных, а в абсолютных единицах — чувствительность алгоритма в базовых операциях модели вычислений, на основе которой можно перейти к оценке разброса времени выполнения (и оценке стабильности по времени программной реализации алгоритма).

Операционная чувствительность на фиксированной длине входа должна показывать границы изменения трудоёмкости алгоритма при заданной исследователем доверительной вероятности и изменение этих границ по длине входа.

КОЛИЧЕСТВЕННАЯ ОЦЕНКА ОПЕРАЦИОННОЙ ЧУВСТВИТЕЛЬНОСТИ (II)

Предлагается использовать оценку $\delta_{IQ}^{*}(\gamma, n)$, прейдя в ней к абсолютным значениям. Обращение формулы для $\delta_{IQ}^{*}(\gamma, n)$ приводит к тому, что значению $x_{\gamma}^{(1)}(n)$ соответствует $f_A^{\vee}(n) + x_{\gamma}^{(1)}(n) \cdot R(n)$, значению $x_{\gamma}^{(2)}(n)$ — $f_A^{\vee}(n) + x_{\gamma}^{(2)}(n) \cdot R(n)$, а длине нормированного сегмента $x_{\gamma}^{(2)}(n) - x_{\gamma}^{(1)}(n)$ — сегмент значений трудоёмкости, связанный с информационной чувствительностью $\delta_{IQ}(\gamma, n)$ соотношением $\delta_{IQ}(\gamma, n) \cdot R(n)$.

Таким образом, вводится оценка операционной чувствительности с обозначением $\delta_{OP}(\gamma, n)$ на основе модифицированной квантильной оценки информационной чувствительности и размаха варьирования:

$$\delta_{OP}(\gamma, n) = \left(f_A^{\vee}(n) + x_{\gamma}^{(1)}(n) \cdot R(n),\ f_A^{\vee}(n) + x_{\gamma}^{(2)}(n) \cdot R(n),\ \delta_{IQ}(\gamma, n) \cdot R(n) \right).$$

Более корректно $\delta_{OP}(\gamma, n)$ есть функционал, отображающий алгоритм и множество входов фиксированной длины в три числовых значения, т.е. отображающий пару (A, D_n) в R^3 (A — алгоритм, D_n — множество входов фиксированной длины) с функциональной зависимостью по n.

МЕТОДИКА ОПРЕДЕЛЕНИЯ ОПЕРАЦИОННОЙ ЧУВСТВИТЕЛЬНОСТИ

Непосредственное применение формулы для $\delta_{OP}(\gamma, n)$ требует построения аппроксимирующей функции плотности для значений трудоёмкости. Более простой метод:

1. Фиксация длины входа и проведение экспериментального исследования программной реализации алгоритма для определения числа операций. Рекомендуемое число экспериментов не менее 10 000.

2. Построение гистограммы относительных частот полученных значений трудоёмкости. При $m = 10\,000$ рекомендуется 100 полусегментов гистограммы.

3. Задание доверительной вероятности γ для операционной чувствительности и определение значения $\alpha = (1-\gamma)/2$.

4. Суммирование относительных частот в левых и правых полусегментах гистограммы вплоть до достижения полученного значения α. Тем самым, мы отбрасываем левый и правый хвосты распределения, имеющие вероятность α. Полученные границы и являются оценками границ в $\delta_{OP}(\gamma, n)$.

МОДЕЛЬНЫЙ ПРИМЕР (I)

Проиллюстрируем предложенную оценку на основе исследования алгоритма Кнута-Морриса-Пратта для поиска подстроки в строке. Алгоритм использует префиксную функцию для предобработки строки поиска. Теоретический анализ показывает, что трудоёмкость в лучшем случае имеет вид $f_A^{\vee}(n,m) = 14n + 17m - 7$, в худшем случае — $f_A^{\wedge}(n,m) = 24n + 17m - 27$, где n — длина строки, m — длина подстроки. Заметим, что функция трудоёмкости для этого алгоритма имеет два аргумента. Для значений $n = 10\,000$, $m = 20$ получаем теоретический минимум и максимум трудоёмкости для однократного вхождения:

$$f_A^{\vee}(10\,000,\, 20) = 140\,333, \quad f_A^{\wedge}(10\,000,\, 20) = 240\,313.$$

Исследование выполнено с генерацией случайных входов с подстрокой однократного вхождения. Гистограмма относительных частот W значений трудоёмкости построена по 20 000 экспериментам. Поведение значений трудоёмкости как ограниченной случайной величины имеет для этого алгоритма и данных параметров входа ярко выраженную левую асимметрию.

МОДЕЛЬНЫЙ ПРИМЕР (II)

