Информатика и её применения

Том 8 Выпуск 2 Год 2014

СОДЕРЖАНИЕ

Технический редактор *Л. Кокушкина* Художественный редактор *М. Седакова*
Сдано в набор 20.04.14. Подписано в печать 25.06.14. Формат 60 x 84 / 8
Бумага офсетная. Печать цифровая. Усл.-печ. л. 17,9. Уч.-изд. л. 16,0. Тираж 100 экз.
Заказ № 1059
Издательство «ТОРУС ПРЕСС», Москва 121614, ул. Крылатская, 29-1-43
Отпечатано в Академиздатцентре «Наука» РАН с готовых файлов
Москва 121099, Шубинский пер., д. 6.

ИНФОРМАТИКА И ЕЁ ПРИМЕНЕНИЯ, 2014. Т. 8. Вып. 2. С. 2–14

АНАЛИТИЧЕСКОЕ МОДЕЛИРОВАНИЕ РАСПРЕДЕЛЕНИЙ С ИНВАРИАНТНОЙ МЕРОЙ В НЕГАУССОВСКИХ ДИФФЕРЕНЦИАЛЬНЫХ И ПРИВОДИМЫХ К НИМ ЭРЕДИТАРНЫХ СТОХАСТИЧЕСКИХ СИСТЕМАХ*

И. Н. Синицын[1]

Аннотация: Представлены методы и алгоритмы аналитического моделирования одно- и многомерных распределений с инвариантной мерой в дифференциальных и интегродифференциальных (эредитарных) стохастических системах (СтС), описываемых уравнениями Ито в конечномерных пространствах с винеровскими и пуассоновскими шумами. Сначала в разд. 2 рассматриваются интегродифференциальные уравнения Пугачева для распределений процессов в дифференциальных СтС (ДСтС). Применительно к ДСтС с гладкими и разрывными регулярными правыми частями найдены условия сохранения инвариантной меры для нестационарных и стационарных процессов. Сформулированы 4 теоремы, определяющие точные алгоритмы аналитического моделирования распределений с инвариантной мерой в ДСтС общего вида. В разд. 3 дан краткий обзор приближенных методов аналитического моделирования в ДСтС, основанных на параметризации распределений. Особое внимание уделено методам нормальной аппроксимации и статистической линеаризации для приближенного определения одно- и двумерных распределений с инвариантной мерой. Получены условия устойчивости алгоритмов. Сформулированы две теоремы, определяющие приближенные алгоритмы аналитического моделирования в ДСтС. Раздел 4 посвящен методам и алгоритмам аналитического моделирования распределений с инвариантной мерой в интегродифференциальных эредитарных СтС (ЭСтС), приводимых к дифференциальным. Представлены нелинейные стохастические интегродифференциальные уравнения Ито с винеровскими и пуассоновскими шумами. Для затухающих физически возможных эредитарных ядер рассматривается два способа их аппроксимации (на основе линейных операторных уравнений и вырожденных ядер). Рассмотрены три теоремы, определяющие точные и приближенные алгоритмы приведения ЭСтС к ДСтС для гладких и разрывных регулярных правых частей. В приложении даны тестовые примеры для разрабатываемого в ИПИ РАН инструментального программного обеспечения «ID StS» в среде MATLAB. Заключение содержит основные выводы и возможные обобщения. Рассмотрено применение результатов разд. 2–4 к задачам эквивалентности гауссовских и негауссовских ДСтС и ЭСтС.

Ключевые слова: аналитическое моделирование; гауссовская (нормальная) стохастическая система; дифференциальная стохастическая система; инструментальное программное обеспечение «ID StS»; метод нормальной аппроксимации; метод статистической линеаризации; негауссовская (с винеровскими и пуассоновскими шумами) стохастическая система; распределение с инвариантной мерой; сингулярное (вырожденное) ядро; стохастическое дифференциальное уравнение Ито; система, приводимая к дифференциальной; эредитарное ядро

DOI: 10.14375/19922264140201

1 Введение

Вопросам разработки методов, алгоритмов и инструментальных программных средств для анализа и моделирования распределений процессов в гауссовских (нормальных) ДСтС с инвариантной мерой посвящена обширная литература (см., например, [1–18]). Методы анализа и моделирования распределений процессов с инвариантной мерой в интегродифференциальных ЭСтС подробно изложены в [1, 11, 19–22].

Для ДСтС и ЭСтС, приводимых к ДСтС, с винеровскими и пуассоновскими шумами соответствующие точные и приближенные методы аналитического моделирования распределений с инвариантной мерой не разработаны. Особое внимание уделяется приближенным методам аналитического моделирования распределений с инвариантной мерой, основанным на нормальной аппроксимации одно- и двумерных распределений. Приводятся тестовые примеры.

* Работа выполнена при финансовой поддержке программы «Интеллектуальные информационные технологии, системный анализ и автоматизация» (проект 1.7).

[1] Институт проблем информатики Российской академии наук, sinitsin@dol.ru

2 Уравнения для распределений процессов с инвариантной мерой в дифференциальных стохастических системах

Как известно [1, 11], для ДСтС в конечномерных пространствах используется дифференциальное стохастическое уравнение Ито следующего вида:

$$dY = a(Y,t)\,dt + b(Y,t)\,dW_0 + \int\limits_{R_0^q} c(Y,t,v)\,dP^0(t,dv)\,. \quad (1)$$

Здесь Y — p-мерный вектор состояния, $Y \in \Delta^y$ (Δ^y — многообразие состояний); $a = a(y,t)$ и $b = b(y,t)$ — известные $(p\times 1)$-мерная и $(p\times r)$-мерная функции вектора Y и времени t; $W_0 = W_0(t)$ — r-мерный винеровский случайный процесс интенсивности $\nu_0 = \nu_0(t)$; $c(y,t,v)$ — $(p\times 1)$-мерная функция y, t и вспомогательного $(q\times 1)$-мерного параметра v; $\int\limits_{\Delta_t} dP^0(t,A)$ — центрированная пуассоновская мера, удовлетворяющая условию:

$$\int\limits_{\Delta_t} dP^0(t,A) = \int\limits_{\Delta_t} dP(t,A) - \int\limits_{\Delta_t} \nu_P(t,A)\,dt\,, \quad (2)$$

где $\int\limits_{\Delta_t} dP(t,A)$ — число скачков пуассоновского процесса $P(t,A)$ в интервале времени Δ; $\nu_P(t,A)$ — интенсивность пуассоновского процесса $P(t,A)$; A — некоторое борелевское множество пространства R_0^q с выколотым началом координат.

Интеграл (1) в общем случае распространяется на все пространство R_0^q с выколотым началом координат. Начальное значение Y_0 вектора Y представляет собой случайную величину, не зависящую от приращений винеровского процесса $W_0(t)$ и пуассоновского процесса $P(t,A)$ на интервалах времени $\Delta_t = (t_1,t_2]$, следующих за t_0, $t_0 \le t_1 \le t_2$, для любого множества A.

В случае, когда подынтегральная функция $c(y,t,v)$ в уравнении (1) допускает представление $c(y,t,v) = b(y,t)c'(v)$, уравнение (1) приводится к виду:

$$\dot Y = a(Y,t) + b(Y,t)V\,,$$

если принять

$$V = \dot W\,;\ \ W(t) = W_0(t) + \int\limits_{R_0^q} c'(v)P^0(t,dv)\,.$$

Пусть существуют одно- и n-мерные плотности $f_1 = f_1(z;t)$ и $f_n = f_n(z_1,\dots,z_n;t_1,\dots,t_n)$ и характеристические функции $g_1 = g_1(\lambda;t)$ и $g_n = g_n(\lambda_1,\dots,\lambda_n;t_1,\dots,t_n)$ $(n \ge 2)$, удовлетворяющие интегродифференциальным уравнениям Пугачева [1, 11]:

$$\frac{\partial f_1(z;t)}{\partial t} + \frac{\partial^T}{\partial z}\,[a(z,t)f_1(z;t)] = \\ = \frac{1}{(2\pi)^p}\int\limits_{-\infty}^{\infty}\int\limits_{-\infty}^{\infty}\chi^f(\lambda,\zeta,t)e^{i\lambda^{\mathrm T}(\zeta-z)}f_1(z;t)\,d\zeta d\lambda\,; \quad (3)$$

$$f_1(z;t_0) = f_0(z)\,; \quad (4)$$

$$\frac{\partial f_n(z_1,\dots,z_n;t_1,\dots,t_n)}{\partial t_n} + \\ + \frac{\partial^{\mathrm T}}{\partial z_n}\,[a(z_n,t_n)f_n(z_1,\dots,z_n;t_1,\dots,t_n)] = \\ = \frac{1}{(2\pi)^{pn}}\int\limits_{-\infty}^{\infty}\int\limits_{-\infty}^{\infty}\chi_n^f(\lambda_n,\zeta_n,t_n)\times \\ \times\exp\left\{i\sum_{l=1}^{n}\lambda_l^{\mathrm T}(\zeta_l - z_l)\right\}f_n(\zeta_1,\dots,\zeta_n;t_1,\dots,t_n)\times \\ \times\,d\zeta_1\cdots d\zeta_n d\lambda_1\cdots d\lambda_n\,;$$

$$f_n(z_1,\dots,z_{n-1},z_n;t_1,\dots,t_{n-1},t_{n-1}) = \\ = f_{n-1}(z_1,\dots,z_{n-1};t_1,\dots,t_{n-1})\delta(z_n - z_{n-1})\,, \\ t_1 \le t_2 \le \cdots \le t_n\,,\ \ n = 2,3,\dots;$$

$$\frac{\partial g_1(\lambda;t)}{\partial t} - \\ - \frac{1}{(2\pi)^p}\int\limits_{-\infty}^{\infty}\int\limits_{-\infty}^{\infty} i\lambda^{\mathrm T}a(z,t)e^{i(\lambda^{\mathrm T}-\mu^{\mathrm T})z}g_1(\mu;t)\,d\mu dz = \\ = \frac{1}{(2\pi)^k}\int\limits_{-\infty}^{\infty}\int\limits_{-\infty}^{\infty}\chi^g(\lambda,z,t)e^{i(\lambda^{\mathrm T}-\mu^{\mathrm T})z}\times \\ \times\,g_1(\mu;t)\,d\mu dz\,; \quad (5)$$

$$g_1(\lambda;t_0) = g_0(\lambda)\,; \quad (6)$$

$$\frac{\partial g_n(\lambda_1,\dots,\lambda_n;t_1,\dots,t_n)}{\partial t_n} - \frac{1}{(2\pi)^{pn}}\times \\ \times\int\limits_{-\infty}^{\infty}\cdots\int\limits_{-\infty}^{\infty} i\lambda^{\mathrm T}a(z_n,t_n)\exp\left[i\sum_{k=1}^{n}(\lambda_k^{\mathrm T} - \mu_k^{\mathrm T})z_k\right]\times \\ \times g_n(\mu_1,\dots,\mu_n;t_1,\dots,t_n)d\mu_1\dots d\mu_n dz_1\dots dz_n = \\ = \frac{1}{(2\pi)^{pn}}\int\limits_{-\infty}^{\infty}\cdots\int\limits_{-\infty}^{\infty}\chi^n(\lambda_n,z_n,t_n)\times \\ \times\exp\left[i\sum_{k=1}^{n}(\lambda_k^{\mathrm T} - \mu_k^{\mathrm T})z_k\right]\times \\ \times g_n(\mu_1,\dots,\mu_n;t_1,\dots,t_n)\,d\mu_1\cdots d\mu_n dz_1\cdots dz_n\,;$$

$$g_n(\lambda_1, \ldots, \lambda_n; t_1, \ldots, t_{n-1}, t_{n-1}) = \\ = g_{n-1}(\lambda_1, \ldots, \lambda_{n-2}, \lambda_{n-1} + \lambda_n; t_1, \ldots, t_{n-1}) \\ t_1 \le t_2 \le \cdots \le t_n\,, \quad n = 2, 3, \ldots$$

Здесь приняты следующие обозначения:

$$\chi^f(\lambda, \zeta, t) = -\frac{1}{2}\,\lambda^{\mathrm{T}} b(\zeta, t)\nu_0(t) b(\zeta, t)^{\mathrm{T}}\lambda + \\ + \int\limits_{R_0^q} \left\{\exp\left[i\lambda^{\mathrm{T}} c(\zeta, t, v)\right] - 1 - \right. \\ \left. - i\lambda^{\mathrm{T}} c(\zeta, t, v)\right\} \nu_P(t, dv)\,;$$

$$\chi_n^f(\lambda_n, \zeta_n, t_n) = -\frac{1}{2}\,\lambda_n^{\mathrm{T}} b(\zeta_n, t)\nu_0(t) b(\zeta_n, t)^{\mathrm{T}}\lambda + \\ + \int\limits_{R_0^q} \left\{\exp\left[i\lambda_n^{\mathrm{T}} c(\zeta_n, t_n, v)\right] - 1 - \right. \\ \left. - i\lambda_n^{\mathrm{T}} c(\zeta_n, t_n, v)\right\} \nu_P(t_n, dv)\,;$$

$$\chi^g(\lambda, z, t) = -\frac{1}{2}\,\lambda^{\mathrm{T}} b(z, t)\nu_0(t) b(z, t)^{\mathrm{T}}\lambda + \\ + \int\limits_{R_0^q} \left\{\exp\left[i\lambda^{\mathrm{T}} c(z, t, v)\right] - 1 - \right. \\ \left. - i\lambda^{\mathrm{T}} c(z, t, v)\right\} \nu_P(t, dv)\,; \quad (7)$$

$$\chi_n^g(\lambda_n, z_n, t_n) = -\frac{1}{2}\,\lambda_n^{\mathrm{T}} b(z_n, t)\nu_0(t) b(z_n, t)^{\mathrm{T}}\lambda + \\ + \int\limits_{R_0^q} \left\{\exp\left[i\lambda_n^{\mathrm{T}} c(z_n, t_n, v)\right] - 1 - \right. \\ \left. - i\lambda_n^{\mathrm{T}} c(z_n, t_n, v)\right\} \nu_P(t_n, dv)\,.$$

При этом одно- и n-мерные плотности и характеристические функции связаны между собой известными соотношениями:

$$f_1(z; t) = \frac{1}{(2\pi)^p} \int\limits_{-\infty}^{\infty} e^{-i\mu^{\mathrm{T}} z} g_1(\mu; t)\, d\mu\,;$$

$$g_1(\lambda; t) = \int\limits_{-\infty}^{\infty} e^{i\lambda^{\mathrm{T}} z} f_1(z; t)\, dz\,;$$

$$f_n(z_1, \ldots, z_n; t_1, \ldots, t_n) = \\ = \frac{1}{(2\pi)^{pn}} \int\limits_{-\infty}^{\infty} \cdots \int\limits_{-\infty}^{\infty} \exp\left\{-i\sum_{l=1}^{n} \lambda_l^{\mathrm{T}} z_l\right\} \times \\ \times\, g_n(\lambda_1, \ldots, \lambda_n; t_1, \ldots, t_n)\, d\lambda_1 \cdots \lambda_n\,;$$

$$g_n(\lambda_1, \ldots, \lambda_n; t_1, \ldots, t_n) = \\ = \int\limits_{-\infty}^{\infty} \cdots \int\limits_{-\infty}^{\infty} \exp\left\{i\sum_{l=1}^{n} \lambda_l^{\mathrm{T}} z_l\right\} \times \\ \times\, f_n(z_1, \ldots, z_n; t_1, \ldots, t_n)\, dz_1 \cdots dz_n\,.$$

Для нахождения одномерных плотностей $f_1(z, t) = f_1^*(z)$ и характеристических функций $g_1(\lambda; t) = g_1^*(\lambda)$ стохастических процессов в стационарных ДСтС (1), когда

$$\left.\begin{aligned} a(z, t) = a^*(z)\,; \quad b(z, t) = b^*(z)\,; \\ \chi(\mu; t) = \chi^f(\mu, \zeta, t) = \chi^{*f}(\mu, \zeta)\,, \end{aligned}\right\} \quad (8)$$

в (3) и (5) следует положить $\partial f_1/\partial t = 0$ и $\partial g_1/\partial t = 0$. В результате получим соответственно следующие интегродифференциальные уравнения:

$$\frac{\partial^{\mathrm{T}}}{\partial z}\left[a^*(z) f_1^*(z)\right] = \\ = \frac{1}{(2\pi)^p} \int\limits_{-\infty}^{\infty} \int\limits_{-\infty}^{\infty} \chi^{*f}(\lambda, \zeta) e^{i\lambda^{\mathrm{T}}(\zeta - z)} f_1^*(\zeta)\, d\zeta d\lambda\,;$$

$$- \frac{1}{(2\pi)^p} \int\limits_{-\infty}^{\infty} \int\limits_{-\infty}^{\infty} i\lambda^{\mathrm{T}} a^*(z) e^{i(\lambda^{\mathrm{T}} - \mu^{\mathrm{T}})z} g_1^*(\mu)\, d\mu dz = \\ = \frac{1}{(2\pi)^p} \int\limits_{-\infty}^{\infty} \int\limits_{-\infty}^{\infty} \chi^{*g}(\lambda, z) e^{i(\lambda^{\mathrm{T}} - \mu^{\mathrm{T}})z} g_1^*(\mu)\, d\mu dz\,.$$

Пусть функция a в ДСтС (1) допускает представление

$$a = a(z, t) = a_1(z, t) + a_2(z, t) \quad (9)$$

такое, что функция $f_1 = f_1(z; t)$ является плотностью инвариантной меры не возмущенной шумами системы, описываемой векторным обыкновенным дифференциальным уравнением вида

$$\dot{z} = a_1(z, t)\,, \quad (10)$$

т. е. удовлетворяет следующему условию сохранения инвариантной меры:

$$\frac{\partial f_1(z; t)}{\partial t} + \frac{\partial^{\mathrm{T}}}{\partial z}\left[a_1(z, t) f_1(z; t)\right] = 0\,. \quad (11)$$

Для гладких функций $a_1 = a_1(z, t)$ вопросы существования и основные свойства интегральных инвариантов и инвариантных мер изучены в [23, 24]. При этом функция $a_2 = a_2(z, t)$ в (9) определяется путем решения следующего интегродифференциального уравнения:

$$\frac{\partial^{\mathrm{T}}}{\partial z}\left[a_2(z, t) f_1(z; t)\right] = \\ = \frac{1}{(2\pi)^k} \int\limits_{-\infty}^{\infty} \int\limits_{-\infty}^{\infty} \chi^f(\lambda, \zeta, t) e^{i\lambda^{\mathrm{T}}(\zeta - z)} f_1(\zeta; t)\, d\zeta d\lambda\,. \quad (12)$$

Для стационарных ДСтС, когда выполнены условия (8), уравнения (9)–(11) имеют вид:

$$a(z) = a_1(z) + a_2(z)\,; \tag{13}$$

$$\dot z = a_1(z)\,; \tag{14}$$

$$\frac{\partial^{\mathrm T}}{\partial z}\,[a_2^*(z) f_1^*(z)] = 0\,, \tag{15}$$

$$\frac{\partial^{\mathrm T}}{\partial z}\,[a_2^*(z) f_1^*(z)] =$$

$$= \frac{1}{(2\pi)^p} \int\limits_{-\infty}^{\infty}\int\limits_{-\infty}^{\infty} \chi^{*f}(\lambda,\zeta) e^{i\lambda^{\mathrm T}(\zeta - z)} f_1^*(\zeta)\, d\zeta d\lambda\,. \tag{16}$$

Условия сохранения инвариантной меры можно представить в следующем развернутом виде:

$$\left.\begin{gathered} \frac{\partial f_1(z;t)}{\partial t} + A_a f_1(z;t) = 0\,; \\ A_a f_1(z;t) = \frac{\partial^{\mathrm T}}{\partial z}\,[a_1(z,t) f_1(z;t)] = \operatorname{div} \pi(z;t)\,; \\ A_a^* f_1^*(z) = 0\,; \\ A_a^* f_1(z) = \frac{\partial^{\mathrm T}}{\partial z}\,[a_1^*(z) f_1^*(z)] = \operatorname{div} \pi^*(z)\,; \end{gathered}\right\} \tag{17}$$

$$\left.\begin{gathered} \frac{\partial g_1(\lambda;t)}{\partial t} - B_a g_1(\lambda;t) = 0\,; \\ B_a g_1(\lambda;t) = \\ = \frac{1}{(2\pi)^p} \int\limits_{-\infty}^{\infty}\int\limits_{-\infty}^{\infty} i\lambda^{\mathrm T} a_1(z,t) e^{i(\lambda^{\mathrm T} - \mu^{\mathrm T})z} \times \\ \times\, g_1(\mu;t)\, d\mu dz = \\ = \int\limits_{-\infty}^{\infty} i\lambda^{\mathrm T} a(z,t) e^{i\lambda^{\mathrm T} z} f_1(z;t)\, dz = \\ = \int\limits_{-\infty}^{\infty} e^{i\lambda^{\mathrm T} z} i\lambda^{\mathrm T} \pi(z;t)\, dz\,; \end{gathered}\right\} \tag{17}$$

$$B_a^* g_1^*(\lambda) = 0\,; \quad B_a^* g_1^*(\lambda) =$$

$$= \frac{1}{(2\pi)^p} \int\limits_{-\infty}^{\infty} i\lambda^{\mathrm T} a_1^*(z) e^{i(\lambda^{\mathrm T} - \mu^{\mathrm T})z} g_1^*(\mu)\, d\mu dz =$$

$$= \int\limits_{-\infty}^{\infty} i\lambda^{\mathrm T} a_1^*(z) e^{i\lambda^{\mathrm T} z} f_1^*(z)\, dz =$$

$$= \int\limits_{-\infty}^{\infty} e^{i\lambda^{\mathrm T} z} i\lambda^{\mathrm T} \pi^*(z)\, dz\,. \tag{18}$$

Для разрывных функций $a_1(z,t)$ в терминах характеристических функций соотношения (11) и (15) могут быть записаны в виде (17) и (18). При этом для составляющих $a_2(z,t)$ и $a_2^*(z)$ имеют место уравнения:

$$B_{a_2} g_1(\lambda;t) =$$

$$= \frac{1}{(2\pi)^p} \int\limits_{-\infty}^{\infty}\int\limits_{-\infty}^{\infty} \chi^g(\lambda,z,t) e^{i(\lambda^{\mathrm T} - \mu^{\mathrm T})z} g_1(\mu;t)\, d\mu dz; \tag{19}$$

$$B_{a_2}^* g_1^*(\lambda) =$$

$$= \frac{1}{(2\pi)^p} \int\limits_{-\infty}^{\infty}\int\limits_{-\infty}^{\infty} \chi^{*g}(\lambda,z) e^{i(\lambda^{\mathrm T} - \mu^{\mathrm T})z} g_1^*(\mu)\, d\mu dz. \tag{20}$$

Отсюда вытекают точные алгоритмы аналитического моделирования распределений с инвариантной мерой. В их основе лежат следующие теоремы.

Теорема 1. *Функция $f_1 = f_1(z;t)$ будет решением* (3) *и* (4) *тогда и только тогда, когда $a = a(z,t)$ допускает представление* (9) *такое, что $f_1 = f_1(z;t)$ является плотностью инвариантной меры обыкновенного дифференциального уравнения* (10)*, т. е. удовлетворяет условию* (11)*. При этом составляющая a_2 определяется из решения интегродифференциального уравнения* (12).

Теорема 2. *Функция $f_1^* = f_1^*(z)$ будет решением* (3) *тогда и только тогда, когда $a^* = a^*(z)$ допускает представление* (13) *такое, что $f_1^* = f_1^*(z)$ является плотностью инвариантной меры* (14)*. При этом составляющая a_2^* определяется из решения уравнения* (16).

Теорема 3. *Функция $g_1 = g_1(\lambda;t)$ будет решением* (5)*,* (6) *тогда и только тогда, когда недифференцируемая функция $a = a(z,t)$ допускает представление* (9) *такое, что $g_1 = g_1(\lambda;t)$ является характеристической функцией инвариантной меры уравнения* (10)*, т. е. удовлетворяет условию* (16)*. При этом составляющая a_2 определяется из уравнения* (19).

Теорема 4. *Функция $g_1^* = g_1^*(\lambda)$ будет решением* (15) *тогда и только тогда, когда недифференцируемая функция $a^* = a^*(z)$ допускает представление* (13) *такое, что g_1^* является характеристической функцией инвариантной меры уравнения* (10)*. При этом a_2^* определяется из решения* (20).

Теоремы 1–4 легко обобщаются на случай многомерных распределений с инвариантной мерой.

3 Приближенные методы и алгоритмы аналитического моделирования распределений процессов с инвариантной мерой в дифференциальных стохастических системах

Пусть нелинейная ДСтС (1) допускает применение метода нормальной аппроксимации (МНА) [1, 11]. Тогда одно- и двумерные нормальные плотности $f_1^{\text{МНА}}$, $f_2^{\text{МНА}}$ и характеристические функции $g_1^{\text{МНА}}$, $g_2^{\text{МНА}}$, а также вектор математического ожидания $m_t = \mathsf{M}^{\text{МНА}} Z(t)$, ковариационная матрица $K_t = \mathsf{M}^{\text{МНА}} Z^{0T} Z^0(t)$ $(Z^0(t) = Z(t) - m_t)$ и матрица ковариационных функций $K(t_1, t_2) = \mathsf{M}^{\text{МНА}} Z^{0T}(t_1) Z^0(t_2)$ $(t_1 < t_2)$ определяются следующими уравнениями:

$$f_1^{\text{МНА}} = f_1^{\text{МНА}}(z; t, m_t, K_t) = [(2\pi)^p |K_t|]^{-1/2} \times \times \exp\left\{-\frac{1}{2}(z^{\mathrm{T}} - m_t^{\mathrm{T}}) K_t^{-1} (z - m_t)\right\}; \quad (21)$$

$$f_2^{\text{МНА}} = \\ = f_2^{\text{МНА}}(z_1, z_2; t_1, t_2, m_{t_1}, m_{t_2}, K_{t_1}, K_{t_2}, K(t_1, t_2)) = \\ = \left[(2\pi)^p |\bar{K}_2|\right]^{-1/2} \exp\left(-\left(\left[z_1^{\mathrm{T}} z_2^{\mathrm{T}}\right] - \bar{m}_2^{\mathrm{T}}\right) \times \times \bar{K}_2^{-1} \left(\left[z_1^{\mathrm{T}} z_2^{\mathrm{T}}\right]^{\mathrm{T}} - \bar{m}_2\right)\right);$$

$$g_1^{\text{МНА}}(\lambda; t) = \exp\left\{i\lambda^{\mathrm{T}} m - \frac{1}{2}\lambda^{\mathrm{T}} K_t \lambda\right\}; \quad (22)$$

$$g_2^{\text{МНА}}(\lambda_1, \lambda_2; t_1, t_2) = \\ = \exp\left\{i\bar{\lambda}^{\mathrm{T}} \bar{m}_2 - \frac{1}{2}\bar{\lambda}^{\mathrm{T}} \bar{K}_2 \bar{\lambda}\right\}, \ \bar{\lambda} = \left[\lambda_1^{\mathrm{T}} \lambda_2^{\mathrm{T}}\right]^{\mathrm{T}}; \quad (23)$$

$$\dot{m}_t = \Phi_1(t, m_t, K_t) = \\ = \int_{-\infty}^{\infty} a(z, t) f_1^{\text{МНА}}(z; t, m_t, K_t)\, dz; \quad (24)$$

$$\dot{K}_t = \Phi_2(t, m_t, K_t) = \Phi_{21} + \Phi_{12} + \Phi_{22} = \\ = \left[\int_{-\infty}^{\infty} a(z, t)(z^{\mathrm{T}} - m_t^{\mathrm{T}}) + (z - m_t) a^{\mathrm{T}}(z, t) + \right. \\ \left. + \bar{\sigma}(z, t)\right] f_1^{\text{МНА}}(z; t, m_t, K_t)\, dz, \quad (25)$$

$$\frac{\partial K(t_1, t_2)}{\partial t_2} = \\ = \Phi_3(t_1, t_2, m_{t_1}, m_{t_2}, K_{t_1}, K_{t_2}, K(t_1, t_2)) = \\ = \left[(2\pi)^{2p} |\bar{K}_2|\right]^{-1/2} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (z_1 - m_{t_1})\, a(z_2, t_2) \times \\ \times \exp\left\{-\left(\left[z_1^{\mathrm{T}} z_2^{\mathrm{T}}\right] - \bar{m}_2^{\mathrm{T}}\right) \bar{K}_2^{-1} \times \\ \times \left(\left[z_1^{\mathrm{T}} z_2^{\mathrm{T}}\right] - \bar{m}_2\right)\right\} dz_1 dz_2 = \\ = K(t_1, t_2) K(t_2)^{-1} \Phi_{21}(m(t_2), K(t_2), t_2)^{\mathrm{T}}. \quad (26)$$

Здесь введены следующие обозначения:

$$\left.\begin{aligned} z_1 = z_{t_1};\ z_2 = z_{t_2};\ \bar{m}_2 = \left[m_{t_1}^{\mathrm{T}} m_{t_2}^{\mathrm{T}}\right]^{\mathrm{T}}; \\ \bar{K}_2 = \begin{bmatrix} K(t_1, t_1)\ K(t_1, t_2) \\ K(t_2, t_1)\ K(t_2, t_2) \end{bmatrix}; \end{aligned}\right\} \quad (27)$$

$$\left.\begin{aligned} \bar{\sigma}(z, t) = \\ = \sigma(z, t) + \int_{R_0^q} c(z, t, v) c(z, t, v)^{\mathrm{T}} \nu_P(t, dv); \\ \sigma(z, t) = b(z, t) \nu_0(t) b(z, t)^{\mathrm{T}}. \end{aligned}\right\} \quad (28)$$

Для стационарных ДСтС при $\dot{m}^* = 0$, $\dot{K}^* = 0$, $K(t_1, t_2) = k(\tau)$ $(\tau = t_1 - t_2)$ соотношения (24)–(28) принимают вид:

$$\Phi_1^*(m^*, K^*) = 0; \quad (29)$$

$$\Phi_2^*(m^*, K^*) = 0; \quad (30)$$

$$\frac{dk(\tau)}{d\tau} = \Lambda k(\tau); \ \Lambda = \Phi_{21}(m^*, K^*) K^{*-1} k(\tau); \quad (31)$$

$$k(\tau) = k(-\tau)^{\mathrm{T}}, \ k(0) = K.$$

Из (31) следует, что алгоритм МНА будет устойчивым, если матрица $\Lambda^* = \Lambda(m^*, K^*) = \Phi_{21}(m^*, K^*) K^{*-1}$ будет асимптотически устойчива.

Уравнения метода статистической линеаризации (МСЛ) в нелинейных ДСтС при аддитивных шумах, когда $b(z, t) = b_0(t)$, $b^*(z) = b_0^*$ получаются из (24)–(26) и (29)–(31) как частный случай.

Условия наличия нормального распределения с инвариантной мерой, если заменить $a(z, t)$ статистически линеаризованным выражением вида:

$$a(Z, t) \approx a_{10}^{\text{МНА}}(t, m_t, K_t) + a_{11}^{\text{МНА}}(t, m_t, K_t)(Z - m_t),$$

где

$$a_{10}^{\text{МНА}} = a_{10}^{\text{МНА}}(t, m_t, K_t);$$

$$a_{11}^{\text{МНА}} = a_{11}^{\text{МНА}}(t, m_t, K_t) = \\ = \left[\int_{-\infty}^{\infty} a(z, t)(z^{\mathrm{T}} - m_t^{\mathrm{T}}) f_1^{\text{МНА}}(z; t, m_t, K_t)\, dz\right] K_t^{-1} = \\ = \left[\frac{\partial}{\partial m_t} \left(a_{10}^{\text{МНА}}\right)^{\mathrm{T}}\right]^{\mathrm{T}},$$

приводят к следующим соотношениям:

$$\frac{\partial f_1^{\text{МНА}}(z;t,m_t,K_t)}{\partial t} + \\ + \frac{\partial^{\text{T}}}{\partial z}\left\{\left[a_{10}^{\text{МНА}}(t,m_t,K_t) + a_{11}^{\text{МНА}}(t,m_t,K_t)\times\right.\right. \\ \left.\left.\times\,(z-m_t)\right] f_1^{\text{МНА}}(z;t,m_t,K_t)\right\} = 0\,; \quad (32)$$

$$\frac{\partial^{\text{T}}}{\partial z}\left\{\left[a_{10}^{*\text{МНА}}(m^*,K^*) + a_{11}^{*\text{МНА}}(m^*,K^*)\times\right.\right. \\ \left.\left.\times\,(z-m^*)\right] f_1^{*\text{МНА}}(z;m^*,K^*)\right\} = 0\,, \quad (33)$$

где

$$f_1^{*\text{МНА}}(z;m^*,K^*) = [(2\pi)^p|K^*|]^{-1/2}\times \\ \times\exp\left\{-\frac{1}{2}\left(z^{\text{T}}-m^{*T}\right)(K^*)^{-1}(z-m^*)\right\}.$$

Аналогично в развернутом виде выписываются условия (17) и (18):

$$\frac{\partial g_1^{\text{МНА}}(\lambda;t)}{\partial t} - \int\limits_{-\infty}^{\infty} i\lambda^{\text{T}}\left[a_{10}^{\text{МНА}}(m_t,K_t,t) + \right. \\ \left. + a_{11}^{\text{МНА}}(m_t,K_t,t)(z-m_t)\right]\times \\ \times\, e^{i\lambda^{\text{T}}z} f_1^{\text{МНА}}(z;m_t,K_t,t)dz = 0\,, \quad (34)$$

$$\int\limits_{-\infty}^{\infty} i\lambda^{\text{T}}\left[a_{10}^{*\text{МНА}}(m^*,K^*) + a_{11}^{*\text{МНА}}(m^*,K^*)(z-m^*)\right]\times \\ \times\, e^{i\lambda^{\text{T}}z} f_1^{*\text{МНА}}(z;m^*,K^*)dz = 0\,. \quad (35)$$

Отсюда вытекают следующие теоремы, лежащие в основе приближенных нелинейных методов.

Теорема 5. *Если существуют одно- и двумерные плотности стохастического процесса, а матрица* $a_{11}^{\text{МНА}}$ *коэффициентов статистической линеаризации асимптотически устойчива, то приближенный алгоритм аналитического моделирования МНА нестационарных стохастических процессов в ДСтС* (1) *с инвариантной мерой определяется выражениями* (21)–(26) *и* (32).

Теорема 6. *Если существуют стационарные одно- и двумерные плотности стохастического процесса, а матрица* $a_{11}^{*\text{МНА}}$ *коэффициентов статистической линеаризации асимптотически устойчива, то приближенный алгоритм аналитического моделирования стационарных стохастических процессов с инвариантной мерой в стационарной ДСтС* (1) *определяется выражениями* (29)–(31) *и* (33).

Как известно [1, 11], одно- и двумерные нормальные распределения определяют и все n-мерные распределения ($n > 3$). Поэтому МНА и МСЛ при $b(Y,t) = b_0(t)$, $c(Y,t,z) = c_0(t,v)$ дают приближенные алгоритмы для любых многомерных плотностей стохастических процессов, если они существуют. Аналогично формулируются теоремы 3.3 и 3.4 в терминах характеристических функций на основе условий (34) и (35).

Обобщением МНА являются различные приближенные методы, основанные на параметризации распределений [1, 11]. Аппроксимируя одномерную характеристическую функцию $g_1(\lambda;t)$ и соответствующую плотность $f_1(z,t)$ известными функциями $g_1^*(\lambda;\theta)$ и $f_1^*(z;\theta)$, зависящими от конечномерного векторного параметра θ, можно свести задачу приближенного определения одномерного распределения к выводу из уравнения для характеристических функций обыкновенных дифференциальных уравнений, определяющих θ как функцию времени. Это относится и к остальным многомерным распределениям.

При аппроксимации многомерных распределений целесообразно выбирать последовательности функций $\{f_n^*(z_1,\ldots,z_n;\theta_n)\}$ и $\{g_n^*(\lambda_1,\ \ldots\ ,\lambda_n;\theta_n)\}$, каждая пара которых находилась бы в такой зависимости от векторного параметра θ_n, чтобы при любом n множество параметров, образующих вектор θ_n, включало в качестве подмножества множество параметров, образующих вектор θ_{n-1}. Тогда при аппроксимации n-мерного распределения придется определять только те координаты вектора θ_n, которые не были определены ранее при аппроксимации функций $f_1, g_1, \ldots, f_{n-1}$, g_{n-1}. В зависимости от того, что представляют собой параметры, от которых зависят функции $f_n^*(z_1,\ \ldots\ ,z_n;\theta_n)$ и $g_n^*(\lambda_1,\ \ldots\ ,\lambda_n;\theta_n)$, аппроксимирующие неизвестные многомерные плотности $f_n(z_1,\ldots,z_n;t_1,\ \ldots\ ,t_n)$ и характеристические функции $g_n(\lambda_1,\ \ldots\ ,\lambda_n;t_1,\ldots,t_n)$, используются различные приближенные методы решения уравнений, определяющих многомерные распределения вектора состояния системы X_t, в частности методы моментов, семиинвариантов, ортогональных разложений, квазимоментов и др. [1, 11] и в условиях сохранения инвариантной меры.

4 Анализ и моделирование распределений с инвариантной мерой в эредитарных стохастических системах, приводимых к дифференциальным

Рассмотрим ЭСтС, описываемую интегродифференциальным уравнением Ито следующего вида [22]:

$$dX = \left[a(X,t) + \int_{t_0}^{t} a_1(X(\tau),\tau,t)\,d\tau\right] dt + \left[b(X,t) + \int_{t_0}^{t} b_1(X(\tau),\tau,t)\,d\tau\right] dW_0 + \int_{R_0^q} \left[c(X,t,v) + \int_{t_0}^{t} c_1(X(\tau),\tau,t,v)\right] dP^0(t,dv) \quad (36)$$

с начальным условием $X(t_0) = X_0$.

В (36) приняты следующие обозначения и допущения: $X = X(t)$ — p-мерный вектор состояния; W_0 — r-мерный винеровский процесс интенсивности $\nu_0 = \nu_0(t)$; $\int\limits_{\Delta_t} dP^0(t,A)$ — центрированная пуассоновская мера, удовлетворяющая условию (2).

Функции $a = a(X,t)$, $a_1 = a_1(X(\tau),\tau,t)$, $b = b(X,t)$, $b_1 = b_1(X(\tau),\tau,t)$, $c = c(X,t,v)$ и $c_1 = c_1(X(\tau),\tau,t,v)$ имеют соответственно размерности $p\times 1$, $p\times 1$, $p\times r$, $p\times r$, $p\times 1$ и $p\times 1$ и допускают представления следующего вида:

$$\left.\begin{aligned} a_1 &= A(t,\tau)\varphi(X(\tau),\tau)\,;\\ b_1 &= B(t,\tau)\psi(X(\tau),\tau)\,;\\ c_1 &= C(t,\tau)\chi(X(\tau),\tau,v)\,. \end{aligned}\right\} \quad (37)$$

Здесь эредитарные ядра $A = A(t,\tau) = [A_{ij}(t,\tau)]$ $(i,j = \overline{1,p})$, $B = B(t,\tau) = [B_{il}(t,\tau)]$ $(i = \overline{1,p},\ l = \overline{1,r})$ и $C = C(t,\tau) = [C_{ij}(t,\tau)]$ $(i,j = \overline{1,p})$ имеют соответственно размерности $p\times p$, $p\times r$ и $p\times p$. Они удовлетворяют следующим условиям физической реализуемости и асимптотического затухания:

$$\left.\begin{aligned} A_{ij}(t,\tau) = 0\,;\ \ B_{il}(t,\tau) = 0\,;\\ C_{ij}(t,\tau) = 0\ \ \forall\tau > t\,; \end{aligned}\right\} \quad (38)$$

$$\left.\begin{aligned} \int_{-\infty}^{\infty} |A_{ij}(t,\tau)|\,d\tau < \infty\,;\\ \int_{-\infty}^{\infty} |B_{il}(t,\tau)|\,d\tau < \infty\,;\\ \int_{-\infty}^{\infty} |C_{ij}(t,\tau)|\,d\tau < \infty\,. \end{aligned}\right\} \quad (39)$$

При этом нелинейные в общем случае функции $\varphi = \varphi(X(\tau),\tau)$, $\psi = \psi(X(\tau),\tau)$ и $\chi = \chi(X(\tau),\tau,v)$ имеют размерности $p\times 1$, $p\times p$ и $p\times 1$ соответственно.

В случае если эредитарные ядра A, B, C удовлетворяют условиям

$$\begin{aligned} A_{ij}(t,\tau) &= \tilde{A}_{ij}(u)\,;\\ B_{il}(t,\tau) &= \tilde{B}_{il}(u)\,;\\ C_{ij}(t,\tau) &= \tilde{C}_{ij}(u)\ \ (u = t-\tau)\,, \end{aligned}$$

то говорят об ЭСтС со стационарным затуханием.

Важный класс ядер представляют собой сингулярные (вырожденные) ядра, когда имеют место представления:

$$\left.\begin{aligned} A_{ij}(t,\tau) &= A_{ij}^{+}(t)A_{ij}^{-}(\tau)\,;\\ B_{il}(t,\tau) &= B_{il}^{+}(t)B_{il}^{-}(\tau)\,;\\ C_{ij}(t,\tau) &= C_{ij}^{+}(t)C_{ij}^{-}(\tau) \end{aligned}\right\} \quad (40)$$

$(i,l = \overline{1,p};\ j = \overline{1,r})$.

В случае, когда подынтегральные функции $c(X,t,v)$ и $c_1(X(\tau),\tau,v)$ в (36) допускают представления

$$\begin{aligned} c(X,t,v) &= b(X,t)c'(v)\,;\\ c_1(X(\tau),\tau,v) &= b(X(\tau),\tau)c'(v)\,, \end{aligned}$$

ЭСтС (36) приводится к виду:

$$\dot{X} = a(X,t) + \int_{t_0}^{t} a_1(X(\tau),\tau,t)\,d\tau + \left[b(X,t) + \int_{t_0}^{t} b_1(X(\tau),\tau,t)\,d\tau\right] V\,, \quad (41)$$

если принять

$$V = \dot{W}\,;\ \ W(t) = W_0(t) + \int_{R_0^q} c'(v)P^0(t,dv)\,.$$

В [22] решена задача приведения ЭСтС (36) при условиях (37)–(39) и (37)–(39), (40) к ДСтС (1), а также установлены следующие утверждения.