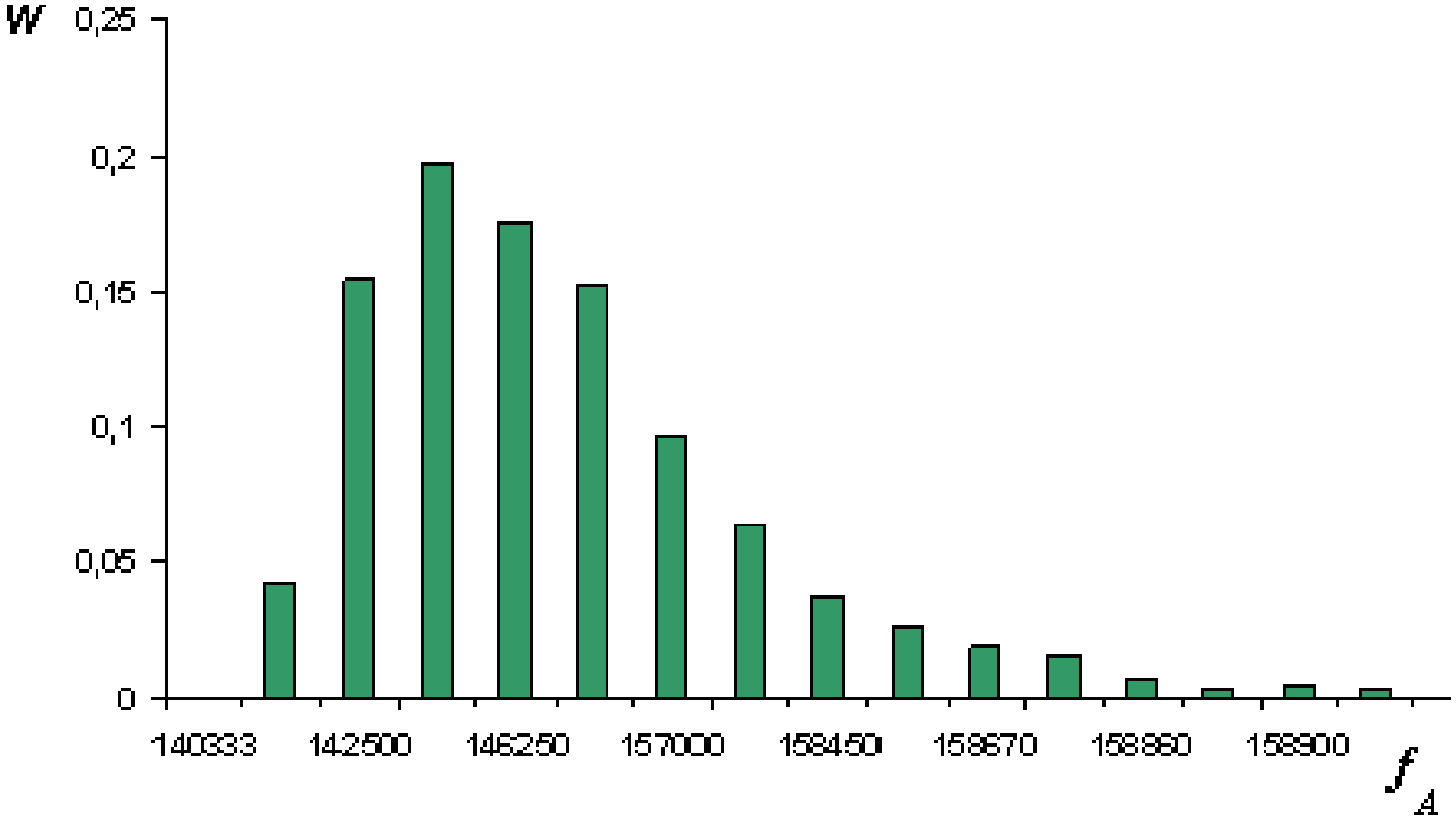

Рис 2. Ненулевая часть гистограммы относительных частот значений функции трудоёмкости для алгоритма Кнута-Морриса-Пратта ($n = 10\,000$, $m = 20$).

МОДЕЛЬНЫЙ ПРИМЕР (III)

Нормирование значений трудоёмкости в сегмент $[0,1]$ проведено по теоретическим границам. Была выдвинута гипотеза H_0 об аппроксимации гистограммы нормированных относительных частот функцией плотности бета-распределения. Методом моментов определены параметры бета-распределения: $\alpha = 1{,}42;\ \beta = 18{,}47$. Для вероятности $\gamma = 0{,}95$ вычислено значение $\delta_{IQ}(\gamma) = 0{,}2104$, при этом границы сегмента, соответствующего $1/2 \pm \gamma/2$-квантилям распределения, оказались равными: $x_1 = 0{,}0049$, $x_2 = 0{,}2153$. Операционная чувствительность:

$$\delta_{OP}(0.95,\ 10\ 000,\ 20) = (140\ 823,\ 163\ 349,\ 21\ 526).$$

Таким образом, для данного модельного примера операционная чувствительность при доверительной вероятности $\gamma = 0{,}95$ составляет 21 526 базовых операций (менее 22 % размаха варьирования). Положение сегмента операционной чувствительности близко к теоретическому лучшему случаю. Переходя от обобщённых операций к временным оценкам можно получить значения временной чувствительности.

Библиографический список

1. **Петрушин В.Н, Ульянов М.В.** Информационная чувствительность компьютерных алгоритмов. — М.: ФИЗМАТЛИТ, 2010. — 224 с.

2. **Ульянов М. В.** Ресурсно-эффективные компьютерные алгоритмы. Разработка и анализ. М.: ФИЗМАТЛИТ, 2008. 304 с.

3. **Ульянов М. В., Петрушин В. Н., Кривенцов А. С.** Доверительная трудоёмкость — новая оценка качества алгоритмов // Информационные технологии и вычислительные системы. 2009. №2. С.23–37.

4. **Головешкин В.А., Ульянов М.В., Выборнов А. Н** Операционная чувствительность алгоритмов // Автоматизация и современные технологии. 2015. № 8. С. 41–46.

БЛАГОДАРЮ ЗА ВНИМАНИЕ!