Рассмотрим сначала ЭСтС (36) при условиях (37)–(39). Будем считать, что эредитарные ядра $A(t,\tau)$, $B(t,\tau)$, $C(t,\tau)$ удовлетворяют следующим нестационарным линейным операторным уравнениям:

$$\begin{aligned} F^{At}A(t,\tau) &= H^{At}\delta(t-\tau)\,;\\ F^{Bt}B(t,\tau) &= H^{Bt}\delta(t-\tau)\,;\\ F^{Ct}C(t,\tau) &= H^{Ct}\delta(t-\tau)\,;\\ A(t,\tau) &= A'(t,\tau)^{\mathrm{T}}(H^{A*\tau})^{\mathrm{T}}\,;\\ A'(t,\tau)^{\mathrm{T}}(F^{A*\tau})^{\mathrm{T}} &= I_h^A\delta(t-\tau)\,;\\ B(t,\tau) &= B'(t,\tau)^{\mathrm{T}}(H^{B*\tau})^{\mathrm{T}}\,;\\ B'(t,\tau)^{\mathrm{T}}(F^{B*\tau})^{\mathrm{T}} &= I_h^B\delta(t-\tau)\,;\\ C(t,\tau) &= C'(t,\tau)^{\mathrm{T}}(H^{C*\tau})^{\mathrm{T}}\,;\\ C'(t,\tau)^{\mathrm{T}}(F^{C*\tau})^{\mathrm{T}} &= I_h^C\delta(t-\tau)\,. \end{aligned}$$

Здесь F^A, H^A, F^B, H^B, F^C и H^C — известные матричные дифференциальные операторы размерности $h_A \times h_A$, $h_B \times h_B$, $h_C \times h_C$ порядка n_A, m_A, n_B, m_B, n_C, m_C, ($n_A > m_A$, $n_B > m_B$, $n_C > m_C$) соответственно:

$$\left.\begin{aligned} F^A = F^A(t,D) &= \sum_{l=0}^{n_A} \alpha_l^A(t) D^l\,; \\ H^A = H^A(t,D) &= \sum_{l=0}^{m_A} \beta_l^A(t) D^l\,; \\ F^B = F^B(t,D) &= \sum_{l=0}^{n_B} \alpha_l^B(t) D^l\,; \\ H^B = H^B(t,D) &= \sum_{l=0}^{m_B} \beta_l^B(t) D^l\,; \\ F^C = F^C(t,D) &= \sum_{l=0}^{n_C} \alpha_l^C(t) D^l\,; \\ H^C = H^C(t,D) &= \sum_{l=0}^{m_C A} \beta_l^C(t) D^l\,; \end{aligned}\right\} \qquad (42)$$

индекс t у операторов означает, что оператор действует на функцию от t при фиксированном τ; звездочкой обозначен символ сопряжения оператора; I_h^A, I_h^B, I_h^C — единичные $(h \times h)$-матрицы. Введем h^A-, h^B-, h^C-мерные векторы посредством соотношений:

$$Z_1' = U' = \int_{t_0}^{t} A(t,\tau)\varphi(X(\tau),\tau)\,d\tau\,;$$

$$Z_1'' = U'' = \int_{t_0}^{t} B(t,\tau)\psi(X(\tau),\tau)\,d\tau\,;$$

$$Z_1''' = U''' = \int_{t_0}^{t} C(t,\tau)\chi(X(\tau),\tau,v)\,d\tau\,.$$

Эти переменные Z', Z'', Z''' будут удовлетворять следующим линейным дифференциальным уравнениям:

$$\begin{aligned} F^A(t,D)Z_1' &= H^A(t,D)\varphi(X,t)\,; \\ F^B(t,D)Z_1'' &= H^B(t,D)\psi(X,t)\,; \\ F^C(t,D)Z_1''' &= H^C(t,D)\chi(X,t,v)\,. \end{aligned}$$

Тогда ЭСтС (36) приводится к искомой ДСтС для расширенного вектора состояния $Z = \left[X^{\mathrm{T}} Z_1'^{\mathrm{T}}\, Z_1''^{\mathrm{T}}\, Z_1'''^{\mathrm{T}}\right]^{\mathrm{T}}$:

$$dZ = a_1^z(Z,t)\,dt + b_1^z(Z,t)\,dW_0 + \int_{R_0^q} c_1^z(Z,t,v)\,dP^0(t,dv)\,. \qquad (43)$$

Для случая $h_A = h_B = h_C = h$, $n_A = n_B = n_C = n$, $m_A = m_B = m_C = m$ в подробной записи функции $a_1^z(Z,t)$, $b_1^z(Z,t)$, $c_1^z(Z,t,v)$ имеют следующий вид:

$$\left.\begin{aligned} a_1^z(Z,t) &= \begin{bmatrix} a(X,t) + Z_1' \\ a'(t)Z_1' \\ a''(t)Z_1'' \\ a'''(t)Z_1''' \end{bmatrix}; \\ b_1^z(Z,t) &= \begin{bmatrix} b(X,t) + Z_1'' \\ b''(t)Z_1'' \\ 0 \\ 0 \end{bmatrix}; \\ c_1^z(Z,t,v) &= \begin{bmatrix} c(X,t,v) + Z_1''' \\ c'''(t)Z_1''' \\ 0 \\ 0 \end{bmatrix}. \end{aligned}\right\} \qquad (44)$$

При условии существования обратных матриц $(\alpha_n^A)^{-1}$, $(\alpha_n^B)^{-1}$, $(\alpha_n^C)^{-1}$ входящие в (44) переменные и коэффициенты допускают следующую запись:

$$\left.\begin{aligned} Z_{j+1}' &= \dot{Z}_j' - q_j'\varphi(X,t)\,; \\ Z_{j+1}'' &= \dot{Z}_j'' - q_j''\psi(X,t)\,; \\ Z_{j+1}''' &= \dot{Z}_j''' - q_j'''\chi(X,t,v)\ \ (j = \overline{1,(n-1)})\,; \end{aligned}\right\} \qquad (45)$$

$$a'(t) = \begin{bmatrix} I_h & 0 & \cdots & 0 \\ 0 & I_h & \ddots & 0 \\ \vdots & \ddots & \ddots & \cdots \\ 0 & 0 & \cdots & I_h \\ -(\alpha_n^A)^{-1}\alpha_0^A & -(\alpha_n^A)^{-1}\alpha_1^A & \cdots & -(\alpha_n^A)^{-1}\alpha_{n-1}^A \end{bmatrix}; \qquad (46)$$

$$a''(t) = \begin{bmatrix} I_h & 0 & \cdots & 0 \\ 0 & I_h & \ddots & 0 \\ \vdots & \ddots & \ddots & \cdots \\ 0 & 0 & \cdots & I_h \\ -(\alpha_n^B)^{-1}\alpha_0^B & -(\alpha_n^B)^{-1}\alpha_1^B & \cdots & -(\alpha_n^B)^{-1}\alpha_{n-1}^B \end{bmatrix}; \qquad (47)$$

$$a'''(t) = \begin{bmatrix} I_h & 0 & \cdots & 0 \\ 0 & I_h & \ddots & 0 \\ \vdots & \ddots & \ddots & \cdots \\ 0 & 0 & \cdots & I_h \\ -(\alpha_n^C)^{-1}\alpha_0^C & -(\alpha_n^C)^{-1}\alpha_1^C & \cdots & -(\alpha_n^C)^{-1}\alpha_{n-1}^C \end{bmatrix}; \qquad (48)$$

$$q_j' = (\alpha_n^A)^{-1}\left[\beta_{n-j}^A - \sum_{i=0}^{j-1}\sum_{l=0}^{j-i} \mathcal{C}_{n-j-l}^{n-j}\alpha_{n-j+i+l}^A q_i'^{(l)}\right]; \qquad (49)$$

$$q'_n = (\alpha_n^A)^{-1}\left[\beta_0^A - \sum_{i=0}^{n-1}\sum_{l=0}^{n-i}\alpha_{i+l}^A q_i'^{(l)}\right]; \quad (50)$$

$$q''_j = (\alpha_n^B)^{-1}\left[\beta_{n-j}^B - \right.$$
$$\left. - \sum_{i=0}^{j-1}\sum_{l=0}^{j-i}\mathcal{C}_{n-j-l}^{n-j}\alpha_{n-j+i+l}^A q_i''^{(l)}\right]; \quad (51)$$

$$q''_n = (\alpha_n^B)^{-1}\left[\beta_0^B - \sum_{i=0}^{n-1}\sum_{l=0}^{n-i}\alpha_{i+l}^A q_i''^{(l)}\right]; \quad (52)$$

$$q'''_j = (\alpha_n^C)^{-1}\left[\beta_{n-j}^C - \right.$$
$$\left. - \sum_{i=0}^{j-1}\sum_{l=0}^{j-i}\mathcal{C}_{n-j-l}^{n-j}\alpha_{n-j+i+l}^C q_i'''^{(l)}\right]; \quad (53)$$

$$q'''_j = (\alpha_n^C)^{-1}\left[\beta_0^C - \sum_{i=0}^{n-1}\sum_{l=0}^{n-i}\alpha_{i+l}^C q_i'''^{(l)}\right]. \quad (54)$$

Здесь $C_m^n = n!/(m!(n-m)!)$; индекс l означает, что суммирование проводится по всем индексам, исключая l.

Таким образом, справедливо следующее утверждение.

Теорема 7. *Пусть ядра $A(t,\tau)$, $B(t,\tau)$, $C(t,\tau)$ в ЭСтС* (1) *удовлетворяют условиям* (37)–(39) *или* (44)*, причем матрицы α_n^A, α_n^B, α_n^C в* (42) *обратимы, а функции φ, ψ, χ дифференцируемы по переменным расширенного вектора состояния достаточное число раз. Тогда ЭСтС* (36) *приводится к ДСтС* (43) *на основе* (44)–(54).

Замечание 1. Векторное уравнение (43) всегда линейно относительно Z'_1, Z''_1, Z'''_1, но в общем случае нелинейно относительно X.

В том случае, когда выполнены условия (37)–(39), а функции φ, ψ, χ не дифференцируемы по переменным расширенного вектора состояния, целесообразна аппроксимация вырожденными ядрами (40). В этом случае имеют место следующие соотношения:

$$dZ = a_2^z(Z,t)\,dt + b_2^z(Z,t)\,dW_0 + \int\limits_{R_0^q} c_2^z(Z,t,v)\,dP^0(t,dv); \quad (55)$$

$$Z = \left[X^{\mathrm{T}}Y'^{\mathrm{T}}Y''^{\mathrm{T}}Y'''^{\mathrm{T}}\right]^{\mathrm{T}}; \quad (56)$$

$$\left.\begin{aligned}
&\int\limits_{t_0}^{t} A(t,\tau)\varphi(X(\tau),\tau)d\tau = A^+Y';\\
&\int\limits_{t_0}^{t} B(t,\tau)\psi(X(\tau),\tau)d\tau = B^+Y'';\\
&\int\limits_{t_0}^{t} C(t,\tau)\chi(X(\tau),\tau,v)d\tau = C^+Y''';
\end{aligned}\right\} \quad (57)$$

$$\dot{Y}' = A^-\varphi;\ Y'(t_0) = 0;\ \dot{Y}'' = B^-\psi;$$
$$Y''(t_0) = 0;\ \dot{Y}''' = C^-\chi;\ Y'''(t_0) = 0;$$

$$\left.\begin{aligned}
a_2^z(Z,t) &= \begin{bmatrix} a(X,t) + A^+\varphi \\ A^-\varphi \\ B^-\psi \\ C^-\chi \end{bmatrix};\\
b_2^z(Z,t) &= \begin{bmatrix} b(X,t) + B^+\psi \\ 0 \\ 0 \\ 0 \end{bmatrix};\\
c_2^z(Z,t,v) &= \begin{bmatrix} c(X,t,v) + C^+\chi \\ 0 \\ 0 \\ 0 \end{bmatrix}.
\end{aligned}\right\} \quad (58)$$

Таким образом, имеем следующий результат.

Теорема 8. *Пусть эредитарные ядра $A(t,\tau)$, $B(t,\tau)$, $C(t,\tau)$ в ЭСтС* (36) *удовлетворяют условиям* (38), (39) *и* (40)*, а функции φ, ψ, χ не дифференцируемы по переменным расширенного вектора состояния. Тогда ЭСтС* (36) *приводится к ДСтС* (55) *на основе* (56)–(58).

Замечание 2. Векторное уравнение (55) для Y', Y'', Y''' относится к числу так называемых приводимых к линейным уравнениям [11].

Аналогичные теоремы устанавливаются для ЭСтС (41).

Следовательно, если выполнены условия теорем 7 и 8, то ЭСтС (36) приводится к ДСтС (43) или (55) и могут быть использованы точные и приближенные методы анализа и моделирования распределений с инвариантной мерой (см. разд. 2 и 3).

Таким образом, получены следующие утверждения, лежащие в основе точных и приближенных методов для ЭСтС, приводимых к ДСтС.

Теорема 9. *В условиях теоремы* 7 *для гладких функций a, a_1, b, b_1, c, c_1 одномерные нестационарные и стационарные распределения с инвариантной мерой определяются уравнениями теорем* 1 *и* 2.

Теорема 10. *В условиях теоремы* 8 *для разрывных функций a, a_1, b, b_1, c, c_1 одномерные нестационарные*

и стационарные распределения с инвариантной мерой определяются уравнениями теорем 3 *и* 4.

Теорема 11. *В условиях теорем* 7 *и* 8 *приближенный алгоритм аналитического моделирования нестационарных процессов с инвариантной мерой по МНА определяется теоремой* 5*, а стационарных процессов — теоремой* 6.

5 Заключение

Получено обобщение точных и приближенных (основанных на параметризации распределений) методов и алгоритмов моделирования стационарных и нестационарных процессов с инвариантной мерой в негауссовских ДСтС и ЭСтС с винеровскими и пуассоновскими шумами, приводимых к ДСтС, для случаев гладких и разрывных регулярных правых частей уравнений.

Особое внимание уделено приближенным МНА и МСЛ для нахождения распределений процессов с инвариантной мерой в ДСтС и ЭСтС, приводимых к ДСтС.

Аналогично [11, 15, 25], результаты допускают обобщение на случай ЭСтС, приводимых к ДСтС, с автокоррелированными шумами.

Разработан комплекс тестовых примеров для инструментального программного обеспечения в «ID StS» в среде MATLAB (см. приложение).

Аналогично [2–8] может быть рассмотрено применение представленных методов в задачах эквивалентности гауссовских и негауссовских ДСтС и ЭСтС. В частности, соотношения (7) и (28) позволяют заменять в ДСтС и ЭСтС стационарные и нестационарные негауссовские шумы гауссовскими. Часто оказывается полезным заменить p-мерную негауссовскую ДСтС или ЭСтС эквивалентной системой из p_1 независимых ДСтС меньшей размерности ($p_1 < p$). В этом случае следует учесть дополнительные связи на $K_{ij}(t)$, вытекающие из аналитической природы рассматриваемой задачи.

Приложение

Тестовые примеры

Пример 1. В условиях примера 6 [21], когда

$$\ddot{X} + \omega^2 X - \mu X^3 = -\delta\dot{X} + \gamma + V^{\mathrm{OP}} - \\ - \int\limits_{t_0}^{t} \left[\omega_1 X(\tau) - \delta_1 \dot{X}(\tau) + \mu_1 X^3(\tau)\right] e^{-\lambda|t-\tau|}\, d\tau\,, \\ X(t_0) = X_0\,,\ \ \dot{X}(t_0) = \dot{X}_0\,,$$

для обобщенного пуассоновского белого шума интенсивности $\nu = \nu^{\mathrm{OP}}$ уравнения для математических ожиданий, дисперсий и ковариаций в силу теоремы 11 имеют следующий вид:

$$\left.\begin{aligned} \dot{m}_1 &= m_2; \\ \dot{m}_2 &= -\omega_{0\text{э}}^2 m_1 - \delta m_2 - \lambda^{-1} m_3 + \gamma\,; \\ \dot{m}_3 &= \lambda(\omega_{1\text{э}} m_1 + \delta_1 m_2 - m_3)\,; \end{aligned}\right\} \qquad (\text{П1})$$

$$\left.\begin{aligned} \dot{K}_{11} &= 2K_{12}\,; \\ \dot{K}_{12} &= K_{22} - (\omega_{0\text{э}}^2 K_{11} + \delta K_{12} + \lambda^{-1} K_{13})\,; \\ \dot{K}_{13} &= K_{23} + \lambda\omega'_{1\text{э}} K_{11} + \lambda\delta_1 K_{12} - \lambda K_{13}\,; \\ \dot{K}_{22} &= -2(\omega_{0\text{э}}^2 K_{12} + \delta K_{22} + \lambda^{-1} K_{23}) + \nu^{\mathrm{OP}}\,; \\ \dot{K}_{23} &= -(\omega_{0\text{э}}^2 K_{13} + \lambda^{-1} K_{33}) + \lambda\omega'_{1\text{э}} K_{12} + \\ &\qquad + \lambda\delta_1 K_{22} - (\delta + \lambda) K_{23}\,; \\ \dot{K}_{33} &= 2(\lambda\omega'_{1\text{э}} K_{13} + \lambda\delta_1 K_{23} - \lambda K_{33})\,. \end{aligned}\right\} \qquad (\text{П2})$$

Здесь приняты следующие обозначения:

$$Z_1 = X\,;\ \ Z_2 = \dot{X}_3\,;$$

$$Z_3 = \int\limits_{t_0}^{t} \left[\omega_1 Z_1^{(\tau)} + \delta_1 Z_2(\tau) - \mu_1 Z_1^3(\tau)\right] e^{-\lambda|t-\tau|}\, d\tau\,;$$

$$m_i = \mathsf{M} Z_i\ (i = 1,2,3)\,;\ \ K_{ij} = \mathsf{M} Z_i^0 Z_j^0\ (i,j = 1,2,3)\,;$$

$$\omega_{0\text{э}}^2 = \omega_{0\text{э}}^2(m_1, D_1) = \omega^2\left[1 - \mu\frac{(m_1^2 + 3D_1)}{\omega^2}\right];$$

$$\omega_{1\text{э}} = \omega_{1\text{э}}(m_1, D_1) = \omega_1\left[1 - \frac{\mu_1(m_1^2 + 3D_1)}{\omega_1}\right];$$

$$\omega'_{1\text{э}} = \omega'_{1\text{э}}(m_1, D_1) = \omega_1\left[1 - \frac{3\mu(m_1^2 + 3D_1)}{\omega_1}\right].$$

Приравнивая в (П1) и (П2) правые части нулю, получим уравнения для стационарных значений m_i^* и K_{ij}^*. Для устойчивости (в среднем квадратическом) стационарных колебаний необходима асимптотическая устойчивость матрицы

$$\Lambda = \begin{bmatrix} 0 & 1 & 0 \\ -\omega_{0\text{э}}^2(m_1, D_1) & -\delta & -\lambda^{-1} \\ \lambda\omega'_{1\text{э}}(m_1, D_1) & \lambda\delta_1 & -\lambda \end{bmatrix}$$

в (31).

Пример 2. Для релейного осциллятора

$$\ddot{X} + \alpha\,\mathrm{sgn}\, X = -\delta\dot{X} + \gamma + V^{\mathrm{OP}} - \\ - \int\limits_{t_0}^{t} \left[\alpha_1 \mathrm{sgn}, X(\tau) + \delta_1 \dot{X}(\tau)\right] e^{-\lambda|t-\tau|}\, d\tau\,; \\ X(t_0) = X_0\,;\ \ \dot{X}(t_0) = \dot{X}\ \ (\alpha > 0)$$

в случае обобщенного пуассоновского белого шума V^{OP} интенсивности $\nu = \nu^{\mathrm{OP}}$ в силу теоремы 11 имеем:

$$\left.\begin{aligned} \dot{m}_1 &= m_2\,; \\ \dot{m}_2 &= \alpha k_0 m_1 - \delta m_2 - \lambda^{-1} m_3 + \gamma\,; \\ \dot{m}_3 &= \lambda(\alpha_1 k_0 m_1 + \delta_1 m_2 - m_3)\,; \end{aligned}\right\} \qquad (\text{П3})$$

$$\left.\begin{aligned}
\dot K_{11} &= 2K_{12}\,;\\
\dot K_{12} &= K_{22} - \alpha k_1 K_{11} - \delta K_{12} - \lambda^{-1} K_{13}\,;\\
\dot K_{13} &= K_{23} + \lambda\alpha_1 k_1 K_{12} + \lambda\delta_1 K_{22} - \lambda K_{13}\,;\\
\dot K_{22} &= -2\alpha k_1 K_{12} - 2\delta K_{22} - 2\lambda^{-1} K_{23} + \nu^{\mathrm{OP}}\,;\\
\dot K_{23} &= -\alpha k_1 K_{13} - (\delta+\lambda) K_{23} - \lambda^{-1} K_{33} + \\
&\qquad + \lambda\alpha_1 k_1 K_{12} + \lambda\delta_1 K_{22}\,;\\
\dot K_{33} &= 2\lambda(\alpha_1 k_1 K_{13} + \delta_1 K_{23} - K_{33})\,.
\end{aligned}\right\} \qquad (\text{П4})$$

Здесь введены следующие обозначения:

$$Z_1 = X\,; \quad Z_2 = \dot X\,;$$

$$Z_3 = \lambda \int\limits_{t_0}^{t} \left[\alpha_1 \operatorname{sgn} X(\tau) + \delta_1 \dot X(\tau)\right] e^{-\lambda|t-\tau|}\, d\tau\,;$$

$$k_0 = k_0(m_1, D_1) = 2\Phi\left(\frac{m_1}{\sqrt{D_1}}\right)\,;$$

$$k_1 = k_1(m_1, D_1) = \frac{2}{\sqrt{2\pi D_1}} \exp\left[-\frac{1}{2}\left(\frac{m_1}{\sqrt{D_1}}\right)^2\right]\,.$$

Приравнивая правые части (П3) и (П4) нулю, получим уравнения для стационарных значений m_i^* и K_{ij}^*. Устойчивость стационарных колебаний определяется асимптотической устойчивостью матрицы

$$\Lambda = \begin{bmatrix} 0 & 1 & 0 \\ -\alpha k_1(m_1, D_1) & -\delta & -\lambda^{-1} \\ \lambda\alpha_1 k_1(m_1, D_1) & \lambda\delta_1 & -\lambda \end{bmatrix}$$

в (31).

Литература

1. *Пугачев В. С., Синицын И. Н.* Стохастические дифференциальные системы. Анализ и фильтрация. — М.: Наука, 1990. 632 с. [Англ. пер. Stochastic differential systems. Analysis and filtering. — Chichester, N.Y.: Jonh Wiley, 1987. 549 p.].
2. *Moshchuk N. K., Sinitsyn I. N.* On stationary distributions in nonlinear stochastic differential systems. — Coventry, UK: University of Warwick, Mathematics Institute, 1989. Preprint. 15 p.
3. *Moshchuk N. K., Sinitsyn I. N.* On stochastic nonholonomic systems. — Coventry, UK: University of Warwick, Mathematics Institute, 1989. Preprint. 32 p.
4. *Мощук Н. К., Синицын И. Н.* О стохастических неголономных системах // Прикладная механика и математика, 1990. Т. 54. Вып. 2. С. 213–223.
5. *Moshchuk N. K., Sinitsyn I. N.* On stationary distributions in nonlinear stochastic differential systems // Quart. J. Mech. Appl. Math., 1991. Vol. 44. Pt. 4. P. 571–579.
6. *Мощук Н. К., Синицын И. Н.* О стационарных и приводимых к стационарным режимах в нормальных стохастических системах // Прикладная механика и математика, 1991. Т. 55. Вып. 6. С. 895–903.
7. *Мощук Н. К., Синицын И. Н.* Распределения с инвариантной мерой в механических стохастических нормальных системах // Докл. АН СССР, 1992. Т. 322. № 4. С. 662–667.
8. *Синицын И. Н.* Конечномерные распределения с инвариантной мерой в стохастических механических системах // Докл. РАН, 1993. Т. 328. № 3. С. 308–310.
9. *Soize C.* The Fokker–Plank equation for stochastic dynamical systems and its explicit steady state solutions. — Singapore: World Scientific, 1994. 321 p.
10. *Синицын И. Н.* Конечномерные распределения с инвариантной мерой в стохастических нелинейных дифференциальных системах. — М.: Диалог–МГУ, 1997. С. 129–140.
11. *Пугачев В. С., Синицын И. Н.* Теория стохастических систем. — М.: Логос, 2000; 2004. 1000 с. [Англ. пер. Stochastic systems. Theory and applications. — Singapore: World Scientific, 2001. 908 p.].
12. *Синицын И. Н., Корепанов Э. Р., Белоусов В. В.* Точные методы расчета стационарных режимов с инвариантной мерой в стохастических системах управления // Кибернетика и высокие технологии XXI века: Сб. докл. II Междунар. науч.-технич. конф. — Воронеж: Саквоее, 2002. С. 124–131.
13. *Синицын И. Н., Корепанов Э. Р., Белоусов В. В.* Точные аналитические методы в статистической динамике нелинейных информационно-управляющих систем // Системы и средства информатики. Спец. вып. Математическое и алгоритмическое обеспечение информационно-телекоммуникационных систем. — М.: Наука, 2002. С. 112–121.
14. *Синицын И. Н.* Развитие методов аналитического моделирования распределений с инвариантной мерой в стохастических системах // Современные проблемы прикладной математики, информатики, автоматизации, управления: Мат-лы Междунар. семинара. — Севастополь: СевНТУ, 2012. С. 24–35.
15. *Синицын И. Н.* Аналитическое моделирование распределений с инвариантной мерой в стохастических системах с автокоррелированными шумами // Информатика и её применения, 2012. Т. 6. Вып. 4. С. 4–8.
16. *Синицын И. Н.* Аналитическое моделирование распределений с инвариантной мерой в стохастических системах с разрывными характеристиками // Информатика и её применения, 2013. Т. 7. Вып. 1. С. 3–11.
17. *Синицын И. Н.* Параметрическое статистическое и аналитическое моделирование распределений в нелинейных стохастических системах на многообразиях // Информатика и её применения, 2013. Т. 7. Вып. 2. С. 4–16.
18. *Синицын И. Н., Синицын В. И.* Лекции по нормальной и эллипсоидальной аппроксимации распределений в стохастических системах. — М.: ТОРУС ПРЕСС, 2013. 488 с.
19. *Синицын И. Н.* Stochastic hereditary control systems // Проблемы управления и теории информации, 1986. Т. 15. № 4. С. 287–298.
20. *Синицын И. Н.* Конечномерные распределения процессов в стохастических интегральных и интегродифференциальных системах // Preprints of the 2nd IFAC

Symposium on Stochastic Control. Vol. 1. — Zurich: Pergamon Press, 1987. P. 144–153.

21. *Синицын И. Н., Синицын В. И., Корепанов Э. Р., Белоусов В. В., Сергеев И. В., Басилашвили Д. А.* Опыт моделирования эредитарных стохастических систем // Кибернетика и высокие технологии XXI века: Сб. докл. XIII Междунар. науч.-технич. конф. — Воронеж: Саквоее, 2012. Т. 2. С. 346–357.
22. *Синицын И. Н.* Анализ и моделирование распределений в эредитарных стохастических системах // Информатика и её применения, 2014. Т. 8. Вып. 1. С. 2–11.
23. *Немыцкий В. В., Степанов В. В.* Качественная теория дифференциальных уравнений. — М.–Л.: Гостехиздат, 1949. 448 с.
24. *Козлов В. В.* О существовании интегрального инварианта гладких динамических систем // ПММ, 1987. № 1. С. 538–545.
25. *Синицын И. Н.* Фильтры Калмана и Пугачева. — 2-е изд. — М.: Логос, 2007. 776 с.

Поступила в редакцию 16.01.14

ANALYTICAL MODELING OF DISTRIBUTIONS WITH INVARIANT MEASURE IN NON-GAUSSIAN DIFFERENTIAL AND REDUCABLE TO DIFFERENTIAL HEREDITARY STOCHASTIC SYSTEMS

I. N. Sinitsyn

Institute of Informatics Problems, Russian Academy of Sciences, 44-2 Vavilov Str., Moscow 119333, Russian Federation

Abstract: Exact and approximate methods and algorithms of one- and multidimensional distributions with invariant measure for analytical modeling in differential non-Gaussian (with Wiener and Poisson noises) stochastic systems (StS) and hereditary StS (HStS) reducible to differential are presented. Four theorems giving exact methods of analysis modeling in differential StS (DStS) of general type are proved. Approximate methods based on distributions parametrization in DStS are disscused. Special attention is paid to the methods of normal approximation (MNA) and statistical linearization (MSL) for one- and dimensional distributions in DStS. Stability conditions are presented. three theorems giving exact and approximate analytical modeling in HStS resucible to DStS with asymptotically dying kernels are given. Some equavalency applications of DStS and HStS are considered. Test examples for software tools "ID StS" are given.

Keywords: analytical modeling; differential stochastic system; distribution with invariant measure; Gaussian (normal) stochastic system; hereditary kernel; hereditary stochastic system; hereditary system reducible to differential; Ito stochastic differential equation; method of statistical linearization; non-Gaussian (with Wiener and Poisson noises) stochastic system; normal approximation method; singular kernel; software tools "ID StS"

DOI: 10.14375/19922264140201

Acknowledgments

The work was financially supported by the Program "Intelligent information technology, system analysis, and automation" (project 1.7).

References

1. Pugachev, V. S., and I. N. Sinitsyn. 1987. *Stochastic differential systems. Analysis and filtering.* Chichester, New York: Jonh Wiley. 549 p.
2. Moshchuk, N. K., and I. N. Sinitsyn. 1989. *On stationary distributions in nonlinear stochastic differential systems*. Coventry, UK: University of Warwick, Mathematics Institute. Preprint. 15 p.
3. Moshchuk, N. K., and I. N. Sinitsyn. 1989. *On stochastic nonholonomic systems*. Coventry, UK: University of Warwick, Mathematics Institute. Preprint. 32 p.
4. Moshchuk, N. K., and I. N. Sinitsyn. 1990. O stokhasticheskikh negolonomnykh sistemakh [About stochastic nonholonomial systems]. *Prikladnaya Mekhanika i Matematika* [*Appl. Mech. Math.*]. 54(2):213–223.
5. Moshchuk, N. K., and I. N. Sinitsyn. 1991. On stationary distributions in nonlinear stochastic differential systems. *Quart. J. Mech. Appl. Math.* 44(4):571–579.
6. Moshchuk, N. K., and I. N. Sinitsyn. 1991. O statsionarnykh i privodimykh k statsionarnym rezhimakh v normal'nykh stokhasticheskikh sistemakh [About stationary and reducible to stationary regimes in normal stochastic systems]. *Prikladnaya Mekhanika i Matematika* [*Appl. Mech. Math.*]. 55(6):895–903.
7. Moshchuk, N. K., and I. N. Sinitsyn. 1992. Raspredeleniya s invariantnoy meroy v mekhanicheskikh stokhasticheskikh normal'nykh sistemakh [Distributions with

invariant measure in mechanical stochastic normal systems]. *Dokl. AN SSSR* 322(4):662–667.

8. Sinitsyn, I. N. 1993. Konechnomernye raspredeleniya s invariantnoy meroy v stokhasticheskikh mekhanicheskikh sistemakh [Finite dimensional distributions with invariant measure in stochastic mechanical systems]. *Dokl. RAN* 328(3):308–310.
9. Soize, C. 1994. *The Fokker–Plank equation for stochastic dynamical systems and its explicit steady state solutions*. Singapore: World Scientific. 321 p.
10. Sinitsyn, I. N. 1997. Konechnomernye raspredeleniya s invariantnoy meroy v stokhasticheskikh nelineynykh differentsial'nykh sistemakh [Finite dimensional distributions with invariant measure in stochastic nonlinear differential systems]. Moscow: Dialog–MGU. 129–140.
11. Pugachev, V. S., and I. N. Sinitsyn. 2001. *Stochastic systems. Theory and applications*. Singapore: World Scientific. 908 p.
12. Sinitsyn, I. N., E. R. Korepanov, and V. V. Belousov. 2002. Tochnye metody rascheta statsionarnykh rezhimov s invariantnoy meroy v stokhasticheskikh sistemakh upravleniya [Exact analysis of stationary with invariant measure regimes in stochastic control systems]. *Kibernetika i Tekhnologii XXI veka: Tr. II Mezhdunar. Nauch.-Tekhnich. Konf.* [Cybernetics and High Technologies of XXI Century. 2nd International Science and Technology Conference Proceedings] C&T'2002. Voronezh: Sakvoe. 124–131.
13. Sinitsyn, I. N., E. R. Korepanov, and V. V. Belousov. 2002. Tochnye analiticheskie metody v statisticheskoy dinamike nelineynykh informatsionno-upravlyayushchikh sistem [Exact analytical methods in statistical dynamics of nonlinear informational and control issue]. *Sistemy i Sredstva Informatiki — Systems and Means of Informatics*. Spets. Vyp. Matematicheskoe i algoritmicheskoe obespechenie informatsionno-telecommunikatsionnykh sistem [Mathematical sofware for information and telecommunication systems]. Moscow: Nauka. 112–121.
14. Sinitsyn, I. N. 2012. Razvitie metodov analiticheskogo modelirovaniya raspredeleniy s invariantnoy meroy v stokhasticheskikh sistemakh [Development of analytical modeling methods for distributions with invariant measure in stochastic systems]. *Sovremennye Problemy Prikladnoy Matematiki, Informatiki, Avtomatizatsii, Upravleniya: Materialy Mezhdunar. Seminara* [Modern Problems of Applied Mathematics Informatics, Atomization and Control: Seminar (International) Proceedings]. Sevastopol': SevNTU. 24–35.
15. Sinitsyn, I. N. 2012. Analiticheskoe modelirovanie raspredeleniy s invariantnoy meroy v stokhasticheskikh sistemakh s avtokorrelirovannymi shumami [Analytical modeling of distributions with invariant measure in stochastic systems with autocorrelated noise]. *Informatika i ee Primeneniya — Inform. Appl.* 6(4):4–8.
16. Sinitsyn, I. N. 2013. Analiticheskoe modelirovanie raspredeleniy s invariantnoy meroy v stokhasticheskikh sistemakh s razryvnymi kharakteristikami [Analytical modeling of distributions with invariant measure in stochastic systems with discontinuous nonlinearities]. *Informatika i ee Primeneniya — Inform. Appl.* 7(1):3–11.
17. Sinitsyn, I. N. 2013. Parametricheskoe statisticheskoe i analiticheskoe modelirovanie raspredeleniy v nelineynykh stokhasticheskikh sistemakh na mnogoobraziyakh [Parametric statistical and analytical modeling of distributions in stochastic systems on manifolds]. *Informatika i ee Primeneniya — Inform. Appl.* 7(2):4–16.
18. Sinitsyn, I. N., and V. I. Sinitsyn. 2013. *Lektsii po normal'noy i ellipsoidal'noy approksimatsii raspredeleniy v stokhasticheskikh sistemakh* [Lectures on normal and ellipsoidal approximation of distributions in stochastic systems]. Moscow: TORUS PRESS. 488 p.
19. Sinitsyn, I. N. 1986. Stochastic hereditary control systems. *Problems Control Inform. Theory* 15(4):287–298.
20. Sinitsyn, I. N. 1987. Konechnomernye raspredeleniya protsessov v stokhasticheskikh integral'nykh i integrodifferentsial'nykh sistemakh [Finite dimensional distributions of processes in stochastic integral and integrodifferential systems]. *2nd Symposium (International) IFAC on Stochastic Control*. Preprints. Vilnius, 1986. Pergamon Press. 1:144–153.
21. Sinitsyn, I. N., V. I. Sinitsyn, E. R. Korepanov, V. V. Belousov, I. V. Sergeev, and D. A. Basilashvili. 2012. Opyt modelirovaniya ereditarnykh stokhasticheskikh sistem [Experience of modeling in hereditary stochastic systems]. *Kibernetika i Vysokie Tekhnologii XXI Veka: Sbornik dokladov XIII Mezhdunar. Nauch.-Tekhnich. Konf.* [Cybernatics and High Technologies of XXI Century. 13th Scientific and Technological Conference (International) Proceedings]. Voronezh: Sakvoee. 2:346–357.
22. Sinitsyn, I. N. 2014. Analiz i modelirovanie raspredeleniy v ereditarnykh stokhasticheskikh sistemakh [Analysis and modeling of distributions in hereditary stochastic systems]. *Informatika i ee Primeneniya — Inform. Appl.* 8(1):2–11.
23. Nemytskiy, V. V., and V. V. Stepanov. 1949. *Kachestvennaya teoriya differentsial'nykh uravneniy* [Analytical theory of differential equations]. Moscow–Leningrad: Gostekhizdat. 448 p.
24. Kozlov, V. V. 1987. O sushchestvovanii integral'nogo invarianta gladkikh dinamicheskikh sistem [Existence of integral invariants in oblique dynamical systems]. *Prikladnaya Mekhanika i Matematika* [Appl. Mech. Math.] 1: 538–545.
25. Sinitsyn, I. N. 2007. *Fil'try Kalmana i Pugacheva* [Kalman and Pugachev filters]. 2nd ed. Moscow: Logos. 776 p.

Received January 16, 2014

Contributor

Sinitsyn Igor N. (b. 1940) — Doctor of Science in technology, professor, Honored scientist of RF, Head of Department, Institute of Informatics Problems, Russian Academy of Sciences, 44-2 Vavilov Str., Moscow 119333, Russian Federation; sinitsin@dol.ru

ИНФОРМАТИКА И ЕЁ ПРИМЕНЕНИЯ, 2014. Т. 8. Вып. 2. С. 15–27

СИСТЕМА Geo/Geo/1/R С ГИСТЕРЕЗИСНОЙ ПОЛИТИКОЙ*

А. В. Печинкин[1], Р. В. Разумчик[2]

Аннотация: Пороговое управление нагрузкой является одним из основных средств предотвращения перегрузок в сетях связи. Его разновидности применяются при обнаружении перегрузок как в сетях общеканальной сигнализации № 7, так и в сетях связи следующего поколения, где основой сигнализации служит протокол установления сессий (SIP — session initiation protocol). В работе рассматривается функционирующая в дискретном времени система массового обслуживания (СМО) Geo/Geo/1/R с двухпороговой гистерезисной политикой, представляющая собой одну из возможных математических моделей SIP-сервера с управлением нагрузкой. Предложены методы нахождения совместного стационарного распределения числа заявок в системе и состояния системы, распределения времен выхода системы из множества состояний нормальной работы и множества состояний перегрузки и блокировки и их моментов, стационарного распределения времени ожидания начала обслуживания заявки. Приведены примеры численных расчетов, проведенных с помощью полученных аналитических соотношений.

Ключевые слова: система массового обслуживания; дискретное время; гистерезисное управление нагрузкой; показатели функционирования системы

DOI: 10.14375/19922264140202

1 Введение и описание системы

Исследованию СМО с гистерезисным управлением посвящено много работ. Достаточно полный обзор результатов по гистерезисному управлению можно найти в работах [1–9]. Как видно из этих работ, большинство исследований относится к изучению СМО, функционирующих в непрерывном времени.

Однако, как давно было замечено, дискретные СМО позволяют учитывать дискретность функционирования телекоммуникационных систем и дискретный характер передаваемой информации. В настоящей работе представлен подробный анализ стационарных вероятностных и временны́х характеристик системы Geo/Geo/1 конечной емкости с двухпороговым гистерезисным управлением. Некоторые результаты исследования этой модели приведены в [10]. Однако предлагаемый здесь подход, а также используемые методы являются несколько отличными от представленных в [10] и позволяют проводить более глубокий анализ рассматриваемой системы.

Рассмотрим функционирующую в дискретном времени однолинейную СМО с накопителем ограниченной емкости, входящим геометрическим потоком заявок (вероятность поступления заявки на такте $a = a_1 + a_2$) и геометрическим распределением времени обслуживания (вероятность окончания обслуживания заявки на любом такте b). Простейший вариант гистерезисной политики, который здесь рассматривается, заключается в следующем. Пусть имеется три числа: R (максимальное число находящихся в системе заявок, или емкость системы), L (нижняя граница) и H (средняя граница), причем $L < H < R$. Каждая поступающая в систему заявка может быть одного из двух типов. С вероятностью a_1 поступает заявка первого типа, а с вероятностью a_2 — второго типа, причем поступление заявки одного типа исключает поступление заявки другого типа. Пусть в начальный момент система свободна. Тогда до того такта, после которого в системе станет H заявок, в нее принимаются все заявки. Но как только число заявок достигнет H, прекращается прием заявок второго типа. Далее заявки второго типа не принимаются до того такта, после которого число заявок в системе станет равным или $L-1$, или R. В первом случае в систему снова начинают приниматься заявки обоих типов. Во втором случае прекращается прием заявок первого типа, прибор занят только обслуживанием заявок, и так происходит до того такта, после которого число заявок станет H, после чего возобновляется прием заявок первого типа.

Для определенности будем считать, что если на некотором такте одновременно оканчивается обслуживание заявки на приборе и в систему поступает новая заявка, то число заявок в системе не изменяется. Кроме того, если к этому такту, кроме

*Работа выполнена при поддержке РФФИ (проекты №№ 12-07-00108, 13-07-00223 и 13-07-00665).

[1]Институт проблем информатики Российской академии наук; Российский университет дружбы народов, apechinkin@ipiran.ru

[2]Институт проблем информатики Российской академии наук, rrazumchik@ieee.org

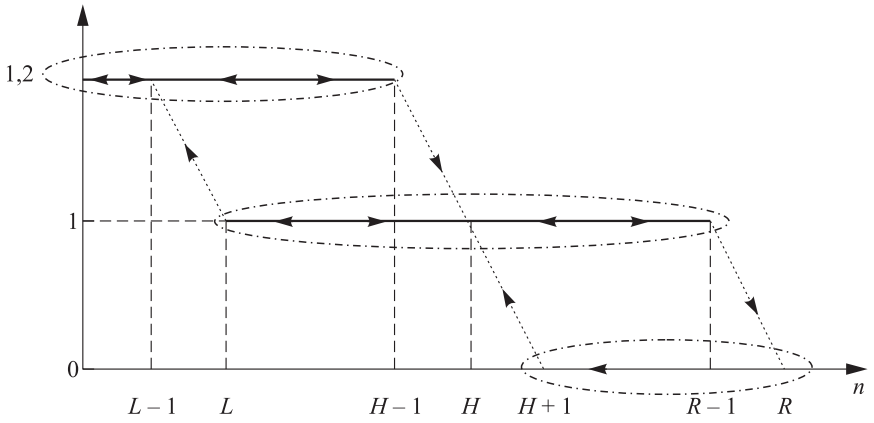


Рис. 1 Схема двухпорогового гистерезисного управления перегрузками

заявки на приборе, в системе не было других заявок, то вновь поступившая заявка сразу же начинает обслуживаться.

Описанная процедура обслуживания, носящая название гистерезисной политики, графически изображена на рис. 1, где на оси абсцисс отложено число заявок в системе. Уровень 1,2 соответствует состояниям, когда в систему принимаются заявки обоих типов, уровень 1 — принимаются заявки только первого типа, уровень 0 — в систему не принимаются заявки обоих типов. Возможные переходы отмечены стрелками.

Далее с целью сокращения записи для любой вероятности дополнительную вероятность будем снабжать чертой сверху. Например, $\overline{a} = 1 - a$, $\overline{b} = 1 - b$.

2 Цепь Маркова

Введем цепь Маркова $\{\nu(t),\ t \geq 0\}$, описывающую функционирование системы. Множество состояний цепи Маркова $\nu(t)$ имеет вид:

$$\mathcal{X} = \mathcal{X}_1 \cup \mathcal{X}_{20} \cup \mathcal{X}_{21} \cup \mathcal{X}_{31} \cup \mathcal{X}_{32}\,.$$

Здесь подмножество $\mathcal{X}_1$ включает состояния (n), $n = \overline{0, L-1}$, причем пребывание в состоянии (n) означает, что в системе находится n заявок и принимаются заявки обоих типов. Подмножество $\mathcal{X}_{20}$ содержит состояния $(n, 0)$, $n = \overline{L, H-1}$, где пребывание в состоянии $(n, 0)$ означает, что в системе находится n заявок и принимаются заявки обоих типов. Подмножество $\mathcal{X}_{21}$ состоит из состояний $(n, 1)$, $n = \overline{L, H-1}$, и при этом пребывание в состоянии $(n, 1)$ означает, что в системе находится n заявок и принимаются заявки только первого типа. Подмножество $\mathcal{X}_{31}$ включает состояния $(n, 1)$, $n = \overline{H, R-1}$, причем пребывание в состоянии $(n, 1)$ означает, что в системе находится n заявок и принимаются заявки только первого типа. Наконец, подмножество $\mathcal{X}_{32}$ содержит состояния $(n, 2)$, $n = \overline{H+1, R}$, а пребывание в состоянии $(n, 2)$ означает, что в системе находится n заявок и не принимаются заявки любого типа.

Будем считать, что значение цепи Маркова $\nu(t)$ образуют состояния системы непосредственно после такта t.

3 Вспомогательные функции

Определим вспомогательные функции, которые понадобятся в дальнейшем.

Пусть в начальный момент цепь Маркова $\nu(t)$ находится в состоянии $(n, 1) \in \mathcal{X}_{31}$, $n = \overline{H+1, R-1}$ (т. е. в системе имеется n заявок, причем в систему принимаются заявки только первого типа). Обозначим через c_n вероятность того, что $\nu(t)$ в состояние $(n-1, 1)$ попадет раньше, чем в состояние $(R, 2)$ (т. е. до того момента, когда в системе впервые останется $(n-1)$ заявок, в системе никогда не будет R заявок).

В состояние $(R-2, 1)$ из состояния $(R-1, 1)$ с вероятностью $\overline{a}_1 b$ можно попасть на первом же такте. Кроме того, из $(R-1, 1)$ в $(R-2, 1)$ раньше, чем в $(R, 2)$, можно попасть также с вероятностью $a_1 b + \overline{a}_1 \overline{b}$, оставшись на первом такте в состоянии $(R-1, 1)$, а затем уже с вероятностью c_{R-1} перейти из $(R-1, 1)$ в $(R-2, 1)$. Следовательно,

$$c_{R-1} = \overline{a}_1 b + (a_1 b + \overline{a}_1 \overline{b}) c_{R-1}\,. \qquad (1)$$

Из состояния $(n, 1)$, $n = \overline{H+1, R-2}$, в состояние $(n-1, 1)$ можно попасть, как и в предыдущем случае, или после первого же шага, или оставшись после первого такта в этом состоянии, а затем уже перейти в $(n-1, 1)$. Кроме того, из $(n, 1)$ в $(n-1, 1)$ раньше,

чем в $(R,2)$, можно попасть, перейдя на первом шаге с вероятностью $a_1\overline{b}$ в состояние $(n+1,1)$, затем, не заходя в $(R,2)$, с вероятностью c_{n+1} вернуться в состояние $(n,1)$ и, наконец, снова, не заходя в $(R,2)$, с вероятностью c_n попасть в $(n-1,1)$. Поэтому

$$c_n = \overline{a}_1 b + (a_1 b + \overline{a}_1\overline{b})c_n + a_1\overline{b}c_{n+1}c_n\,, \quad n = \overline{H+1, R-2}\,. \quad (2)$$

Пусть в начальный момент цепь Маркова $\nu(t)$ находится в состоянии $(n,0)\in\mathcal{X}_{20}$, $n=\overline{L+1,H-1}$ (т. е. в системе имеется n заявок, причем в систему принимаются заявки обоих типов). Обозначим через c_n вероятность того, что $\nu(t)$ попадет в состояние $(n-1,0)$ раньше, чем в состояние $(H,1)$ (т. е. до того момента, когда в системе впервые останется $(n-1)$ заявок, в системе никогда не будет H заявок). Следующие формулы получаются аналогично (1) и (2):

$$c_{H-1} = \overline{a}b + (ab + \overline{a}\overline{b})c_{H-1}\,; \quad (3)$$

$$c_n = \overline{a}b + (ab + \overline{a}\overline{b})c_n + a\overline{b}c_{n+1}c_n\,, \quad n = \overline{L+1, H-2}\,. \quad (4)$$

Пусть в начальный момент цепь Маркова $\nu(t)$ находится в состоянии $(n,1)\in\mathcal{X}_{21}$, $n=\overline{L,H-1}$ (т. е. в системе имеется n заявок, причем в систему принимаются заявки только первого типа). Обозначим через c_n^* вероятность того, что $\nu(t)$ попадет в состояние $(n+1,1)$ раньше, чем в состояние $(L-1)$ (т. е. до того момента, когда в системе впервые окажется $(n+1)$ заявок, в системе никогда не будет меньше L заявок). Поскольку из состояния $(L,1)$ на каждом такте можно попасть только в состояния $(L+1,1)$, $(L,1)$ и $(L-1)$, причем с вероятностями $a_1\overline{b}$, $a_1 b+\overline{a}_1\overline{b}$ и $\overline{a}_1 b$, то

$$c_L^* = a_1\overline{b} + (a_1 b + \overline{a}_1\overline{b})c_L^*\,. \quad (5)$$

Из состояния $(n,1)$, $n=\overline{L+1,H-1}$, в состояние $(n+1,1)$, не заходя в $(L-1)$, можно попасть с вероятностью $a_1\overline{b}$ после первого шага, а также оставшись после первого шага с вероятностью $a_1 b+\overline{a}_1\overline{b}$ в состоянии $(n,1)$, а затем уже с вероятностью c_n^* перейти из $(n,1)$ в $(n+1,1)$. Кроме того, не заходя в $(L-1)$, в это же состояние можно попасть, перейдя на первом шаге с вероятностью $\overline{a}_1 b$ в состояние $(n-1,1)$, затем, не заходя в $(L-1)$, с вероятностью c_{n-1}^* вернуться в состояние $(n,1)$ и, наконец, снова, не заходя в $(L-1)$, с вероятностью c_n^* попасть в $(n+1,1)$. Значит,

$$c_n^* = a_1\overline{b} + (a_1 b + \overline{a}_1\overline{b})c_n^* + \overline{a}_1 b c_{n-1}^* c_n^*\,, \quad n = \overline{L+1, H-1}\,. \quad (6)$$

Соотношения (1)–(6) позволяют получить рекуррентный алгоритм вычисления c_n и c_n^*. Действительно, c_n при $n=\overline{H+1,R-1}$ вычисляются последовательно от $n=R-1$ до $H+1$ по формулам (1) и (2):

$$c_{R-1} = \frac{\overline{a}_1 b}{a_1\overline{b} + \overline{a}_1 b}\,;$$

$$c_n = \frac{\overline{a}_1 b}{a_1\overline{b} + \overline{a}_1 b - a_1\overline{b}c_{n+1}}\,, \quad n = \overline{H+1, R-2}\,,$$

а при $n=\overline{L+1,H-1}$ — последовательно от $n=H-1$ до $L+1$ по формулам (3) и (4):

$$c_{H-1} = \frac{\overline{a}b}{a\overline{b} + \overline{a}b}\,;$$

$$c_n = \frac{\overline{a}b}{a\overline{b} + \overline{a}b - a\overline{b}c_{n+1}}\,, \quad n = \overline{L+1, H-2}\,.$$

Аналогично c_n^* вычисляются последовательно от $n=L$ до $n=H-1$ по формулам (5) и (6):

$$c_L^* = \frac{a_1\overline{b}}{a_1\overline{b} + \overline{a}_1 b}\,;$$

$$c_n^* = \frac{a_1\overline{b}}{a_1\overline{b} + \overline{a}_1 b - \overline{a}_1 b c_{n-1}^*}\,, \quad n = \overline{L+1, H-1}\,.$$

4 Стационарные вероятности состояний

Рассмотрим стационарный режим функционирования системы.

Введем обозначения (напомним, что состояние системы определяется после окончания такта):

- p_n, $n=\overline{0,L-1}$, — стационарная вероятность того, что система находится в состоянии (n) (в системе имеется n заявок);
- p_n, $n=\overline{L,H-1}$, — стационарная вероятность того, что система находится в состоянии $(n,0)$ (в системе имеется n заявок и к обслуживанию принимаются все заявки);
- p'_n, $n=\overline{L,R-1}$, — стационарная вероятность того, что система находится в состоянии $(n,1)$ (в системе имеется n заявок и к обслуживанию принимаются только заявки первого типа);
- p''_n, $n=\overline{H+1,R}$, — стационарная вероятность того, что система находится в состоянии $(n,2)$ (в системе имеется n заявок и новые заявки к обслуживанию не принимаются).

Выпишем систему уравнений равновесия (СУР).

Рассматривая потоки вероятностей переходов из состояний $(n-1)$ в (n), $n=\overline{1,L-1}$, и обратно, получаем

$$p_n\overline{a}b = p_{n-1}a\overline{b}\,,\quad n=\overline{1,L-1}\,. \tag{7}$$

Для нахождения p_n, $n=\overline{L,H-2}$, применим метод исключения состояний. С этой целью исключим все состояния $(i,0)$, $i=\overline{n+1,H-1}$ (в системе находится i, $i=\overline{n+1,H-1}$, заявок и к обслуживанию принимаются все заявки). После такого исключения из состояния $(n,0)$ можно перейти в состояние $(n-1,0)$ (с вероятностью $\overline{a}b$) и в состояние $(H,1)$ (с вероятностью $a\overline{b}\overline{c}_{n+1}$), а в состояние $(n,0)$ можно попасть из состояния $(n-1)$ (с вероятностью $a\overline{b}$) и из самого этого состояния (с вероятностью $ab+\overline{a}\overline{b}+a\overline{b}c_{n+1}$). Из уравнения глобального баланса для состояния $(n,0)$ имеем:

$$p_n = p_{n-1}a\overline{b} + p_n(ab+\overline{a}\overline{b}+a\overline{b}c_{n+1})\,,\quad n=\overline{L,H-2}\,. \tag{8}$$

Из уравнения глобального баланса для состояния $(H-1,0)$ (в системе находится $(H-1)$ заявок и к обслуживанию принимаются все заявки) находим

$$p_{H-1} = p_{H-2}a\overline{b} + p_{H-1}(ab+\overline{a}\overline{b})\,. \tag{9}$$

Снова обратимся к методу исключения состояний. Исключим состояния $(i,1)$, $i=\overline{H+1,R-1}$ (в системе находится более H заявок и к обслуживанию принимаются только заявки первого типа), состояния $(i,1)$, $i=\overline{L,H-1}$ (в системе находится менее H заявок и к обслуживанию принимаются только заявки первого типа), и состояния $(i,2)$, $i=\overline{H+1,R}$ (в системе находится более H заявок и заявки к обслуживанию не принимаются). После этого исключения из состояния $(H,1)$ можно будет перейти только в состояние $(L-1)$ (с вероятностью $\overline{a}_1b\overline{c}^*_{H-1}$), а в состояние $(H,1)$ можно попасть из состояния $(H-1,0)$ (с вероятностью $a\overline{b}$) и из самого состояния $(H,1)$ (с вероятностью $a_1b+\overline{a}_1\overline{b}+a_1\overline{b}+\overline{a}_1bc^*_{H-1}$). Поэтому из уравнения глобального баланса для этого состояния получаем

$$p'_H = p_{H-1}a\overline{b} + p'_H(a_1b+\overline{a}_1\overline{b}+a_1\overline{b}+\overline{a}_1bc^*_{H-1})\,. \tag{10}$$

Исключая состояния $(i,1)$, $i=\overline{n+1,R-1}$ (в системе находится i, $i=\overline{n+1,R-1}$, заявок и к обслуживанию принимаются только заявки первого типа), из уравнения глобального баланса для состояния $(n,1)$, $n=\overline{H+1,R-1}$, имеем:

$$p'_n = p'_{n-1}a_1\overline{b} + p'_n(a_1b+\overline{a}_1\overline{b}+a_1\overline{b}c_{n+1})\,,\quad n=\overline{H+1,R-2}\,, \tag{11}$$

$$p'_{R-1} = p'_{R-2}a_1\overline{b} + p'_{R-1}(a_1b+\overline{a}_1\overline{b})\,. \tag{12}$$

Из уравнения глобального баланса для состояния $(R,2)$ (в системе R заявок) находим:

$$bp''_R = p'_{R-1}a_1\overline{b}\,. \tag{13}$$

Аналогично из уравнения глобального баланса для состояния $(n,2)$, $n=\overline{H+1,R-1}$ (в системе находится n, $n=\overline{H+1,R-1}$, заявок и новые заявки в систему не принимаются), получаем:

$$p''_n = p''_{n+1}\,,\quad n=\overline{H+1,R-1}\,. \tag{14}$$

Наконец, исключая состояния i, $i=\overline{L,n-1}$ (в системе находится i, $i=\overline{L,n-1}$, заявок и к обслуживанию принимаются только заявки первого типа), из уравнения глобального баланса для состояния $(n,1)$, $n=\overline{L,H-1}$, имеем:

$$p'_n = p'_{n+1}\overline{a}_1b + p'_n(a_1b+\overline{a}_1\overline{b}+\overline{a}_1bc^*_{n-1})\,,\quad n=\overline{L+1,H-1}\,; \tag{15}$$

$$p'_L = p'_{L+1}\overline{a}_1b + p'_L(a_1b+\overline{a}_1\overline{b})\,. \tag{16}$$

Вероятность p_0 определяется из условия нормировки:

$$\sum_{n=0}^{H-1} p_n + \sum_{n=L}^{R-1} p'_n + \sum_{n=H+1}^{R} p''_n = 1\,. \tag{17}$$

Приведем алгоритм решения СУР.

Сначала по формулам (7)–(9) последовательно по n от $n=1$ до $(H-1)$ через p_0 вычисляются вероятности p_n:

$$p_n = \begin{cases} \dfrac{a\overline{b}}{\overline{a}b}p_{n-1}\,, & n=\overline{1,L-1}\,; \\ \dfrac{a\overline{b}}{a\overline{b}+\overline{a}b-a\overline{b}c_{n+1}}\,p_{n-1}\,, & n=\overline{L,H-2}\,; \end{cases}$$

$$p_{H-1} = \frac{a\overline{b}}{a\overline{b}+\overline{a}b}\,p_{H-2}\,.$$

Затем по формуле (10) находится p'_H:

$$p'_H = \frac{a\overline{b}}{\overline{a}_1b-\overline{a}_1bc^*_{H-1}}\,p_{H-1}\,,$$

а по формулам (11), (12) последовательно по n от $(H+1)$ до $(R-1)$ и по формулам (15), (16) последовательно по n от $(H-1)$ до L определяются p'_n:

$$p'_n = \frac{a_1\overline{b}}{a_1\overline{b} + \overline{a}_1 b - a_1\overline{b}c_{n+1}}\, p'_{n-1}\,, \qquad n = \overline{H+1, R-2}\,;$$

$$p'_{R-1} = \frac{a_1\overline{b}}{a_1\overline{b} + \overline{a}_1 b} p'_{R-2}\,;$$

$$p'_n = \frac{\overline{a}_1 b}{a_1\overline{b} + \overline{a}_1 b - \overline{a}_1 b c^*_{n-1}}\, p'_{n+1}\,, \qquad n = \overline{L+1, H-1}\,;$$

$$p'_L = \frac{\overline{a}_1 b}{a_1\overline{b} + \overline{a}_1 b}\, p'_{L+1}.$$

Далее по формулам (13) и (14) последовательно по n от $(R-1)$ до $(H+1)$ находятся p''_n:

$$p''_R = \frac{a_1\overline{b}}{b} p'_{R-1}\,,\ \ p''_n = p''_{n+1}\,,\ \ n = \overline{H+1, R-1}\,.$$

Наконец, вычисляется p_0 из условия нормировки (17).

В заключение этого раздела приведем выражение для стационарного среднего числа N заявок в системе:

$$N = \sum_{n=0}^{H-1} n p_n + \sum_{n=L}^{R-1} n p'_n + \sum_{n=H+1}^{R} n p''_n\,.$$

5 Время выхода из множества состояний

В этом разделе решим следующую задачу: найти распределения времен выхода цепи Маркова $\nu(t)$ из некоторых множеств состояний. Задача будет решена двумя способами. Первый способ заключается в последовательном вычислении вероятностей. Удобство этого способа применительно к настоящей системе заключается в том, что переходы цепи Маркова из каждого состояния происходят не более чем в три других состояния, что снижает число операций при вычислении вероятностей выхода на каждом шаге с n^2 операций до $3n$ операций. Второй способ — определение времен выхода в терминах производящих функций (ПФ). Такой способ обычно применяется при вычислении моментов случайных величин.

В качестве множеств, вероятности выхода из которых будут найдены, выбраны множество $\mathcal{Y}_0 = \mathcal{X}_0 \cup \mathcal{X}_{20}$ нормального функционирования системы и множество $\mathcal{Y}_1 = \mathcal{X}_{21} \cup \mathcal{X}_{31} \cup \mathcal{X}_{32}$, при пребывании системы в котором либо принимаются заявки только первого типа, либо вообще не принимаются никакие заявки, хотя используемые здесь методы применимы и для вычисления распределения времени до момента первого достижения из каждого состояния (или множества состояний) любого другого состояния (или множества состояний). Выбор множеств $\mathcal{Y}_0$ и $\mathcal{Y}_1$ связан с техническим приложением задачи к анализу качества управления SIP-сервером с помощью вариантов гистерезисной политики.

5.1 Последовательное вычисление вероятностей

Пусть в начальный момент система находится в состоянии n, $n = \overline{0, L-1}$, или в состоянии $(n, 0)$, $n = \overline{L, H-1}$ (в обоих этих случаях в систему принимаются заявки любых типов). Обозначим через $t_{n,i}$, $n = \overline{0, H-1}$, $i \geq 1$, вероятность того, что первый переход из множества состояний $\mathcal{X}_0 \cup \mathcal{X}_{20}$ в состояние H (т. е. в множество $\mathcal{X}_{21} \cup \mathcal{X}_{31} \cup \mathcal{X}_{32}$) произойдет на i-м шаге, через $\vec{t}_i = (t_{0,i}, \ldots, t_{H-1,i})^{\mathrm{T}}$, $i \geq 1$, — вектор размерности H с координатами $t_{n,i}$ и через $P = (p_{n,m})_{n,m=\overline{0,H-1}}$ — квадратную матрицу порядка H с ненулевыми элементами:

$$p_{n,n} = \begin{cases} \overline{a}\,, & n = 0\,, \\ \overline{a}\overline{b} + ab\,, & n = \overline{1, H-1}\,, \end{cases}$$

$$p_{n,n+1} = \begin{cases} a\,, & n = 0\,, \\ a\overline{b}\,, & n = \overline{1, H-2}\,, \end{cases}$$

$$p_{n,n-1} = \overline{a}b\,,\ \ n = \overline{1, H-1}\,.$$

Заметим теперь, что вектор $\vec{t}_1$ имеет вид:

$$\vec{t}_1 = (0, \ldots, 0, a\overline{b})^{\mathrm{T}}\,, \tag{18}$$

а векторы $\vec{t}_i$ при $i \geq 2$ определяются рекуррентной формулой

$$\vec{t}_i = P\vec{t}_{i-1}\,,\ \ i \geq 2\,. \tag{19}$$

Далее, пусть в начальный момент система находится в состоянии $(n, 1)$, $n = \overline{L, R-1}$ (в систему принимаются заявки только первого типа). Обозначим через $t'_{n,i}$, $n = \overline{L, R-1}$, $i \geq 1$, вероятность того, что первый переход из множества состояний $\mathcal{X}_{21} \cup \mathcal{X}_{31}$ в состояние $(L-1)$ (т. е. в множество $\mathcal{X}_0 \cup \mathcal{X}_{20}$) произойдет на i-м шаге.

Наконец, пусть в начальный момент система находится в состоянии $(n, 2)$, $n = \overline{H+1, R}$ (в систему не принимаются заявки любого типа). Обозначим через $t''_{n,i}$, $n = \overline{H+1, R}$, $i \geq 1$, вероятность того, что первый переход из множества состояний $\mathcal{X}_{32}$ в состояние $(L-1)$ (т. е. в множество $\mathcal{X}_0 \cup \mathcal{X}_{20}$) произойдет на i-м шаге.

Введем теперь вектор

$$\vec{t}^{\,*}_i = (t^*_{L,i}, \ldots, t^*_{2R-H-1,i})^{\mathrm{T}} = \\ = (t'_{L,i}, \ldots, t'_{R-1,i}, t''_{H+1,i}, \ldots, t''_{R,i})^{\mathrm{T}}\,,\ i \geq 1\,,$$

размерности $R-L+R-H=2R-L-H$, первые $(R-L)$ координат которого образуют вероятности $t'_{n,i}$, а остальные $(R-H)$ — вероятности $t''_{n,i}$, и квадратную матрицу

$$P^*=(p^*_{n,m})_{n,m=\overline{L,2R-H-1}}$$

порядка $2R-L-H$ с ненулевыми элементами:

$$p^*_{n,n}=\begin{cases}\overline{a}_1\overline{b}+a_1b\,, & n=\overline{L,R-1}\,;\\ \overline{b}\,, & n=\overline{R,2R-H-1}\,;\end{cases}$$

$$p^*_{n,n+1}=\begin{cases}a_1\overline{b}\,, & n=\overline{L,R-2}\,;\\ 0\,, & n=\overline{R-1,2R-H-2}\,;\end{cases}$$

$$p^*_{R-1,2R-H-1}=a_1\overline{b}\,;$$

$$p^*_{n,n-1}=\begin{cases}\overline{a}_1b\,, & n=\overline{L+1,R-1}\,;\\ b\,, & n=\overline{R+1,2R-H-1}\,;\end{cases}$$

$$p^*_{R,H}=b\,.$$

Заметим теперь, что вектор $\vec{t}^{\,*}_1$ имеет вид:

$$\vec{t}^{\,*}_1=(\overline{a}_1b,0,\ldots,0)^{\mathrm{T}}\,, \tag{20}$$

а векторы $\vec{t}^{\,*}_i$ при $i\geq 2$ определяются рекуррентной формулой

$$\vec{t}^{\,*}_i=P^*\vec{t}^{\,*}_{i-1}\,,\ \ i\geq 2\,. \tag{21}$$

Соотношения (18)–(21) позволяют вычислять вероятности моментов выхода из рассматриваемого множества состояний.

5.2 Применение производящих функций

Введем ПФ

$$T(z|n)=\sum_{i=1}^{\infty}z^it_{n,i}\,,\ \ n=\overline{0,H-1}\,;$$

$$T'(z|n)=\sum_{i=1}^{\infty}z^it'_{n,i}\,,\ \ n=\overline{L,R-1}\,;$$

$$T''(z|n)=\sum_{i=1}^{\infty}z^it''_{n,i}\,,\ \ n=\overline{H+1,R}\,.$$

Для вычисления ПФ времен выхода из состояний множества $\mathcal{X}_0\cup\mathcal{X}_{20}$ и из состояний множества $\mathcal{X}_{21}\cup\mathcal{X}_{31}\cup\mathcal{X}_{32}$ удобно воспользоваться одним из двух следующих способов.

Первый способ заключается в решении однородного разностного уравнения второго порядка с постоянными коэффициентами.

Второй способ аналогичен методу, примененному в разд. 4 для нахождения стационарных вероятностей состояний, и основан на исключении определенных состояний.

Чтобы продемонстрировать оба способа, решим задачу вычисления ПФ времени выхода из состояния множества $\mathcal{X}_0\cup\mathcal{X}_{20}$ первым способом, а из состояния множества $\mathcal{X}_{21}\cup\mathcal{X}_{31}\cup\mathcal{X}_{32}$ — вторым способом.

Вероятности $t_{n,i}$, $n=\overline{0,H-1}$, $i\geq 1$, удовлетворяют соотношениям

$$\begin{aligned}
t_{n,1}&=0\,,\ \ n=\overline{0,H-2}\,;\\
t_{H-1,1}&=a\overline{b}\,;\\
t_{0,i}&=\overline{a}t_{0,i-1}+at_{1,i-1}\,,\ \ i\geq 2\,;\\
t_{n,i}&=\overline{a}bt_{n-1,i-1}+(ab+\overline{a}\overline{b})t_{n,i-1}+a\overline{b}t_{n+1,i-1}\,,\\
&\qquad\qquad n=\overline{1,H-2}\,,\ \ i\geq 2\,;\\
t_{H-1,i}&=\overline{a}bt_{H-2,i-1}+(ab+\overline{a}\overline{b})t_{H-1,i-1}\,,\ \ i\geq 2\,,
\end{aligned}$$

откуда, переходя к ПФ, получаем

$$T(z|0)=\overline{a}zT(z|0)+azT(z|1)\,; \tag{22}$$

$$\begin{aligned}
T(z|n)=\overline{a}bzT(z|n-1)+(ab+\overline{a}\overline{b})zT(z|n)+\\
+a\overline{b}zT(z|n+1)\,,\ \ n=\overline{1,H-2}\,;
\end{aligned} \tag{23}$$

$$\begin{aligned}
T(z|H-1)=a\overline{b}z+\overline{a}bzT(z|H-2)+\\
+(ab+\overline{a}\overline{b})zT(z|H-1)\,.
\end{aligned} \tag{24}$$

Решение системы уравнений (22)–(24) имеет вид:

$$T(z|n)=C_1u_1^n+C_2u_2^n\,,\ \ n=\overline{0,H-1}\,,$$

где $u_1=u_1(z)$ и $u_2=u_2(z)$ — решения уравнения

$$u=\overline{a}bz+(ab+\overline{a}\overline{b})zu+a\overline{b}zu^2\,,$$

т. е.

$$\begin{aligned}
u_{1,2}&=\\
&=\frac{1-(ab+\overline{a}\overline{b})z\pm\sqrt{[1-(ab+\overline{a}\overline{b})z]^2-4a\overline{a}b\overline{b}z^2}}{2a\overline{b}z}\,.
\end{aligned}$$

Для вычисления коэффициентов $C_1=C_1(z)$ и $C_2=C_2(z)$ воспользуемся равенствами (22) и (24). Тогда

$$\begin{aligned}
C_1+C_2&=\overline{a}z(C_1+C_2)+az(C_1u_1+C_2u_2)\,;\\
C_1u_1^{H-1}+C_2u_2^{H-1}&=a\overline{b}z+\overline{a}bz(C_1u_1^{H-2}+C_2u_2^{H-2})+\\
&\quad+(ab+\overline{a}\overline{b})z(C_1u_1^{H-1}+C_2u_2^{H-1})\,.
\end{aligned}$$

Из этих равенств получаем, что коэффициенты C_1 и C_2 имеют вид:

$$\begin{aligned}
C_1=a\overline{b}z(\overline{a}z+azu_2-1)\big/\big((\overline{a}z+azu_2-1)\left[u_1-\right.\\
\left.-\overline{a}bz-(ab+\overline{a}\overline{b})zu_1\right]u_1^{H-2}+(\overline{a}z+azu_1-1)\times\\
\times\left[\overline{a}bz+(ab+\overline{a}\overline{b})zu_2-u_2\right]u_2^{H-2}\big)\,;
\end{aligned}$$

$$C_2 = a\overline{b}z(1-\overline{a}z-azu_1)/\left((\overline{a}z+azu_2-1)\left[u_1-\right.\right.$$
$$\left.-\overline{a}bz-(ab+\overline{a}\overline{b})zu_1\right]u_1^{H-2}+$$
$$\left.+(\overline{a}z+azu_1-1)\left[\overline{a}bz+(ab+\overline{a}\overline{b})zu_2-u_2\right]u_2^{H-2}\right).$$

Вычисление ПФ $T'(z|n)$, $n=\overline{L,R-1}$, и $T''(z|n)$, $n=\overline{H+1,R}$, начнем с введения вспомогательных ПФ:

$G^*(z|n)$, $n=\overline{H+1,R-1}$, — ПФ момента первого достижения состояния $(n-1,1)$ и вероятность того, что до этого момента система не попадет в состояние $(R,2)$, при условии, что в начальный момент система находилась в состоянии $(n,1)$;

$g^*(z|n)$, $n=\overline{H+1,R-1}$, — ПФ момента первого достижения состояния $(n+1,1)$ (или состояния $(R,2)$, если $n=R-1$) и вероятность того, что до этого момента система не попадет в состояние $(H,1)$, при условии, что в начальный момент система находилась в состоянии $(n,1)$;

$\tilde{G}(z|n)$, $n=\overline{H+1,R-1}$, — ПФ момента первого достижения состояния $(H,1)$ и вероятность того, что до этого момента система не попадет в состояние $(R,2)$, при условии, что в начальный момент система находилась в состоянии $(n,1)$;

$\tilde{g}(z|n)$, $n=\overline{H+1,R-1}$, — ПФ момента первого достижения состояния $(R,2)$ и вероятность того, что до этого момента система не попадет в состояние $(H,1)$, при условии, что в начальный момент система находилась в состоянии $(n,1)$;

$G'(z|n)$, $n=\overline{H+1,R-1}$, — ПФ момента первого достижения состояния $(H,1)$ при условии, что в начальный момент система находилась в состоянии $(n,1)$;

$G''(z|n)$, $n=\overline{H+1,R}$, — ПФ момента первого достижения состояния $(H,1)$ при условии, что в начальный момент система находилась в состоянии $(n,2)$.

Очевидно,

$$G''(z|n)=\left(\frac{bz}{1-\overline{b}z}\right)^{n-H},\quad n=\overline{H+1,R}.$$

Далее,

$$G^*(z|n)=z\left[\overline{a}_1b+(a_1b+\overline{a}_1\overline{b})G^*(z|n)+\right.$$
$$\left.+a_1\overline{b}G^*(z|n+1)G^*(z|n)\right],\quad n=\overline{H+1,R-1};\quad (25)$$

$$g^*(z|n)=z\left[\overline{a}_1bg^*(z|n-1)g^*(z|n)+\right.$$
$$\left.+(a_1b+\overline{a}_1\overline{b})g^*(z|n)+a_1\overline{b}\right],$$
$$n=\overline{H+1,R-1},\quad (26)$$

где положено

$$G^*(z|R)=0;\quad (27)$$
$$g^*(z|H)=0.\quad (28)$$

Решая уравнения (25) и (26), получаем

$$G^*(z|n)=z\frac{\overline{a}_1b}{1-(a_1b+\overline{a}_1\overline{b})z-a_1\overline{b}zG^*(z|n+1)},$$
$$n=\overline{H+1,R-1};\quad (29)$$

$$g^*(z|n)=z\frac{a_1\overline{b}}{1-\overline{a}_1bzg^*(z|n-1)-(a_1b+\overline{a}_1\overline{b})z},$$
$$n=\overline{H+1,R-1}.\quad (30)$$

Согласно формулам (27) и (29) функции $G^*(z|n)$, $n=\overline{H+1,R-1}$, являются дробно-рациональными функциями $G^*(z|n)=zG_n^*(z)/H_n^*(z)$, где полиномы $G_n^*(z)$ и $H_n^*(z)$ степеней $(R-n-1)$ и $(R-n)$ вычисляются по рекуррентным соотношениям:

$$G_{R-1}^*(z)=\overline{a}_1b;$$
$$H_{R-1}^*(z)=1-(a_1b+\overline{a}_1\overline{b})z;$$

$$G_n^*(z)=\overline{a}_1bH_{n+1}^*(z),\quad n=\overline{H+1,R-2};\quad (31)$$
$$H_n^*(z)=\left[1-(a_1b+\overline{a}_1\overline{b})z\right]H_{n+1}^*(z)-$$
$$-a_1\overline{b}z^2G_{n+1}^*(z),\quad n=\overline{H+1,R-2}.$$

Аналогично в соответствии с формулами (28) и (30) функции $g^*(z|n)$, $n=\overline{H+1,R-1}$, также являются дробно-рациональными функциями $g^*(z|n)=zg_n^*(z)/h_n^*(z)$ с полиномами $g_n^*(z)$ и $h_n^*(z)$ степеней $(n-H-1)$ и $(n-H)$, вычисляемыми рекуррентно следующим образом:

$$g_{H+1}^*(z)=a_1\overline{b};$$
$$h_{H+1}^*(z)=1-(a_1b+\overline{a}_1\overline{b})z;$$

$$g_n^*(z)=a_1\overline{b}h_{n-1}^*(z),\quad n=\overline{H+2,R-1};\quad (32)$$
$$h_n^*(z)=\left[1-(a_1b+\overline{a}_1\overline{b})z\right]h_{n+1}^*(z)-\overline{a}_1bz^2g_{n-1}^*(z),$$
$$n=\overline{H+2,R-1}.$$

Из приведенных соотношений получаем выражения для $\tilde{G}(z|n)$ и $\tilde{g}(z|n)$, с учетом (31) и (32) имеющие вид:

$$\tilde{G}(z|H+1)=G^*(z|H+1)=z\frac{G_{H+1}^*(z)}{H_{H+1}^*(z)};$$

$$\tilde{G}(z|n)=G^*(z|n)\tilde{G}(z|n-1)=$$
$$=(\overline{a}_1bz)^{n-H-1}z\frac{G_n^*(z)}{H_{H+1}^*(z)},\quad n=\overline{H+2,R-1};$$

$$\tilde{g}(z|R-1)=g^*(z|R-1)=z\frac{g_{R-1}^*(z)}{h_{R-1}^*(z)};$$

$$\tilde{g}(z|n) = g^*(z|n)\tilde{G}(z|n+1) = \\ = (a_1\overline{b}z)^{R-n-1} z \frac{g_n^*(z)}{h_{R-1}^*(z)}\,,\ \ n = \overline{H+1, R-2}\,.$$

Вспомогательная функция $G'(z|n)$ определяется формулой:

$$G'(z|n) = \tilde{G}(z|n) + \tilde{g}(z|n)G''(z|R) = z\frac{G'_n(z)}{H'_{H+1}(z)}\,, \\ n = \overline{H+1, R-1}\,,$$

где полиномы $G'_n(z)$ и $H'_{H+1}(z)$ степеней $(2R-2H-2)$ и $(2R-2H-1)$ задаются равенствами:

$$G'_n(z) = (1-\overline{b}z)^{R-H}(\overline{a}_1 bz)^{n-H-1}G_n^*(z) + \\ + (bz)^{R-H}(a_1\overline{b}z)^{R-n-1}g_n^*(z)\,,\ \ n = \overline{H+1, R-1}\,;$$

$$H'_{H+1}(z) = (1-\overline{b}z)^{R-H}H_{H+1}^*(z)\,.$$

Последняя необходимая вспомогательная функция $G'(z|n)$, $n = \overline{L, H}$, представляющая собой ПФ момента первого достижения состояния $(n-1, 1)$ при условии, что в начальный момент система находилась в состоянии $(n, 1)$, удовлетворяет уравнению

$$G'(z|n) = z\left[\overline{a}_1 b + (a_1 b + \overline{a}_1\overline{b})G'(z|n) + \right. \\ \left. + a_1\overline{b}G'(z|n+1)G'(z|n)\right]\,,\ \ n = \overline{L, H}\,,$$

т. е. определяется равенством:

$$G'(z|n) = \frac{\overline{a}_1 bz}{1-(a_1 b + \overline{a}_1\overline{b})z - a_1\overline{b}zG'(z|n+1)} = \\ = z\frac{G'_n(z)}{H'_n(z)}\,,\ \ n = \overline{L, H}\,,$$

где полиномы $G'_n(z)$, $n = \overline{L, H}$, и $H'_n(z)$, $n = \overline{L, H}$, вычисляются по формулам:

$$G'_n(z) = \overline{a}_1 bzH'_{n+1}(z)\,,\ \ n = \overline{L, H}\,; \\ H'_n(z) = \left[1-(a_1 b + \overline{a}_1\overline{b})z\right]H'_{n+1}(z) - \\ - a_1\overline{b}z^2 G'_{n+1}(z)\,,\ \ n = \overline{L, H}\,.$$

Теперь можно привести выражения для ПФ $T'(z|n)$, $n = \overline{L, R-1}$, и $T''(z|n)$, $n = \overline{H+1, R}$:

$$T'(z|L) = G'(z|L) = z\frac{G'_L(z)}{H'_L(z)}\,;$$

$$T'(z|n) = \begin{cases} G'(z|n)T'(z|n-1) = \\ = (\overline{a}_1 bz)^{n-L} z\dfrac{G'_n(z)}{H'_L(z)}\,, \quad n = \overline{L+1, H}; \\ G'(z|n)T'(z|H) = (\overline{a}_1 bz)^{H-L+1} z\dfrac{G'_n(z)}{H'_L(z)}\,, \\ \qquad\qquad n = \overline{H+1, R-1}; \end{cases}$$

$$T''(z|n) = G''(z|n)T'(z|H) = \\ = (\overline{a}_1 bz)^{H-L+1} z\frac{G'_H(z)}{H'_L(z)}\left(\frac{bz}{1-\overline{b}z}\right)^{n-H}\,, \\ n = \overline{H+1, R}\,,$$

решающие задачу вычисления распределения моментов выхода из рассматриваемого множества состояний в терминах ПФ.

6 Средние времена

Пусть в начальный момент система находится в состоянии $(n, 1)$, $n = \overline{L, R-1}$ (в системе имеется n заявок, причем к обслуживанию принимаются заявки первого типа), или в состоянии $(n, 2)$, $n = \overline{H+1, R}$ (в системе имеется n заявок и к обслуживанию новые заявки не принимаются). Вычислим M_n, $n = \overline{L, R-1}$, и M_n^*, $n = \overline{H+1, R}$, — средние времена первого достижения состояния $(L-1)$ (в системе впервые останется $L-1$ заявок) для первого и второго вариантов. При этом не будем пользоваться результатами предыдущего раздела, дифференцируя соответствующие ПФ, а выведем для средних простые рекуррентные соотношения.

Введем сначала следующие величины:

m_n, $n = \overline{L, H}$, — среднее время до того момента, когда система впервые попадет в состояние $(n-1, 1)$ (в системе впервые останется $n-1$ заявок), при условии, что в начальный момент система находилась в состоянии $(n, 1)$ (в системе было n заявок и принимались заявки только первого типа);

m_n^*, $n = \overline{H+1, R}$, — среднее время до того момента, когда система впервые попадет в состояние $(n-1, 2)$ (в системе впервые останется $n-1$ заявок), при условии, что в начальный момент система находилась в состоянии $(n, 2)$ (в системе было n заявок и новые заявки не принимались);

m_n, $n = \overline{H+1, R-1}$, — среднее время до того момента, когда система впервые попадет или в состояние $(n-1, 1)$ (в системе впервые останется $n-1$ заявок), причем всегда принимались заявки первого типа, или в состояние $(H, 1)$ (в системе впервые останется H заявок), причем с некоторого момента заявки в систему не принимались, при условии, что в начальный момент система находилась в состоянии$(n, 1)$ (в системе было n заявок и принимались заявки первого типа);

$\tilde{m}_n$, $n = \overline{H+1, R-1}$, — среднее время до того момента, когда система впервые попадет в состояние $(H,1)$ (в системе впервые останется H заявок), при условии, что в начальный момент система находилась в состоянии $(n,1)$ (в системе было n заявок и принимались заявки первого типа).

Поскольку из состояния $(n,2)$, $n = \overline{H+1, R}$, система может попасть только в состояние $(n-1,2)$ (в случае $n = H+1$ только в состояние $(H,1)$), причем за геометрически распределенное с параметром b время, то

$$m_n^* = \frac{1}{b}\,, \quad n = \overline{H+1, R}\,.$$

Далее, из состояния $(R-1,1)$ за один шаг можно попасть с вероятностью $\overline{a}_1 b$ в состояние $(R-2,1)$, с вероятностью $a_1\overline{b}$ в состояние $(R,2)$ и, наконец, остаться в этом состоянии с вероятностью $a_1 b + \overline{a}_1\overline{b}$. При попадании в состояние $(R,2)$ система не может попасть в состояние $(R-2,1)$ раньше, чем в состояние $(H,1)$, причем, как нетрудно видеть, среднее время перехода из состояния $(R,2)$ в состояние $(H,1)$ равно $(R-H)/b$. Оставшись в состоянии $(R-1,1)$, система впервые попадет или в состояние $(n-1,1)$, или в состояние $(H,1)$ за среднее время m_{R-1}. Поэтому

$$m_{R-1} = 1 + a_1\overline{b}\frac{R-H}{b} + (a_1 b + \overline{a}_1\overline{b})m_{R-1}\,,$$

или

$$m_{R-1} = \frac{1}{a_1\overline{b} + \overline{a}_1 b}\left(1 + a_1\overline{b}\frac{R-H}{b}\right)\,.$$

Из состояния $(n,1)$, $n = \overline{H+1, R-2}$, так же, как и прежде, за один шаг можно попасть с вероятностью $\overline{a}_1 b$ в состояние $(n-1,1)$, с вероятностью $a_1\overline{b}$ — в состояние $(n+1,1)$ и, наконец, остаться в этом состоянии с вероятностью $a_1 b + \overline{a}_1\overline{b}$. Однако, в отличие от предыдущего случая, при попадании в состояние $(n+1,1)$ система за среднее время m_{n+1} может впервые или вернуться в состояние $(n,1)$, или попасть в состояние $(H,1)$. При этом она с вероятностью c_{n+1} возвратится в состояние $(n,1)$, и тогда до момента первого попадания или в состояние $(n-1,1)$, или в состояние $(H,1)$ пройдет дополнительно среднее время m_n. Таким образом,

$$m_n = 1 + a_1\overline{b}(m_{n+1} + c_{n+1}m_n) + (a_1 b + \overline{a}_1\overline{b})m_n\,, \quad n = \overline{H+1, R-2}\,,$$

и

$$m_n = \frac{1}{a_1\overline{b} + \overline{a}_1 b - a_1\overline{b}c_{n+1}}\left(1 + a_1\overline{b}m_{n+1}\right)\,, \quad n = \overline{H+1, R-2}\,.$$

Заметим теперь, что, попав из состояния $(n,1)$, $n = \overline{L, H}$, в состояние $(n+1,1)$, система за среднее время m_{n+1} обязательно возвратится в состояние $(n,1)$. Поступая, как и раньше, имеем:

$$m_n = 1 + a_1\overline{b}(m_{n+1} + m_n) + (a_1 b + \overline{a}_1\overline{b})m_n\,, \quad n = \overline{L, H}\,,$$

т. е.

$$m_n = \frac{1}{a_1\overline{b} + \overline{a}_1 b - a_1\overline{b}}\left(1 + a_1\overline{b}m_{n+1}\right)\,, \quad n = \overline{L, H}\,.$$

Вычислим $\tilde{m}_n$, $n = \overline{H+1, R-1}$. Очевидно,

$$\tilde{m}_{H+1} = m_{H+1}\,.$$

В остальных случаях среднее время первого достижения состояния $(H,1)$ состоит из среднего времени первого попадания или в состояние $(n-1,1)$, или в состояние $(H,1)$. Кроме того, в случае попадания в состояние $(n-1,1)$ (с вероятностью c_n) нужно добавить еще среднее время $\tilde{m}_{n-1}$ первого попадания в состояние $(H,1)$ из состояния $(n-1,1)$. Значит,

$$\tilde{m}_n = m_n + c_n\tilde{m}_{n-1}\,, \quad n = \overline{H+2, R-1}\,.$$

Формулы для вычисления M_n и M_n^* в силу их очевидности приводятся здесь без пояснений:

$$M_L = m_L\,;$$
$$M_n = m_n + M_{n-1}\,, \quad n = \overline{L+1, H}\,;$$
$$M_n = \tilde{m}_n + M_H\,, \quad n = \overline{H+1, R-1}\,;$$
$$M_n^* = \frac{n-H}{b} + M_H\,, \quad n = \overline{H+1, R}\,.$$

Полученные результаты позволяют получить простой рекуррентный алгоритм вычисления средних времен M_n и M_n^*, который в силу его элементарности здесь не приводится.

7 Некоторые стационарные характеристики

В этом разделе найдем выражения для некоторых важных стационарных показателей функционирования системы. При этом будем предполагать, что заявки обоих типов образуют общую очередь и обслуживаются в порядке поступления.

Проще всего вычисляются вероятности π_1 и π_2 потерь заявок первого и второго типа. Поскольку заявка первого типа теряется только в том случае, когда после предыдущего такта система находится в

одном из состояний $(n,2)$, $n=\overline{H+1,R}$ (в системе имеется от $H+1$ до R заявок и новые заявки не принимаются), то

$$\pi_1=\sum_{n=H+1}^{R}p''_n\,.$$

Аналогично

$$\pi_2=\sum_{n=L}^{R-1}p'_n+\sum_{n=H+1}^{R}p''_n\,.$$

Обратимся теперь к стационарным вероятностям состояний по моментам поступления заявок в систему.

Введем обозначения:

p^*_n, $n=\overline{0,R-1}$, — стационарная вероятность того, что поступившая (принятая к обслуживанию) заявка первого типа застанет перед собой сразу же после поступления n других заявок (любого типа);

$p^{*\prime}_n$, $n=\overline{0,H-1}$, — стационарная вероятность того, что поступившая (принятая к обслуживанию) заявка второго типа застанет перед собой сразу же после поступления n других заявок (любого типа).

Поскольку перед поступившей заявкой первого типа будут отсутствовать другие заявки, только если она поступает в свободную систему (с вероятностью p_0) или если перед ее поступлением в системе была заявка (с вероятностью p_1), которая на данном такте обслужилась (с вероятностью b), то с учетом условия принятия заявки в систему (вероятность $1-\pi_1$) имеем:

$$p^*_0=\frac{1}{1-\pi_1}(p_0+p_1b)\,.$$

Подобные рассуждения для случая $n=\overline{1,L-2}$ дают

$$p^*_n=\frac{1}{1-\pi_1}(p_n\overline{b}+p_{n+1}b)\,,\ \ n=\overline{1,L-2}\,.$$

Далее, перед поступившей заявкой первого типа будет $L-1$ других заявок, если или перед ее поступлением в системе было $L-1$ заявок (с вероятностью p_{L-1}) и ни одна из них не обслужилась (с вероятностью $\overline{b}$), или если перед ее поступлением в системе было L заявок (с вероятностью $p_L+p^*_L$, поскольку заявки второго типа могут как приниматься, так и не приниматься в систему) и одна обслужилась (с вероятностью b). Поэтому

$$p^*_{L-1}=\frac{1}{1-\pi_1}(p_{L-1}\overline{b}+p_Lb+p'_Lb)\,.$$

Незначительные изменения в рассуждениях приводят к формулам:

$$p^*_n=\frac{1}{1-\pi_1}(p_n\overline{b}+p_{n+1}b+p'_n\overline{b}+p'_{n+1}b)\,,\qquad n=\overline{L,H-2}\,;$$

$$p^*_{H-1}=\frac{1}{1-\pi_1}\left(p_{H-1}\overline{b}+p'_Hb+p'_{H-1}\overline{b}\right)\,;$$

$$p^*_H=\frac{1}{1-\pi_1}\left(p'_H\overline{b}+p'_{H+1}b+p''_{H+1}b\right)\,;$$

$$p^*_n=\frac{1}{1-\pi_1}\left(p'_n\overline{b}+p'_{n+1}b\right)\,,\ \ n=\overline{H+1,R-2}\,;$$

$$p^*_{R-1}=\frac{1}{1-\pi_1}p'_{R-1}\overline{b}\,.$$

Следующие формулы для $p^{*\prime}_n$ приводятся без пояснений:

$$p^{*\prime}_0=\frac{1}{1-\pi_2}\left(p_0+p_1b\right)\,;$$

$$p^{*\prime}_n=\frac{1}{1-\pi_2}\left(p_n\overline{b}+p_{n+1}b\right)\,,\ n=\overline{1,H-2}\,,\qquad n\neq L-1\,;$$

$$p^{*\prime}_{L-1}=\frac{1}{1-\pi_2}\left(p_{L-1}\overline{b}+p_Lb+p'_Lb\right)\,;$$

$$p^{*\prime}_{H-1}=\frac{1}{1-\pi_2}\,p_{H-1}\overline{b}\,.$$

Поскольку распределение времени ожидания начала обслуживания заявки, перед которой в очереди находится n других заявок, имеет распределение Паскаля с параметрами b и n, то ПФ $\omega_1(z)$ и $\omega_2(z)$ стационарных распределений времен ожидания начала обслуживания заявок первого и второго типа определяются выражениями:

$$\omega_1(z)=\sum_{n=0}^{R-1}\left(\frac{bz}{1-\overline{b}z}\right)^n p^*_n\,;$$

$$\omega_2(z)=\sum_{n=0}^{H-1}\left(\frac{bz}{1-\overline{b}z}\right)^n p^{*\prime}_n\,.$$

В частности, стационарные средние времена w_1 и w_2 ожидания начала обслуживания заявок первого и второго типа имеют вид:

$$w_1=\omega'_1(1)=\frac{1}{b}\sum_{n=0}^{R-1}np^*_n=\frac{1}{b}\,N^*_1\,;$$

$$\omega_2(z)=\omega'_2(1)=\frac{1}{b}\sum_{n=0}^{H-1}np^{*\prime}_n=\frac{1}{b}\,N^*_2\,,$$

где N^*_1 и N^*_2 — стационарные средние числа заявок, которых застают в очереди поступающие в систему заявки первого и второго типа.

Наконец, используя формулу Литтла (см., например, [11]), получаем выражения для стационарных средних N_1 и N_2 чисел заявок первого и второго типа в очереди

$$N_1 = a_1(1-\pi_1)w_1\,;$$
$$N_2 = a_2(1-\pi_2)w_2\,.$$

8 Примеры расчетов

На основе полученных результатов была написана программа, позволяющая вычислять распределения времен выхода из множества состояний, совместное стационарное распределение числа заявок в системе и состояния системы и связанные с ним характеристики, а также исследовать поведение рассматриваемой СМО в зависимости от значений определяющих ее исходных параметров.

Приведем лишь некоторые из результатов расчетов. Всюду в дальнейшем предполагается, что $R = 100$, $H = 70$, $L = 40$, а вероятность обслуживания заявки на такте $b = 0{,}2$. Загрузка системы обозначается через $\rho = (a_1 + a_2)/b$.

Так как простое выписывание всех совместных стационарных вероятностей состояний занимает очень много места, представим их графически (рис. 2). На рис. 2, *а*, при загрузке $\rho = 1{,}5$ и вероятности поступления заявки первого типа, равной $a_1 = 0{,}21$, изображено поведение вероятностей p_n, p'_n и p''_n как функций от числа заявок в системе n.

На рис. 2, *б* изображено поведение вероятностей p_n, p'_n и p''_n как функций от n при загрузке $\rho = 1{,}5$ и вероятности поступления заявки первого типа, равной $a_1 = 0{,}15$.

С точки зрения практического применения подобных систем к моделированию SIP-серверов от-

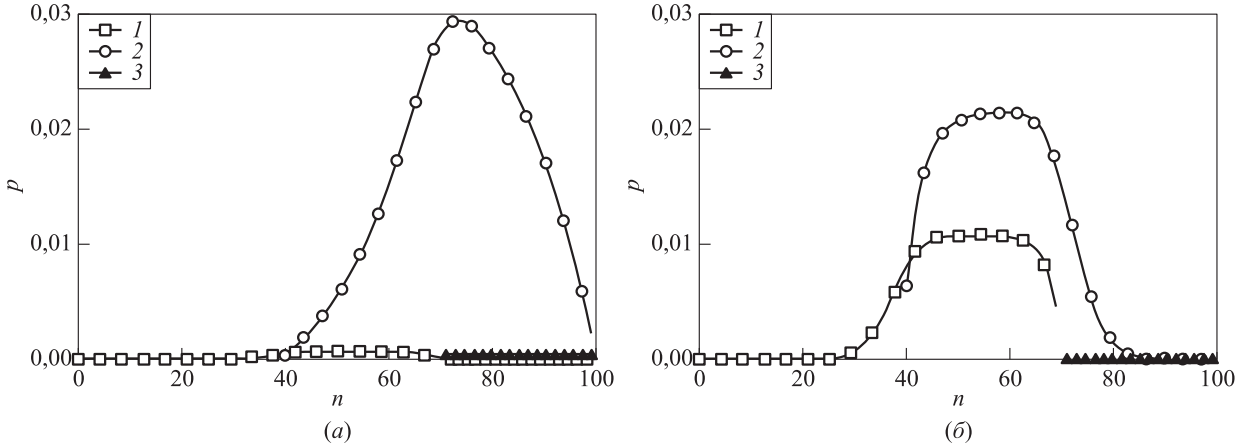


Рис. 2 Поведение вероятностей p_n (*1*), p'_n (*2*) и p''_n (*3*) как функций от n ($b = 0{,}2$): (*а*) $a_1 = 0{,}21$; $a_2 = 0{,}09$; (*б*) $a_1 = 0{,}15$; $a_2 = 0{,}15$

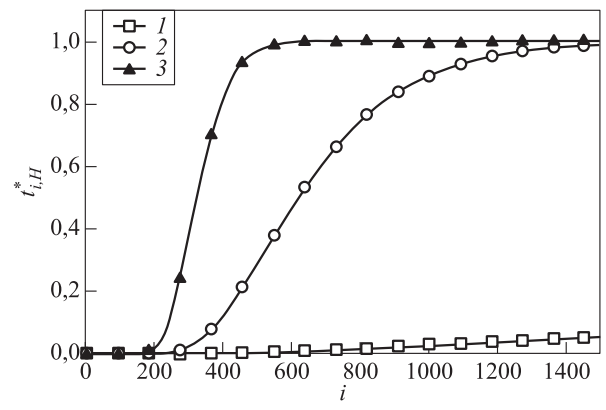


Рис. 3 Распределение времени (числа тактов i) до первого достижения состояния $(L-1)$ из состояния H ($b = 0{,}2$, $\rho = 1{,}5$, $a_2 = 0{,}3 - a_1$): *1* — $a_1 = 0{,}21$; *2* — $0{,}15$; *3* — $a_1 = 0{,}09$

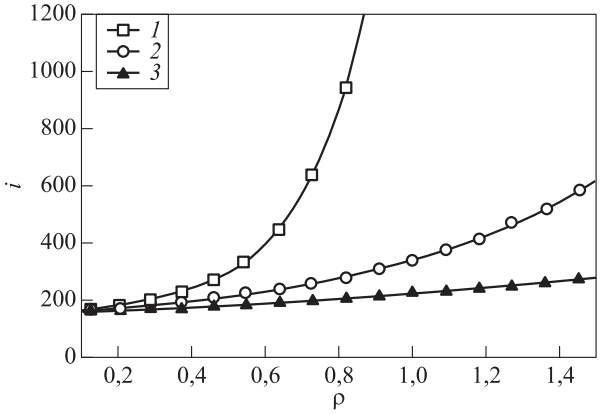


Рис. 4 Среднее время (число тактов) до первого достижения состояния $(L-1)$ из состояния H в зависимости от загрузки системы ρ ($b = 0{,}2$): *1* — $a_1 = 2{,}3a_2$; *2* — $a_1 = a_2$; *3* — $a_1 \approx 0{,}4a_2$

дельный интерес представляют временны́е характеристики работы системы. Они позволяют выявлять эффективность работы гистерезисного управления и подлежат оптимизации. Как упоминается в [6], одной из таких характеристик является время выхода системы из режима перегрузки. Моментом входа в режим перегрузки можно считать момент, когда число заявок в системе впервые стало равным H. Поэтому обратимся к распределению времени (числа тактов i) до первого достижения состояния $(L-1)$ из состояния H и его среднему значению (т. е. $t^*_{i,H}$, $i \geq 0$, и M_H в обозначениях разд. 5 и 6 соответственно). На рис. 3 представлено распределение времени (числа тактов i) достижения состояния $(L-1)$ из H при загрузке $\rho = 1{,}5$ для различных вероятностей поступления заявки первого типа ($a_1 = 0{,}21$, $0{,}15$ и $0{,}09$).

Наконец, на рис. 4 представлено поведение среднего времени (числа тактов) до первого достижения состояния $(L-1)$ из состояния H в зависимости от загрузки системы ρ. Рассмотрено 3 соотношения для потоков первого и второго типа, а именно: рассмотрен случай, когда заявки первого типа поступают в систему чаще, чем заявки второго типа ($a_1 \approx 2{,}3\, a_2$), случай, когда в среднем в систему поступает одинаковое число заявок обоих типов ($a_1 = a_2$), и случай, когда заявки второго поступают в систему чаще, чем заявки первого типа ($a_1 \approx 0{,}4\, a_2$).

Полученные аналитические результаты были проверены путем сравнения с результатами работы имитационной модели, разработанной на языке GPSS (General Purpose Simulation System). Сравнения показали высокую точность расчетов, проведенных на основе аналитических соотношений.

Литература

1. *Gebhart R. F.* A queuing process with bilevel hysteretic service-rate control // Nav. Res. Logist. Q., 1967. Vol. 14. No. 1. P. 55–67.
2. *Golubchik L., Lui J. C. S.* Bounding of performance measures for a threshold-based queueing system with hysteresis // Newsl. ACM SIGMETRICS Performance Evaluation Rev., 1997. Vol. 25. No. 1. P. 147–157.
3. *Dshalalow J. H.* Queueing systems with state dependent parameters // Frontiers in queueing: Models and applications in science and engineering. — Boca Raton: CRC Press, 1997. P. 61–116.
4. *Roughan M., Pearce C.* A martingale analysis of hysteretic overload control // Adv. Performance Anal. J. Teletraffic Theory Performance Anal. Comm. Syst. Networks, 2000. Vol. 3. No. 1. P. 1–30.
5. *Bekker R.* Queues with Levy input and hysteretic control // Queueing Syst., 2009. Vol. 63. No. 1. P. 281–299.
6. *Abaev P. O., Gaidamaka Yu. V., Pechinkin A. V., Razumchik R. V., Shorgin S. Ya.* Simulation of overload control in SIP Server Networks // 26th European Conference on Modelling and Simulation (ECMS 2012) Proceedings. — Koblenz: Digitaldruck Pirrot GmbH, 2012. P. 533–539.
7. *Жерновый К. Ю., Жерновый Ю. В.* Система $M^\theta/G/1$ с гистерезисным переключением интенсивности обслуживания // Информационные процессы, 2012. Т. 12. № 3. С. 176–190.
8. *Abaev P., Pechinkin A., Razumchik R.* Analysis of queueing system with constant service time for SIP server hop-by-hop overload control // Modern Probab. Meth. Anal. Telecommunication Networks Comm. Comp. Information Sci., 2013. Vol. 356. P. 1–10. doi:10.1007/978-3-642-35980-4_1.
9. *Shorgin S., Samouylov K., Gaidamaka Yu., Etezov Sh.* Polling system with threshold control for modeling of SIP server under overload // 18th Conference (International) on Systems Science (ICSS 2013) Proceedings. Advances in intelligent systems and computing ser., 2014. Vol. 240. P. 97–107.
10. *Разумчик Р. В., Абаев П. О., Корабельников Д. М., Пяткина Д. А.* Моделирование SIP-сервера с гистерезисным управлением нагрузкой с помощью системы массового обслуживания в дискретном времени // Научные труды ФГУП ЦНИИС: Сб. статей. — М.: ЦНИИС, 2011. С. 119–127.
11. *Bocharov P. P., D'Apice C., Pechinkin A. V., Salerno S.* Queueing theory. — Utrecht, Boston: VSP, 2004.

Поступила в редакцию 2.04.14

PERFORMANCE CHARACTERISTICS OF Geo/Geo/1/R QUEUE WITH HYSTERETIC LOAD CONTROL

A. V. Pechinkin[1,2] and R. V. Razumchik[1]

[1]Institute of Informatics Problems, Russian Academy of Sciences, 44-2 Vavilov Str., Moscow 119333, Russian Federation

[2]Peoples' Friendship University of Russia, 6 Miklukho-Maklaya Str., Moscow 117198, Russian Federation

Abstract: Threshold load control is one of the key techniques to prevent overloads in telecommunication networks. Its variants are used for overload detection in signalling system No. 7 as well as in next generation networks where

session initiation protocol (SIP) is the main signalling protocol. Consideration is given to discrete-time Geo/Geo/1/R queueing system which is one of the possible mathematical models of the SIP proxy-server. Bi-level hysteretic load control is implemented in the system. The methods that allow one to obtain a stationary joint probability distribution of the number of customers in a system and system's state, stationary waiting and sojourn time distributions, and distribution of first passage times from different system's states are presented. A numerical example based on obtained analytical expressions is given.

Keywords: queueing system; discrete time; hysteretic load control; performance characteristics

DOI: 10.14375/19922264140202

Acknowledgments

The research was supported by the Russian Foundation for Basic Research (projects Nos. 12-07-00108, 13-07-00223, and 13-07-00665).

References

1. Gebhart, R. F. 1967. A queuing process with bilevel hysteretic service-rate control. *Nav. Res. Logist. Q.* 14(1):55–68.
2. Golubchik, L., and C. S. Lui. 1997. Bounding of performance measures for a threshold-based queueing system with hysteresis. *Newsl. ACM SIGMETRICS Performance Evaluation Rev.* 25(1):147–157.
3. Dshalalow, J. H. 1997. Queueing systems with state dependent parameters. *Frontiers in queueing: Models and applications in science and engineering*. Boca Raton: CRC Press. 61–116.
4. Roughan, M., and C. Pearce. 2000. A martingale analysis of hysteretic overload control. *Adv. Performance Anal. J. Teletraffic Theory Performance Anal. Comm. Syst. Networks* 3:1–30.
5. Bekker, R. 2009 Queues with Levy input and hysteretic control. *Queueing Syst.* 63(1):281–299.
6. Abaev, P. O., Yu. V. Gaidamaka, A. V. Pechinkin, R. V. Razumchik, and S. Ya. Shorgin. 2012. Simulation of overload control in SIP Server Networks. *26th European Conference on Modelling and Simulation Proceedings*. Koblenz: Digitaldruck Pirrot GmbH. 533–539.
7. Zhernovyi, K., and Yu. Zhernovyi. 2012. Sistema $M^{\theta}/G/1$ s gisterezisnym pereklyucheniem intensivnosti obsluzhivaniya [The system $M^{\theta}/G/1$ with a hysteretic switching intensity of service]. *Informatsionnye Protsessy* [Information Processes] 12(3):176–190.
8. Abaev, P., A. Pechinkin, and R. Razumchik. 2013. Analysis of queueing system with constant service time for SIP server hop-by-hop overload control. *Modern Probab. Meth. Anal. Telecommunication Networks Comm. Comp. Information Sci.* 356:1–10. doi:10.1007/978-3-642-35980-4_1.
9. Shorgin, S., K. Samouylov, Yu. Gaidamaka, and Sh. Etezov. 2013. Polling system with threshold control for modeling of SIP server under overload. *18th Conference (International) on Systems Science Proceedings*. Wroclaw. 240:97–107. doi:10.1007/978-3-319-01857-7_10.
10. Razumchik, R. V., P. O. Abaev, D. M. Korabelnikov, and D. A. Pyatkina. 2012. Modelirovanie servera s gisterezisnym upravleniem nagruzkoy s pomoshch'yu sistemy massovogo obsluzhivaniya v diskretnom vremeni [Modeling of SIP-server with hysteric overload control as discrete time queueing system]. *Nauchnie Trudy ZNIIS* [Proceedings of ZNIIS]. 119–127.
11. Bocharov, P. P., C. D'Apice, A. V. Pechinkin, and S. Salerno. 2004. *Queueing theory*. Utrecht, Boston: VSP. 446 p.

Received April 2, 2014

Contributors

Pechinkin Alexander V. (b. 1946) — Doctor of Science in physics and mathematics; principal scientist, Institute of Informatics Problems, Russian Academy of Sciences, 44-2 Vavilov Str., Moscow 119333, Russian Federation; professor, Peoples' Friendship University of Russia, 6 Miklukho-Maklaya Str., Moscow 117198, Russian Federation; apechinkin@ipiran.ru

Razumchik Rostislav V. (b. 1984) — Candidate of Science (PhD) in physics and mathematics, senior research fellow, Institute of Informatics Problems, Russian Academy of Sciences, 44-2 Vavilov Str., Moscow 119333, Russian Federation; rrazumchik@ieee.org

INFORMATICS AND APPLICATIONS 2014. Vol. 8. Iss. 2. P. 28–38

ON THE OVERFLOW PROBABILITY ASYMPTOTICS IN A GAUSSIAN QUEUE

O. V. Lukashenko[1,2], E. V. Morozov[1,2], and M. Pagano[3]

Abstract: Gaussian processes are a powerful tool in network modeling since they permit to capture the long memory property of actual traffic flows. In more detail, under realistic assumptions, fractional Brownian motion (FBM) arise as the limit process when a huge number of on-off sources (with heavy-tailed sojourn times) are multiplexed in backbone networks. This paper studies fluid queuing systems with a constant service rate fed by a sum of independent FBMs, corresponding to the aggregation of heterogeneous traffic flows. For such queuing systems, logarithmic asymptotics of the overflow probability, an upper bound for the loss probability in the corresponding finite-buffer queues, are derived, highlighting that the FBM with the largest Hurst parameter dominates in the estimation. Finally, asymptotic results for the workload maximum in the more general case of a Gaussian input with slowly varying at infinity variance are given.

Keywords: Gaussian fluid system; overflow probability; logarithmic asymptotics

DOI: 10.14375/19922264140203

1 Introduction

Gaussian processes are well-recognized models to describe the traffic dynamics of a wide class of modern telecommunication networks. The main motivation to apply these models is their ability of capturing, in a simple and parsimonious way, the properties of self-similarity and long-range dependence, which are inherent in multimedia network traffic [1, 2]. Self-similarity means that the distribution of the process remains unchanged under suitable scaling of time and space, while long-range dependence implies a slow decay of the autocorrelation function. These properties make difficult the probabilistic analysis and, as a consequence, to obtain key performance characteristics, crucial to evaluate the Quality of Service (QoS) provided by the considered networks, in an explicit form.

The FBM is one of the most studied self-similar long-range dependent Gaussian processes. Its use as a traffic model is supported by the following theoretical analysis [3]: the sum of an increasing number of the so-called on-off inputs, with either on-times or off-times having a heavy-tailed distribution with infinite variance, converges weakly to an FBM, after an appropriate time scaling. If an FBM is the input to a queueing system, then let call it fractional Brownian (FB) input.

One of the main characteristics of the queueing systems is the *overflow probability*, i. e., the probability that the workload process exceeds a finite threshold.

In Gaussian queueing systems with infinite buffer, the analysis of the overflow probability (closely related to the workload maximum) is reduced to the analysis of the extremes of Gaussian processes [4].

There are no explicit expressions for the overflow probability in queueing systems with general Gaussian input (including FB input), while a few asymptotic results are available. In this regard, let mention the following key works [5–7]. It is important to stress that in a general setting, the asymptotic analysis of the overflow probability is based on a number of the assumptions which sometimes are difficult to verify. The loss probability in the Gaussian queues with the FB input and a finite buffer is studied in [8–12]. Also, let mention closely related works [13–17], where the maximum of the workload process is studied. Since explicit analysis is unavailable in general case, the numerical analysis of the overflow probability presented in [18–22] plays an important role in the studying of the Gaussian queueing systems. Note that analysis of the systems with the Brownian input is much easier because this process has independent increments. This property allows to obtain the tail probability for the maximum of the Brownian motion [23] in an explicit form. In turn, this result is directly connected with the overflow probability in the queueing system fed by the Brownian input.

In this paper, the authors first present the asymptotic analysis of the overflow probability in the queueing system where the input is a sum of the independent FBMs. Thus, they extend the result which has been proved in a seminal work [5] for only FB input. The present authors follow mainly the approach developed

[1] Institute of Applied Mathematical Research, Karelian Research Center, Russian Academy of Sciences, 11 Pushkinskaya Str., Petrozavodsk 185910, Russian Federation

[2] Petrozavodsk State University, 33 Lenin Str., Petrozavodsk 185910, Russian Federation

[3] University of Pisa, 43 Lungarno Pacinotti, Pisa 56126, Italy

in [5] and discuss in brief inevitable differences in the proofs. For this reason, the proof is straightforward and more transparent than that can be extracted from the related works [6, 7], where generalizations of the basic model from [5] are studied. In particular, the proof in [7] is based on a number of rather complicated assumptions some of which, as was mentioned, are not easy to verify for the specific models.

In summary, the present work reviews the main results on the asymptotics of the overflow probability in Gaussian queueing systems and also discusses some new results on the workload maximum.

The paper is organized as follows. Section 2 contains the description of Gaussian queueing systems, while section 3 presents the proof of the overflow probability asymptotics for the superposition of independent identically distributed (i.i.d.) FB inputs. Finally, in section 4, the results concerning the workload maximum, when the input process belongs to a wide class of the Gaussian processes, are analyzed.

2 Theoretical Background

First of all, the authors motivate their interest to Gaussian queueing systems. To this aim, they consider N i.i.d. *on-off sources*, modelling the traffic flows generated by independent connections. Each source k is described by the process $\{I_k(t),\, t \geq 0\}$, $k = 1, \ldots, N$, where

$$I_k(t) = \begin{cases} 1\,, & t \in \text{ on-period}\,; \\ 0\,, & t \in \text{ off-period}. \end{cases}$$

During an *on-period*, a source is active, while it keeps silence (inactive) during the following *off-period*. The on-off periods are i.i.d. and form an *alternating renewal process*. Furthermore, the processes formed by different sources are assumed to be independent. As a result, the aggregated traffic (cumulative workload) generated by all N sources during time interval $[0, t]$ is given by

$$A_N(t) := \int\limits_0^t \left(\sum_{k=1}^{N} I_k(u) \right) du\,.$$

It is assumed that there are M types of sources , and N_i is the number of the ith type sources, $i = 1, \ldots, M$, so $\sum\limits_{i=1}^{M} N_i = N$. The statistical behavior of the cumulative workload crucially depends on the distribution of on-off periods. Let F^i_{on}, F^i_{off} be the distribution of on- and off-period, respectively. Let assume that the following conditions hold:

$$\left.\begin{aligned} &1 - F^i_{\text{on}}(x) \sim \ell^i_{\text{on}} x^{-\alpha^i_{\text{on}}} L^i_{\text{on}}(x)\,; \\ &1 - F^i_{\text{off}}(x) \sim \ell^i_{\text{off}} x^{-\alpha^i_{\text{off}}} L^i_{\text{off}}(x), \quad x \to \infty\,, \end{aligned}\right\} \qquad (1)$$

where ℓ^i_{on} and ℓ^i_{off} are the positive constants; exponents $\alpha^i_{\text{on}}, \alpha^i_{\text{off}} \in (1,\, 2)$; and functions L^i_{on} and L^i_{off} are slowly varying at infinity, i. e., for any $t > 0$,

$$\lim_{x\to\infty} \frac{L^i(tx)}{L^i(x)} = 1\,, \quad i = 1, \ldots, M\,.$$

(Relation $a \sim b$ means that $a/b \to 1$.) Indeed, conditions (1) mean that the distributions F^i_{on} and F^i_{off} are *heavy-tailed*. For each i, denote by μ^i_{on}, μ^i_{off} the mean length of on- and off-period, respectively (note that μ^i_{on} and $\mu^i_{\text{off}} < \infty$ because α^i_{on} and $\alpha^i_{\text{off}} > 1$). It has been shown in [3] that the scaled cumulative workload arrived during period $[0,\, Tt]$ converges weakly to a sum of independent FBMs provided that:

(i) $N_i \to \infty$ such that $\lim\limits_{N\to\infty} N_i/N > 0$ for each i; and

(ii) the scaling factor $T \to \infty$.

This functional limit theorem leads to the following approximation:

$$A(tT) \approx T \left(\sum_{i=1}^{M} N_i \frac{\mu^i_{\text{on}}}{\mu^i_{\text{on}} + \mu^i_{\text{off}}} \right) t + \sum_{i=1}^{M} T^{H_i} \sqrt{L_i(T) N_i}\, c_i B_{H_i}(t)$$

where c_i are the positive constants; L_i are the slowly varying at infinity functions (expressed in the terms of given parameters); and B_{H_i} are the independent FBMs with the Hurst parameters H_i given by

$$H_i = \frac{3 - \min(\alpha^i_{\text{on}},\, \alpha^i_{\text{off}})}{2} \in \left(\frac{1}{2},\, 1 \right)\,, \quad i = 1, \ldots, M\,.$$

Thus, the aggregated traffic generated by a large number of i.i.d. heavy-tailed on-off sources is approximated by a superposition of the independent FBMs with a linear drift. This result gives a motivation to consider a queueing system fed by a sum of independent FBMs as a suitable model for a wide class of modern telecommunication systems.

Now, let describe a *fluid queue* with a constant service rate C driven by the input process $\{A(t),\, t \geq 0\}$ which is defined as follows:

$$A(t) = mt + X(t)$$

where $m > 0$ is the mean input rate and the process $X := \{X(t)\}$ is the sum of M independent FB inputs such that the ith summand has the Hurst parameter $H_i \in (1/2, 1)$. Obviously, $A(t)$ describes the amount of data (workload) arrived into a communication node

within time interval $[0, t]$. Thus, the variance of the input process in interval $[0, t]$ is

$$v(t) = \sum_{i=1}^{M} t^{H_i}.$$

Introduce the parameter $r := C - m$. Denote $W(t) = X(t) - rt$ and let $Q(t)$ be the current workload at instant t. If $Q(0) = 0$, then the workload $Q(t)$ satisfies the following equation [4]:

$$\begin{aligned} Q(t) &=_d \sup_{0\le s\le t} (A(t) - A(s) - C(t-s)) \\ &= \sup_{0\le s\le t} (X(t) - X(s) - r(t-s)) \\ &= \sup_{0\le s\le t} (W(t) - W(s)) \end{aligned}$$

where symbol $=_d$ stands for the equality in distribution. If, moreover, $r > 0$, then the system is stable and a stationary workload process Q exists such that [24]:

$$\begin{aligned} Q &=_d \sup_{t\in\mathbb{T}} (A(t) - Ct) \\ &= \sup_{t\in\mathbb{T}} W(t) \quad (\mathbb{T} = \mathbb{Z}_+ \text{ or } \mathbb{T} = \mathbb{R}_+). \end{aligned} \tag{2}$$

Hence, for an arbitrary threshold $b \in [0, \infty)$, the *overflow probability* is defined as

$$\mathbb{P}(Q > b) = \mathbb{P}\left(\sup_{t\in\mathbb{T}} W(t) > b\right).$$

It is worth mentioning that some nonasymptotic upper bounds for the overflow probability have been proposed. For instance, in case of ordinary FB input (H is the Hurst parameter) and $\mathbb{T} = \mathbb{Z}_+$, it was shown that [25, 26]

$$\mathbb{P}(Q > b) \le \frac{\Gamma(1/(2\beta))}{2\beta(-\log\eta)^{1/(2\beta)}}$$

where Γ denotes the Gamma-function, $\beta \in (0, 1-H)$ is the free parameter and

$$\eta = \exp\left(-\frac{1}{2}\left(\frac{C-m}{H+\beta}\right)^{2(H+\beta)} \times \left(\frac{b}{1-(H+\beta)}\right)^{2-2(H+\beta)}\right).$$

This result can be extended to general Gaussian inputs, but the value of η can be estimated only by numerical methods (see [27]).

3 Asymptotics of the Overflow Probability for a Superposition of Fractional Brownian Inputs

The following result shows that the FB input with the largest Hurst parameter dominates in the asymptotic analysis of the overflow probability. Recall that the presented proof is mainly based on the technique developed in [5] where a system with single FB input process has been analyzed.

Theorem 3.1. *For the stationary workload* (2)*, the following asymptotic holds*:

$$\lim_{b\to\infty} b^{2H-2}\log\mathbb{P}(Q > b) = -\frac{r^{2H}}{2H^{2H}(1-H)^{2(1-H)}} := -\Theta$$

where $H = \max(H_1, \ldots, H_M)$.

Proof. Consider the following relations:

$$\begin{aligned} \mathbb{P}(W(t)/t > x) &= \mathbb{P}\left(\mathcal{N}(0,1) > \frac{(x+r)t}{\sqrt{v(t)}}\right) \\ &= \Psi\left(\frac{(x+r)t}{\sqrt{v(t)}}\right) \end{aligned} \tag{3}$$

where

$$\Psi(x) := \frac{1}{\sqrt{2\pi}}\int_x^\infty e^{-y^2/2}dy$$

is the tail distribution of the standard normal variable $\mathcal{N}(0,1)$. Function Ψ satisfies the following inequalities [24] for $x > 0$:

$$\frac{1-x^{-2}}{x\sqrt{2\pi}}e^{-x^2/2} \le \Psi(x) \le \frac{1}{x\sqrt{2\pi}}e^{-x^2/2},$$

which, in turn, imply the approximation

$$\log\Psi(x) \sim -\frac{x^2}{2}, \quad x\to\infty. \tag{4}$$

Denote

$$\nu(t) = \frac{t^2}{v(t)}$$

and note that $\nu(t) \sim t^{2-2H} \to \infty$ as $t \to \infty$. It now follows from (4) and (3) that the following limit exists:

$$\lim_{t\to\infty}\frac{1}{\nu(t)}\log\mathbb{P}(W(t)/t > x) = -\frac{1}{2}(x+r)^2 := -\lambda(x). \tag{5}$$

It is easy to check that

$$\inf_{c>0} c^{2H-2}\lambda(c) = \Theta.$$

Let emphasize that, in contrast to this straightforward analysis, the proof of (5) in [5, 7] (obtained for general non-Gaussian case) is based on a large deviation

principle and some technical conditions placed on the logarithmic moment generating function.

To prove the statement of theorem 3.1, it suffices to establish the following lower and upper bounds:

$$\liminf_{b\to\infty} \frac{\log \mathbb{P}(Q>b)}{\nu(b)} \geq -\inf_{c>0} c^{2H-2}\lambda(c)\,; \quad (6)$$

$$\limsup_{b\to\infty} \frac{\log \mathbb{P}(Q>b)}{\nu(b)} \leq -\inf_{c>0} c^{2H-2}\lambda(c)\,. \quad (7)$$

First, let note that for each $c>0$,

$$\liminf_{b\to\infty} \frac{\log \mathbb{P}(Q>b)}{\nu(b)} \geq \liminf_{b\to\infty} \frac{\log \mathbb{P}\left(W(b/c)>b\right)}{\nu(b)} = \lim_{t\to\infty} \frac{\log \mathbb{P}(W(t)/t>c)}{\nu(t\,c)} = -c^{2H-2}\lambda(c)$$

and, thus, the lower bound follows.

The proof of the upper bound is more challenging. First, let consider the discrete time case $\mathbb{T}=\mathbb{Z}_+$ and then verify some technical conditions to extend this result to continuous time case $\mathbb{T}=\mathbb{R}_+$.

Consider an arbitrary $d>0$; then, one obtains

$$\mathbb{P}(Q>b) \leq \mathbb{P}\left(\sup_{n<b/d} W(n)>b\right) + \mathbb{P}\left(\sup_{n\geq b/d} W(n)>b\right) \leq \frac{b}{d}\sup_{c>d}\mathbb{P}(W(b/c)>b) + \sum_{n\geq b/d}\mathbb{P}(W(n)>b). \quad (8)$$

Denote

$$\left.\begin{aligned} f_1(b) &= \frac{b}{d}\sup_{c>d}\mathbb{P}(W(b/c)>b)\,; \\ f_2(b) &= \sum_{n\geq b/d}\mathbb{P}(W(n)>b)\,. \end{aligned}\right\} \quad (9)$$

It is easy to check that if $(b+rk)/\sqrt{v(k)}>1$, then

$$\mathbb{P}(W(k)>b) = \Psi\left(\frac{b+rk}{\sqrt{v(k)}}\right) \leq \exp\left(-\frac{(b+rk)^2}{2v(k)}\right) \leq \exp\left(-\frac{r^2}{2}\nu(k)\right). \quad (10)$$

Also, note that for $t\geq 1$,

$$\nu(t) = \frac{t^2}{\sum\limits_{i=1}^{M} t^{H_i}} \geq \frac{1}{M}t^{2-2H}\,. \quad (11)$$

Denote

$$a = 2-2H>0\,,\ \ \gamma = \frac{r^2}{2M}\,.$$

It then follows from (10) and (11) that

$$\limsup_{b\to\infty}\frac{\log f_2(b)}{\nu(b)} \leq M\limsup_{b\to\infty}\frac{1}{b^a}\log\left[\sum_{k=\lfloor b/d\rfloor}^{\infty} e^{-\gamma k^a}\right]. \quad (12)$$

Note that if $k\geq\lfloor b/d\rfloor$ and $k-1\leq x\leq k$, then the inequality $e^{-\gamma k^a}\leq e^{-\gamma x^a}$ holds. Hence,

$$\sum_{k=\lfloor b/d\rfloor}^{\infty} e^{-\gamma k^a} \leq \int\limits_{\lfloor b/d\rfloor-1}^{\infty} e^{-\gamma x^a}dx \leq \int\limits_{b/d-2}^{\infty} e^{-\gamma x^a}dx\,. \quad (13)$$

It follows from (12) and (13) that

$$\limsup_{b\to\infty}\frac{\log f_2(b)}{\nu(b)} \leq M\limsup_{b\to\infty}\frac{1}{b^a}\log\left[\int\limits_{b/d-2}^{\infty} e^{-\gamma x^a}dx\right] \quad (14)$$

and by applying the L'Hôpital's rule twice, one obtains the following limit:

$$\lim_{b\to\infty}\frac{1}{b^a}\log\int\limits_{b/d-2}^{\infty} e^{-\gamma x^a}dx = -\frac{1}{da}\lim_{b\to\infty}\left[e^{-\gamma(b/d-2)^a}b^{1-a}\frac{1}{\int\limits_{b/d-2}^{\infty} e^{-\gamma x^a}dx}\right] = -\frac{\gamma}{d}\lim_{b\to\infty}\left(\frac{b/d-2}{b}\right)^{a-1} = -\gamma d^{-a}\,.$$

Now, let choose $d\in(0,\,((1-H)r)/H)$ such that

$$-\gamma d^{-a} \leq -\inf_{c>0} c^{-a}\lambda(c)\,.$$

It then follows from (14) that

$$\limsup_{b\to\infty}\frac{\log\left[\sum\limits_{n\geq b/d}\mathbb{P}(W(n)>b)\right]}{\nu(b)} \leq -\inf_{c>0}\frac{\lambda(c)}{c^a}. \quad (15)$$

Consider the term f_1 from (9) and note that

$$\limsup_{b\to\infty}\frac{1}{\nu(b)}\log f_1(b) = \limsup_{b\to\infty}\frac{1}{\nu(b)}\log\left[\frac{b}{d}\sup_{c>d}\mathbb{P}\left(W(b/c)>b\right)\right]$$

$$= \limsup_{b\to\infty} \frac{1}{\nu(b)} \log\left[\sup_{c>d} \mathbb{P}\left(W(b/c) > b\right)\right]$$
$$= \limsup_{b\to\infty} \sup_{c>d} \frac{1}{\nu(b)} \log \mathbb{P}\left(W(b/c) > b\right)$$
$$= \limsup_{n\to\infty} \sup_{c>d} \frac{1}{\nu(nc)} \log \mathbb{P}(W(n)/n > c)\,.$$

By (5), for any given $\delta > 0$, let choose sufficiently large n such that

$$\log \mathbb{P}(W(n)/n > x) \le \nu(n)(\delta - \lambda(x))\,.$$

Using the last inequality,

$$\limsup_{b\to\infty} \frac{\log f_1(b)}{\nu(b)} \le \limsup_{n\to\infty} \sup_{c>d} \frac{\nu(n)}{\nu(cn)} [\delta - \lambda(c)]$$
$$= \limsup_{n\to\infty} \sup_{c>d} \left[\frac{\nu(n)\delta}{h(cn)} - \frac{\lambda(c)\nu(n)}{\nu(cn)}\right]$$
$$\le \limsup_{n\to\infty} \left[\sup_{c>d} \frac{\nu(n)\delta}{h(cn)} - \inf_{c>d} \frac{\lambda(c)\nu(n)}{\nu(cn)}\right]$$
$$= -\limsup_{n\to\infty} \inf_{c>d} \frac{\lambda(c)\nu(n)}{\nu(cn)} + \limsup_{n\to\infty} \frac{\nu(n)\delta}{\nu(dn)}$$
$$= -\limsup_{n\to\infty} \inf_{c>d} \frac{\lambda(c)\nu(n)}{\nu(cn)} + \frac{\delta}{d^a}\,. \quad (16)$$

Consider the following function:

$$f(x) := \frac{1}{2}\left(x + r\right)^2 x^{2H-2}\,,\ \ x \ge 0\,.$$

It is easy to check that $\min f(x)$ is attained at the point $x = (1-H)r/H > 0$. Thus, for any $d \in (0,\, ((1-H)r)/H)$,

$$\inf_{c>0} f(c) = \inf_{c>d} f(c)\,.$$

Moreover,

$$\inf_{c>d} \frac{\lambda(c)\nu(t)}{\nu(ct)} = \inf_{c>d} \frac{(c+r)^2 \sum_{i=1}^{M} c^{2H_i-2} t^{2H_i}}{2\sum_{i=1}^{M} t^{2H_i}}$$
$$\ge \frac{\sum_{i=1}^{M} \inf_{c>d} t^{2H_i}(c+r)^2 c^{2H_i-2}}{2\sum_{i=1}^{M} t^{2H_i}}$$
$$= \frac{\sum_{i=1}^{M} t^{2H_i} f\left(((1-H_i)r)/H_i\right)}{\sum_{i=1}^{M} t^{2H_i}}\,.$$

It now easily follows as $t \to \infty$ that

$$\frac{\sum_{i=1}^{M} t^{2H_i} f\left(((1-H_i)r)/H_i\right)}{\sum_{i=1}^{M} t^{2H_i}} \to \frac{1}{2} f\left(\frac{(1-H)r}{H}\right)$$
$$= \inf_{c>d} c^{2H-2}\lambda(c)\,. \quad (17)$$

Thus, let obtain from (16) that

$$\limsup_{b\to\infty} \frac{\log f_1(b)}{\nu(b)} \le -\inf_{c>0} c^{2H-2}\lambda(c) + \frac{\delta}{d^a}\,,$$

and because δ is arbitrary,

$$\limsup_{b\to\infty} \frac{\log f_1(b)}{\nu(b)} \le -\inf_{c>0} c^{2H-2}\lambda(c)\,. \quad (18)$$

Now, let take into account the following inequality

$$\limsup_{b\to\infty} \frac{\log\left(f_1(b) + f_2(b)\right)}{h(b)}$$
$$\le \limsup_{b\to\infty} \frac{\log\left(\max(f_1(b), f_2(b))\right) + \log 2}{h(b)}\,. \quad (19)$$

Then, let combine (19) with (8), (15), and (18) to obtain (7). In turn, inequalities (6) and (7) imply

$$\lim_{b\to\infty} \frac{\log \mathbb{P}(Q > b)}{\nu(b)} = -\inf_{c>0} c^{2H-2}\lambda(c)\,,$$

and the proof for $\mathbb{T} = \mathbb{Z}_+$ is completed.

To consider the case $\mathbb{T} = \mathbb{R}_+$, let define the process $\{W^*(n),\, n \in \mathbb{Z}_+\}$ as

$$W^*(n) = \sup_{0\le s\le 1} W(n+s)\,.$$

Note that

$$\mathbb{P}\left(\sup_{t\in\mathbb{R}_+} W(t) > b\right) = \mathbb{P}\left(\sup_{n\in\mathbb{Z}_+} W^*(n) > b\right)\,.$$

Thus, the asymptotics for the process $\{W^*(n),\, n \in \mathbb{Z}_+\}$ implies the asymptotics for the continuous-time process $\{W(t),\, t \in \mathbb{R}_+\}$. The lower bound follows from the obvious inequality:

$$\mathbb{P}\left(\sup_{n\in\mathbb{Z}_+} W^*(n) > b\right) \ge \mathbb{P}\left(\sup_{n\in\mathbb{Z}_+} W(n) > b\right)\,.$$

Denote

$$Y(s) = W(n+s) - W(n)\,,\ \ s \ge 0\,,\ \ n \in \mathbb{Z}_+\,,$$

and note that $\mathbb{E}Y(s) = -rs$. Applying Borell–Sudakov–Tsirelson inequality [28, 29], one obtains

$$\mathbb{P}\left(\sup_{0\le s\le 1} Y(s) \ge u\right) \le \mathbb{P}\left(\sup_{0\le s\le 1} (Y(s)+rs) \ge u\right) \le 2\mathbb{P}\left(\mathcal{N}(a,\sigma^2) > u\right) \quad (20)$$

where

$$a := \operatorname{med}\left(\sup_{0\le s\le 1} (Y(s)+rs)\right);$$
$$\sigma^2 := \sup_{0\le s\le 1} \mathbb{D}Y(s).$$

Denote $\varphi(n) = \theta\nu(n)/n$. It now follows from (20) that

$$\mathbb{E}\exp\left[\varphi(n)\left(W^*(n)-W(n)\right)\right] = \mathbb{E}\exp\left[\varphi(n)\sup_{0\le s\le 1} Y(s)\right] \le 2\mathbb{E}\exp\left[\varphi(n)\mathcal{N}(a,\sigma^2)\right] = \exp\left[\frac{\sigma^2\varphi^2(n)+2\varphi(n)a}{2}\right].$$

Thus, one obtains

$$\limsup_{n\to\infty}\frac{1}{\nu(n)}\log\mathbb{E}e^{\theta\nu(n)(W^*(n)-W(n))/n} \le \limsup_{n\to\infty}\frac{\sigma^2\varphi^2(n)+2\varphi(n)a}{2\nu(n)} = \limsup_{n\to\infty}\frac{\sigma^2\theta^2\nu(n)+2\theta an}{2n^2} = 0. \quad (21)$$

Note that the statement

$$\limsup_{n\to\infty}\frac{1}{\nu(n)}\log\mathbb{E}e^{\theta\nu(n)(W^*(n)-W(n))/n} = 0 \quad (22)$$

is used as an assumption in the paper [5], while above, a detailed proof of (22) is given for the system with Gaussian input. By the Hölder's inequality, one obtains for $1 < p, g < \infty$, $1/p + 1/q = 1$ that

$$\begin{aligned}\lambda_1(\theta) &:= \limsup_{n\to\infty}\frac{1}{\nu(n)}\log\mathbb{E}e^{\theta\nu(n)W^*(n)/n} = \limsup_{n\to\infty}\frac{1}{\nu(n)} \times \log\mathbb{E}e^{\theta\nu(n)(W^*(n)-W(n))/n}e^{\theta\nu(n)W(n)/n} \\ &\le \limsup_{n\to\infty}\frac{1}{\nu(n)}\log\left\{\left[\mathbb{E}e^{\theta q\nu(n)(W^*(n)-W(n))/n}\right]^{1/q} \times \left[\mathbb{E}e^{\theta p\nu(n)W(n)/n}\right]^{1/p}\right\} \\ &= \frac{1}{q}\limsup_{n\to\infty}\frac{1}{\nu(n)}\log\mathbb{E}e^{(\theta q\nu(n)(W^*(n)-W(n)))/n} + \frac{1}{p}\limsup_{n\to\infty}\frac{1}{\nu(n)}\log\mathbb{E}e^{(\theta p\nu(n)W(n))/n} \quad (23)\end{aligned}$$

$$\le \limsup_{n\to\infty}\frac{1}{\nu(n)}\log\mathbb{E}e^{\theta p\nu(n)W(n)/n} = \frac{1}{2}(\theta p)^2 - \theta pr$$

where the first term in (23) equals zero by (21). Taking $p \to 1$, one gets

$$\lambda_1(\theta) \le \frac{1}{2}\theta^2 - \theta r.$$

On the other hand,

$$\liminf_{n\to\infty}\frac{1}{\nu(n)}\log\mathbb{E}e^{\theta\nu(n)W^*(n)/n} \ge \liminf_{n\to\infty}\frac{1}{\nu(n)}\log\mathbb{E}e^{\theta\nu(n)W(n)/n} = \frac{1}{2}\theta^2 - \theta r.$$

Thus, the following limit exists:

$$\lim_{n\to\infty}[\nu(n)]^{-1}\log\mathbb{E}e^{\theta\nu(n)W^*(n)/n} = \frac{1}{2}\theta^2 - \theta r = \lambda_1(\theta).$$

It now follows by the Gärtner–Ellis theorem [30] that the sequence $\{W^*(n)/n,\ \nu(n)\}$ satisfies a large deviation principle with the following rate function λ which is the Fenchel–Legendre transform of function λ_1:

$$\sup_{\theta\in\mathbb{R}}\{\theta x - \lambda_1(\theta)\} = \sup_{\theta\in\mathbb{R}}\left\{-\frac{\theta^2}{2} + (x+r)\theta\right\} = \frac{1}{2}(x+r)^2 := \lambda(x).$$

(For more details, see [30].) As a consequence, one obtains

$$\lim_{n\to\infty}\frac{\mathbb{P}(W^*(n)/n > x)}{\nu(n)} = -\lambda(x).$$

Thus, equation (5) for the process $\{W^*(n)\}$ is proved.

To establish the upper bound, let repeat steps (8)–(19) with $W(n)$ replaced by $W^*(n)$. The proof remains unchanged with exception of the point, where another arguments are used to come to the upper bound (10). More exactly, at this point, Chernoff's inequality is applied which gives, for any $\varepsilon > 0$,

$$\mathbb{P}\left(W^*(k) > b\right) \le e^{-\sup_\theta\left(\theta b - \log\mathbb{E}e^{\theta W^*(k)}\right)} \le e^{-\theta\nu(k)b/k}e^{(\lambda_1(\theta)+\varepsilon)\nu(k)} \le e^{(\lambda_1(\theta)+\varepsilon)\nu(k)}$$

where θ is chosen in such a way that $\lambda_1(\theta) + \varepsilon < 0$. □

4 Closely Related Results

In this section, a number of results for Gaussian queueing systems are discussed in brief which are closely connected with the asymptotics of the overflow probability analyzed above. More exactly, the asymptotics for the workload maximum are obtained in the more general case when the variance v of the input process X is regu-

larly varying at infinity function with index $0 < V < 2$, i. e., for any $y > 0$,

$$\lim_{t\to\infty} \frac{v(yt)}{v(t)} = y^V.$$

The asymptotic has the following form [6, 7]

$$\lim_{b\to\infty} \frac{v(b)}{b^2} \log \mathbb{P}(Q > b) = -\Theta$$

where

$$\Theta = \frac{2}{(2-V)^{2-V}} \left(\frac{r}{V}\right)^V. \tag{24}$$

It is well-known that every regularly varying at infinity function can be represented as

$$v(t) = t^V L(t) \tag{25}$$

where function $L(t)$ is slowly varying (as $t \to \infty$) and index $V \in (0, 2)$ [31]. Denote $\beta = (2-V)^{-1}$ and take (arbitrary) $\varepsilon \in (0, 2-V)$. Moreover, it is assumed that the following conditions hold as $t \to \infty$:

$$L(tL^\beta(t)) \sim L(t); \tag{26}$$

function $L(t)$ is twice differentiable on $\mathbb{R}_+$ and

$$L''(t) = o\left(\frac{1}{t^{V+\varepsilon}}\right). \tag{27}$$

It follows from (27) that

$$v''(t) \log t \to 0, \ \ t \to \infty. \tag{28}$$

Also, recall that a stationary version $Q^*(t)$ of the workload process $Q(t)$ exists [32]. Let

$$\gamma(t) = L[(\log t)^\beta] \log t;$$
$$M(t) := \max_{0\le s\le t} Q(s); \ \ M^*(t) := \max_{0\le s\le t} Q^*(s).$$

The following asymptotic result for the workload maximum has been established in [15] (see also [16, 17]).

Theorem 4.1. *If the variance $v(t)$ of the Gaussian component X satisfies conditions* (25)–(27) *and $r > 0$, then*

$$\frac{M^*(t)}{\gamma^\beta(t)} \Rightarrow \left(\frac{1}{\Theta}\right)^\beta; \ \ \frac{M(t)}{\gamma^\beta(t)} \Rightarrow \left(\frac{1}{\Theta}\right)^\beta, \ \ t \to \infty, \tag{29}$$

where the constant Θ is given by expression (24)*, and $\Rightarrow$ stands for convergence in probability.*

Proof. Let give a sketch of the proof which mainly follows the technique developed in [13]. It is sufficient to prove that for any $\delta > 0$,

$$\mathbb{P}\left(\frac{M^*(t)}{\gamma^\beta(t)} \ge \left(\frac{1-\delta}{\theta}\right)^\beta\right) \to 1, \ \ t \to \infty; \tag{30}$$

$$\mathbb{P}\left(\frac{M^*(t)}{\gamma^\beta(t)} \ge \left(\frac{1+\delta}{\theta}\right)^\beta\right) \to 0, \ \ t \to \infty. \tag{31}$$

To prove (30), let fix $\Delta \in (0, t)$ and note that

$$Q^*(t) \ge W(t) - W(t-\Delta).$$

Denote

$$Y_k^{(\Delta)} = W(k\Delta) - W((k-1)\Delta),$$

then for each t, one has

$$M^*(t) \ge \max_{1\le k\le \lfloor t/\Delta\rfloor} Y_k^{(\Delta)}.$$

Thus, the original problem reduces to the analysis of the extremes of a stationary normal sequence $\{Z_k\}$ (with $Z_k =_d \mathcal{N}(0,1)$) since

$$\mathbb{P}\left(\frac{M^*(t)}{\gamma^\beta(t)} \ge \left(\frac{1-\delta}{\theta}\right)^\beta\right) \ge \mathbb{P}\left(\max_{i=1,\dots,\lfloor t/\Delta\rfloor} Z_i \ge u(t)\right) := \mathbb{P}_1(t),$$

where

$$u(t) := \frac{\alpha(t) + r\Delta(t)}{\sqrt{v(\Delta(t))}}, \ \ \alpha(t) := \left(\frac{1-\delta}{\theta}\,\gamma(t)\right)^\beta.$$

If $\Delta := \Delta(t) = A\gamma^\beta(t)$, where $A > 0$ is the constant, is chosen, then it is possible to show that

$$\mathbb{P}_1(t) \to 1, \ \ t \to \infty,$$

as required. To establish (31), let consider another stationary sequence:

$$Y_i := \sup_{s\in[i-1,i)} Q^*(s), \ \ i = 1, \dots, \lceil t\rceil,$$

and, thus,

$$M^*(t) \le \max\{Y_i: \ \ i = 1, \dots, \lceil t\rceil\}.$$

This immediately implies

$$\mathbb{P}\left(M^*(t) \ge \left(\frac{1+\delta}{\theta}\right)^\beta \gamma^\beta(t)\right) \le \mathbb{P}\left(\max_{i=1,\dots,\lceil t\rceil} Y_i \ge \left(\frac{1+\delta}{\theta}\right)^\beta \gamma^\beta(t)\right) \le \lceil t\rceil \mathbb{P}\left(Y \ge \left(\frac{1+\delta}{\theta}\right)^\beta \gamma^\beta(t)\right) := \mathbb{P}_2(t).$$

A careful analysis allows to conclude that

$$\mathbb{P}_2(t) \to 0, \ \ t \to \infty,$$

that completes the proof. (More detailed analysis can be found in [15].) □

Under slightly less general assumptions, the previous result can be generalized to the convergence in the space L_p for any $p \ge 1$.

Theorem 4.2. *Let conditions of theorem 4.1 hold. If, moreover,*

$$\liminf_{t\to\infty} L(t) > 0\,;\ \limsup_{t\to\infty} L(t) < \infty\,, \qquad (32)$$

then convergence in (29) *holds in the space* L_p *for any* $p \in [1, \infty)$.

To prove this statement, it is sufficient to establish the uniform integrability of the following sequence: $\left\{\left(M^*(t)/\gamma^\beta(t)\right)^p,\ t \geq T\right\}$ where T is the finite positive constant. Indeed, it is shown in [17] that

$$\sup_{t\geq T} \mathbb{E}\left[\frac{M^*(t)}{\gamma^\beta(t)}\right]^{p+1} < \infty\,.$$

Remark. If the limit

$$\lim_{t\to\infty} L(t) = A \in (0,\ \infty)$$

exists, then conditions (32) are automatically fulfilled.

Theorems 4.1 and 4.2 can be applied to some specific processes. First, let consider the following input:

$$X(t) = \sum_{i=1}^{n} B_{H_i}(t)\,,\ \ t \geq 0\,,$$

where B_{H_i} are the independent FBMs with the Hurst parameters $H_i \in (0,\ 1)$ and $H_1 > \max\limits_{i>1} H_i$. Then, the variance $v(t)$ of the process $X(t)$ satisfies condition (25), and the statements of theorems 4.1 and 4.2 hold with $V = 2H_1$. This is an extension of the results derived in [13].

The second example is the integrated Gaussian process X, that is,

$$X(t) = \int_0^t Z(s)\,ds$$

where Z is the centered stationary Gaussian process with the covariance function $R(u) := \mathbb{C}\mathrm{ov}(Z(0), Z(u))$. Such models have been considered in [33, 34]. It is easy to check that the variance $v(t)$ of $X(t)$ can be written as

$$v(t) = 2\int_0^t \int_0^s R(u)\,duds\,. \qquad (33)$$

Then $v''(t) = 2R(t)$ and condition (28) is equivalent to

$$R(t)\log t \to 0\,,\ \ t \to \infty\,.$$

If, in addition, $A \in (0,\ \infty)$ and exists $V \in (0, 2)$ such that

$$\frac{\int_0^t \int_0^s R(u)\,duds}{t^V} \to A\,,\ \ t \to \infty\,, \qquad (34)$$

then conditions of theorem 4.2 are satisfied as well. For example, let Z be the Ornstein–Uhlenbeck process with $R(t) = \lambda e^{-\alpha t}$ and parameters $\lambda, \alpha > 0$. It then follows from (33) that

$$v(t) = \frac{2\lambda}{\alpha}\,t + \frac{2\lambda}{\alpha^2}\left(e^{-\alpha t} - 1\right)$$

and, hence, condition (34) is satisfied with $V = 1$ and $A = \lambda/\alpha$.

Note that the integrated Ornstein–Uhlenbeck process is the Gaussian counterpart of the well-known Anick–Mitra–Sondi fluid model [35] (see also [36]), and its relevance for the modeling of the traffic in communications systems is motivated in [33].

Theorem 4.1 yields the asymptotics for another important characteristic, called hitting time, the time required to reach a threshold b,

$$T(b) = \inf\left\{t \geq 0 :\ Q^*(t) \geq b\right\}\,.$$

The analysis of $T(b)$ is based on the following relation between $M^*(t)$ and $T(b)$:

$$\left\{T(b) \leq t\right\} = \left\{M^*(t) \geq b\right\}\,.$$

Finally, the following result proved in [17] has been got as well.

Theorem 4.3. *If conditions of theorem* 4.1 *hold and function* $\gamma(t)$ *monotonically increases in* $[t_0, \infty)$*, for some* $t_0 < \infty$*, then*

$$\frac{\gamma(T(b))}{b^{1/\beta}} \Rightarrow \Theta\,,\ \ b \to \infty\,.$$

Acknowledgments

This work is done under financial support of the Program of Strategy Development of Petrozavodsk State University in the framework of the research activity.

References

1. Leland, W. E., M. S. Taqqu, W. Willinger, and D. V. Wilson. 1994. On the self-similar nature of ethernet traffic (extended version). *IEEE/ACM Trans. Networking* 2(1):1–15.
2. Willinger, W., M. S. Taqqu, W. E. Leland, and D. Wilson. 1995. Self-similarity in high-speed packet traffic: Analysis and modeling of Ethernet traffic measurements. *Stat. Sci.* 10(1):67–85.
3. Taqqu, M. S., W. Willinger, and R. Sherman. 1997. Proof of a fundamental result in self-similar traffic modeling. *Comp. Comm. Rev.* 27:5–23.
4. Reich, E. 1958. On the integrodifferential equation of Takacs I. *Ann. Math. Stat.* 29:563–570.

5. Duffield, N., and N. O'Connell. 1995. Large deviations and overflow probabilities for the general single server queue, with applications. *Proc. Cambridge Phil. Soc.* 118:363–374.
6. Debicki, K. 1999. A note on LDP for supremum of Gaussian processes over infinite horizon. *Stat. Probab. Lett.* 44:211–220.
7. Duffy, K., J. T. Lewis, and W. G. Sullivan. 2003. Logarithmic asymptotics for the supremum of a stochastic process. *Ann. Appl. Probab.* 13(2):430–445.
8. Kim, H. S., and N. B. Shroff. 2001. Loss probability calculations and asymptotic analysis for finite buffer multiplexers. *IEEE/ACM Trans. Networking* 9:755–768.
9. Kim, H. S., and N. B. Shroff. 2001. On the asymptotic relationship between the overflow probability and the loss ratio. *Adv. Appl. Probab.* 33:836–863.
10. Goricheva, R. S., O. V. Lukashenko, E. V. Morozov, and M. Pagano. 2010. Regenerative analysis of a finite buffer fluid queue. *2010 Congress (International) on Ultra Modern Telecommunications and Control Systems and Workshops (ICUMT) Proceedings*. 1132–1136.
11. Lukashenko, O. V., E. V. Morozov, and M. Pagano. 2011. Estimation of loss probability in Gaussian queues. *Conference (International) "Modern Probabilistic Methods for Analysis and Optimization of Information and Telecommunication Networks" Proceedings*. 142–147.
12. Lukashenko, O. V., E. V. Morozov, R. S. Nekrasova, and M. Pagano. 2013. Performance evaluation of finite buffer queues through regenerative simulation. *Comm. Comp. Inform. Sci. BWWQT 2013.* 356:131–139.
13. Zeevi, A., and P. Glynn. 2000. On the maximum workload in a queue fed by fractional Brownian motion. *Ann. Appl. Probab.* 10:1084–1099.
14. Hüsler, J., and V. I. Piterbarg. 2004. Limit theorem for maximum of the storage process with fractional Brownian as input. *Stochastic Proc. Their Appl.* 114:231–250.
15. Lukashenko, O. V., and E. V. Morozov. 2012. Asymptotics of the maximum workload for a class of Gaussian queues. *Informatics and Its Applications — Inform. Appl.* 6(3):81–89.
16. Lukashenko O. V., and E. V. Morozov. 2012. On the maximum workload for a class of Gaussian queues. *Conference (International) "Probability Theory and Its Applications" in Commemoration of the Centennial of B. V. Gnedenko*. 231–232.
17. Lukashenko, O. V., and E. V. Morozov. 2013. On convergence in the L_p space of the workload maximum for a class of Gaussian queueing systems. *Informatics and Its Applications — Inform. Appl.* 7(1):36–43.
18. Dieker, A. B., and M. Mandjes. 2005. On asymptotically efficient simulation of large deviation probabilities. *Adv. Appl. Probab.* 37:539–552.
19. Giordano, S., M. Gubinelli, and M. Pagano. 2005. Bridge Monte-Carlo: A novel approach to rare events of Gaussian processes. *5th St. Petersburg Workshop on Simulation Proceedings*. St. Petersburg, Russia. 281–286.
20. Dieker, A. B., and M. Mandjes. 2006. Fast simulation of overflow probabilities in a queue with Gaussian input. *ACM Trans. Model. Comput. Simul.* 16(2):119–151.
21. Giordano, S., M. Gubinelli, and M. Pagano. 2007. Rare events of Gaussian processes: A performance comparison between Bridge Monte-Carlo and Importance Sampling. *Next generation teletraffic and wired/wireless advanced networking*. St. Petersburg, Russia. 269–280.
22. Lukashenko, O. V., E. V. Morozov, and M. Pagano. 2012. Performance analysis of bridge Monte-Carlo estimator. *Trans. KarRC RAS* 3:54–60.
23. Takacs, L. 1967. *Combinatorial methods in the theory of stochastic processes*. John Wiley&Sons. 262 p.
24. Mandjes, M. 2007. *Large deviations of Gaussian queues*. Chichester: Wiley. 339 p.
25. Rizk, A., and M. Fidler. 2010. Sample path bounds for long memory fbm traffic. *29th Conference on Information Communications, INFOCOM'10 Proceedings*. Piscataway, NJ, USA: IEEE Press. 61–65.
26. Rizk, A., and M. Fidler. 2012. Non-asymptotic end-to-end performance bounds for networks with long range dependent FBM cross traffic. *Comp. Networks*. 56(1):127–141.
27. Lukashenko, O., E. Morozov, and M. Pagano. 2014. On the effective envelopes for fluid queues with Gaussian input. *Comm. Comp. Inform. Sci. DCCN 2013.* 279:178–189.
28. Adler, R. J. 1990. *An introduction to continuity, extrema, and related topics for general Gaussian processes*. Hayward, CA: Institute of Mathematical Statistics. 160 p.
29. Debicki, K. 2004. Gaussian processes. *Encyclopedia of actuarial sciences* 2:752–757.
30. Dembo, A., and O. Zeitouni. 1998. *Large deviations techniques and applications*. Springer. 396 p.
31. Seneta, E. 1985. *Regularly varying functions*. Springer. 116 p.
32. Konstantopoulos, T., M. Zazanis, and G. De Veciana. 1996. Conservation laws and reflection mappings with application to multiclass mean value analysis for stochastic fluid queues. *Stochastic Proc. Their Appl.* 65:139–146.
33. Kulkarni, V., and T. Rolski. 1994. Fluid model driven by an Ornstein–Uhlenbeck process. *Prob. Eng. Inform. Sci.* 8:403–417.
34. Debicki, K., and T. Rolski. 1995. A Gaussian fluid model. *Queueing Syst.* 20:433–452.
35. Anick, D., D. Mitra, and M. M. Sondhi. 1982. Stochastic theory of a data handling system with multiple resources. *Bell Syst. Techn. J.* 61:1871–1894.
36. Addie, R., P. Mannersalo, and I. Norros. 2002. Most probable paths and performance formulae for buffers with Gaussian input traffic. *Eur. Trans. Telecomm.* 13:183–196.

Received March 8, 2014

Contributors

Lukashenko Oleg V. (b. 1986) — Candidate of Science (PhD) in physics and mathematics, junior scientist, Institute of Applied Mathematical Research of Karelian Research Center, Russian Academy of Sciences; lecturer, Petrozavodsk State University; lukashenko-oleg@mail.ru

Morozov Evsei V. (b. 1947) — Doctor of Science in physics and mathematics, professor, leading scientist, Institute of Applied Mathematical Research of Karelian Research Center, Russian Academy of Sciences, 11 Pushkinskaya Str., Petrozavodsk 185910, Republic of Karelia, Russian Federation; professor, Petrozavodsk State University, 33 Lenin Str., Petrozavodsk 185910, Republic of Karelia, Russian Federation; emorozov@karelia.ru

Pagano Michele (b. 1968) — PhD in electronics engineering, associate professor, University of Pisa, 43 Lungarno Pacinotti, Pisa 56126, Italy; m.pagano@iet.unipi.it

ОБ АСИМПТОТИКЕ ВЕРОЯТНОСТИ ПЕРЕПОЛНЕНИЯ ГАУССОВСКОЙ ОЧЕРЕДИ*

О. В. Лукашенко[1], Е. В. Морозов[2], М. Пагано[3]

[1]Институт прикладных математических исследований КарНЦ РАН, Россия, Республика Карелия, г. Петрозаводск 185910, ул. Пушкинская 11; Петрозаводский государственный университет, Россия, Республика Карелия, г. Петрозаводск 185910, пр. Ленина 33; lukashenko-oleg@mail.ru
[2]Институт прикладных математических исследований КарНЦ РАН, Россия, Республика Карелия, г. Петрозаводск 185910, ул. Пушкинская 11; Петрозаводский государственный университет, Россия, Республика Карелия, г. Петрозаводск 185910, пр. Ленина 33; emorozov@karelia.ru
[3]Университет г. Пиза, Италия; m.pagano@iet.unipi.it

Аннотация: Гауссовские процессы являются мощным инструментом в моделировании сетей, так как они позволяют описать эффект долгой памяти реальных сетевых потоков. Более точно, при реалистичных предположениях, дробное броуновское движение (ДБД) возникает как предельный процесс, когда огромное число on-off источников (с тяжелыми хвостами) мультиплексируются в магистральных сетях. В данной работе изучается жидкостная система массового обслуживания с постоянной скоростью обслуживания, с суммой независимых ДБД на входе, что соответствует агрегации гетерогенных сетевых потоков. Для таких систем массового обслуживания получена логарифмическая асимптотика вероятности переполнения, которая является верхней границей вероятности потери в соответствующих очередях с конечным буфером и которая показывает, что в оценке доминирует ДБД с наибольшим параметром Херста. Наконец, приведены асимптотические результаты для максимума нагрузки в более общем случае гауссовского входного процесса с дисперсией, которая правильно меняется на бесконечности

Ключевые слова: гауссовские жидкостные системы; вероятность переполнения; логарифмические асимптотики

DOI: 10.14375/19922264140203

Литература

1. *Leland W. E., Taqqu M. S., Willinger W., Wilson D. V.* On the self-similar nature of ethernet traffic (extended version). IEEE/ACM Transactions of Networking, 1994. Vol. 2. No. 1. P. 1–15.
2. *Willinger W., Taqqu M. S., Leland W. E., Wilson D.* Self-similarity in high-speed packet traffic: Analysis and modeling of Ethernet traffic measurements // Stat. Sci., 1995. Vol. 10. No. 1. P. 67–85.
3. *Taqqu M. S., Willinger W., Sherman R.* Proof of a fundamental result in self-similar traffic modeling // Comp. Comm. Rev., 1997. Vol. 27. P. 5–23.
4. *Reich E.* On the integrodifferential equation of Takacs I // Ann. Math. Stat., 1958. Vol. 29. P. 563–570.
5. *Duffield N., O'Connell N.* Large deviations and overflow probabilities for the general single server queue, with applications // Proc. Cambridge Phil. Soc., 1995. Vol. 118. P. 363–374.

*Работа проводится при финансовой поддержке Программы стратегического развития Петрозаводского государственного университета в рамках научно-исследовательской деятельности.

6. *Debicki K.* A note on LDP for supremum of Gaussian processes over infinite horizon // Stat. Probab. Lett., 1999. Vol. 44. P. 211–220.
7. *Duffy K., Lewis J. T., Sullivan W. G.* Logarithmic asymptotics for the supremum of a stochastic process // Ann. Appl. Probab., 2003. Vol. 13. No. 2. P. 430–445.
8. *Kim H. S., Shroff N. B.* Loss probability calculations and asymptotic analysis for finite buffer multiplexers // IEEE/ACM Trans. Networking, 2001. Vol. 9. P. 755–768.
9. *Kim H. S., Shroff N. B.* On the asymptotic relationship between the overflow probability and the loss ratio // Adv. Appl. Probab., 2001. Vol. 33. P. 836–863.
10. *Goricheva R. S., Lukashenko O. V., Morozov E. V., Pagano M.* Regenerative analysis of a finite buffer fluid queue // *2010 Congress (International) on Ultra Modern Telecommunications and Control Systems and Workshops (ICUMT) Proceedings*, 2010. P. 1132–1136.
11. *Lukashenko O. V., Morozov E. V., Pagano M.* Estimation of loss probability in Gaussian queues // Conference (International) "Modern Probabilistic Methods for Analysis and Optimization of Information and Telecommunication Networks" Proceedings, 2011. P. 142–147.
12. *Lukashenko O. V., Morozov E. V., Nekrasova R. S., Pagano M.* Performance evaluation of finite buffer queues through regenerative simulation // Comm. Comp. Inform. Sci. BWWQT 2013, 2013. Vol. 356. P. 131–139.
13. *Zeevi A., Glynn P.* On the maximum workload in a queue fed by fractional Brownian motion // Ann. Appl. Probab., 2000. Vol. 10. P. 1084–1099.
14. *Hüsler J., Piterbarg V. I.* Limit theorem for maximum of the storage process with fractional Brownian as input // Stochastic Proc. Their Appl., 2004. Vol. 114. P. 231–250.
15. *Lukashenko O. V., Morozov E. V.* Asymptotics of the maximum workload for a class of Gaussian queues // Информатика и её применения, 2012. Т. 6. Вып. 3. С. 81–89.
16. *Lukashenko O. V., Morozov E. V.* On the maximum workload for a class of Gaussian queues // Conference (International) "Probability Theory and Its Applications" in Commemoration of the Centennial of B. V. Gnedenko, 2012. P. 231–232.
17. *Lukashenko O. V., Morozov E. V.* On convergence in the L_p space of the workload maximum for a class of Gaussian queueing systems // Информатика и её применения, 2013. Т. 7. Вып. 1. С. 36–43.
18. *Dieker A. B., Mandjes M.* On asymptotically efficient simulation of large deviation probabilities // Adv. Appl. Probab., 2005. Vol. 37. P. 539–552.
19. *Giordano S., Gubinelli M., Pagano M.* Bridge Monte-Carlo: A novel approach to rare events of Gaussian processes // 5th St. Petersburg Workshop on Simulation Proceedings. St. Petersburg, Russia, 2005. P. 281–286.
20. *Dieker, A. B., Mandjes M.* Fast simulation of overflow probabilities in a queue with Gaussian input // ACM Trans. Model. Comput. Simul., 2006. Vol. 16. No. 2. P. 119–151.
21. *Giordano S., Gubinelli M., Pagano M.* Rare events of Gaussian processes: A performance comparison between Bridge Monte-Carlo and Importance Sampling // Next Generation Teletraffic and Wired/Wireless Advanced Networking. St. Petersburg, Russia, 2007. P. 269–280.
22. *Lukashenko O. V., Morozov E. V., Pagano M.* Performance analysis of bridge Monte-Carlo estimator // Trans. KarRC RAS, 2012. Vol. 3. P. 54–60.
23. *Takacs L.* Combinatorial methods in the theory of stochastic processes. — John Wiley&Sons, 1967. 262 p.
24. *Mandjes M.* Large deviations of Gaussian queues. Chichester: Wiley, 2007. 339 p.
25. *Rizk A., Fidler M.* Sample path bounds for long memory fbm traffic // 29th Conference on Information Communications, INFOCOM'10 Proceedings. — Piscataway, NJ, USA: IEEE Press, 2010. P. 61–65.
26. *Rizk A., Fidler M.* Non-asymptotic end-to-end performance bounds for networks with long range dependent fbm cross traffic // Computer Networks, 2012. Vol. 56. No. 1. P. 127–141.
27. *Lukashenko O., Morozov E., Pagano M.* On the effective envelopes for fluid queues with Gaussian input // Comm. Comp. Inform. Sci. DCCN 2013, 2014. Vol. 279. P. 178–189.
28. *Adler R. J.* An introduction to continuity, extrema, and related topics for general Gaussian processes. — Hayward, CA: Institute of Mathematical Statistics, 1990. 160 p.
29. *Debicki K.* Gaussian processes // Encyclopedia of actuarial sciences, 2004. Vol. 2. P. 752–757.
30. *Dembo A., Zeitouni O.* Large deviations techniques and applications. Springer, 1998. 396 p.
31. *Seneta E.* Regularly varying functions. — Springer, 1985. 116 p.
32. *Konstantopoulos T., Zazanis M., De Veciana G.* Conservation laws and reflection mappings with application to multiclass mean value analysis for stochastic fluid queues // Stochastic Proc. Their Appl., 1996. Vol. 65. P. 39–146.
33. *Kulkarni V., Rolski T.* Fluid model driven by an Ornstein–Uhlenbeck process // Prob. Eng. Inform. Sci., 1994. Vol. 8. P. 403–417.
34. *Debicki K., Rolski T.* A Gaussian fluid model // Queueing Syst., 1995. Vol. 20. P. 433–452.
35. *Anick D., Mitra D., Sondhi M. M.* Stochastic theory of a data handling system with multiple resources // Bell Syst. Techn. J., 1982. Vol. 61. P. 1871–1894.
36. *Addie R., Mannersalo P., Norros I.* Most probable paths and performance formulae for buffers with Gaussian input traffic // Eur. Trans. Telecommunications, 2002. Vol. 13. P. 183–196.

Поступила в редакцию 08.03.2014

ИНФОРМАТИКА И ЕЁ ПРИМЕНЕНИЯ, 2014. Т. 8. Вып. 2. С. 39–47

ОБОБЩЕННАЯ ЗАДАЧА РАСПРЕДЕЛЕНИЯ РЕСУРСОВ ПРОГРАММНОЙ СИСТЕМЫ

А. В. Босов[1]

Аннотация: Представлены постановка и решение задачи оптимизации динамической системы с линейным выходом по квадратичному критерию качества. Неопределенность системы описывается наблюдаемым случайным процессом второго порядка. В качестве практического обоснования рассматриваются потребности оптимизации распределения ресурсов некоторой программной системы. В такой интерпретации неопределенность системы описывает пользовательскую активность, а выход — число выполняемых запросов или объем запрашиваемой памяти. Цель оптимизации формализуется квадратичным критерием качества общего вида. Критерий, в частности, обобщает две задачи распределения ресурсов программной системы, рассмотренные ранее. Целевой функционал позволяет, в частности, ставить задачи выделения достаточного объема программных ресурсов (нитей, памяти и т. п.), штрафуя за их неограниченное расходование. Для решения задачи используется метод динамического программирования. Оптимальная стратегия находится в виде линейной комбинации выхода и прогнозов состояния вплоть до горизонта управления. В связи с вычислительной трудоемкостью оптимальной стратегии обсуждается возможность ее упрощения и применения локально-оптимальной стратегии.

Ключевые слова: программная система; стохастическая система наблюдения; квадратичный критерий; динамическое программирование

DOI: 10.14357/19922264140204

1 Введение

Задачи оптимального распределения ресурсов могут возникать в самых разных приложениях — от индустриальных до финансовых. Наиболее популярны постановки, имеющие экономический контекст, например, связанные с инвестированием [1]. Однако в связи с повсеместным распространением информационных технологий (ИТ) традиционные постановки оптимального распределения ресурсов нашли новую область применения.

Надо отметить, что понятие ресурса благодаря ИТ приобрело множество специфических оттенков. Так, ИТ-специалист под ресурсом скорее будет понимать веб-сайт или базу данных, чем сырье, средства производства или финансовые инструменты. Для таких ресурсов применяется термин «информационный ресурс» [2]. Близким, но не синонимичным ему является термин «вычислительный ресурс» [3]. Развитие телекоммуникационных сетей, распределенных систем, центров обработки данных и технологий виртуализации обогатило терминологию множеством вариаций на эту тему. Часть из них носит вполне материальный характер (например, сервер, процессорное время, объем памяти, пропускная способность сети). Часть — виртуальны по своей сути (например, программа, сайт, банк данных, запрос). Задачи оптимального распределения ИТ-ресурсов применительно к ресурсам первого типа выглядят вполне традиционными. Но представляется перспективным рассматривать такие же задачи применительно и к ресурсам второго типа. Объединяет виртуальные ресурсы то, что все они имеют отношение к программным системам (сами являются программами или ими обслуживаются).

Примеры программных продуктов, в которых так или иначе алгоритмизируются процедуры управления ресурсами, хорошо известны. Достаточно упомянуть менеджеры задач, имеющиеся в любой операционной системе [4], и оптимизаторы запросов, функционирующие в любой системе управления базами данных [5]. Можно упомянуть еще задачи балансировки нагрузки и управления потоками заданий, а также постановки, связанные с управляемым протоколом TCP (transmission control protocol) [6–8].

Модель и отчасти критерий оптимальности, рассмотренные в данной работе, возникли именно из задачи оптимизации функционирования программной системы. Предложенная постановка, по-видимому, может представлять интерес и в какой-то экономической интерпретации, и в иных приложениях, но главным приложением рассмотренной далее задачи является область программирования (точнее, проектирования программных систем). По

[1] Институт проблем информатики Российской академии наук, AVBosov@ipiran.ru

крайней мере, именно это дает обоснование предложенному далее виду целевого функционала.

2 Постановка задачи

Далее используются следующие обозначения:

- $\triangleq$ — равенство по определению;
- $\mathbb{R}^p$ — p-мерное евклидово пространство;
- $\mathbb{R}^{p\times q}$ — пространство матриц размерности $p\times q$;
- $\mathbb{M}[x], \mathbb{M}[x|\mathfrak{J}]$ — безусловное математическое ожидание случайного вектора $x\in\mathbb{R}^p$ и условное математическое ожидание x относительно σ-алгебры $\mathfrak{J}$;
- x^{T} — операция транспонирования вектора (матрицы) x;
- α^+ — операция псевдообращения матрицы α;
- $\|x\|_\alpha \triangleq \sqrt{x^{\mathrm{T}}\alpha x}$ — вес (норма при $\alpha>0$) вектора $x\in\mathbb{R}^p$, заданный симметричной ($\alpha=\alpha^{\mathrm{T}}$) неотрицательно определенной ($\alpha\geq 0$) матрицей $\alpha\in\mathbb{R}^{p\times p}$; $\|x\|^2\triangleq\|x\|^2_E$, где E — единичная матрица соответствующей размерности;
- $\mathrm{col}\,(x_1,\ldots,x_q)\triangleq(x_1^{\mathrm{T}},\ldots x_q^{\mathrm{T}})^{\mathrm{T}}$ — вектор-столбец, составленный из векторов $x_1,\ldots,x_q$;
- $\mathfrak{J}_t^y\triangleq\sigma\{y_\tau,\tau\leq t\}$ — σ-алгебра, порожденная случайной последовательностью y_τ, $\tau\leq t$.

Пусть задано распределение случайной последовательности $y_t\in\mathbb{R}^q$, $t=0,1,\ldots,N+1$, второго порядка: $\mathbb{M}[\|y_t\|^2]<\infty$. Будем предполагать, что последовательность y_t наблюдается, порождает σ-алгебру $\mathfrak{J}_t^y$ и выход $z_t\in\mathbb{R}^p$, описываемый уравнением

$$z_{t+1}=a_ty_t+b_tz_t+c_tu_t\,,\qquad t=0,1,\ldots,N\,,\quad z_0=0\,,\quad(1)$$

где $u_t\in\mathbb{R}^r$ — управляющее воздействие, формируемое по входной последовательности y_τ, $\tau\leq t$, с целью минимизации следующего целевого функционала:

$$\left.\begin{aligned}&J(U_N)=\\&=\sum_{t=0}^{N}\mathbb{M}\Big[\|P_ty_{t+1}+Q_tz_{t+1}+R_tu_t+S_t\|^2_{\alpha_t}+\\&\qquad+\|u_t-u_{t-1}\|^2_{\beta_t}+\|u_t\|^2_{\gamma_t}\Big]\,;\\&V_N=\mathrm{col}\,(u_0,\ldots,u_t,\ldots,u_N)\in\mathbb{R}^{r\times(N+1)}\end{aligned}\right\}\quad(2)$$

Матрицы $a_t\in\mathbb{R}^{p\times q}$, $b_t\in\mathbb{R}^{p\times p}$, $c_t\in\mathbb{R}^{p\times r}$, $P_t\in\in\mathbb{R}^{s\times q}$, $Q_t\in\mathbb{R}^{s\times p}$, $R_t\in\mathbb{R}^{s\times r}$, $\alpha_t\in\mathbb{R}^{s\times s}$, $\beta_t\in\mathbb{R}^{r\times r}$, $\gamma_t\in\mathbb{R}^{r\times r}$ и векторы $S_t\in\mathbb{R}^s$, $t=0,1,\ldots N$, предполагаются известными, $\beta_0=0$. Допустимыми будем считать управляющие воздействия u_t, являющиеся $\mathfrak{J}_t^y$-измеримыми случайными последовательностями второго порядка: $\mathbb{M}\left[\|u_t\|^2\right]<\infty$. Задачей оптимизации является поиск управляющего воздействия u_t^*, минимизирующего целевой функционал $J(U_N)$:

$$\left.\begin{aligned}&U_N^*\in\operatorname*{argmin}_{U_N}J(U_N)\,;\\&U_N^*=\mathrm{col}\,(u_0^*,\ldots,u_t^*,\ldots,u_N^*)\,.\end{aligned}\right\}\quad(3)$$

Отметим, что процесс, описываемый уравнением (1), в общем случае не предполагается гауссовским и не обладает марковским свойством.

3 Обсуждение целевого функционала

Функционал (2) сформирован в результате обобщения двух задач оптимизации функционирования определенной программной системы. Эта программная система — Информационный веб-портал — реализация одного из вариантов организации распределенных информационных систем, взаимодействующих на принципах федеративности в среде Интернета [9]. В связи с функционированием этой системы возникли две задачи: оптимизации расходования «внутренних» [10] и «внешних» [11] ресурсов. Основной функцией портала является обеспечение промежуточного слоя взаимодействия пользователя, формирующего поисковые запросы к порталу, и информационных источников портала: пользовательский запрос преобразуется в набор команд, исполняемых взаимодействующими с порталом системами-источниками, результаты работы которых консолидируются и возвращаются пользователю в качестве ответа на запрос. Соответственно, портал функционирует в условиях неопределенности пользовательской активности, описываемой процессом y_t.

Для взаимодействия с источниками программными средствами портала поддерживается некоторый пул (множество нитей), используемый для организации параллельного исполнения команд источниками. Определение размера пула и составляет задачу оптимизации расходования «внутренних» ресурсов. При отсутствии какой-либо стратегии управления пулом число поддерживаемых им нитей может, вообще говоря, расти неограниченно. Эта задача по сути сводится к отслеживанию управляющей переменной u_t (числа нитей) выхода z_t (числа выполняемых команд). В [10] цель управле-

ния пулом формализуется слагаемым $\|z_{t+1} - u_t\|^2$, а штраф за его рост — слагаемым $\|u_t\|^2$.

Далее для обеспечения собственного функционирования портал использует «внешние» (предоставляемые обслуживающими подсистемами) ресурсы, в частности, активно изменяющееся дисковое пространство. Оптимизация запросов на выделение/освобождение дискового пространства сводится к задаче отслеживания выходом z_t (общим объемом портального хранилища) определенной траектории, «близкой» к траектории фазовой переменной y_t (размера хранимых пользовательских данных). В [11] цель управления, сформулированная как поддержание хранилища достаточного объема, формализуется слагаемым $\|y_{t+1} - z_{t+1} - S_t\|^2$, где S_t — объем свободного места, штраф за применение управляющего воздействия — слагаемым $\|u_t\|^2$.

Обобщение целей оптимизации этих двух задач в функционале (2) представляет первое слагаемое. Во-первых, пары слагаемых $P_t y_{t+1}, R_t u_t$ и $Q_t z_{t+1}, R_t u_t$ позволяют ставить задачи отслеживания управляющим воздействием u_t траекторий y_t или z_t соответственно, формализуя таким образом задачу оптимизация распределения «внутренних» ресурсов. Во-вторых, пара слагаемых $P_t y_{t+1}, Q_t z_{t+1}$ позволяет отслеживать выходом z_t фазовую переменную y_t. В-третьих, слагаемое S_t позволяет ставить задачи формирования траектории выхода, «близкой» к заданной опорной. Кроме того, слагаемое $Q_t z_{t+1}$ позволяет ставить традиционную задачу минимизации нормы выходной переменной. Наконец, очевидно, что перечисленные цели оптимизации могут произвольно комбинироваться.

Довольно специфическим является второе слагаемое из (2). Оно вносит в целевой функционал штраф за изменение управляющего воздействия. Объяснение этому слагаемому дает заданная предметная область. Его уместно охарактеризовать как плату за реконфигурацию программной системы. Очевидно, что такие действия, как постоянное подключение/отключение хостов, процессоров, дисков и т. п., служат очевидным поводом к сокращению их времени службы, а выделение/освобождение памяти, создание/удаление нитей, указателей и т. п. в программах — источником порождения ошибок как самой программой, так и обеспечивающими подсистемами. Кроме того, частое реконфигурирование существенно труднее реализовать, и оно само по себе расходует значительные ресурсы. Соответственно, назначение этого слагаемого — сократить число подобных реконфигураций.

Последнее, третье, слагаемое вносит в целевой функционал традиционный штраф за размер управляющего воздействия.

Легкость в формализации целей оптимизации является несомненным преимуществом целевого функционала. Понятный физический смысл перечисленных выше слагаемых — это достоинство функционала (2). Однако очевидны и его недостатки. Впрочем, это недостатки, свойственные любому квадратичному критерию. В дискуссии о «нефизичности» квадратичного функционала качества участвовал еще Р. Беллман, а в попытках решить «обратную» задачу, связав параметры квадратичного функционала с динамическими характеристиками процессов, — например, Р. Калман. В исследованиях на эту тему участвовали многие ученые, обозревать эту историю в данной работе неуместно. В рассматриваемой же задаче с учетом предметной области проблему выбора весовых коэффициентов α_t, β_t и γ_t, а также горизонта управления N представляется возможным возложить на этап реализации программной системы. Тем более, что в большинстве случаев создаваемые программы вполне допускают возможность многократного проведения недорогостоящих экспериментов и эмпирический анализ качества оптимизированной программной системы по другим критериям (например, мнению эксперта, пользователя, заказчика). В пользу же квадратичного критерия говорит его «технологичность» — развитый аппарат исследования и возможность получения конечных аналитических решений.

4 Решение задачи оптимизации

Теорема. Решение задачи оптимизации (3) выхода (1), формируемого случайной последовательностью y_t, существует. Оптимальное решение $U_N^* = \operatorname{col}(u_0^*, \ldots, u_N^*)$ с минимальной нормой $\|u_t^*\|^2$ определяется следующими выражениями:

$$u_t^* = L_t^+ \left(\beta_t u_{t-1}^* - K_t^z z_t - K_t\right),\ t = 0, 1, \ldots, N, \quad (4)$$

где

$$\left.\begin{aligned} K_t &\triangleq \sum_{j=0}^{N-t+1} K_t^j \overline{y}(t+j, t) + K_t^s\,, \\ \overline{y}(t+j, t) &\triangleq \mathbb{M}\left[y_{t+j} | \mathfrak{I}_t^y\right]\,; \end{aligned}\right\} \quad (5)$$

$$K_t^z = (Q_t c_t + R_t)^{\mathrm{T}} \alpha_t Q_t b_t + \beta_{t+1} L_{t+1}^+ K_{t+1}^z b_t + \\ + c_t^{\mathrm{T}} (M_{t+1}^z - \left(K_{t+1}^z\right)^{\mathrm{T}} L_{t+1}^+ K_{t+1}^z) b_t\,; \quad (6)$$

$$K_t^s = (Q_t c_t + R_t)^{\mathrm{T}} \alpha_t S_t + \beta_{t+1} L_{t+1}^+ K_{t+1}^s + \\ + c_t^{\mathrm{T}} (M_{t+1}^s - \left(K_{t+1}^z\right)^{\mathrm{T}} L_{t+1}^+ K_{t+1}^s)\,; \quad (7)$$

$$K_t^0 = (Q_t c_t + R_t)^{\mathrm{T}} \alpha_t Q_t a_t + \beta_{t+1} L_{t+1}^+ K_{t+1}^z a_t + \\ + c_t^{\mathrm{T}} \left(M_{t+1}^z - \left(K_{t+1}^z\right)^{\mathrm{T}} L_{t+1}^+ K_{t+1}^z\right) a_t\,; \quad (8)$$

$$K_t^1 = (Q_t c_t + R_t)^{\mathrm{T}} \alpha_t P_t + \beta_{t+1} L_{t+1}^+ K_{t+1}^0 + \\ + c_t^{\mathrm{T}} \left(M_{t+1}^0 - \left(K_{t+1}^z\right)^{\mathrm{T}} L_{t+1}^+ K_{t+1}^0 \right), \quad (9)$$

$$K_t^j = \beta_{t+1} L_{t+1}^+ K_{t+1}^{j-1} + \\ + c_t^{\mathrm{T}} \left(M_{t+1}^j - \left(K_{t+1}^z\right)^{\mathrm{T}} L_{t+1}^+ K_{t+1}^{j-1} \right), \\ j = 2, \ldots, N-t+1, \;\; t = 0, 1, \ldots, N-1; \quad (10)$$

$$\left.\begin{aligned} K_N^z &= (Q_N c_N + R_N)^{\mathrm{T}} \alpha_N Q_N b_N; \\ K_N^s &= (Q_N c_N + R_N)^{\mathrm{T}} \alpha_N S_N; \\ K_N^0 &= (Q_N c_N + R_N)^{\mathrm{T}} \alpha_N Q_N a_N; \\ K_N^1 &= (Q_N c_N + R_N)^{\mathrm{T}} \alpha_N P_N; \end{aligned}\right\} \quad (11)$$

$$\left.\begin{aligned} L_t = (Q_t c_t + R_t)^{\mathrm{T}} \alpha_t (Q_t c_t + R_t) + \beta_t + \gamma_t + \\ + \beta_{t+1} + c_t^{\mathrm{T}} M_{t+1}^z c_t - \left(\beta_{t+1} - c_t^{\mathrm{T}} (K_{t+1}^z)^{\mathrm{T}}\right)^{\mathrm{T}} \times \\ \times L_{t+1}^+ \left(\beta_{t+1} - c_t^{\mathrm{T}} (K_{t+1}^z)^{\mathrm{T}}\right); \\ L_N = (Q_N c_N + R_N)^{\mathrm{T}} \alpha_N (Q_N c_N + R_N) + \\ + \beta_N + \gamma_N; \end{aligned}\right\} \quad (12)$$

$$\left.\begin{aligned} M_t^z &= b_t^{\mathrm{T}} Q_t^{\mathrm{T}} \alpha_t Q_t b_t + \\ &\quad + b_t^{\mathrm{T}} \left(M_{t+1}^z - (K_{t+1}^z)^{\mathrm{T}} L_{t+1}^+ K_{t+1}^z\right) b_t; \\ M_t^s &= b_t^{\mathrm{T}} Q_t^{\mathrm{T}} \alpha_t S_t + \\ &\quad + b_t^{\mathrm{T}} \left(M_{t+1}^s - (K_{t+1}^z)^{\mathrm{T}} L_{t+1}^+ K_{t+1}^s\right); \\ M_t^0 &= b_t^{\mathrm{T}} Q_t^{\mathrm{T}} \alpha_t Q_t a_t + \\ &\quad + b_t^{\mathrm{T}} \left(M_{t+1}^z - (K_{t+1}^z)^{\mathrm{T}} L_{t+1}^+ K_{t+1}^z\right) a_t; \\ M_t^1 &= b_t^{\mathrm{T}} Q_t^{\mathrm{T}} \alpha_t P_t + \\ &\quad + b_t^{\mathrm{T}} \left(M_{t+1}^0 - (K_{t+1}^z)^{\mathrm{T}} L_{t+1}^+ K_{t+1}^0\right); \\ M_t^j &= b_t^{\mathrm{T}} \left(M_{t+1}^{j-1} - (K_{t+1}^z)^{\mathrm{T}} L_{t+1}^+ K_{t+1}^{j-1}\right), \\ &\quad j = 2, \ldots, N-t+1, \;\; t = 0, 1, \ldots, N-1; \end{aligned}\right\} \quad (13)$$

$$\left.\begin{aligned} M_N^z = b_N^{\mathrm{T}} Q_N^{\mathrm{T}} \alpha_N Q_N b_N; \;\; M_N^s = b_N^{\mathrm{T}} Q_N^{\mathrm{T}} \alpha_N S_N; \\ M_N^0 = b_N^{\mathrm{T}} Q_N^{\mathrm{T}} \alpha_N Q_N a_N; \;\; M_N^1 = b_N^{\mathrm{T}} Q_N^{\mathrm{T}} \alpha_N P_N. \end{aligned}\right\} \quad (14)$$

Доказательство. Дополнительно к (5) обозначим

$$M_t \triangleq \sum_{j=0}^{N-t+1} M_t^j \overline{y}(t+j, t) + M_t^s.$$

Для доказательства утверждения воспользуемся методом динамического программирования [12]. Обозначим функцию Беллмана:

$$W_t \triangleq \min_{u_t, \ldots, u_N} \sum_{\tau=t}^{N} \mathbb{M}\left[\|P_\tau y_{\tau+1} + Q_\tau z_{\tau+1} + R_\tau u_\tau + \right. \\ \left. + S_\tau\|_{\alpha_\tau}^2 + \|u_\tau - u_{\tau-1}\|_{\beta_\tau}^2 + \|u_\tau\|_{\gamma_\tau}^2 | \mathfrak{J}_t^y \right].$$

При $t = N$ с учетом (1) имеем:

$$W_N = \min_{u_N} \mathbb{M}\Big[\|P_N y_{N+1} + Q_N a_N y_N + \\ + Q_N b_N z_N + (Q_N c_N + R_N) u_N + S_N\|_{\alpha_N}^2 + \\ + \|u_N - u_{N-1}\|_{\beta_N}^2 + \|u_N\|_{\gamma_N}^2 \Big| \mathfrak{J}_N^y \Big] = \\ = \min_{u_N} \mathbb{M}\Big[u_N^{\mathrm{T}} \Big((Q_N c_N + R_N)^{\mathrm{T}} \alpha_N (Q_N c_N + R_N) + \\ + \beta_N + \gamma_N \Big) u_N - 2 u_N^{\mathrm{T}} \Big(\beta_N u_{N-1} - \\ - (Q_N c_N + R_N)^{\mathrm{T}} \alpha_N (P_N y_{N+1} + Q_N a_N y_N + \\ + Q_N b_N z_N + S_N) \Big) + \|P_N y_{N+1} + Q_N a_N y_N + \\ + Q_N b_N z_N + S_N\|_{\alpha_N}^2 + \|u_{N-1}\|_{\beta_N}^2 \Big| \mathfrak{J}_N^y \Big] = \\ = \min_{u_N} \big(u_N^{\mathrm{T}} L_N u_N - 2 u_N^{\mathrm{T}} (\beta_N u_{N-1} - K_N^z z_N - \\ - K_N^0 \overline{y}(N, N) - K_N^1 \overline{y}(N+1, N) - K_N^s) \big) + \\ + \|u_{N-1}\|_{\beta_N}^2 + \mathbb{M}\Big[\|P_N y_{N+1} + Q_N a_N y_N + \\ + Q_N b_N z_N + S_N\|_{\alpha_N}^2 \Big| \mathfrak{J}_N^y \Big].$$

Отсюда с учетом обозначения

$$K_N \triangleq K_N^0 \overline{y}(N, N) + K_N^1 \overline{y}(N+1, N) + K_N^s$$

и в предположении $L_N \geq 0$ вытекает, что (см. (4))

$$u_N^* = L_N^+ (\beta_N u_{N-1} - K_N^z z_N - K_N)$$

и u_N^* имеет минимальную евклидову норму [13]. Здесь были использованы обозначения, введенные в (11) и (12). Кроме того, здесь и далее учитывается очевидное равенство $\overline{y}(t, t) \triangleq y_t$.

Подставляя u_N^* в полученное выражение для W_N, получаем:

$$W_N = -(\beta_N u_{N-1} - K_N^z z_N - K_N)^{\mathrm{T}} L_N^+ (\beta_N u_{N-1} - \\ - K_N^z z_N - K_N) + \|u_{N-1}\|_{\beta_N}^2 + \mathbb{M}\Big[\|P_N y_{N+1} + \\ + Q_N a_N y_N + Q_N b_N z_N + S_N\|_{\alpha_N}^2 \Big| \mathfrak{J}_N^y \Big] = \\ = -(\beta_N u_{N-1} - K_N^z z_N - K_N)^{\mathrm{T}} L_N^+ (\beta_N u_{N-1} - \\ - K_N^z z_N - K_N) + \|u_{N-1}\|_{\beta_N}^2 + \\ + z_N^{\mathrm{T}} b_N^{\mathrm{T}} Q_N^{\mathrm{T}} \alpha_N Q_N b_N z_N + \\ + 2 z_N^{\mathrm{T}} b_N^{\mathrm{T}} Q_N^{\mathrm{T}} \alpha_N (P_N \overline{y}(N+1, N) + Q_N a_N \overline{y}(N, N) + \\ + S_N) + \mathbb{M}\Big[\|P_N y_{N+1} + Q_N a_N y_N + S_N\|_{\alpha_N}^2 \Big| \mathfrak{J}_N^y \Big] =$$

$$= -(\beta_N u_{N-1} - K_N^z z_N - K_N)^{\mathrm{T}} L_N^+ (\beta_N u_{N-1} -$$
$$- K_N^z z_N - K_N) + \|u_{N-1}\|^2_{\beta_N} + z_N^{\mathrm{T}} M_N^z z_N +$$
$$+ 2z_N^{\mathrm{T}}(M_N^0 \overline{y}(N,N) + M_N^1 \overline{y}(N+1,N) + M_N^s) +$$
$$+ \mathbb{M}\left[\|P_N y_{N+1} + Q_N a_N y_N + S_N\|^2_{\alpha_N} \middle| \mathfrak{J}_N^y\right].$$

Здесь были использованы обозначения, введенные в (14). Группируя далее слагаемые при u_{N-1} и z_N, окончательно получаем:

$$W_N = u_{N-1}^{\mathrm{T}}(\beta_N - \beta_N L_N^+ \beta_N) u_{N-1} +$$
$$+ 2u_{N-1}^{\mathrm{T}} \beta_N L_N^+ K_N^z z_N + 2u_{N-1}^{\mathrm{T}} \beta_N L_N^+ K_N +$$
$$+ 2z_N^{\mathrm{T}}(M_N - (K_N^z)^{\mathrm{T}} L_N^+ K_N) +$$
$$+ z_N^{\mathrm{T}}(M_N^z - (K_N^z)^{\mathrm{T}} L_N^+ K_N^z) z_N + \mathbb{M}\Big[\|P_N y_{N+1} +$$
$$+ Q_N a_N y_N + S_N\|^2_{\alpha_N} - \|K_N\|^2_{L_N^+} \Big| \mathfrak{J}_N^y\Big].$$

Здесь учтена $\mathfrak{J}_N^y$-измеримость K_N из (5).

Предположим теперь, что полученное выражение для W_N имеет место и для W_t, т. е.

$$W_t = u_{t-1}^{\mathrm{T}}(\beta_t - \beta_t L_t^+ \beta_t) u_{t-1} + 2u_{t-1}^{\mathrm{T}} \beta_t L_t^+ K_t^z z_t +$$
$$+ 2u_{t-1}^{\mathrm{T}} \beta_t L_T^+ K_t + 2z_t^{\mathrm{T}} \left(M_t - (K_t^z)^{\mathrm{T}} L_t^+ K_t\right) +$$
$$+ z_t^{\mathrm{T}} \left(M_t^z - (K_t^z)^{\mathrm{T}} L_t^+ K_t^z\right) z_t + \mathbb{M}\left[A_t \middle| \mathfrak{J}_t^y\right], \quad (15)$$

где обозначено:

$$A_t = \|P_t y_{t+1} + Q_t a_t y_t + S_t\|^2_{\alpha_t} - \|K_t\|^2_{L_t^+} +$$
$$+ 2y_t^{\mathrm{T}} a_t^{\mathrm{T}} \left(M_t - (K_t^z)^{\mathrm{T}} L_t^+ K_t\right) + y_t^{\mathrm{T}} a_t^{\mathrm{T}} \left(M_{t+1}^z -$$
$$- \left(K_{t+1}^z\right)^{\mathrm{T}} L_{t+1}^+ K_{t+1}^z\right) a_t y_t + A_{t+1};$$
$$A_N = \|P_N y_{N+1} + Q_N a_N y_N + S_N\|^2_{\alpha_N} - \|K_N\|^2_{L_N^+}.$$

Для доказательства (15) и (4) для $t-1$ запишем уравнение Беллмана для W_{t-1}:

$$W_{t-1} = \min_{u_{t-1}} \mathbb{M}\Big[\|P_{t-1} y_t + Q_{t-1} z_t + R_{t-1} u_{t-1} +$$
$$+ S_{t-1}\|^2_{\alpha_{t-1}} + \|u_{t-1} - u_{t-2}\|^2_{\beta_{t-1}} +$$
$$+ \|u_{t-1}\|^2_{\gamma_{t-1}} + W_t \Big| \mathfrak{J}_{t-1}^y\Big].$$

С учетом (1) и (15) имеем:

$$W_{t-1} = \min_{u_{t-1}} \mathbb{M}\Big[\|P_{t-1} y_t + Q_{t-1} a_{t-1} y_{t-1} +$$
$$+ Q_{t-1} b_{t-1} z_{t-1} + + (Q_{t-1} c_{t-1} + R_{t-1}) u_{t-1} +$$
$$+ S_{t-1}\|^2_{\alpha_{t-1}} + \|u_{t-1} - u_{t-2}\|^2_{\beta_{t-1}} +$$
$$+ \|u_{t-1}\|^2_{\gamma_{t-1}} + u_{t-1}^{\mathrm{T}} \left(\beta_t - \beta_t L_t^+ \beta_t\right) u_{t-1} +$$
$$+ 2u_{t-1}^{\mathrm{T}} \beta_t L_t^+ K_t^z \left(a_{t-1} y_{t-1} + b_{t-1} z_{t-1} +\right.$$
$$\left. + c_{t-1} u_{t-1}\right) + 2u_{t-1}^{\mathrm{T}} \beta_t L_t^+ K_t + 2\left(a_{t-1} y_{t-1} +\right.$$
$$\left. + b_{t-1} z_{t-1} + c_{t-1} u_{t-1}\right)^{\mathrm{T}} \left(M_t - (K_t^z)^{\mathrm{T}} L_t^+ K_t\right) +$$
$$+ \left(a_{t-1} y_{t-1} + b_{t-1} z_{t-1} + c_{t-1} u_{t-1}\right)^{\mathrm{T}} \left(M_t^z -\right.$$
$$\left. - (K_t^z)^{\mathrm{T}} L_t^+ K_t^z\right) \left(a_{t-1} y_{t-1} + b_{t-1} z_{t-1} + c_{t-1} u_{t-1}\right) +$$
$$+ A_t \Big| \mathfrak{J}_{t-1}^y\Big] =$$
$$= \min_{u_{t-1}} \mathbb{M}\Big[u_{t-1}^{\mathrm{T}}\Big(\left(Q_{t-1} c_{t-1} + R_{t-1}\right)^{\mathrm{T}} \alpha_{t-1} \times$$
$$\times \left(Q_{t-1} c_{t-1} + R_{t-1}\right) + \beta_{t-1} + \gamma_{t-1} + \beta_t - \beta_t L_t^+ \beta_t +$$
$$+ 2\beta_t L_t^+ K_t^z c_{t-1} + c_{t-1}^{\mathrm{T}} \left(M_t^z - (K_t^z)^{\mathrm{T}} L_t^+ K_t^z\right) c_{t-1}\Big) \times$$
$$\times u_{t-1} - 2u_{t-1}^{\mathrm{T}} \Big(\beta_{t-1} u_{t-2} - \left(Q_{t-1} c_{t-1} + R_{t-1}\right)^{\mathrm{T}} \times$$
$$\times \alpha_{t-1} \left(P_{t-1} y_t + Q_{t-1} a_{t-1} y_{t-1} + Q_{t-1} b_{t-1} z_{t-1} +\right.$$
$$\left. + S_{t-1}\right) - \beta_t L_t^+ K_t^z \left(a_{t-1} y_{t-1} + b_{t-1} z_{t-1}\right) -$$
$$- \beta_t L_t^+ K_t - c_{t-1}^{\mathrm{T}} \left(M_t - (K_t^z)^{\mathrm{T}} L_t^+ K_t\right) -$$
$$- c_{t-1}^{\mathrm{T}} \Big(M_t^z - (K_t^z)^{\mathrm{T}} L_t^+ K_t^z\Big)\left(a_{t-1} y_{t-1} + b_{t-1} z_{t-1}\right)\Big) +$$
$$+ \|P_{t-1} y_t + Q_{t-1} a_{t-1} y_{t-1} + Q_{t-1} b_{t-1} z_{t-1} +$$
$$+ S_{t-1}\|^2_{\alpha_{t-1}} + \|u_{t-2}\|^2_{\beta_{t-1}} +$$
$$+ 2\left(a_{t-1} y_{t-1} + b_{t-1} z_{t-1}\right)^{\mathrm{T}} \left(M_t - (K_t^z)^{\mathrm{T}} L_t^+ K_t\right) +$$
$$+ \left(a_{t-1} y_{t-1} + b_{t-1} z_{t-1}\right)^{\mathrm{T}} \left(M_t^z - (K_t^z)^{\mathrm{T}} L_t^+ K_t^z\right) \times$$
$$\times \left(a_{t-1} y_{t-1} + b_{t-1} z_{t-1}\right) + A_t \Big| \mathfrak{J}_{t-1}^y\Big] =$$
$$= \min_{u_{t-1}} \left(u_{t-1}^{\mathrm{T}} L_{t-1} u_{t-1} - 2u_{t-1}^{\mathrm{T}} \left(\beta_{t-1} u_{t-2} -\right.\right.$$
$$\left.\left. - K_{t-1}^z z_{t-1} - K_{t-1}\right)\right) + \|u_{t-2}\|^2_{\beta_{t-1}} +$$
$$+ \mathbb{M}\Big[\|P_{t-1} y_t + Q_{t-1} a_{t-1} y_{t-1} +$$
$$+ Q_{t-1} b_{t-1} z_{t-1} + S_{t-1}\|^2_{\alpha_{t-1}} +$$
$$+ 2\left(a_{t-1} y_{t-1} + b_{t-1} z_{t-1}\right)^{\mathrm{T}} \left(M_t - (K_t^z)^{\mathrm{T}} L_t^+ K_t\right) +$$
$$+ \left(a_{t-1} y_{t-1} + b_{t-1} z_{t-1}\right)^{\mathrm{T}} \left(M_t^z - (K_t^z)^{\mathrm{T}} L_t^+ K_t^z\right) \times$$
$$\times \left(a_{t-1} y_{t-1} + b_{t-1} z_{t-1}\right) + A_t \Big| \mathfrak{J}_{t-1}^y\Big],$$

откуда в предположении $L_{t-1} \geq 0$ вытекает (см. (4))

$$u_{t-1}^* = L_{t-1}^+ \left(\beta_{t-1} u_{t-2} - K_{t-1}^z z_{t-1} - K_{t-1}\right) \quad (16)$$

и u_{t-1}^* имеет минимальную евклидову норму [13].

Выше были использованы обозначения, введенные в (6)–(10) и (12), и формулы полного математического ожидания

$$\mathbb{M}\left[A_t \middle| \mathfrak{J}_t^y \middle| \mathfrak{J}_{t-1}^y\right] = \mathbb{M}\left[A_t \middle| \mathfrak{J}_{t-1}^y\right];$$
$$\mathbb{M}\left[\overline{y}(t+j,t) \middle| \mathfrak{J}_{t-1}^y\right] = \overline{y}(t+j, t-1).$$

Подставляя u_{t-1}^* в полученное выражение для W_{t-1}, получаем:

$$W_{t-1} = -\left(\beta_{t-1}u_{t-2} - K^z_{t-1}z_{t-1} - K_{t-1}\right)^{\mathrm{T}} L^+_{t-1} \times$$
$$\times \left(\beta_{t-1}u_{t-2} - K^z_{t-1}z_{t-1} - K_{t-1}\right) + \|u_{t-2}\|^2_{\beta_{t-1}} +$$
$$+ \mathbb{M}\Big[\, \|P_{t-1}y_t +$$
$$+ Q_{t-1}a_{t-1}y_{t-1} + Q_{t-1}b_{t-1}z_{t-1} + S_{t-1}\|^2_{\alpha_{t-1}} +$$
$$+ 2\left(a_{t-1}y_{t-1} + b_{t-1}z_{t-1}\right)^{\mathrm{T}} \left(M_t - (K^z_t)^{\mathrm{T}} L^+_t K_t\right) +$$
$$+ \left(a_{t-1}y_{t-1} + b_{t-1}z_{t-1}\right)^{\mathrm{T}} \left(M^z_t - (K^z_t)^{\mathrm{T}} L^+_t K^z_t\right) \times$$
$$\times \left(a_{t-1}y_{t-1} + b_{t-1}z_{t-1}\right) + A_t\Big|\mathfrak{J}^y_{t-1}\Big] =$$
$$= -\left(\beta_{t-1}u_{t-2} - K^z_{t-1}z_{t-1} - K_{t-1}\right)^{\mathrm{T}} L^+_{t-1} \times$$
$$\times \left(\beta_{t-1}u_{t-2} - K^z_{t-1}z_{t-1} - K_{t-1}\right) + \|u_{t-2}\|^2_{\beta_{t-1}} +$$
$$+ z^{\mathrm{T}}_{t-1}\left(b^{\mathrm{T}}_{t-1}Q^{\mathrm{T}}_{t-1}\alpha_{t-1}Q_{t-1}b_{t-1} +\right.$$
$$\left. + b^{\mathrm{T}}_{t-1}\left(M^z_t - (K^z_t)^{\mathrm{T}} L^+_t K^z_t\right) b_{t-1}\right) z_{t-1} +$$
$$+ 2z^{\mathrm{T}}_{t-1}\Big(b^{\mathrm{T}}_{t-1}Q^{\mathrm{T}}_{t-1}\alpha_{t-1}\big(P_{t-1}\overline{y}(t,t-1) +$$
$$+ Q_{t-1}a_{t-1}\overline{y}(t-1,t-1) + S_{t-1}\big) +$$
$$+ b^{\mathrm{T}}_{t-1}\left(\mathbb{M}\left[M_t\left|\mathfrak{J}^y_{t-1}\right.\right] - (K^z_t)^{\mathrm{T}} L^+_t \mathbb{M}\left[K_t\left|\mathfrak{J}^y_{t-1}\right.\right]\right) +$$
$$+ b^{\mathrm{T}}_{t-1}\left(M^z_t - (K^z_t)^{\mathrm{T}} L^+_t K^z_t\right) a_{t-1}\overline{y}(t-1,t-1)\Big) +$$
$$+ \mathbb{M}\Big[\, \|P_{t-1}y_t + Q_{t-1}a_{t-1}y_{t-1} + S_{t-1}\|^2_{\alpha_{t-1}} +$$
$$+ 2y^{\mathrm{T}}_{t-1}a^{\mathrm{T}}_{t-1}\left(M_t - (K^z_t)^{\mathrm{T}} L^+_t K_t\right) +$$
$$+ y^{\mathrm{T}}_{t-1}a^{\mathrm{T}}_{t-1}\left(M^z_t - (K^z_t)^{\mathrm{T}} L^+_t K^z_t\right) a_{t-1}y_{t-1} +$$
$$+ A_t\Big|\mathfrak{J}^y_{t-1}\Big] =$$
$$= -\left(\beta_{t-1}u_{t-2} - K^z_{t-1}z_{t-1} - K_{t-1}\right)^{\mathrm{T}} L^+_{t-1} \times$$
$$\times \left(\beta_{t-1}u_{t-2} - K^z_{t-1}z_{t-1} - K_{t-1}\right) +$$
$$+ \|u_{t-2}\|^2_{\beta_{t-1}} + z^{\mathrm{T}}_{t-1}M^z_{t-1}z_{t-1} + 2z^{\mathrm{T}}_{t-1}M_{t-1} +$$
$$+ \mathbb{M}\Big[\, \|P_{t-1}y_t + Q_{t-1}a_{t-1}y_{t-1} + S_{t-1}\|^2_{\alpha_{t-1}} +$$
$$+ 2y^{\mathrm{T}}_{t-1}a^{\mathrm{T}}_{t-1}\left(M_t - (K^z_t)^{\mathrm{T}} L^+_t K_t\right) +$$
$$+ y^{\mathrm{T}}_{t-1}a^{\mathrm{T}}_{t-1}\left(M^z_t - (K^z_t)^{\mathrm{T}} L^+_t K^z_t\right) a_{t-1}y_{t-1} +$$
$$+ A_t\Big|\mathfrak{J}^y_{t-1}\Big].$$

Здесь были использованы обозначения, введенные в (13). Группируя далее слагаемые при u_{t-2} и z_{t-1} и учитывая $\mathfrak{J}^y_{t-1}$-измеримость K_{t-1} из (5), окончательно получаем (15) для W_{t-1}.

Подставляя далее в (16) $u_{t-2} = u^*_{t-2}$ (в том числе учитывая в (16) для u^*_0 условие $\beta_0 = 0$), окончательно получаем (4).

Для завершения доказательства осталось заметить, что выполнение неравенства $L_t \geq 0$ для всех t очевидно. Действительно, в выкладках, проделанных для функции Беллмана W_{t-1}, показано, что ее можно представить в виде минимума квадратичной формы с матрицей L_{t-1} при квадратичных членах u_{t-1}. Неотрицательная определенность L_{t-1} вытекает, таким образом, из неотрицательности W_{t-1}, имеющей место по условию задачи.

Теорема доказана.

Согласно полученному результату решение u^*_t является линейной комбинацией предыдущего управления u^*_{t-1}, последнего выхода z_t и оптимальных в среднем квадратическом прогнозов состояния K_t из (5) вплоть до горизонта времени $N+1$. Соответственно, с каждым шагом по времени число слагаемых в (5) уменьшается на единицу. Коэффициенты для вычисления K_t заданы рекуррентными (в обратном времени) соотношениями (6)–(10) и (13): коэффициент K^j_t, соответствующий прогнозу на j шагов (K^0_t — множитель при y_t), вычисляется через коэффициенты K^{j-1}_{t+1}, M^{j-1}_{t+1}, определенные на предыдущем шаге рекурсии. На каждом шаге, кроме того, вычисляются коэффициенты K^z_t, M^z_t, K^s_t, M^s_t и L_t — соответственно множитель при z_t, свободный член и общий нормировочный коэффициент. Рекуррентное вычисление коэффициентов стартует при $t = N$ по формулам (11) и (14).

Полученный результат объясняет, почему задача изначально не формулировалась как классическая задача линейно-квадратичного управления путем расширения вектора состояния y_t. Действительно, расширив вектор состояния вектором u_t, можно избавиться от слагаемого u_t в уравнении входа z_t, сохранив квадратичность критерия ($c_t = 0$). Но тогда придется рассматривать зависимость прогнозов y_t от управлений u_t, что приведет либо к необходимости учета влияния управлений на точность прогнозирования (дуальное управление), либо ограничиваться линейностью y_t.

Кроме того, надо отметить, что невозможны и другие упрощения в предложенной модели и критерии. Так, не удается за счет расширения вектора выхода z_t избавиться от слагаемого z_{t-1} в уравнении входа z_t ($b_t = 0$). Также нельзя избавиться от слагаемого $\|u_t - u_{t-1}\|^2_{\beta_t}$ в критерии, расширив вектор u_t (такое расширение вообще искажает смысл задачи).

Таким образом, предложенную в данной работе постановку можно рассматривать как вариант обобщения линейно-квадратичной задачи на случай нелинейного состояния y_t: нелинейность может иметь место, но не должна влиять на выбор управления.

Сделанные замечания показывают также и возможность дальнейшего обобщения критерия. На-

пример, целью оптимизации можно объявить отслеживание u_t линейной комбинации выходов z_τ, $t \le \tau \le N+1$, или вместо точечного штрафа за смену управляющего воздействия (слагаемое $\|u_t - u_{t-1}\|^2_{\beta_t}$) задавать интегральный. Представляется, что подобные обобщения будут иметь смысл только в том случае, если найдут обоснование какими-либо прикладными постановками.

5 Вопросы практической реализуемости

Обращаясь к вопросу практического применения полученного результата, в частности реализуемости оптимального алгоритма, еще раз отметим, что целевой функционал сформировался в связи с оптимизацией функционирования программной системы. Логично предположить, что для разных программ вполне могут возникать аналогичные описанным выше цели оптимизации функционирования — распределения ресурсов. При этом более-менее общим, по-видимому, можно считать и физический смысл неопределенности, описываемой в предложенной модели процессом y_t, а именно: y_t характеризует активность пользователей, обращающихся к программной системе. Процесс может описывать как число пользователей, применяющих программную систему, так и число формируемых ими запросов или команд. Конкретные математические модели y_t могут быть довольно разнообразными. Практически от модели y_t требуется только возможность прогнозирования. В любом случае для рассматриваемой задачи оптимизации задачи собственно оптимизации и оценивания фазового процесса разделены, как в традиционной задаче линейно-квадратичного управления, что является еще одним важным достоинством квадратичного функционала качества.

Реализуемость алгоритма оптимизации даже применительно к обозначенной области приложения зависит прежде всего от выбора горизонта управления $N+1$. Понятно, что, выбрав слишком большой горизонт, придется проводить значительный объем вычислений в связи с расчетом прогнозов. Использование малого горизонта не вполне понятно в связи с реальной целью оптимизации — распределять ресурсы эффективно все время, пока программа работает. Рассуждения здесь зависят единственно от характера априорной информации об y_t. Если y_t свойственны какие-то прогнозируемые особенности (например, резкое увеличение пользовательской активности во вполне определенные часы, дни), то их нельзя не учитывать. На практике такую ситуацию действительно можно видеть, и характеризуется она временно́й периодичностью, например суточными или недельными циклами. Это позволяет вполне обоснованно делить весь потенциально бесконечный период функционирования программы на небольшие повторяющиеся части, равные периодам.

Если же процесс y_t не имеет таких особых моментов, например является эргодическим, то можно ограничиться и очень малым горизонтом. Проведенные в работах [10, 11] расчеты показывают, что в действительности можно ограничиться значениями $N = 1$ и даже $N = 0$. Это означает, что вместо интегрального квадратичного критерия (2) можно использовать локально-оптимальный подход [14]. Так, одношаговое локально-оптимальное управление u_t^L минимизирует целевой функционал

$$J^L(u_t) = \mathbb{M}\left[\|P_t y_{t+1} + Q_t z_{t+1} + R_t u_t + S_t\|^2_{\alpha_t} + \right.$$
$$\left. + \|u_t - u^L_{t-1}\|^2_{\beta_t} + \|u_t\|^2_{\gamma_t}\right],$$

т. е. одно слагаемое исходного функционала (2). Соответственно, в двухшаговое локально-оптимальное управление надо включить два слагаемых. В расчетах надо сравнивать значения (2), доставляемые субоптимальными (локально-оптимальными) и оптимальным алгоритмами. Примеры вычислений, выполненные в работах [10, 11], показывают, что потери при этом составляют 2%–5%, что с учетом ограниченной адекватности любой модели можно отнести к статистической погрешности.

Собственно, вывод о целесообразности практического применения именно локально-оптимального подхода и является основным результатом обсуждения практической реализуемости. Значение же оптимального алгоритма, как чаще всего и бывает, заключается в определении некоторого эталона, близость к которому и является основанием для применения субоптимального алгоритма.

6 Заключение

Задача, сформулированная в данной работе как задача оптимизации выхода стохастической системы, очевидно, может быть обобщена на случай косвенных наблюдений. Неопределенность, описываемая процессом y_t, при этом будет считаться состоянием стохастической динамической системы наблюдения, а выход z_t — косвенными наблюдениями. При этом в уравнение (1) логичным будет добавить шум, описывающий ошибку наблюдения. Вторым направлением для дальнейших исследований является постановка аналогичной задачи для системы с непрерывным временем. В общем случае

здесь выход z_t будет описываться стохастическим дифференциальным уравнением.

Литература

1. *Гитман Л. Дж., Джонк М. Д.* Основы инвестирования / Пер. с англ. — М.: Дело, 1997. 810 с. (*Gitman L. J., Joehnk M. D.* Fundamentals of investing. — 4th ed. — N.Y.: Harper & Row, 1990. 1008 p.)
2. ГОСТ 7.70-2003. СИБИД. Описание баз данных и машиночитаемых информационных ресурсов. Состав и обозначение характеристик. — М.: Изд-во стандартов, 2003. 11 с.
3. ГОСТ 28195-89. Оценка качества программных средств. Общие положения. — М.: Изд-во стандартов, 2001. 39 с.
4. *Таненбаум Э. С., Вудхалл А. С.* Операционные системы. Разработка и реализация / Пер. с англ. — 3-е изд. — СПб.: Питер, 2007. 704 с. (*Tanenbaum A. S., Woodhull A. S.* Operating systems: Design and implementation. — 3rd ed. — Upper Saddle River, NJ: Prentice Hall, 2006. 1080 p.)
5. *Дейт К. Дж.* Введение в системы баз данных / Пер. с англ. — 8-е изд. — М.: Вильямс, 2005. 1328 с. (*Date C. J.* An introduction to database systems. — 8th ed. — Reading, MA: Addison-Wesley, 2004. 1024 p.)
6. *Elsässer R., Monien B., Preis R.* Diffusion schemes for load balancing on heterogeneous networks // Theory Comput. Syst., 2002. Vol. 35. No. 3. P. 305–320.
7. *Low S. H., Paganini F., Doyle J. C.* Internet congestion control // IEEE Control Syst. Magazine, 2002. Vol. 22. No. 1. P. 28–43.
8. *Welzl M.* Network congestion control. — N.Y.: Wiley, 2005. 263 p.
9. *Босов А. В., Иванов А. В.* Программная инфраструктура Информационного web-портала РАН // Информатика и её применения, 2007. Т. 1. Вып. 2. С. 39–53.
10. *Босов А. В.* Задачи анализа и оптимизации для модели пользовательской активности. Часть 2. Оптимизация внутренних ресурсов // Информатика и её применения, 2012. Т. 6. Вып. 1. С. 18–25.
11. *Босов А. В.* Задачи анализа и оптимизации для модели пользовательской активности. Часть 3. Оптимизация внешних ресурсов // Информатика и её применения, 2012. Т. 6. Вып. 2. С. 15–22.
12. *Бертсекас Д., Шрив С.* Стохастическое оптимальное управление: случай дискретного времени / Пер. с англ. — М.: Наука, 1985. 279 с. (*Bertsekas D. P., Shreve S. E.* Stochastic optimal control: The discrete-time case. — N.Y.: Academic Press, 1978. 323 p.)
13. *Алберт А.* Регрессия, псевдоинверсия и рекуррентное оценивание / Пер. с англ. — М.: Наука, 1977. 224 с. (*Albert A.* Regression and the Moore–Penrose pseudoinverse. — N.Y.: Academic Press, 1972. 179 p.)
14. *Коган М. М., Неймарк Ю. И.* Адаптивное локально-оптимальное управление // Автоматика и телемеханика, 1987. № 8. С. 126–136.

Поступила в редакцию 6.03.14

THE GENERALIZED PROBLEM OF SOFTWARE SYSTEM RESOURCES DISTRIBUTION

A. V. Bosov

Institute of Informatics Problems, Russian Academy of Sciences, 44-2 Vavilov Str., Moscow 119333, Russian Federation

Abstract: The paper presents the statement and the solution of the optimization problem for a dynamic system with a linear output and the quadratic performance criterion. System uncertainty is described by the observed second-order stochastic process. The need to optimize resource distribution of software systems gives practical justification to the problem. In such interpretation, the uncertainty of a system describes user activity and the output describes running queries or the volume of the requested memory. The goals of optimization are formalized by the quadratic performance criterion of the general form. The criterion, in particular, summarizes two problems of resource distribution of software systems discussed earlier. The objective functional makes it possible, in particular, to state the problem of adequate program resources allocation (of threads, memory, etc.), penalizing for unlimited spending. To solve the problem, the method of dynamic programming is used. The optimal strategy is a linear combination of the output and state predictions up to the control horizon. In the context of computational complexity of the optimal strategy, the possibility of its simplicity and of using the locally-optimal strategy is discussed.

Keywords: software system; stochastic observation system; quadratic criterion; dynamic programming

DOI: 10.14357/19922264140204

References

1. Gitman, L. J., and M. D. Joehnk. 1990. *Fundamentals of investing*. 4th ed. N.Y.: Harper & Row. 1008 p.
2. GOST 7.70-2003 SIBID. 2003. Opisanie baz dannykh i mashinochitaemykh informatsionnykh resursov. Sostav i oboznachenie kharakteristik [Description of data bases and information resources. The composition and characteristics of the designation]. Moscow: Standardinform Publs. 11 p.
3. GOST 28195-89. 2001. Otsenka kachestva programmnykh sredstv. Obshchie polozheniya [Assessment of the quality of software. General provisions]. Moscow: Standardinform Publs. 39 p.
4. Tanenbaum, A. S., and A. S. Woodhull. 2006. *Operating systems: Design and implementation*. 3rd ed. Upper Saddle River, NJ: Prentice Hall. 1080 p.
5. Date, C. J. 2004. *An introduction to database systems*. 8th ed. Reading, MA: Addison-Wesley. 1024 p.
6. Elsässer, R., B. Monien, and R. Preis. 2002. Diffusion schemes for load balancing on heterogeneous networks. *Theory Comput. Syst.* 35(3):305–320.
7. Low, S. H., F. Paganini, and J. C. Doyle. 2002. Internet congestion control. *IEEE Control Syst. Magazine* 22(1):28–43.
8. Welzl, M. 2005. *Network congestion control*. N.Y.: Wiley. 263 p.
9. Bosov, A. V., and A. V. Ivanov. 2007. Programmnaya infrastruktura Informatsionnogo web-portala RAN [RAS Informational web-portal software infrastructure]. *Informatika i ee Primeneniya — Inform. Appl.* 2(1):39–53.
10. Bosov, A. V. 2012. Zadachi analiza i optimizatsii dlya modeli pol'zovatel'skoy aktivnosti. Chast' 2. Optimizatsiya vnutrennikh resursov [Analysis and optimization problems for some users activity model. Part 2. Internal resources optimization]. *Informatika i ee Primeneniya — Inform. Appl.* 6(1):18–25.
11. Bosov, A. V. 2012. Zadachi analiza i optimizatsii dlya modeli pol'zovatel'skoy aktivnosti. Chast' 3. Optimizatsiya vneshnikh resursov [Analysis and optimization problems for some users activity model. Part 3. External resources optimization]. *Informatika i ee Primeneniya — Inform. Appl.* 6(2):15–22.
12. Bertsekas, D. P., and S. E. Shreve. 1978. *Stochastic optimal control: The discrete-time case*. N.Y.: Academic Press. 323 p.
13. Albert, A. 1972. *Regression, and the Moore–Penrose pseudoinverse*. N.Y.: Academic Press. 179 p.
14. Kogan, M. M., and Ju. I. Nejmark. 1987. Adaptivnoe lokal'no-optimal'noe upravlenie [Locally-optimal adaptive control]. *Avtomatika i Telemehanika* [Automation and Remote Control] 8:126–136.

Received March 6, 2014

Contributor

Bosov Alexey V. (b. 1969) — Doctor of Science in technology, Head of Department, Institute of Informatics Problems, Russian Academy of Sciences, 44-2 Vavilov Str., Moscow 119333, Russian Federation; AVBosov@ipiran.ru

ИНФОРМАТИКА И ЕЁ ПРИМЕНЕНИЯ, 2014. Т. 8. Вып. 2. С. 48–54

БАЙЕСОВСКАЯ РЕКУРРЕНТНАЯ МОДЕЛЬ РОСТА НАДЕЖНОСТИ: БЕТА-РАСПРЕДЕЛЕНИЕ ПАРАМЕТРОВ*

Ю. В. Жаворонкова[1], А. А. Кудрявцев[2], С. Я. Шоргин[3]

Аннотация: Одной из актуальных задач современной прикладной математики является задача прогнозирования надежности сложных модифицируемых информационных систем. Любая впервые созданная сложная система, предназначенная для переработки или передачи информационных потоков, как правило, не обладает требуемой надежностью. Такие системы подвергаются модификациям в ходе разработки, опытной эксплуатации и штатного функционирования. Целью таких модификаций является увеличение надежности информационных систем. В связи с этим возникает необходимость формализации понятия надежности модифицируемых информационных систем и разработки методов и алгоритмов оценивания и прогнозирования различных надежностных характеристик. Одним из подходов к определению надежности системы является вычисление вероятности того, что на сигнал, поданный на вход системы в определенный момент времени, система отреагирует корректно. В статье рассматривается экспоненциальная рекуррентная модель роста надежности, в которой вероятность надежности системы представляется как линейная комбинация параметров «дефективности» и «эффективности» средства, исправляющего недостатки системы. Предполагается, что исследователь не имеет точных сведений об исследуемой системе, а лишь знаком с характеристиками класса, из которого берется данная система. В рамках байесовского подхода предполагается, что показатели «дефективности» и «эффективности» имеют бета-распределение. Вычисляется средняя предельная надежность системы. Приводятся численные результаты для модельных примеров.

Ключевые слова: модифицируемые информационные системы; теория надежности; байесовский подход; бета-распределение

DOI: 10.14375/19922264140205

1 Введение

Задача прогнозирования надежности сложных модифицируемых информационных систем была сформулирована в [1], а в дальнейшем более подробно рассмотрена в [2].

В статье [3] подробно описана байесовская рекуррентная модель роста надежности. Ниже соответствующая постановка задачи будет сформулирована вкратце.

Любой впервые созданный более или менее сложный агрегат, предназначенный для переработки или передачи информационных потоков, например новая программная система для компьютера, новая информационно-вычислительная сеть или новая административно-информационная система, как правило, не обладает требуемой надежностью. Для единства терминологии впредь будет говориться о сложных информационных системах. Такие системы подвергаются периодическим изменениям (модификациям), целью которых является увеличение надежности информационных систем. В связи с этим возникает необходимость формализации понятия надежности модифицируемых информационных систем и разработки методов и алгоритмов оценивания и прогнозирования различных надежностных характеристик.

В книге [2] приводятся общие соображения, являющиеся основой для построения математических моделей, описывающих изменение надежности модифицируемых информационных систем, а затем обсуждаются конкретные аналитические модели роста (изменения) надежности модифицируемых информационных систем.

К числу таких моделей относятся рекуррентные модели роста надежности. Они могут использоваться в случае, когда удобно иметь дело непосредственно с параметром, интерпретируемым как надежность системы.

Рассмотрим произвольную систему, на вход которой подаются некоторые сигналы (например, ко-

* Работа выполнена при поддержке РФФИ (проекты 12-07-00109-а и 12-07-00115-а).

[1] ООО КМ Медиа, juliana-zh@yandex.ru

[2] Московский государственный университет им. М. В. Ломоносова, факультет вычислительной математики и кибернетики, nubigena@hotmail.com

[3] Институт проблем информатики Российской академии наук, sshorgin@ipiran.ru

манды оператора или внешние воздействия). Реакция системы на поданные сигналы может быть либо правильной (корректной), либо неправильной (некорректной). В каждый момент времени t надежность системы можно характеризовать параметром $p(t)$ — вероятностью того, что на сигнал, поданный на вход системы в момент t, система отреагирует правильно. По смыслу такая характеристика надежности ближе всего к традиционно используемому коэффициенту готовности. В случайные моменты времени $0 = Y_0 \le Y_1 \le Y_2 \le \ldots$ система подвергается (мгновенной) модификации, в результате чего изменяется параметр $p(t)$. Следует обратить внимание на то обстоятельство, что ниже рассматривается непрерывное время, без привязки напрямую процесса модифицирования системы к процессу ее тестирования. Предположим, что траектории процесса $p(t)$ непрерывны справа и кусочно-постоянны, так что $p(t) = p(Y_j)$ при $Y_j \le t < Y_{j+1}$.

Задача прогнозирования поведения процесса $p(t)$ чрезвычайно важна. Например, в программировании параметр $p(t)$ можно рассматривать как надежность программного обеспечения, в которое по ходу отладки в моменты $0 = Y_0 \le Y_1 \le Y_2 \le \ldots$ вносятся изменения для исправления замеченных ошибок. Оценивание $p(t)$ и прогнозирование поведения этого параметра здесь важно как для оценивания надежности всего комплекса, составной частью которого является программное обеспечение, так и для прогнозирования продолжительности отладки.

Обозначим $p_j = p(Y_j)$. Рассмотрим поведение p_j в зависимости от изменения j. Другими словами, будем изучать изменение надежности системы в зависимости от номера модификации. В книге [2] рассматривается, в частности, следующая рекуррентная модель роста надежности. Пусть $\{(\theta_j, \eta_j)\}$, $j \ge 1$, — последовательность независимых одинаково распределенных двумерных случайных векторов таких, что $0 < \eta_1 < 1$; $0 < \theta_1 < 1$ почти наверное.

Задав начальную надежность p_0, рассмотрим модель, определяемую рекуррентным соотношением

$$p_{j+1} = \eta_{j+1} p_j + \theta_{j+1}\,(1 - p_j)\ .$$

Эта модель названа дискретной экспоненциальной моделью. В такой модели случайные величины η_j описывают возможное уменьшение надежности из-за некачественных модификаций, в ходе которых вместо исправления существующих дефектов в систему могут быть внесены новые, в то время как величины θ_j описывают повышение надежности за счет исправления дефектов.

Обозначим $\lambda = 1 - \mathsf{E}\theta_1$, $\mu = \mathsf{E}\eta_1$. В [2] доказано, что при условии $\lambda + \mu \ne 1$

$$p = \lim_{j\to\infty} \mathsf{E}p_j = \frac{\mu}{\lambda + \mu}\,.$$

2 Постановка задачи

Изучение предельного значения средней величины $\mathsf{E}p_j$ представляет значительный интерес, поскольку эта величина характеризует асимптотическое значение надежности системы в рамках некоторой рекуррентной модели, задаваемой набором $\{(\theta_j, \eta_j)\}$. Из результатов [2] следует, что это асимптотическое значение зависит только от средних значений величин $\{(\theta_j, \eta_j)\}$, $j \ge 1$.

В [3] исследовалась ситуация, при которой рассматривается целый набор однотипных сложных модифицируемых объектов (МО), каждый из которых обслуживается собственной ремонтной бригадой (РБ). Исследователю хотелось бы определить усредненное значение p по всем МО. Для решения этой задачи в указанной работе предложена так называемая байесовская постановка. Предполагается, что рассматривается целая группа однотипных МО и группа им соответствующих однотипных РБ. Пусть $m = 1, 2, \ldots$ — номера этих объектов. Для каждого МО (вместе с его РБ) существует собственный набор $\{(\theta_j^m, \eta_j^m)\}$ ($j \ge 1$, $m \ge 1$) независимых одинаково распределенных при каждом фиксированном j двумерных случайных векторов таких, что $0 < \eta_1^m < 1$; $0 < \theta_1^m < 1$ почти наверное. Но средние значения величин θ_j^m, η_j^m, $j \ge 1$, $m \ge \ge 1$, не предполагаются известными; более того, они не предполагаются даже одинаковыми. Вводится предположение, что величины $\lambda = 1 - \mathsf{E}\theta_j^m$, $\mu = \mathsf{E}\eta_j^m$ сами по себе являются случайными, т. е. на вероятностном пространстве, в которое в качестве элементарных событий входят все рассматриваемые в рамках данной постановки МО вместе с их РБ, заданы случайные величины λ и μ (которые полагаем независимыми), имеющие смысл $\lambda = 1 - \mathsf{E}\theta_j^m$, $\mu = \mathsf{E}\eta_j^m$, где m — случайный номер МО.

Принимаемые исследователем за основу распределения величин λ и μ будем называть априорными. При этом подлежащие вычислению характеристики такой «рандомизированной» группы МО, естественно, являются рандомизацией аналогичных характеристик «отдельно взятой» МО с учетом априорного распределения параметров λ и μ, взятого исследователем за основу. Наиболее естественной и удобной для изучения характеристикой является усредненное по всем МО значение предельной вероятности надежности, т. е.

$$p_{\text{сред}} = \mathsf{E}p = \mathsf{E}\frac{\mu}{\lambda+\mu}\,,$$

где усреднение ведется по совместному распределению случайных величин (λ, μ).

В рассматриваемой ситуации величины η_1^m и θ_1^m удовлетворяют ограничениям $0 < \eta_1^m < 1, 0 < \theta_1^m < 1$. Значит, и средние значения λ и μ величин $1 - \mathsf{E}\theta_j$ и $\mathsf{E}\eta_j$ также находятся на отрезке $[0, 1]$. Поэтому в качестве априорных распределений параметров λ и μ следует выбирать только распределения, сосредоточенные на $[0, 1]$.

В работе [3] были рассмотрены независимые случайные параметры λ и μ, распределенные равномерно на некоторых (вообще говоря, разных) отрезках, являющихся подмножествами отрезка $[0, 1]$. В настоящей статье исследования байесовской рекуррентной модели роста надежности продолжены для ситуации, когда оба параметра имеют бета-распределение.

3 Основные результаты

Пусть λ и μ имеют соответственно бета-распределения $\beta(m, n)$ и $\beta(k, l)$, $m, n, k, l > 0$. Введем следующие обозначения.

Через $B(m, n)$, $m, n > 0$, будем обозначать бета-функцию. Пусть

$$(\alpha)_i = \alpha\,(\alpha+1)\cdots(\alpha+i-1)\;,\;\; (\alpha)_0 = 1\,.$$

Несмотря на то что $(\alpha)_i$ имеет смысл неполного факториала, нигде далее не требуется, чтобы α было положительным. Рассмотрим классическую гипергеометрическую функцию Гаусса

$$G(\alpha, \beta, \gamma; x) = \sum_{i=0}^{\infty} \frac{(\alpha)_i (\beta)_i}{(\gamma)_i\, i!}\, x^i\,.$$

По аналогии с 9.180.1 и 9.14 п. 1 из [4] введем в рассмотрение обобщенную гипергеометрическую функцию двух переменных

$$G_{s,t}^{p,q}(\alpha, \beta_1, \ldots, \beta_p, \beta_1', \ldots, \beta_q';\; \gamma, \delta_1, \ldots, \delta_s, \delta_1', \ldots, \delta_t'; x, y) = \sum_{i=0}^{\infty}\sum_{j=0}^{\infty} \frac{(\alpha)_{i+j}(\beta_1)_i \cdots (\beta_p)_i (\beta_1')_j \cdots (\beta_q')_j}{(\gamma)_{i+j}(\delta_1)_i \cdots (\delta_s)_i (\delta_1')_j \cdots (\delta_t')_j}\, \frac{x^i y^j}{i!\,j!}.$$

В дальнейшем изложении будет интересен частный случай последней функции

$$G_{1,0}^{2,1}(\alpha, \beta_1, \beta_2, \beta_1'; \gamma, \delta_1; x, y) = \sum_{i=0}^{\infty}\sum_{j=0}^{\infty} \frac{(\alpha)_{i+j}(\beta_1)_i (\beta_2)_i (\beta_1')_j}{(\gamma)_{i+j}(\delta_1)_i}\, \frac{x^i y^j}{i!\,j!}\,. \quad (1)$$

Теорема. *Пусть случайные величины λ и μ независимы и имеют соответственно распределения $\beta(m, n)$ и $\beta(k, l)$, $m, n, k, l > 0$. Тогда*

$$p_{\text{сред}} = \frac{B(k+m, n)}{B(k, l)B(m, n)(k+1)} \times G_{1,0}^{2,1}(k+1, 1-l, k+m, 1; k+2, k+m+n; -1, 1) + \frac{B(k+m, l)}{B(k, l)B(m, n)m} G_{1,0}^{2,1}(m, 1-n, k+m, 1; m+1, k+m+l; -1, 1)\,. \quad (2)$$

Доказательство. Найдем плотность $f_p(x)$ случайной величины p. Имеем

$$f_p(x) = \int_0^1 \frac{y}{(1-x)^2}\, f_\mu\left(\frac{x}{1-x}\, y\right) f_\lambda(y)\, dy\,.$$

Положим $f_p(1/2) = 0$. Согласно формуле 3.197.3 из [4] получаем при $0 < x < 1/2$

$$f_p(x) = \int_0^1 \frac{y}{(1-x)^2} \left(\frac{x}{1-x}\right)^{k-1} \times \frac{y^{k-1}}{B(k, l)} \left(1 - \frac{x}{1-x}\, y\right)^{l-1} \frac{y^{m-1}(1-y)^{n-1}}{B(m, n)}\, dy = \frac{B(k+m, n)}{B(k, l)B(m, n)} \frac{x^{k-1}}{(1-x)^{k+1}} \times G\left(1-l, k+m, k+m+n; \frac{x}{1-x}\right) \equiv S_1(x)\,.$$

Аналогично при $1/2 < x < 1$, используя замену переменной $z = xy/(1-x)$, получаем:

$$f_p(x) = \frac{B(k+m, l)}{B(k, l)B(m, n)} \frac{(1-x)^{m-1}}{x^{m+1}} \times G\left(1-n, k+m, k+m+l; \frac{1-x}{x}\right) \equiv S_2(x)\,.$$

Таким образом,

$$\mathsf{E}p = \int x f_p(x)\, dx = \int_0^{1/2} x S_1(x) dx + \int_{1/2}^1 x S_2(x)\, dx. \quad (3)$$

Вычислим отдельно первый интеграл из правой части (3). Имеем:

$$\int_0^{1/2} xS_1(x)\,dx = \frac{B(k+m,n)}{B(k,l)B(m,n)} \times$$
$$\times \sum_{i=0}^{\infty} \frac{(1-l)_i(k+m)_i}{(k+m+n)_i\, i!} \int_0^{1/2} \frac{x^{k+i}dx}{(1-x)^{k+i+1}} =$$
$$= \frac{B(k+m,n)}{B(k,l)B(m,n)} \times$$
$$\times \sum_{i=0}^{\infty} \frac{(1-l)_i(k+m)_i}{(k+m+n)_i\, i!} \int_0^{1/2} \frac{(1/2-t)^{k+i}dt}{(1/2+t)^{k+i+1}}.$$

Воспользовавшись формулой 3.196.1 из [4], получим:

$$\int_0^{1/2} xS_1(x)\,dx = \frac{B(k+m,n)}{B(k,l)B(m,n)} \times$$
$$\times \sum_{i=0}^{\infty} \frac{(1-l)_i(k+m)_i}{(k+m+n)_i\, i!} \times$$
$$\times \frac{G(1,k+i+1,k+i+2;-1)}{k+i+1} =$$
$$= \frac{B(k+m,n)}{B(k,l)B(m,n)} \times$$
$$\times \sum_{i=0}^{\infty} \frac{(1-l)_i(k+m)_i}{(k+m+n)_i\, i!} \sum_{j=0}^{\infty} \frac{(-1)^j}{k+i+j+1}. \quad (4)$$

Аналогично (4) получаем:

$$\int_{1/2}^{1} xS_2(x)\,dx = \frac{B(k+m,l)}{B(k,l)B(m,n)} \times$$
$$\times \sum_{i=0}^{\infty} \frac{(1-n)_i(k+m)_i}{(k+m+n)_i\, i!} \sum_{j=0}^{\infty} \frac{(-1)^j}{m+i+j}. \quad (5)$$

Заметим, что соотношения (4) и (5) можно преобразовать, что дает возможность получить из (3) для $p_{\text{сред}}$ следующее выражение:

$$p_{\text{сред}} = \frac{B(k+m,n)}{B(k,l)B(m,n)} \times$$
$$\times \sum_{i=0}^{\infty}\sum_{j=0}^{\infty} \frac{(1-l)_i(k+m)_i(k+1)_{i+j}(1)_j(-1)^j 1^i}{(k+m+n)_i(k+2)_{i+j}(k+1)i!j!} +$$
$$+ \frac{B(k+m,l)}{B(k,l)B(m,n)} \times$$
$$\times \sum_{i=0}^{\infty}\sum_{j=0}^{\infty} \frac{(1-n)_i(k+m)_i(m)_{i+j}(1)_j(-1)^j 1^i}{(k+m+l)_i(m+1)_{i+j}\, m\, i!j!}.$$

Из последнего соотношения по определению (1) получаем (2), что завершает доказательство теоремы.

Замечание. Выражение (2) служит для компактной записи $p_{\text{сред}}$. Для практического использования (непосредственного вычисления) имеет смысл представлять $p_{\text{сред}}$ в виде рядов типа (4) и (5), которые несложно вычисляются с любой наперед заданной точностью.

В качестве иллюстрации приведем несколько таблиц со значениями $p_{\text{сред}}$. Для удобства изложения будем использовать индексы, соответствующие номерам таблицы, для случайных величин λ и μ.

Таблица 1 была опубликована в [3] для случая, когда λ_1 и μ_1 имеют равномерное распределение на отрезках $[a_\lambda, b_\lambda]$ и $[a_\mu, b_\mu]$ соответственно.

В табл. 2 приведены значения $p_{\text{сред}}$ для случая, когда λ_2 и μ_2 имеют бета-распределение с параметрами $(m;n)$ и $(k;l)$ соответственно, причем имеют место соотношения $\mathsf{E}\lambda_1 = \mathsf{E}\lambda_2 = m/(m+n)$ и $\mathsf{E}\mu_1 = \mathsf{E}\mu_2 = k/(k+l)$.

В табл. 3 также выполняются соотношения $\mathsf{E}\lambda_1 = \mathsf{E}\lambda_3$ и $\mathsf{E}\mu_1 = \mathsf{E}\mu_3$.

Для табл. 4 и 5 имеет место $\mathsf{E}\lambda_4 = \mathsf{E}\lambda_5$ и $\mathsf{E}\mu_4 = \mathsf{E}\mu_5$.

Таблица 1 Частные значения средней надежности (равномерный случай)

$[a_\lambda, b_\lambda]$	$[a_\mu, b_\mu]$									
	$[0,1/4]$	$[0,1/2]$	$[0,3/4]$	$[0,1]$	$[1/4,1/2]$	$[1/4,3/4]$	$[1/4,1]$	$[1/2,3/4]$	$[1/2,1]$	$[3/4,1]$
$[0,1/4]$	0,50	0,63	0,70	0,74	0,76	0,80	0,83	0,84	0,86	0,88
$[0,1/2]$	0,37	0,50	0,58	0,63	0,63	0,68	0,72	0,73	0,76	0,79
$[0,3/4]$	0,30	0,42	0,50	0,56	0,55	0,60	0,64	0,66	0,69	0,72
$[0,1]$	0,25	0,37	0,44	0,50	0,48	0,54	0,58	0,60	0,63	0,67
$[1/4,1/2]$	0,24	0,37	0,45	0,52	0,50	0,56	0,61	0,63	0,66	0,70
$[1/4,3/4]$	0,20	0,32	0,40	0,46	0,44	0,50	0,55	0,56	0,60	0,64
$[1/4,1]$	0,17	0,28	0,36	0,42	0,39	0,46	0,50	0,51	0,55	0,60
$[1/2,3/4]$	0,16	0,27	0,34	0,40	0,37	0,44	0,49	0,50	0,54	0,58
$[1/2,1]$	0,14	0,24	0,31	0,37	0,34	0,40	0,45	0,46	0,50	0,54
$[3/4,1]$	0,12	0,21	0,28	0,33	0,30	0,36	0,40	0,42	0,46	0,50

Таблица 2 Частные значения средней надежности (случай бета-распределения)

$m;n$	$k;l$									
	1; 7	1; 3	3; 5	1; 1	6; 10	2; 2	10; 6	5; 3	3; 1	7; 1
1; 7	0,47	0,64	0,74	0,80	0,75	0,80	0,84	0,84	0,87	0,88
1; 3	0,35	0,50	0,59	0,66	0,60	0,66	0,71	0,72	0,76	0,79
3; 5	0,22	0,35	0,44	0,51	0,44	0,52	0,59	0,59	0,65	0,70
1; 1	0,23	0,35	0,43	0,49	0,43	0,49	0,53	0,53	0,56	0,60
6; 10	0,21	0,33	0,42	0,50	0,42	0,50	0,58	0,58	0,65	0,70
2; 2	0,20	0,31	0,39	0,45	0,40	0,45	0,50	0,50	0,55	0,60
10; 6	0,15	0,26	0,32	0,35	0,33	0,35	0,37	0,37	0,43	0,52
5; 3	0,15	0,26	0,32	0,36	0,33	0,37	0,40	0,39	0,44	0,51
3; 1	0,14	0,24	0,31	0,35	0,30	0,36	0,38	0,38	0,38	0,40
7; 1	0,12	0,21	0,28	0,33	0,26	0,34	0,35	0,35	0,33	0,29

Таблица 3 Частные значения средней надежности (случай бета-распределения)

$m;n$	$k;l$									
	1; 7	1,1; 3,3	1,2; 2,0	1,3; 1,3	1,4; 2,33	1,5; 1,5	1,6; 0,96	1,7; 1,02	1,8; 0,6	1,9; 0,27
1; 7	0,47	0,61	0,70	0,76	0,71	0,77	0,81	0,82	0,85	0,88
1,1; 3,3	0,34	0,47	0,56	0,63	0,57	0,64	0,69	0,69	0,74	0,78
1,2; 2,0	0,26	0,38	0,46	0,53	0,47	0,53	0,59	0,59	0,64	0,69
1,3; 1,3	0,21	0,32	0,39	0,45	0,40	0,45	0,51	0,51	0,56	0,60
1,4; 2,33	0,25	0,37	0,45	0,52	0,46	0,53	0,59	0,59	0,64	0,69
1,5; 1,5	0,21	0,31	0,38	0,45	0,39	0,45	0,50	0,50	0,55	0,60
1,6; 0,96	0,17	0,26	0,33	0,38	0,34	0,39	0,43	0,43	0,47	0,51
1,7; 1,02	0,17	0,26	0,33	0,38	0,33	0,38	0,43	0,43	0,47	0,51
1,8; 0,6	0,14	0,23	0,29	0,33	0,30	0,34	0,37	0,37	0,38	0,39
1,9; 0,27	0,12	0,20	0,26	0,30	0,27	0,31	0,32	0,32	0,30	0,26

Таблица 4 Частные значения средней надежности (случай бета-распределения)

$m;n$	$k;l$									
	1; 10	2; 9	3; 8	4; 7	5; 6	6; 5	7; 4	8; 3	9; 2	10; 1
1; 10	0,47	0,64	0,74	0,79	0,83	0,86	0,88	0,89	0,91	0,91
2; 9	0,30	0,46	0,56	0,64	0,70	0,74	0,78	0,80	0,82	0,84
3; 8	0,23	0,36	0,45	0,52	0,58	0,64	0,69	0,73	0,75	0,78
4; 7	0,18	0,30	0,37	0,43	0,49	0,55	0,60	0,65	0,69	0,72
5; 6	0,16	0,26	0,33	0,38	0,42	0,46	0,52	0,57	0,62	0,66
6; 5	0,13	0,23	0,30	0,34	0,37	0,40	0,44	0,49	0,55	0,60
7; 4	0,12	0,21	0,28	0,32	0,35	0,36	0,38	0,42	0,47	0,53
8; 3	0,11	0,19	0,26	0,31	0,33	0,35	0,35	0,36	0,39	0,44
9; 2	0,09	0,17	0,24	0,29	0,33	0,34	0,34	0,33	0,32	0,34
10; 1	0,09	0,16	0,22	0,28	0,32	0,35	0,35	0,34	0,29	0,25

Таблица 5 Частные значения средней надежности (случай бета-распределения)

$m;n$	$k;l$									
	0,1; 1,0	0,2; 0,9	0,3; 0,8	0,4; 0,7	0,5; 0,6	0,6; 0,5	0,7; 0,4	0,8; 0,3	0,9; 0,2	1,0; 0,1
0,1; 1,0	0,50	0,56	0,63	0,69	0,74	0,79	0,83	0,86	0,88	0,90
0,2; 0,9	0,44	0,49	0,55	0,60	0,65	0,69	0,74	0,77	0,80	0,83
0,3; 0,8	0,39	0,43	0,48	0,53	0,58	0,62	0,66	0,70	0,74	0,77
0,4; 0,7	0,35	0,39	0,43	0,48	0,52	0,56	0,60	0,64	0,67	0,71
0,5; 0,6	0,30	0,34	0,38	0,42	0,46	0,50	0,54	0,57	0,61	0,65
0,6; 0,5	0,26	0,30	0,34	0,38	0,41	0,44	0,47	0,51	0,54	0,57
0,7; 0,4	0,22	0,26	0,30	0,33	0,37	0,39	0,41	0,44	0,46	0,48
0,8; 0,3	0,18	0,22	0,26	0,29	0,32	0,34	0,36	0,37	0,38	0,38
0,9; 0,2	0,14	0,18	0,22	0,25	0,28	0,29	0,30	0,30	0,29	0,28
1,0; 0,1	0,10	0,14	0,18	0,22	0,24	0,25	0,25	0,24	0,21	0,18

4 Заключение

Полученные результаты могут применяться, например, для вычисления других моментов и построения доверительных интервалов для характеристики p.

В дальнейшем предполагается рассмотреть другие модели, в частности ситуации, когда один из параметров (λ, μ) распределен равномерно, а другой имеет бета-распределение. Предполагается разработать расчетные алгоритмы для вычисления величины $p_{\text{сред}}$, соответствующие программные модели и провести тестовые расчеты.

Литература

1. *Gnedenko B. V., Korolev V. Yu.* Random summation: Limit theorems and applications. — Boca Raton, FL: CRC Press, 1996. 288 p.
2. *Королев В. Ю., Соколов И. А.* Основы математической теории надежности модифицируемых систем. — М.: ИПИ РАН, 2006. 108 с.
3. *Кудрявцев А. А., Соколов И. А., Шоргин С. Я.* Байесовская рекуррентная модель роста надежности: равномерное распределение параметров // Информатика и её применения, 2013. Т. 7. Вып. 2. С. 55–59.
4. *Градштейн И. С., Рыжик И. М.* Таблицы интегралов, сумм, рядов и произведений. — М.: Наука, 1971. 1108 с.

Поступила в редакцию 19.11.13

BAYESIAN RECURRENT MODEL OF RELIABILITY GROWTH: BETA-DISTRIBUTION OF PARAMETERS

Iu. V. Zhavoronkova[1], A. A. Kudryavtsev[2], and S. Ya. Shorgin[3]

[1]KM Media Company, 8/2 Prishvina Str., Moscow 127549, Russian Federation

[2]Department of Mathematical Statistics, Faculty of Computational Mathematics and Cybernetics, M. V. Lomonosov Moscow State University, 1-52 Leninskiye Gory, GSP-1, Moscow 119991, Russian Federation

[3]Institute of Informatics Problems, Russian Academy of Sciences, 44-2 Vavilov Str., Moscow 119333, Russian Federation

Abstract: One of the topical problems of modern applied mathematics is the task of forecasting reliability of modifiable complex information systems. Any first established complex system designed for processing or transmission information flows, as a rule, does not possess the required reliability. Such systems are subject to modifications during development, testing, and regular functioning. The purpose of such modifications is to increase reliability of information systems. In this connection, it is necessary to formalize the concept of reliability of modifiable information systems and to develop methods and algorithms of estimating and forecasting various reliability characteristics. One approach to determine system reliability is to compute the probability that a signal fed to the input of a system at a given point of time will be processed correctly by the system. The article considers the exponential recurrent growth model of reliability, in which the probability of system reliability is represented as a linear combination of the "defectiveness" and "efficiency" parameters of tools correcting deficiencies in a system. It is assumed that the researcher does not have exact information about the system under study and is only familiar with the characteristics of the class from which this system is taken. In the framework of the Bayesian approach, it is assumed that the indicators of "defectiveness" and "efficiency" have beta-distribution. Average marginal system reliability is calculated. Numerical results for model examples are obtained.

Keywords: modifiable information systems; theory of reliability; Bayesian approach; beta-distribution

DOI: 10.14375/19922264140205

Acknowledgments

The work was supported by the Russian Foundation for Basic Research (projects 12-07-00109-a and 12-07-00115-a).

References

1. Gnedenko, B. V., and V. Yu. Korolev. 1996. *Random summation: Limit theorems and applications*. Boca Raton, FL: CRC Press, 1996. 288 p.
2. Korolev, V. Yu., and I. A. Sokolov. 2006. *Osnovy matematicheskoy teorii nadezhnosti modifitsiruemykh system* [Fundamentals of mathematical theory of modified systems reliability]. Moscow.: IPI RAN, 2006. 108 p.
3. Kudryavtsev, A. A., I. A. Sokolov, and S. Ya. Shorgin. 2013. Bayesovskaya rekurrentnaya model' rosta nadezhnosti: Ravnomernoe raspredelenie parametrov [Bayesian recurrent model of reliability growth: Uniform distribution of parameters]. *Informatika i ee Primeneniya — Inform. Appl.* 7(2):55–59.
4. Gradshteyn, I. S., and I. M. Ryzhik. 1971. *Tablitsy integralov, summ, ryadov i proizvedeniy* [Tables of integrals, sums, series, and products]. Moscow: Nauka, 1971. 1108 p.

Received November 19, 2013

Contributors

Zhavoronkova Iuliia V. (b. 1990) — Software Developer, KM Media Company, 8/2 Prishvina Str., Moscow 127549, Russian Federation; juliana-zh@yandex.ru

Kudryavtsev Alexey A. (b. 1978) — Candidate of Science (PhD) in physics and mathematics, associate professor, Department of Mathematical Statistics, Faculty of Computational Mathematics and Cybernetics, M. V. Lomonosov Moscow State University, 1-52 Leninskiye Gory, GSP-1, Moscow 119991, Russian Federation; nubigena@hotmail.com

Shorgin Sergey Ya. (b. 1952) — Doctor of Science in physics and mathematics, professor, Deputy Director, Institute of Informatics Problems, Russian Academy of Sciences, 44-2 Vavilov Str., Moscow 119333, Russian Federation; sshorgin@ipiran.ru

ИНФОРМАТИКА И ЕЁ ПРИМЕНЕНИЯ, 2014. Т. 8. Вып. 2. С. 55–69

МЕТОД ДОКАЗАТЕЛЬСТВА НАБЛЮДАЕМОЙ ЭКВИВАЛЕНТНОСТИ ПРОЦЕССОВ С ПЕРЕДАЧЕЙ СООБЩЕНИЙ

А. М. Миронов[1]

Аннотация: Рассматривается проблема доказательства наблюдаемой эквивалентности для класса вычислительных процессов, называемых процессами с передачей сообщений. Действия, выполняемые такими процессами, заключаются в посылке или приеме сообщений, проверке логических условий и обновлении значений внутренних переменных процессов. Основным результатом статьи является теорема, сводящая задачу доказательства наблюдаемой эквивалентности пары процессов с передачей сообщений к задаче нахождения формул, ассоциированных с парами состояний этих процессов, удовлетворяющих некоторым условиям, которые связаны с переходами этих процессов. Данное сведение является обобщением известного метода Флойда верификации блок-схем, в котором задача верификации блок-схемы сводится к задаче нахождения формул (называемых промежуточными утверждениями), связанных с некоторыми точками в блок-схемах и удовлетворяющих некоторым условиям, соответствующим переходам в блок-схемах. Изложенный метод доказательства наблюдаемой эквивалентности процессов с передачей сообщений иллюстрируется примером верификации протокола скользящего окна.

Ключевые слова: верификация; процессы с передачей сообщений; наблюдаемая эквивалентность; протокол скользящего окна

DOI: 10.14375/19922264140206

1 Введение

1.1 Понятие наблюдаемой эквивалентности процессов

Понятие наблюдаемой эквивалентности процессов было сформулировано Р. Милнером и подробно изучено им в основополагающей работе [1]. Данное понятие зарекомендовало себя как одно из наиболее эффективных средств для решения задач спецификации и верификации различных свойств вычислительных процессов.

В своей исходной формулировке понятие наблюдаемой эквивалентности относится к процессам, определяемым в терминах исчисления взаимодействующих систем Милнера [1]. Одним из наиболее важных достоинств понятия наблюдаемой эквивалентности является возможность сформулировать аналог данного понятия (и разработать основанные на нем методы спецификации и верификации) в самых различных классах вычислительных процессов, в том числе в классах вероятностных процессов [2], процессов реального времени [3] и др.

В настоящей работе рассматривается один из классов вычислительных процессов, называемых процессами с передачей сообщений. Определяется понятие наблюдаемой эквивалентности таких процессов, формулируется достаточное условие наблюдаемой эквивалентности процессов с передачей сообщений. Применение данного понятия иллюстрируется примером спецификации и верификации протокола скользящего окна.

1.2 Мотивация предлагаемого метода

Предлагаемый в настоящей работе метод доказательства наблюдаемой эквивалентности процессов с передачей сообщений предназначен для решения различных проблем, связанных со спецификацией и верификацией вычислительных систем. Проблемы спецификации и верификации систем образуют одно из наиболее актуальных направлений в теоретической информатике. Существует несколько подходов, относящихся к этому направлению, наиболее важными из них являются: исчисление взаимодействующих систем Р. Милнера (CCS — calculus of communicating systems) и π-исчисление [1, 4], теория взаимодействующих процессов А. Хоара (CSP — communicating sequential processes) и ее обобщения [5], темпоральная логика и model checking [6], сети Петри [7], процессные алгебры [8], теория взаимодействующих машин с конечным числом состояний [9]. Основным недостатком современных методов верификации дискретных процессов является их высокая сложность. В частности,

– основным недостатком методов верификации, основанных на подходе model checking, явля-

[1] Институт проблем информатики Российской академии наук, amironov66@gmail.com

ется высокая вычислительная сложность, связанная с проблемой «взрыва числа состояний» (state explosion problem);

– недостатки методов, основанных на доказательстве теорем, связаны с высокой сложностью построения соответствующих теорем и их доказательств, а также со сложностью понимания этих доказательств.

Мотивация предлагаемого подхода к моделированию и верификации дискретных систем заключается в желании упростить и сделать более явными следующие аспекты моделирования и анализа дискретных систем: представление математических моделей анализируемых систем, построение доказательств корректности этих систем и понимание этих доказательств любым, кто не является экспертом в математической теории верификации дискретных систем.

Рассматриваемая в настоящей работе модель дискретных процессов представляет собой синтез Милнеровской модели процессов [1] и модели взаимодействующих машин с конечным числом состояний [9]. Дискретные процессы представляются в данной модели в виде графов, ребра которых помечены операторами. Эти операторы состоят из внутренних действий и действий ввода-вывода. Доказательства корректности процессов представляются множествами формул, связанных с парами состояний анализируемых процессов. Этот метод верификации процессов является синтезом метода Милнера, основанного на понятии наблюдаемой эквивалентности [1], и метода индуктивных утверждений Флойда [10]. Для упрощения анализа процессов вводятся операции упрощения на процессах.

1.3 Преимущества предлагаемого подхода

Предлагаемая математическая модель процессов с передачей сообщений позволяет строить такие математические модели анализируемых систем, которые во многом похожи на исходные описания этих систем на каком-либо императивном языке программирования. В разд. 7 будет приведен пример такой модели, которая соответствует программе на языке C, описывающей функционирование протокола скользящего окна с возвратом на n (эта программа была взята из [11, п. 3.4.2].

Основные преимущества предложенного подхода заключаются в возможности использования операций упрощения моделей анализируемых систем, которые позволяют облегчить решение проблемы верификации этих моделей. Следует отметить, что эти упрощенные модели позволяют более ясно понимать основные особенности анализируемых систем и облегчают построение доказательств корректности анализируемых систем. В разд. 7 будет приведен пример такого упрощения модели для протокола скользящего окна: эта модель может быть упрощена до модели с одним состоянием.

Если анализируемое свойство какой-либо системы имеет вид поведения, которое описывается в виде некоторого процесса, например в том случае, когда анализируемая система представляет собой телекоммуникационный протокол и свойство этой системы представляет собой описание внешнего поведения этого протокола (связанного с его взаимодействием с протоколом вышестоящего уровня), то доказательством корректности такой системы в этой модели является множеством формул, ассоциированных с парами состояний, первое из которых является состоянием анализируемой системы, а второе – состоянием процесса, который описывает свойство этой системы. В разд. 7 будет приведен пример доказательства такого типа, которое представляет собой небольшое множество простых формул. Эти формулы могут быть естественным образом извлечены из упрощенной модели анализируемого протокола.

Другое преимущество предлагаемого подхода заключается в возможности верифицировать на его основе системы с неограниченными множествами состояний.

1.4 Сравнение с другими работами

В этом пункте излагается обзор работ, относящихся к проблеме верификации систем с передачей сообщений, которые наиболее тесно связаны с настоящей статьей.

В статье [12] рассматривается моделирование и верификация процессов, описание которых производится на алгебраическом языке μCRL. Для анализа таких процессов используется система доказательства теорем PVS (Prototype Verification System). В частности, рассматривается пример автоматической верификации протокола передачи сообщений с выборочным повтором. Отметим, что этот протокол может быть верифицирован существенно проще (по сравнению с тем рассуждением, обосновывающим корректность данного протокола, которое приведено в работе [12]) при помощи подхода, излагаемого в настоящей работе.

Существует большое количество работ, относящихся к верификации систем с передачей сообщений, основанных на темпоральной логике и методе верификации model checking. Наиболее близкое отношение к материалу, излагаемому в настоящей статье, имеют работы [13–19]. Главным

недостатком всех этих методов является их ограниченная возможность: они позволяют верифицировать только системы с конечным числом состояний.

Среди других подходов следует отметить те, которые основаны на использовании логики первого порядка и верификации путем построения вспомогательных утверждений [20, 21], а также подходы, связанные с использованием процессной алгебры [22–24].

Наиболее существенным недостатком всех этих подходов является высокая сложность построения доказательств корректности анализируемых систем.

2 Вспомогательные понятия

2.1 Термы

Будем считать, что заданы множество $\mathcal{X}$ **переменных**, множество $\mathcal{D}$ **значений**, множество $\mathcal{C}$ **констант** и множество $\mathcal{F}$ **функциональных символов**. Каждая константа из $\mathcal{C}$ интерпретирована некоторым значением из $\mathcal{D}$ и каждый функциональный символ — некоторой операцией на $\mathcal{D}$. Будем считать, что $\mathcal{C}$ содержит константы 0 и 1, а $\mathcal{F}$ содержит булевы функциональные символы $\wedge, \vee, \rightarrow$, которым соответствуют стандартные булевы операции на $\{0, 1\}$.

Множество $\mathcal{E}$ **термов** определяется стандартным образом. Переменные и константы являются термами. Другие термы имеют вид $f(e_1, \ldots, e_n)$, где $f \in \mathcal{F}$ и $e_1, \ldots, e_n$ — термы. Для каждого $e \in \mathcal{E}$ запись X_e обозначает множество всех переменных, входящих в e.

Если $X \subseteq \mathcal{X}$, то **означиванием** переменных из X называется соответствие ξ, сопоставляющее каждой переменной $x \in X$ некоторое значение из $\mathcal{D}$, которое будем обозначать записью x^ξ. Множество всех означиваний переменных из X будем обозначать записью $X^\bullet$. Для каждого $e \in \mathcal{E}$, каждого $X \supseteq X_e$ и каждого $\xi \in X^\bullet$ запись e^ξ обозначает объект, называемый **значением** e на ξ и определяемый стандартным образом. Будем считать термы e_1 и e_2 равными, если $\forall\, \xi \in (X_{e_1} \cup X_{e_2})^\bullet \;\; e_1^\xi = e_2^\xi$.

Терм e называется **формулой**, если $\forall\, \xi \in X_e^\bullet$ значение e^ξ равно 0 или 1. Символ $\mathcal{B}$ обозначает множество всех формул. Символы $\top$ и $\bot$ обозначают истинную и ложную формулу соответственно. Формулы вида $\wedge(b_1, b_2)$, $\vee(b_1, b_2)$ и т. д. будем записывать в более привычном виде $b_1 \wedge b_2$, $b_1 \vee b_2$ и т. д. Формулы вида $b_1 \wedge \ldots \wedge b_n$ могут записываться также в виде $\left\{ \begin{array}{c} b_1 \\ \vdots \\ b_n \end{array} \right\}$.

2.2 Атомарные операторы

Будем считать, что задано множество $\mathcal{N}$, элементы которого рассматриваются как имена объектов, которые могут посылать или получать процессы.

Атомарный оператор (АО) — это конструкция o одного из трех перечисляемых ниже видов. Каждой паре (o, ξ), где o — АО и ξ — означивание переменных, входящих в o, соответствует некоторое действие o^ξ, неформально определяемое ниже.

1. **Ввод** — это АО вида $\alpha?x$, где $\alpha \in \mathcal{N}$ и $x \in \mathcal{X}$. Действие $(\alpha?x)^\xi$ заключается в получении от другого процесса объекта с именем α вместе с присоединенным к этому объекту сообщением, которое записывается в переменную x.
2. **Вывод** — это АО вида $\alpha!e$, где $\alpha \in \mathcal{N}$ и $e \in \mathcal{E}$. Действие $(\alpha!e)^\xi$ заключается в посылке другому процессу объекта с именем α, к которому присоединено сообщение e^ξ.
3. **Присваивание** — это АО вида $x := e$, где $x \in \in \mathcal{X}$, $e \in \mathcal{E}$. Действие $(x := e)^\xi$ заключается в присваивании переменной x значения e^ξ.

Ниже будем использовать следующие обозначения:

- для каждого АО o запись X_o обозначает множество всех переменных, содержащихся в o;
- если $e \in \mathcal{E}$ и o — присваивание, то запись $o(e)$ обозначает терм, определяемый следующим образом: пусть o имеет вид $(x := e')$, тогда $o(e)$ получается из e заменой всех вхождений в него переменной x на терм e';
- если o — присваивание и $\xi \in X^\bullet$, где $X_o \subseteq \subseteq X \subseteq \mathcal{X}$, то запись $\xi \cdot o$ обозначает означивание из $X^\bullet$, определяемое следующим образом: пусть $o = (x := e)$, тогда $x^{\xi \cdot o} = e^\xi$ и $\forall\, y \in \in X \setminus \{x\} \;\; y^{\xi \cdot o} = y^\xi$.

Нетрудно доказать, что если o — присваивание и $e \in \mathcal{E}$, то для каждого $\xi \in X^\bullet$, где $X_o \cup X_e \subseteq X \subseteq \mathcal{X}$, будет верно равенство $o(e)^\xi = e^{\xi \cdot o}$. Данное равенство доказывается индукцией по структуре терма e.

2.3 Операторы

Оператор — это запись O вида $b[o_1, \ldots, o_n]$, где b — формула, называемая **предусловием** оператора O (данная формула будет обозначаться записью $\langle O \rangle$), и $o_1, \ldots, o_n$ — последовательность АО (данная последовательность будет обозначаться записью $[O]$), среди которых присутствует не более одного ввода или вывода. Последовательность $[O]$ может быть пустой ($[\,]$).

Если в $[O]$ есть ввод (или вывод), то будем называть O **оператором ввода** (или **вывода**) и обозначать записью N_O имя, входящее в O. Если же в $[O]$ нет вводов и выводов, то будем называть O **внутренним** оператором.

Если $\langle O\rangle = \top$, то такое предусловие в записи оператора O можно опускать.

Ниже будем использовать следующие обозначения.

1. Для каждого оператора O запись X_O обозначает множество всех переменных, содержащихся в O.
2. Если O — оператор и $b \in \mathcal{B}$, то запись $O \cdot b$ обозначает объект, который либо является формулой, либо не определен. Данный объект определяется рекурсивно следующим образом. Если $[O]$ пуста, то $O \cdot b \stackrel{\text{def}}{=} \langle O\rangle \wedge b$. Если $[O] = o_1, \ldots, o_n$, где $n \geq 1$, то обозначим записью $O \setminus o_n$ оператор, получаемый из O удалением последнего АО, и
   - если $o_n = \alpha?x$, то
   $$O \cdot b \begin{cases} \text{не определен для } x \in X_b\,; \\ \stackrel{\text{def}}{=} (O \setminus o_n) \cdot b \text{ иначе;} \end{cases}$$
   - если $o_n = \alpha!e$, то $O \cdot b \stackrel{\text{def}}{=} (O \setminus o_n) \cdot b$;
   - если $o_n = (x := e)$, то $O \cdot b \stackrel{\text{def}}{=} (O \setminus o_n) \cdot o_n(b)$.
3. Если O — внутренний оператор и $\xi \in X^\bullet$, где $X_O \subseteq X \subseteq \mathcal{X}$, то запись $\xi \cdot O$ обозначает означивание из $X^\bullet$, определяемое следующим образом: если $[O]$ пуста, то $\xi \cdot O \stackrel{\text{def}}{=} \xi$, и если $[O] = o_1, \ldots, o_n$, где $n \geq 1$, то $\xi \cdot O \stackrel{\text{def}}{=} (\xi \cdot (O \setminus o_n)) \cdot o_n$.

Нетрудно доказать, что если оператор O — внутренний и $b \in \mathcal{B}$, то для каждого $\xi \in X^\bullet$, где $X_O \cup X_b \subseteq X \subseteq \mathcal{X}$, такого что $\langle O\rangle^\xi = 1$, будет верно равенство $(O \cdot b)^\xi = b^{\xi \cdot O}$. Данное равенство доказывается индукцией по длине $[O]$.

2.4 Конкатенация операторов

Пусть заданы операторы O_1 и O_2, причем хотя бы один из них внутренний.

Конкатенацией O_1 и O_2 называется объект, обозначаемый записью $O_1 \cdot O_2$, который либо является оператором, либо не определен. Данный объект определен в том и только в том случае, когда определена формула $O_1 \cdot \langle O_2\rangle$, и в этом случае

$$O_1 \cdot O_2 \stackrel{\text{def}}{=} (O_1 \cdot \langle O_2\rangle)\,[[O_1], [O_2]]\ .$$

Нетрудно доказать, что конкатенация операторов обладает следующими свойствами.

1. Если операторы O_1, O_2 и формула b таковы, что определены все объекты в обеих частях равенства
   $$(O_1 \cdot O_2) \cdot b = O_1 \cdot (O_2 \cdot b)\,,$$
   то данное равенство верно.
2. Если операторы O_1, O_2, O_3 таковы, что определены все объекты в обеих частях равенства
   $$(O_1 \cdot O_2) \cdot O_3 = O_1 \cdot (O_2 \cdot O_3)\,,$$
   то данное равенство верно.

3 Процессы с передачей сообщений

3.1 Понятие процесса с передачей сообщений

Процесс с передачей сообщений (называемый также просто **процессом**) — это четверка P вида

$$P = (S_P, s_P^0, T_P, I_P)\,, \tag{1}$$

компоненты которой имеют следующий смысл:

- S_P — множество **состояний** процесса P;
- $s_P^0 \in S_P$ — **начальное состояние** процесса P;
- T_P — множество **переходов** процесса P, каждый из которых имеет вид $s_1 \stackrel{O}{\to} s_2$, где $s_1, s_2 \in S_P$ и O — оператор;
- $I_P \in \mathcal{B} \setminus \{\bot\}$ — **предусловие** процесса P.

Будем называть переход $s_1 \stackrel{O}{\to} s_2$ **вводом**, **выводом** или **внутренним**, если O является оператором ввода, вывода или внутренним соответственно.

Для каждого процесса P

- запись X_P обозначает множество, состоящее из
  - всех переменных, входящих в какой-либо из переходов из T_P или в I_P, и
  - переменной at_P, не входящей в переходы и в I_P, значениями которой являются состояния из S_P
- запись $\langle P\rangle$ обозначает формулу $(\mathrm{at}_P = s_P^0) \wedge I_P$.

Для каждого перехода $t \in T_P$ записи O_t, $\langle t\rangle$, $\mathrm{start}\,(t)$ и $\mathrm{end}\,(t)$ обозначают оператор, формулу и состояния, определяемые следующим образом: если t имеет вид $s_1 \stackrel{O}{\to} s_2$, то $O_t \stackrel{\text{def}}{=} O$, $\langle t\rangle \stackrel{\text{def}}{=} (\mathrm{at}_P = s_1) \wedge \langle O\rangle$, $\mathrm{start}\,(t) \stackrel{\text{def}}{=} s_1$, $\mathrm{end}\,(t) \stackrel{\text{def}}{=} s_2$. Если t — ввод или вывод, то запись N_t обозначает имя N_{O_t}.

Множество X_P^s **существенных переменных** процесса P определяется как наименьшее (по включению) множество, удовлетворяющее следующим условиям.

– X_P^s содержит все переменные, содержащиеся в предусловиях и выводах в операторах процесса P;
– если P содержит АО $x := e$ и $x \in X_P^s$, то X_P^s содержит все переменные, входящие в e.

Процессу P соответствует помеченный граф (обозначаемый тем же символом P), вершинами которого являются состояния из S_P, а ребрами — переходы из T_P: каждому переходу $s_1 \xrightarrow{O} s_2$ соответствует ребро из s_1 в s_2 с меткой O.

3.2 Действия процессов

Будем называть **действием процесса** (или просто **действием**) запись одного из следующих трех видов:

(i) $\alpha?d$, где $\alpha \in \mathcal{N}$ и $d \in \mathcal{D}$. Действие такого вида называется **получением объекта** с именем α вместе с присоединенным к этому объекту сообщением d;

(ii) $\alpha!d$, где $\alpha \in \mathcal{N}$ и $d \in \mathcal{D}$. Действие такого вида называется **посылкой объекта** с именем α вместе с присоединенным к этому объекту сообщением d;

(iii) τ. Действие такого вида называется **невидимым действием**.

Множество всех действий будем обозначать символом $\mathcal{A}$.

3.3 Выполнение процесса

Выполнение процесса (1) представляет собой обход графа P, начиная с s_P^0, с выполнением АО, входящих в метки проходимых ребер. На каждом шаге $i \geq 0$ этого обхода определены текущее состояние $s_i \in S_P$ и текущее означивание $\xi_i \in X_P^\bullet$. Предполагается, что $s_0 = s_P^0$, $\langle P \rangle^{\xi_0} = 1$ и для каждого шага i этого обхода $\mathrm{at}_P^{\xi_i} = s_i$.

Выполнение процесса P на шаге i неформально описывается следующим образом. Если в T_P нет переходов с началом в s_i, то P заканчивает свою работу, иначе

– P недетерминированно выбирает переход t, удовлетворяющий условиям: $\langle t \rangle^{\xi_i} = 1$, и если t — ввод или вывод, то в текущий момент времени P имеет возможность принять или послать соответственно объект с именем N_t (т. е. в этот же момент времени выполняется еще один процесс, который в этот момент может послать процессу P или принять от P соответственно объект с именем N_t). Если таких переходов нет, то P временно приостанавливает свою работу до того момента, когда появится хотя бы один такой переход, и после возобновления своей работы P недетерминированно выбирает один из таких переходов;
– после последовательного выполнения всех АО, входящих в оператор O_t выбранного перехода t, P переходит в состояние $\mathrm{end}\,(t)$.

Выполнение каждого из АО o, входящих в $[O_t]$, заключается в выполнении некоторого действия $a \in \mathcal{A}$ и замене текущего означивания ξ на означивание ξ', которое будет считаться текущим после выполнения этого АО. Выполнение АО o производится следующим образом:

– если $o = \alpha?x$, то P выполняет действие вида $\alpha?d$ и $x^{\xi'} \stackrel{\mathrm{def}}{=} d$, $\forall\, y \in X_P \setminus \{x\} \quad y^{\xi'} \stackrel{\mathrm{def}}{=} y^{\xi}$;
– если $o = \alpha!e$, то P выполняет действие $\alpha!(e^{\xi})$ и $\xi' \stackrel{\mathrm{def}}{=} \xi$;
– если $o = (x := e)$, то P выполняет действие τ и $x^{\xi'} \stackrel{\mathrm{def}}{=} e^{\xi}$, $\forall\, y \in X_P \setminus \{x\} \quad y^{\xi'} \stackrel{\mathrm{def}}{=} y^{\xi}$.

4 Реализации процессов

4.1 Реализации атомарных операторов и последовательностей атомарных операторов

Реализацией АО o называется тройка вида (ξ, a, ξ'), удовлетворяющая следующим условиям:

– $\xi, \xi' \in X^\bullet$, где $X_o \subseteq X \subseteq \mathcal{X}$, и $a \in \mathcal{A}$;
– если $o = \alpha?x$, то $a = \alpha?(x^{\xi'})$ и $\forall\, y \in X \setminus \{x\}$ $y^{\xi'} = y^{\xi}$;
– если $o = \alpha!e$, то $a = \alpha!(e^{\xi})$ и $\xi' = \xi$;
– если $o = (x := e)$, то $a = \tau$ и $\xi' = \xi \cdot o$.

Пусть $o_1, \ldots, o_n$ — последовательность АО, которая, возможно, пуста и содержит не более одного ввода или вывода. **Реализацией последовательности** $o_1, \ldots, o_n$ называется произвольная тройка вида (ξ, a, ξ'), удовлетворяющая следующим условиям:

– $\xi, \xi' \in X^\bullet$, где $X \subseteq \mathcal{X}$, и $a \in \mathcal{A}$;
– если $n = 0$, то $\xi' = \xi$ и $a = \tau$, иначе существует последовательность

$$(\xi_0, a_1, \xi_1),\ (\xi_1, a_2, \xi_2),\ \ldots,\ (\xi_{n-1}, a_n, \xi_n), \quad (2)$$

где $\xi_0 = \xi$, $\xi_n = \xi'$, $\forall\, i = 1, \ldots, n$ (ξ_{i-1}, a_i, ξ_i) — реализация o_i и $a = \tau$, если все a_i в (2) равны τ, иначе a совпадает с тем из этих a_i, который отличен от τ.

4.2 Реализации переходов

Пусть задан процесс P вида (1) и переход $t \in T_P$.

Реализацией перехода t называется тройка (ξ_1, a, ξ_2), где $\xi_1, \xi_2 \in X_P^\bullet$ и $a \in \mathcal{A}$, такая что $\langle t \rangle^{\xi_1} = 1$ и $(\xi_1 \cdot (\mathrm{at}_P := \mathrm{end}\,(t))\,, a, \xi_2)$ является реализацией $[O_t]$.

Верны следующие свойства.

1. Если переход t — внутренний или вывод, то для каждого $\xi \in X_P^\bullet$ такого, что $\langle t \rangle^\xi = 1$, существуют единственные $\xi' \in X_P^\bullet$ и $a \in \mathcal{A}$ такие, что (ξ, a, ξ') — реализация t. Будем обозначать такое ξ' записью $\xi \cdot t$.

2. Если переход t — ввод, то для каждого $\xi \in X_P^\bullet$ такого, что $\langle t \rangle^\xi = 1$, и каждого $d \in \mathcal{D}$ существует единственное $\xi' \in X_P^\bullet$ такое, что $(\xi, N_t?d, \xi')$ — реализация t. Будем обозначать такое ξ' записью $\xi \cdot t^d$.

4.3 Реализации процессов

Реализацией процесса P называется граф P^r, имеющий следующие компоненты:

– вершинами графа P^r являются означивания из $X_P^\bullet$, а также еще одна вершина, обозначаемая записью P^0;

– для каждой реализации (ξ_1, a, ξ_2) какого-либо перехода из T_P граф P^r имеет ребро из ξ_1 в ξ_2 с меткой a;

– для каждого $\xi \in X_P^\bullet$ такого, что $\langle P \rangle^\xi = 1$, и каждого ребра графа P^r из ξ в ξ' с меткой a данный граф содержит ребро из P^0 в ξ' с меткой a.

Будем обозначать множества вершин и ребер графа P^r записями S_P^r и T_P^r соответственно. Также будем использовать следующие обозначения: для любой пары v, v' вершин графа P^r

– запись $v \xrightarrow{a} v_2$ обозначает ребро из v_1 в v_2 с меткой a;

– запись $v \xrightarrow{\tau^*} v'$ означает, что $v = v'$ или $\exists\, v_0, v_1, \ldots, v_n : \forall\, i = 1, \ldots, n$ граф P^r содержит ребро $v_{i-1} \xrightarrow{\tau} v_i$ и $v_0 = v$, $v_n = v'$;

– запись $v \xrightarrow{\tau^* a \tau^*} v'$ (где $a \in \mathcal{A}$) означает, что $\exists\, v_1, v_2$: граф P^r содержит ребро $v_1 \xrightarrow{a} v_2$ и $v \xrightarrow{\tau^*} v_1$, $v_2 \xrightarrow{\tau^*} v'$.

5 Наблюдаемая эквивалентность процессов

5.1 Понятие наблюдаемой эквивалентности процессов

Будем называть процессы P_1 и P_2 **наблюдаемо эквивалентными**, если P_1^r и P_2^r наблюдаемо эквивалентны в смысле Милнера [1], т. е. существует $\mu \subseteq S_{P_1}^r \times S_{P_2}^r$, удовлетворяющее следующим условиям:

(i) $(P_1^0, P_2^0) \in \mu$;

(ii) если $(v_1, v_2) \in \mu$ и $v_1 \xrightarrow{\tau} v_1'$, то

$$\exists\, v_2' : v_2 \xrightarrow{\tau^*} v_2',\ (v_1', v_2') \in \mu\,;$$

если $(v_1, v_2) \in \mu$ и $v_2 \xrightarrow{\tau} v_2'$, то

$$\exists\, v_1' : v_1 \xrightarrow{\tau^*} v_1',\ (v_1', v_2') \in \mu\,;$$

(iii) если $(v_1, v_2) \in \mu$ и $v_1 \xrightarrow{a} v_1'$, где $a \neq \tau$, то

$$\exists\, v_2' : v_2 \xrightarrow{\tau^* a \tau^*} v_2',\ (v_1', v_2') \in \mu\,;$$

если $(v_1, v_2) \in \mu$ и $v_2 \xrightarrow{a} v_2'$, где $a \neq \tau$, то

$$\exists\, v_1' : v_1 \xrightarrow{\tau^* a \tau^*} v_1',\ (v_1', v_2') \in \mu\,.$$

Большинство проблем, связанных с верификацией дискретных систем, может быть сведено к проблеме доказательства наблюдаемой эквивалентности двух процессов. Как правило, первый из этих процессов является моделью анализируемой системы, а второй — моделью какого-либо свойства этой системы. В разд. 7 будет рассмотрен пример доказательства наблюдаемой эквивалентности двух процессов, первый из которых является моделью протокола скользящего окна, а второй — моделью процесса, изображающего внешнее поведение этого протокола.

5.2 Метод доказательства наблюдаемой эквивалентности процессов

Один из возможных методов доказательства наблюдаемой эквивалентности двух процессов основан на нижеследующей теореме. Для формулировки и доказательства этой теоремы введем вспомогательные понятия и обозначения.

1. Пусть задан процесс P и пара состояний $s, s' \in \in S_P$.

   Составной переход (СП) из s в s' — это последовательность T переходов процесса P вида

$$s = s_0 \xrightarrow{O_1} s_1\,;\ \ s_1 \xrightarrow{O_2} s_2\,; \ldots ; s_{n-1} \xrightarrow{O_n} s_n = s' \quad (3)$$

такая что среди $O_1, \ldots, O_n$ не более одного оператора ввода или вывода и определены все конкатенации в выражении

$$(\cdots(O_1 \cdot O_2)\cdots) \cdot O_n\,. \tag{4}$$

Последовательность (3) может быть пустой, в этом случае $s = s'$.

Если СП T не пуст и имеет вид (3), то запись O_T обозначает значение выражения (4), а если T пуст, то $O_T \stackrel{\text{def}}{=} [\,]$.

Будем использовать для СП те же понятия и обозначения, что и для обычных переходов ($\text{start}\,(T)$, $\text{end}\,(T)$, N_T и т. п.). Будем называть СП T вводом, выводом или внутренним, если O_T — оператор ввода, вывода или внутренний соответственно.

Как и для обычных переходов, для СП можно ввести понятие реализации, которое будет обладать свойствами, аналогичными свойствам, изложенным в п. 4.2, в частности:

(а) если СП T — внутренний или вывод, то для каждого $\xi \in X_P^\bullet$ такого, что $\langle T\rangle^\xi = 1$, существуют единственные $\xi' \in X_P^\bullet$ и $a \in \mathcal{A}$ такие, что (ξ, a, ξ') — реализация T. Будем обозначать такое ξ' записью $\xi \cdot T$;

(б) если СП T — ввод, то для каждого $\xi \in X_P^\bullet$ такого, что $\langle T\rangle^\xi = 1$, и каждого $d \in \mathcal{D}$ существует единственное $\xi' \in X_P^\bullet$ такое, что $(\xi, N_T?d, \xi')$ — реализация T. Будем обозначать такое ξ' записью $\xi \cdot T^d$.

2. Если b и b' — формулы, то запись $b \le b'$ является сокращенной записью утверждения о том, что формула $b \to b'$ истинна.

3. Если O_1 и O_2 — операторы и $b \in \mathcal{B}$, то запись $(O_1, O_2) \cdot b$ обозначает формулу, определяемую излагаемым ниже рекурсивным определением, в котором записи вида $O \setminus o$ и $o(b)$ обозначают оператор и формулу соответственно, определяемые так же, как в п. 2.3.

Пусть $[O_1] = o_1, \ldots, o_n$ и $[O_2] = o'_1, \ldots, o'_m$, тогда формула

$$(O_1, O_2) \cdot b \tag{5}$$

определяется следующим образом:

(а) $\langle O_1\rangle \wedge \langle O_2\rangle \wedge b$, если $n = m = 0$;

(б) $(O_1 \setminus o_n, O_2) \cdot o_n(b)$, если o_n — присваивание;

(в) $(O_1, O_2 \setminus o'_m) \cdot o'_m(b)$, если o'_m — присваивание;

(г) $((O_1 \setminus o_n), (O_2 \setminus o'_m)) \cdot b(z/x, z/y)$, если $o_n = \alpha?x$, $o'_m = \alpha?y$ и $b(z/x, z/y)$ — формула, получаемая из b заменой всех вхождений x и y на новую переменную z (не входящую в O_1, O_2 и b);

(д) $((O_1 \setminus o_n), (O_2 \setminus o'_m)) \cdot ((e_1 = e_2) \wedge b)$, если $o_n = \alpha!e_1$ и $o'_m = \alpha!e_2$;

(е) $\bot$ в остальных случаях.

Теорема 1. *Пусть* $P_i = (S_{P_i}, s^0_{P_i}, T_{P_i}, \langle P_i\rangle)$ $(i = 1, 2)$ — *процессы, причем* $S_{P_1} \cap S_{P_2} = \emptyset$ и $X_{P_1} \cap X_{P_2} = \emptyset$. Процессы P_1 и P_2 *наблюдаемо эквивалентны, если существует совокупность* $\{b_{s_1 s_2} \mid s_i \in S_{P_i}\ (i = 1, 2)\}$ *формул с переменными из* $(X_{P_1} \cup X_{P_2}) \setminus \{\text{at}_{P_1}, \text{at}_{P_2}\}$, *обладающих следующими свойствами*:

(i) $\langle P_1\rangle \wedge \langle P_2\rangle \le b_{s^0_{P_1} s^0_{P_2}}$;

(ii) для каждого перехода $s_1 \stackrel{O}{\to} s'_1$ процесса P_1 и каждого состояния $s_2 \in S_{P_2}$ существует совокупность СП процесса P_2, имеющая вид $\{s_2 \stackrel{T_i}{\to} s^i_2 \mid i \in \Im\}$ и такая, что

$$b_{s_1 s_2} \wedge \langle O\rangle \le \bigvee_{i \in \Im} (O, O_{T_i}) \cdot b_{s'_1 s^i_2};$$

(iii) свойство, симметричное предыдущему свойству: для каждого перехода $s_2 \stackrel{O}{\to} s'_2$ процесса P_2 и каждого состояния $s_1 \in S_{P_1}$ существует совокупность СП процесса P_1, имеющая вид $\{s_1 \stackrel{T_i}{\to} s^i_1 \mid i \in \Im\}$ и такая, что

$$b_{s_1 s_2} \wedge \langle O\rangle \le \bigvee_{i \in \Im} (O_{T_i}, O) \cdot b_{s^i_1 s'_2}\,.$$

Доказательство. Поскольку $X_{P_1} \cap X_{P_2} = \emptyset$, то существует естественная биекция между $X^\bullet_{P_1} \times X^\bullet_{P_2}$ и $(X_{P_1} \cup X_{P_2})^\bullet$. Ниже будем отождествлять эти два множества.

Определим отношение $\mu \subseteq S^\tau_{P_1} \times S^\tau_{P_2}$:

$$\mu \stackrel{\text{def}}{=} \left\{(\xi_1, \xi_2) \in X^\bullet_{P_1} \times X^\bullet_{P_2} \mid b^{(\xi_1,\xi_2)}_{at^{\xi_1}_{P_1} at^{\xi_2}_{P_2}} = 1\right\} \cup \left\{(P^0_1, P^0_2)\right\}\,.$$

Докажем, что μ удовлетворяет условиям из п. 5.1.

Первое из этих условий непосредственно следует из определения μ.

Докажем второе и третье условия.

Пусть $(v_1, v_2) \in \mu$ и $v_1 \stackrel{a}{\to} v'_1$. Требуется доказать, что

$$\exists\, v'_2 : v_2 \stackrel{a_\tau}{\to} v'_2,\ (v'_1, v'_2) \in \mu\,, \tag{6}$$

где

$$a_\tau = \begin{cases} \tau^*, & \text{если } a = \tau\,, \\ \tau^* a \tau^*, & \text{если } a \ne \tau\,. \end{cases}$$

Изложим доказательство лишь для случая $v_1 = P^0_1$ (в случае $v_1 \ne P^0_1$ доказательство выглядит аналогично).

Если $v_1 = P_1^0$, то $v_2 = P_2^0$ и согласно определению P_1^r (см. п. 4.3) $\exists \xi_1 \in X_{P_1}^\bullet : \langle P_1 \rangle^{\xi_1} = 1$ и P_1^r содержит ребро $\xi_1 \stackrel{a}{\to} \xi_1' = v_1'$, т. е. (ξ_1, a, ξ_1') — реализация перехода t вида $s_{P_1}^0 \stackrel{O_1}{\to} s_1'$ из T_{P_1}.

Согласно п. (ii) в формулировке теоремы, существует совокупность $\{s_{P_2}^0 \stackrel{T_i}{\to} s_2^i \mid i \in \Im\}$ СП процесса P_2 такая, что

$$b_{s_{P_1}^0 s_{P_2}^0} \wedge \langle O_1 \rangle \le \bigvee_{i \in \Im} (O_1, O_{T_i}) \cdot b_{s_1' s_2^i} . \qquad (7)$$

Поскольку $\langle P_2 \rangle \ne \bot$, то $\exists \xi_2 \in X_{P_2}^\bullet$: $\langle P_2 \rangle^{\xi_2} = 1$, поэтому

$$1 = \langle P_1 \rangle^{\xi_1} \wedge \langle P_2 \rangle^{\xi_2} = \\ = (\langle P_1 \rangle \wedge \langle P_2 \rangle)^{(\xi_1, \xi_2)} \le b_{s_{P_1}^0 s_{P_2}^0}^{(\xi_1, \xi_2)} \qquad (8)$$

(последнее неравенство верно на основании свойства (i) в формулировке теоремы).

Согласно определению реализации перехода верно равенство $\langle O_1 \rangle^{\xi_1} = 1$, из которого, а также из (7) и (8) следует, что для некоторого $i \in \Im$ верно равенство

$$\Big((O_1, O_{T_i}) \cdot b_{s_1' s_2^i} \Big)^{(\xi_1, \xi_2)} = 1 , \qquad (9)$$

которое в случае $a = \alpha?d$ следует понимать в следующем смысле: для каждого означивания $\xi \in (X_{P_1} \sqcup X_{P_2} \sqcup \{z\})^\bullet$ (где z — переменная, упомянутая в п. 3г определения из п. 5.2, можно считать, что $z \notin (X_{P_1} \sqcup X_{P_2})$), совпадающего с ξ_i на X_{P_i} $(i = 1, 2)$, верно равенство:

$$\Big((O_1, O_{T_i}) \cdot b_{s_1' s_2^i} \Big)^{\xi} = 1 .$$

Рассмотрим возможные виды a.

1. $a = \tau$. В этом случае O_1 — внутренний оператор и

$$\Big((O_1, O_{T_i}) \cdot b_{s_1' s_2^i} \Big)^{(\xi_1, \xi_2)} = b_{s_1' s_2^i}{}^{(\xi_1 \cdot O_1, \xi_2 \cdot O_{T_i})} . \qquad (10)$$

Равенство (10) является аналогом равенства, приведенного в последнем абзаце п. 2.3, и доказывается индукцией по общему числу АО в $[O_1]$ и $[O_2]$.

Из (9) и (10) следует, что

$$b_{s_1' s_2^i}{}^{(\xi_1 \cdot O_1, \xi_2 \cdot O_{T_i})} = 1 . \qquad (11)$$

По определению μ и ξ_2 доказываемое соотношение (6) в рассматриваемом случае ($v_1 = P_1^0$) следует из соотношения:

$$\exists \xi_2' : \xi_2 \stackrel{\tau^*}{\to} \xi_2', \ b^{(\xi_1', \xi_2')}_{\mathrm{at}_{P_1}^{\xi_1'} \mathrm{at}_{P_2}^{\xi_2'}} = 1 . \qquad (12)$$

Определим $\xi_2' \stackrel{\text{def}}{=} (\xi_2 \cdot (\mathrm{at}_{P_2} := s_2^i)) \cdot O_{T_i}$. Поскольку $\mathrm{at}_{P_1}^{\xi_1'} = s_1'$ и $\xi_1' = (\xi_1 \cdot (\mathrm{at}_{P_1} := s_1')) \cdot O_1$, то (12) следует из соотношений

$$\xi_2 \stackrel{\tau^*}{\to} (\xi_2 \cdot (\mathrm{at}_{P_2} := s_2^i)) \cdot O_{T_i} ; \qquad (13)$$

$$b_{s_1' s_2^i}^{((\xi_1 \cdot (\mathrm{at}_{P_1} := s_1')) \cdot O_1, (\xi_2 \cdot (\mathrm{at}_{P_2} := s_2^i)) \cdot O_{T_i})} = 1 . \qquad (14)$$

Соотношение (13) следует из определений понятия СП и конкатенации операторов, а также из соотношений $\mathrm{at}_{P_2}^{\xi_2} = s_{P_2}^0$ и $\langle O_{T_i} \rangle^{\xi_2} = 1$. Первое из этих соотношений является следствием равенства $\langle P_2 \rangle^{\xi_2} = 1$, а второе обосновывается следующим образом. Из определения формул вида $(O_1, O_2) \cdot b$ следует, что соотношение (9) можно переписать в виде:

$$(\langle O_1 \rangle \wedge \langle O_{T_i} \rangle \wedge b)^{(\xi_1, \xi_2)} = 1 , \qquad (15)$$

где b — некоторая формула. Поскольку $X_{P_1} \cap X_{P_2} = \emptyset$, то из (15) следует искомое соотношение $\langle O_{T_i} \rangle^{\xi_2} = 1$.

Соотношение (14) следует из (11) и из предположения о том, что at_{P_1} и at_{P_2} не входят в $b_{s_1' s_2^i}$, O_1 и O_{T_i}.

2. $a = \alpha?d$. В этом случае O_1 — оператор ввода и из (9) следует, что O_{T_i} — тоже оператор ввода и $N_{O_{T_i}} = N_{O_1} = \alpha$.

Используя обозначение, введенное в конце п. 4.2, можно написать, что $\xi_1' = \xi_1 \cdot t^d$.

Определим $\xi_2' \stackrel{\text{def}}{=} \xi_2 \cdot T_i^d$. Нетрудно доказать, что $\xi_2 \stackrel{\tau^* a \tau^*}{\to} \xi_2'$, и доказываемое соотношение (6) следует из равенства

$$b_{s_1' s_2^i}^{(\xi_1 \cdot t^d, \xi_2 \cdot T_i^d)} = 1 , \qquad (16)$$

которое обосновывается следующим образом.

В рассматриваемом случае O_1 и O_{T_i} можно представить как конкатенации вида

$$O_1 = (O_1' \cdot [\alpha?x]) \cdot O_1'' ; \ \ O_{T_i} = (O_{T_i}' \cdot [\alpha?y]) \cdot O_{T_i}'' .$$

Из определения формул вида (5) следует, что

$$(O_1, O_{T_i}) \cdot b_{s_1' s_2^i} = \\ = \big((O_1' \cdot [\alpha?x]) \cdot O_1'', (O_{T_i}' \cdot [\alpha?y]) \cdot O_{T_i}'' \big) \cdot b_{s_1' s_2^i} = \\ = \big(O_1' \cdot [\alpha?x], O_{T_i}' \cdot [\alpha?y] \big) \cdot \Big((O_1'', O_{T_i}'') \cdot b_{s_1' s_2^i} \Big) = \\ = (O_1', O_{T_i}') \cdot \left(\left((O_1'', O_{T_i}'') \cdot b_{s_1' s_2^i} \right) \left(\frac{z}{x}, \frac{z}{y} \right) \right) . \qquad (17)$$

Из (9) и (17) следует, что верно равенство

$$\left(\left((O_1'', O_{T_i}'') \cdot b_{s_1' s_2^i} \right) \left(\frac{z}{x}, \frac{z}{y} \right) \right)^{(\xi_1 \cdot O_1', \xi_2 \cdot O_{T_i}')} = 1 ,$$

частным случаем которого является равенство

$$\left(\left((O''_1, O''_{T_i}) \cdot b_{s'_1 s^i_2}\right)\left(\frac{d}{x}, \frac{d}{y}\right)\right)^{(\xi_1 \cdot O'_1, \xi_2 \cdot O'_{T_i})} = 1\,.$$

Последнее равенство можно переписать в виде:

$$\left((O''_1, O''_{T_i}) \cdot b_{s'_1 s^i_2}\right)^{(\xi_1 \cdot O'_1 \cdot (x:=d), \xi_2 \cdot O'_{T_i} \cdot (y:=d))} = 1\,,$$

откуда следует равенство

$$\left(b_{s'_1 s^i_2}\right)^{(\xi_1 \cdot O'_1 \cdot (x:=d) \cdot O''_1, \xi_2 \cdot O'_{T_i} \cdot (y:=d) \cdot O''_{T_i})} = 1\,. \quad (18)$$

Нетрудно видеть, что левая часть в (18) совпадает с левой частью доказываемого равенства (16).

3. $a = \alpha!d$. В этом случае O_1 — оператор вывода и из (9) следует, что O_{T_i} — тоже оператор вывода и $N_{O_{T_i}} = N_{O_1} = \alpha$.

Определим $\xi'_2 \stackrel{\text{def}}{=} \xi_2 \cdot T_i$. Для доказательства (6) достаточно доказать соотношения

$$\xi_2 \stackrel{\tau^* a \tau^*}{\longrightarrow} \xi'_2\,; \quad (19)$$

$$b^{(\xi_1 \cdot t, \xi_2 \cdot T_i)}_{s'_1 s^i_2} = 1. \quad (20)$$

В рассматриваемом случае O_1 и O_{T_i} можно представить как конкатенации вида:

$$O_1 = (O'_1 \cdot [\alpha!e_1]) \cdot O''_1\,; \quad (21)$$

$$O_{T_i} = (O'_{T_i} \cdot [\alpha!e_2]) \cdot O''_{T_i}\,. \quad (22)$$

Из определения формул вида (5) следует, что

$$\begin{aligned}(O_1, O_{T_i}) \cdot b_{s'_1 s^i_2} &= \\ = \left((O'_1 \cdot [\alpha!e_1]) \cdot O''_1, \left(O'_{T_i} \cdot [\alpha!e_2]\right) \cdot O''_{T_i}\right) \cdot b_{s'_1 s^i_2} &= \\ = \left(O'_1 \cdot [\alpha!e_1], O'_{T_i} \cdot [\alpha!e_2]\right) \cdot \left((O''_1, O''_{T_i}) \cdot b_{s'_1 s^i_2}\right) &= \\ = (O'_1, O'_{T_i}) \cdot \left\{ \begin{array}{c} e_1 = e_2 \\ (O''_1, O''_{T_i}) \cdot b_{s'_1 s^i_2} \end{array} \right\}\,. \quad (23)\end{aligned}$$

Из (9) и (23) следует, что верно равенство

$$\left\{ \begin{array}{c} e_1 = e_2 \\ (O''_1, O''_{T_i}) \cdot b_{s'_1 s^i_2} \end{array} \right\}^{(\xi_1 \cdot O'_1, \xi_2 \cdot O'_{T_i})} = 1\,,$$

из которого следуют равенства

$$e_1^{\xi_1 \cdot O'_1} = e_2^{\xi_2 \cdot O'_{T_i}}\,; \quad (24)$$

$$\left((O''_1, O''_{T_i}) \cdot b_{s'_1 s^i_2}\right)^{(\xi_1 \cdot O'_1, \xi_2 \cdot O'_{T_i})} = 1\,. \quad (25)$$

По предположению, $(\xi_1, \alpha!d, \xi'_1)$ является реализацией перехода $s^0_{P_1} \stackrel{O_1}{\to} s'_1$. Из представления O_1 в виде конкатенации (21) следует, что $d = e_1^{\xi_1 \cdot O'_1}$, откуда согласно (24) получаем равенство $d = e_2^{\xi_2 \cdot O'_{T_i}}$. Из этого равенства и из представления O_{T_i} в виде конкатенации (22) следует, что $(\xi_2, \alpha!d, \xi_2 \cdot T_i)$ является реализацией СП T_i. Поскольку $\xi_2 \cdot T_i = \xi'_2$ и $\alpha!d = a$, то, следовательно, соотношение (19) обосновано.

Соотношение (20) следует из (25).

Условия на μ, симметричные рассмотренным условиям (т. е. вторые части условий на μ, изложенные в пп. (ii) и (iii) разд. 5.1, рассматриваются аналогично. ■

6 Упрощение процессов

Понятие упрощения процессов предназначено для решения проблемы понижения сложности верификации процессов.

Упрощение процесса P представляет собой последовательность преобразований этого процесса, каждое из которых производится согласно какому-либо из излагаемых ниже правил. Каждое из этих преобразований (кроме первого) производится над результатом предыдущего преобразования. **Результатом** упрощения является результат последнего из этих преобразований.

Правила упрощения определяются следующим образом. Пусть задан процесс P.

Правило 1 (удаление состояний). Если $s \in S_P \backslash \{s^0_P\}$ и

- совокупность переходов из T_P с концом s имеет вид $s_1 \stackrel{O_1}{\to} s, \ldots, s_n \stackrel{O_n}{\to} s$;
- совокупность переходов из T_P с началом s имеет вид $s \stackrel{O'_1}{\to} s'_1, \ldots, s \stackrel{O'_m}{\to} s'_m$, и если все эти переходы внутренние, то $\langle O'_i \rangle \wedge \langle O'_j \rangle = \bot$ при $i \neq j$;
- $s \notin \{s_1, \ldots, s_n, s'_1, \ldots, s'_m\}$;
- $\forall\, i = 1, \ldots, n, \forall\, j = 1, \ldots, m\ \exists\, O_i \cdot O'_j$,

то из P удаляются состояние s и все переходы, началом или концом которых является s, и добавляются переходы

$$s_i \stackrel{O_i \cdot O'_j}{\longrightarrow} s'_j\ \ (\forall\, i = 1, \ldots, n,\ \forall\, j = 1, \ldots, m)\,.$$

Правило 2 (склейка). Если P содержит пару переходов вида $s_1 \stackrel{O}{\to} s_2$, $s_1 \stackrel{O}{\to} s_2$ и $[O] = [O']$, то эта пара заменяется на один переход из s_1 в s_2 с оператором $(\langle O \rangle \vee \langle O' \rangle)[O]$.

Правило 3 (удаление несущественных присваиваний). Если P содержит присваивание $(x := e)$, где $x \notin X^s_P$, то данный АО удаляется из P.

Теорема 2. *Если P' является упрощением P, то $P' \approx P$.*

7 Пример: верификация протокола скользящего окна

В этом разделе излагается пример использования теоремы 1 для верификации протокола скользящего окна.

Протокол скользящего окна обеспечивает передачу сообщений от одного агента другому через среду, в которой сообщения могут искажаться или пропадать. В этом разделе рассматривается двунаправленный протокол скользящего окна, в котором агенты могут посылать и принимать сообщения друг от друга. Здесь не приводится детальное описание этого протокола, его можно найти в [11, п. 3.4.2] (протокол с возвратом на n).

7.1 Структура протокола

Протокол скользящего окна является системой, состоящей из нескольких взаимодействующих компонентов, в том числе

- компонентов, которые осуществляют формирование, посылку, получение, обработку сообщений (эти компоненты называются **агентами**, а сообщения, посылаемые одними агентами другим агентам, называются **кадрами**);
- среды, через которую пересылаются кадры (эта среда называется **каналом**).

Связь между этими компонентами представляется в виде потокового графа (рис. 1).

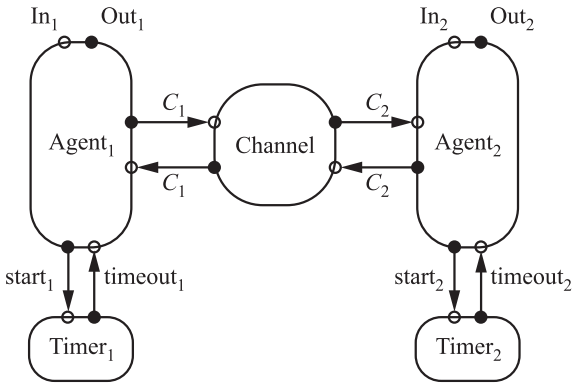


Рис. 1 Потоковый граф

7.2 Кадры

Каждый кадр f, пересылаемый каким-либо из агентов, содержит пакет x и два числа:

(1) $s \in \mathbf{Z}_n \stackrel{\text{def}}{=} \{0, 1, \ldots, n-1\}$ (где n — фиксированное натуральное число), s ассоциировано с пакетом x и с кадром f;

(2) $r \in \mathbf{Z}_n$, r ассоциировано с последним полученным неискаженным кадром.

Для построения кадров используется функция φ: каждый кадр имеет вид $\varphi(x, s, r)$.

Для извлечения компонентов x, s, r из кадра $\varphi(x, s, r)$ используются функции info, seq и ack, эти функции имеют следующие свойства:

$$\text{info}\,(\varphi(x,s,r)) = x\,;\ \ \text{seq}\,(\varphi(x,s,r)) = s\,;$$
$$\text{ack}\,(\varphi(x,s,r)) = r\,.$$

7.3 Окно

Агент содержит среди своих переменных массив $x[n]$, в компонентах которого могут содержаться отправленные, но еще не подтвержденные пакеты. Совокупность компонентов массива x, в которых содержатся такие пакеты в текущий момент времени, называется **окном**.

С окном связаны три переменных этого агента: b (нижняя граница окна); s (верхняя граница окна); w (количество пакетов в окне). Значения переменных b, s и w принадлежат множеству $\mathbf{Z}_n$. В начальный момент времени значения переменных b, s и w равны 0.

В любой момент времени окно может быть пустым (если $b = s$) или не пустым (если $b \neq s$). В последнем случае окно состоит из элементов массива x с индексами из $[b, s[$, где $[b, s[$ обозначает множество

- $\{b, b+1, \ldots, s-1\}$, если $b < s$;
- $\{b, b+1, \ldots, n\} \cup \{0, 1, \ldots, s-1\}$, если $s < b$.

Добавление нового пакета к окну происходит путем выполнения следующих действий: данный пакет записывается в компоненту $x[s]$, s увеличивается на 1 по модулю n (т. е. новое значение s полагается равным $s+1$, если $s < n-1$, и 0, если $s = n-1$) и w увеличивается на 1. Удаление пакета из окна происходит путем выполнения следующих действий: b уменьшается на 1 по модулю n, w уменьшается на 1 (т. е. удаляется тот пакет, номер которого равен нижней границе окна).

Если агент получает кадр, третья компонента r которого (т. е. номер подтверждения) такова, что $r \in [b, s[$, то все пакеты в окне с номерами из $[b, r[$ рассматриваются как подтвержденные и удаляются из окна (даже если их подтверждения не получены).

7.4 Таймеры

Каждая компонента $x[i]$ массива x связана с соответствующим таймером, который определяет продолжительность ожидания подтверждения от другого агента получения им пакета, содержащегося в компоненте $x[i]$. Совокупность этих таймеров

рассматривается как процесс Timer, который имеет массив $t[n]$ булевых переменных. Процесс Timer имеет одно состояние и переходы, помеченные следующими операторами:

- $[\text{start}?i, t[i] := 1]$;
- $[\text{stop}?i,\ t[i] := 0]$;
- $(t[j] = 1)[\text{timeout}!\, j, t[j] := 0]$, где $j = 0, \ldots, n-1$.

Предусловие имеет вид $t = (0, \ldots, 0)$.

Если агент получает объект с именем timeout от таймера, то этот агент посылает еще раз все пакеты из своего окна.

7.5 Агенты

Поведение обоих агентов описывается одним и тем же процессом, сочетающим функции отправителя и получателя. Это поведение представляется блок-схемой (рис. 2), где

- $\forall i \in \{0, n-2\}\ \ i \underset{n}{+} 1 \stackrel{\text{def}}{=} i+1$ и $(n-1) \underset{n}{+} 1 \stackrel{\text{def}}{=} 0$,
- send обозначает список АО

$$\begin{pmatrix} C\,!\,\varphi(x[s], s, r \underset{n}{-} 1) \\ \text{start}\,!\,s \\ s := s \underset{n}{+} 1 \end{pmatrix};$$

- $\forall i, j \in \{0, n-1\}\ \ i \underset{n}{-} j \stackrel{\text{def}}{=} i-j$, если $i-j \in \{0, n-1\}$, и $n+i-j$ иначе;
- символ $*$ обозначает искаженное сообщение,
- значение переменной enable равно 1, если агент имеет возможность получать новые пакеты от сетевого уровня (т. е. $w < n-1$), и 0 иначе.

Процессы Agent_1 и Agent_2 получаются путем несложного преобразования этой блок-схемы с добавлением соответствующего индекса (1 или 2) к их переменным и именам.

7.6 Спецификация

Внешние действия описанного выше протокола (т. е. действия, которые связаны со взаимодействием с сетевым уровнем) имеют вид $\text{In}_1?d$, $\text{In}_2?d$, $\text{Out}_1!d$ и $\text{Out}_2!d$. Предположим, что учитываются только внешние действия $\text{In}_1?d$ и $\text{Out}_2!d$ и игнорируются другие внешние действия (т. е. рассматривается передача только в одном направлении — слева направо). Докажем, что такое поведение эквивалентно поведению процесса B_{n-1}, который называется «буфер, вмещающий не более $n-1$ кадров» и определяется следующим образом:

- переменными B_{n-1} являются
  - массив $(x[0], \ldots, x[n-1])$, тип элементов которого совпадает с типом кадров протокола;
  - переменные r, s и u, значения которых принадлежат $\mathbf{Z}_n$ и имеют следующий смысл: в каждый момент времени
    - значение u равно числу кадров, содержащихся в буфере;
    - значения r и s могут быть интерпретированы как нижняя и верхняя границы той части массива x, где содержатся полученные кадры, которые пока еще не были выведены из буфера;
- B_{n-1} имеет одно состояние и два перехода с метками

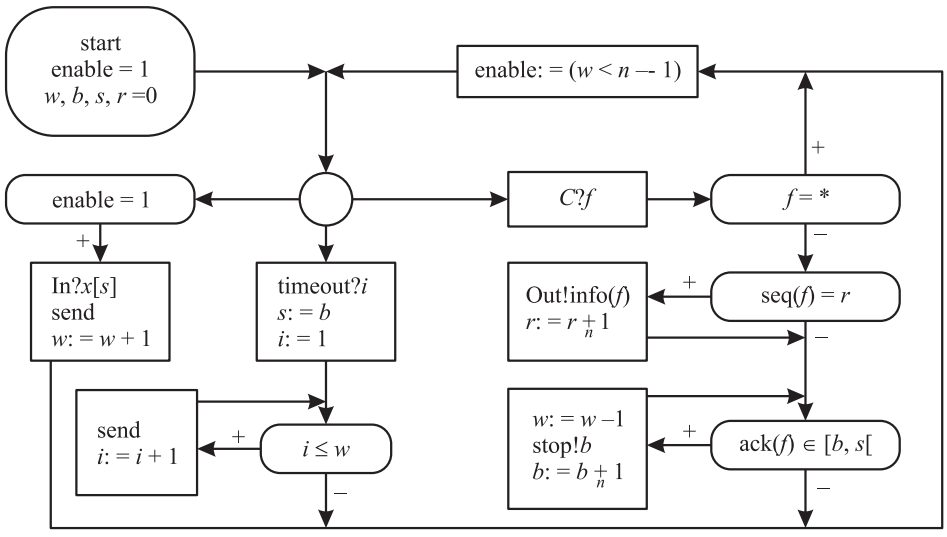


Рис. 2 Блок-схема

$$(u < n-1)\,[In?x[s],\ s := s \underset{n}{+} 1,\ u := u+1]\,;$$

$$(u > 0)\,[\mathrm{Out}\,!\,x[r],\ r := r \underset{n}{+} 1,\ u := u-1]\,;$$

– предусловие имеет вид $r = s = u = 0$.

7.7 Процесс, соответствующий протоколу

Процесс, описывающий поведение этого протокола с учетом указанного выше ограничения (при котором игнорируются действия вида $\mathrm{In}_2?d$ и $\mathrm{Out}_1!d$), определяется как параллельная композиция процессов, соответствующих компонентам этого протокола, с удалением АО, относящихся к игнорируемым взаимодействиям.

Определение параллельной композиции процессов будет изложено лишь для пары процессов (параллельная композиция произвольного числа процессов определяется аналогично).

Пусть P_1 и P_2 — процессы, такие что $S_1 \cap S_2 = \emptyset$ и $X_{P_1} \cap X_{P_2} = \emptyset$. **Параллельной композицией** процессов P_1 и P_2 называется процесс $P = (S_P, s_P^0, T_P, I_P)$, определяемый следующим образом:

$$S_P \stackrel{\text{def}}{=} S_1 \times S_2\,;\ \ s_P^0 \stackrel{\text{def}}{=} (s_1^0, s_2^0)\,,\ \ I_P \stackrel{\text{def}}{=} I_1 \wedge I_2\,,$$

и T_P состоит из следующих переходов:

– для каждого перехода $s_1 \stackrel{O}{\to} s_1'$ процесса P_1 и каждого состояния s процесса P_2 процесс P содержит переход $(s_1, s) \stackrel{O}{\to} (s_1', s)$;

– для каждого перехода $s_2 \stackrel{O}{\to} s_2'$ процесса P_2 и каждого состояния s процесса P_1 процесс P содержит переход $(s, s_2) \stackrel{O}{\to} (s, s_2')$;

– для каждой пары переходов вида

$$\begin{cases} s_1 \stackrel{O_1}{\to} s_1' \in T_{P_1}\,; \\ \\ s_2 \stackrel{O_2}{\to} s_2' \in T_{P_2}\,, \end{cases}$$

где один из операторов O_1, O_2 имеет вид $O_1' \cdot [\alpha?x] \cdot O_1''$, а другой — $O_2' \cdot [\alpha!e] \cdot O_2''$, P имеет переход $(s_1, s_2) \stackrel{O}{\to} (s_1', s_2')$, где $\langle O \rangle = \langle O_1 \rangle \wedge \langle O_2 \rangle$ и

$$[O] = O_1' \cdot O_2' \cdot [x := e] \cdot O_1'' \cdot O_2''\,.$$

7.8 Верификация

С использованием упрощающих преобразований, описанных в разд. 6, можно преобразовать процесс, соответствующий протоколу, в процесс P с одним состоянием и переходами, помеченными следующими операторами:

– $(w < n-1)\ [\mathrm{In}?x[s],\ M_1 := M_1 \cdot \varphi(x[s], s, \ldots),\ s := s \underset{n}{+} 1,\ w := w+1]$;

– $(M_1 \neq \varepsilon) \wedge \left(\mathrm{seq}\left(\hat{M}_1\right) = r\right)\ \left[\mathrm{Out}\,!\,\mathrm{info}\left(\hat{M}_1\right),\ r := r \underset{n}{+} 1,\ M_1 := M_1'\right]$;

– $(M_2 \neq \varepsilon) \wedge \left(\mathrm{ack}\left(\hat{M}_2\right) \in [b, s[\right)\ \left[b := \mathrm{ack}\left(\hat{M}_2\right) \underset{n}{+} 1,\ w := s \underset{n}{-} b,\ M_2 := M_2'\right]$;

– $[M_1 := M_1 \cdot \varphi(x[b], b, \ldots), \ldots\ , M_1 := M_1 \cdot \varphi(x[s \underset{n}{-} 1], s \underset{n}{-} 1, \ldots)]$;

– $(M_1 \neq \varepsilon)\ [M_1 := M_1']$;

– $(M_2 \neq \varepsilon)\ [M_2 := M_2']$;

– $[M_2 := M_2 \cdot \varphi(\ldots, \ldots, r \underset{n}{-} 1)]$,

где многоточия обозначают те компоненты термов, которые являются несущественными, и символы M_i, $\hat{M}_i$, M_i', $\cdot$ и ε имеют следующий смысл:

– M_1 и M_2 — переменные процесса Channel, значения этих переменных являются списками кадров, полученных процессом Channel (M_i содержит кадры, полученные от агента Agent_i), каждый полученный кадр добавляется в конец соответствующего списка;

– $\hat{M}_i$ $(i = 1, 2)$ — терм, значение которого равно первому компоненту списка M_i;

– M_i' $(i = 1, 2)$ — терм, значение которого равно списку M_i, из которого удален первый элемент;

– $\cdot$ — функция добавления кадра в конец списка;

– ε — константа, значением которой является пустой список.

Для доказательства того, что процесс P наблюдаемо эквивалентен процессу B_{n-1}, определим формулу $b_{s_1 s_2}$, где s_1 — единственное состояние процесса P и s_2 — единственное состояние процесса B_{n-1}, как конъюнкцию следующих формул:

– $(M_1 \neq \varepsilon) \wedge \left(\mathrm{seq}\left(\hat{M}\right) = r\right)\ \Rightarrow\ u > 0$;

– $\forall\, f \in M_1\ \mathrm{info}\,(f) = x[\mathrm{seq}\,(f)]$;

– $\forall\, f \in M_2\quad \mathrm{ack}\,(f) \in [b \underset{n}{-} 1, r[$;

– $[r, s[\ \subseteq\ [b, s[$;

– $w = s \underset{n}{-} b \leq n-1$;

– $u = s \underset{n}{-} r \leq w$;

– если значение M_2 равно $f_1 \cdots f_k$, то последовательность $\mathrm{ack}\,(f_1), \ldots, \mathrm{ack}\,(f_k)$ является монотонно возрастающей (mod n) подпоследовательностью последовательности $[b \underset{n}{-} 1, r[$

(последняя запись не является формулой, но может быть представлена в виде формулы). Нетрудно проверить, что $b_{s_1 s_2}$ удовлетворяет условиям теоремы 1, что доказывает наблюдаемую эквивалентность процессов P и B_{n-1}. ■

8 Заключение

Понятие процесса с передачей сообщений, изложенное в настоящей работе, может рассматриваться как формальная модель взаимодействующих нерекурсивных программ. В статье было изложено достаточное условие наблюдаемой эквивалентности процессов с передачей сообщений. Следующими шагами исследований в этом направлении могут быть, например, следующие.

– нахождение необходимых и достаточных условий наблюдаемой эквивалентности процессов с передачей сообщений;
– обобщение введенного понятия до такого понятия процесса с передачей сообщений, которое может служить формальной моделью взаимодействующих рекурсивных программ, и нахождение необходимых и достаточных условий наблюдаемой эквивалентности таких процессов.

Литература

1. *Milner R.* A calculus of communicating systems. Lecture notes in computer science ser. — Berlin – Heidelberg – New York: Springer-Verlag, 1980. Vol. 92. 172 p.
2. *Larsen K. G., Skou A.* Bisimulation through probabilistic testing // Inform. Comput., 1991. Vol. 94. No. 1. P. 1–28.
3. *Larsen K. G., Wang Y.* Time-abstracted bisimulation: Implicit specifications and decidability // Inform. Comput., 1997. Vol. 134. No. 2. P. 75–101.
4. *Milner R.* Communicating and mobile systems: The π-calculus. — Cambridge: Cambridge University Press, 1999. 162 p.
5. *Hoare C. A. R.* Communicating sequential processes. — Prentice Hall, 1985. 256 p.
6. *Clarke E. M., Grumberg O., Peled D.* Model checking. — MIT Press, 1999. 314 p.
7. *Petri C. A.* Introduction to general net theory // Net theory and applications. Lecture notes in computer science ser. / Ed. W. Brauer. — Berlin–Heidelberg: Springer-Verlag, 1980. Vol. 84. P. 1–19.
8. *Handbook of process algebra* / Eds. J. A. Bergstra, A. Ponse, S. A. Smolka. — Amsterdam: North-Holland, 2001. 1357 p.
9. *Brand D., Zafiropulo P.* On communicating finite-state machines // J. ACM, 1983. Vol. 30. No. 2. P. 323–342.
10. *Floyd R. W.* Assigning meaning to programs // Mathematical Aspects of Computer Science: Symposium on Applied Mathematics Proceedings / Ed. J. T. Schwartz. — American Mathematical Society, 1967. Vol. 19. P. 19–32.
11. *Tanenbaum A.* Computer networks. — 4th ed. — Prentice Hall, 2002. 674 p.
12. *Badban B., Fokkink W. J., van de Pol J. C.* Mechanical verification of a two-way sliding window protocol (full version including proofs). — Twente: University of Twente, Centre for Telematics and Information Technology, 2008. Internal Report TR-CTIT-08-45. 55 p.
13. *Hailpern B.* Verifying concurrent processes using temporal logic. Lecture notes in computer science ser. — Berlin–Heidelberg: Springer-Verlag, 1982. Vol. 129. 216 p.
14. *Holzmann G.* Design and validation of computer protocols. — Prentice Hall, 1991. 558 p.
15. *Holzmann G.* The model checker Spin // IEEE Trans. Software Eng., 1991. Vol. 23. No. 5. P. 279–295.
16. *Kaivola R.* Using compositional preorders in the verification of sliding window protocol // Computer aided verification. Lecture notes in computer science ser. / Ed. O. Grumberg. — Berlin–Heidelberg: Springer-Verlag, 1997. Vol. 1254. P. 48–59.
17. *Godefroid P., Long D.* Symbolic protocol verification with queue BDDs // Formal Methods Syst. Design, 1999. Vol. 14. No. 3. P. 257–271.
18. *Stahl K., Baukus K., Lakhnech Y., Steffen M.* Divide, abstract, and model-check // Theoretical and practical aspects of SPIN model cheking. Lecture notes in computer science ser. / Eds. D. Dams, R. Gerth, S. Leue, M. Massink. — Berlin–Heidelberg: Springer-Verlag, 1999. Vol. 1680. P. 57–76.
19. *Latvala T.* Model checking LTL properties of high-level Petri nets with fairness constraints // Applications and theory of Petri nets. Lecture notes in computer science ser. / Eds. J.-M. Colom, M. Koutny. — Berlin–Heidelberg: Springer-Verlag, 2001. Vol. 2075. P. 242–262.
20. *Schoone A.* Assertional verification in distributed computing. — Utrecht: Utrecht University, 1991. Ph.D. Thesis. 191 p.
21. *Chkliaev D., Hooman J., de Vink E.* Verification and improvement of the sliding window protocol // Tools and algorithms for the constriction and analysis of system. Lecture notes in computer science ser. / Eds. H. Garavel, J. Hatcliff. — Berlin–Heidelberg: Springer-Verlag, 2003. Vol. 2619. P. 113–127.
22. *Vaandrager F.* Verification of two communication protocols by means of process algebra. — Amsterdam: Centrum voor Wiskunde en Informatica, 1986. Technical Report CS-R8608. 76 p.
23. *Van Wamel J.* A study of a one bit sliding window protocol in ACP. — Amsterdam: University of Amsterdam, 1992. Technical Report P9212. 59 p.
24. *Bezem M., Groote J.* A correctness proof of a one bit sliding window protocol in μCRL // The Computer J., 1994. Vol. 37. No. 4. P. 289–307.

Поступила в редакцию 4.02.14

A METHOD OF PROVING THE OBSERVATIONAL EQUIVALENCE OF PROCESSES WITH MESSAGE PASSING

A. M. Mironov

Institute of Informatics Problems, Russian Academy of Sciences, Moscow 119333, 44-2 Vavilov Str., Russian Federation

Abstract: The article deals with the problem of proving observational equivalence for the class of computational processes called the processes with message passing. These processes can execute actions of the following forms: sending or receiving the messages, checking the logical conditions, and updating the values of internal variables of the processes. The main result is the theorem that reduces the problem of proving observational equivalence of a pair of processes with message passing to the problem of finding formulas associated with pairs of states of these processes, satisfying certain conditions that are associated with transitions of these processes. This reduction is a generalization of Floyd's method of flowchart verification, which reduces the problem of verification of flowcharts to the problem of finding formulas (called intermediate assertions) associated with points in the flowcharts and satisfying conditions, corresponding to transitions in the flowcharts. The method of proving the observational equivalence of processes with message passing is illustrated by an example of sliding window protocol verification.

Keywords: verification; processes with message passing; observational equivalence; sliding window protocol

DOI: 10.14375/19922264140206

References

1. Milner, R. 1980. *A calculus of communicating systems*. Lecture notes in computer science ser. Berlin – Heidelberg – New York: Springer-Verlag. 92. 172 p.
2. Larsen, K. G., and A. Skou. 1991. Bisimulation through probabilistic testing. *Inform. Comput.* 94(1):1–28.
3. Larsen, K. G., and Y. Wang. 1997. Time-abstracted bisimulation: Implicit specifications and decidability. *Inform. Comput.* 134(2):75–101.
4. Milner, R. 1999. *Communicating and mobile systems: The π-calculus*. Cambridge: Cambridge University Press. 162 p.
5. Hoare, C. A. R. 1985. *Communicating sequential processes*. Prentice Hall. 256 p.
6. Clarke, E. M., O. Grumberg, and D. Peled. 1999. *Model checking*. MIT Press. 314 p.
7. Petri, C. A. 1980. *Introduction to general net theory*. Lecture notes in computer science ser. Ed. W. Brauer. Berlin–Heidelberg: Springer-Verlag. 84:1–19.
8. Bergstra, J. A., A. Ponse, and S. A. Smolka, eds. 2001. *Handbook of process algebra*. North-Holland, Amsterdam. 1357 p.
9. Brand, D., and P. Zafiropulo. 1983. On communicating finite-state machines. *J. ACM* 30(2):323–342.
10. Floyd, R. W. 1967. Assigning meanings to programs. *Mathematical Aspects of Computer Science: Symposium on Applied Mathematics Proceedings*. Ed. J. T. Schwartz. American Mathematical Society. 19:19–32.
11. Tanenbaum, A. 2002. *Computer networks*. 4th ed. Prentice Hall. 674 p.
12. Badban, B., W. J. Fokkink, and J. C. van de Pol. 2008. Mechanical verification of a two-way sliding window protocol (full version including proofs). Twente: Centre for Telematics and Information Technology, University of Twente. Internal Report TR-CTIT-08-45. 55 p.
13. *Hailpern, B.* 1982. *Verifying concurrent processes using temporal logic*. Lecture notes in computer science ser. Berlin–Heidelberg: Springer-Verlag. 129. 216 p.
14. Holzmann, G. 1991. *Design and validation of computer protocols*. Prentice Hall. 558 p.
15. Holzmann, G. 1991. The model checker Spin. *IEEE Trans. Software Eng.* 23(5):279–295.
16. Kaivola, R. 1997. Using compositional preorders in the verification of sliding window protocol. *Computer aided verification*. Lecture notes in computer science ser. Ed. O. Grumberg. Berlin–Heidelberg: Springer-Verlag. 1254:48–59.
17. Godefroid, P., and D. Long. 1999. Symbolic protocol verification with Queue BDDs. *Formal Methods Syst. Design* 14(3):257–271.
18. Stahl, K., K. Baukus, Y. Lakhnech, and M. Steffen. 1999. Divide, abstract, and model-check. *Theoretical and practical aspects of SPIN model checking*. Lecture notes in computer science ser. Eds. D. Dams, R. Gerth, S. Leue, and M. Massink. Berlin–Heidelberg: Springer-Verlag. 1680:57–76.
19. Latvala, T. 2001. Model checking LTL properties of high-level Petri nets with fairness constraints. *Applications and theory of Petri nets*. Lecture notes in computer science ser. Eds. J.-M. Colom and M. Koutny. Berlin–Heidelberg: Springer-Verlag. 2075:242–262.
20. Schoone, A. 1991. Assertional verification in distributed computing. Utrecht University. Ph.D. Thesis. 191 p.

21. Chkliaev, D., J. Hooman, and E. de Vink. 2003. Verification and improvement of the sliding window protocol. *Tools and algorithms for the constriction and analysis of systems*. Lecture notes in computer science ser. Eds. H. Garavel and J. Hatcliff. — Berlin–Heidelberg: Springer-Verlag. 2619:113–127.
22. Vaandrager, F. 1986. Verification of two communication protocols by means of process algebra. Amsterdam: Centrum voor Wiskunde en Informatica. Technical Report CS-R8608. 76 p.
23. Van Wamel, J. 1992. A study of a one bit sliding window protocol in ACP. Amsterdam: University of Amsterdam. Technical Report P9212. 59 p.
24. Bezem, M., and J. Groote. 1994. A correctness proof of a one bit sliding window protocol in μCRL. *The Computer J.* 37(4):289–307.

Received February 4, 2014

Contributor

Mironov Andrew M. (b. 1966) — Candidate of Science (PhD) in physics and mathematics, senior scientist, Institute of Informatics Problems, Russian Academy of Sciences, 44-2 Vavilova Str., Moscow 119333, Russian Federation; amironov66@gmail.com

ИНФОРМАТИКА И ЕЁ ПРИМЕНЕНИЯ, 2014. Т. 8. Вып. 2. С. 70–76

О ПОЛИНОМИАЛЬНОЙ РАЗРЕШИМОСТИ УЛЬТРАМЕТРИЧЕСКИХ ВЕРСИЙ НЕКОТОРЫХ NP-ТРУДНЫХ ЗАДАЧ

М. Г. Адигеев[1]

Аннотация: Статья посвящена анализу важных частных случаев задачи коммивояжера и задачи Штейнера. Обе эти задачи являются NP-трудными даже в метрическом случае, т. е. для графов, у которых функция стоимости ребер удовлетворяет неравенству треугольника. Более строгим ограничением является **усиленное** неравенство треугольника: $\forall x, y, z \in X \quad c(x, z) \leq \max\{c(x, y), c(y, z)\}$. Метрические функции, удовлетворяющие такому условию, называются **ультраметрическими**. В статье на основе анализа графов с ультраметрической функцией стоимости ребер разработан алгоритм, позволяющий построить для такого графа гамильтонов цикл минимальной стоимости за время $O(n^2)$, где n — число вершин графа. Для задачи Штейнера показано, что при ультраметрической функции стоимости ребер минимальное дерево Штейнера содержит только терминальные вершины и поэтому также может быть построено за полиномиальное время как минимальное остовное дерево на подграфе исходного графа.

Ключевые слова: ультраметрическая функция; усиленное неравенство треугольника; задача коммивояжера; дерево Штейнера; полиномиальные алгоритмы

DOI: 10.14375/19922264140207

1 Введение

В статье рассматривается вопрос о полиномиальной разрешимости частных случаев известных вычислительно сложных задач — задачи коммивояжера и задачи Штейнера. Известно [1], что в общем случае обе эти задачи являются NP-трудными, т. е. в настоящее время не известны алгоритмы, находящие точное решение этих задач за полиномиальное время (и, более того, есть основания полагать, что таких алгоритмов не существует в принципе). Эти задачи остаются NP-трудными даже при наложении значительных ограничений на исходные данные — в том числе при условии, что функция стоимости ребер удовлетворяет неравенству треугольника. Более строгим ограничением является **усиленное** неравенство треугольника: $\forall x, y, z \in \in X \quad c(x, z) \leq \max\{c(x, y), c(y, z)\}$. Метрические функции, удовлетворяющие такому условию, называются **ультраметрическими**. Например, подобные функции возникают при решении одной из ключевых задач вычислительной биологии — вычислении филогенетического дерева, отражающего эволюционные связи между современными видами [2–5], а также в задачах управления кэшированием данных [6], разработки и анализа нейроподобных сетей [7].

В работе Д. Гасфилда [3] приведен алгоритм, имеющий временну́ю сложность $O(n^2)$ (n — число вершин графа) и строящий цепь на графе с ультраметрической функцией стоимости. Эта цепь оказывается минимальной гамильтоновой цепью, хотя данный факт не используется в дальнейших построениях Гасфилда. Таким образом, из полученных в [3] результатов следует полиномиальная разрешимость **незамкнутой** ультраметрической задачи коммивояжера, однако приведенный алгоритм не обобщается на ультраметрические версии других NP-трудных задач.

В данной работе, отталкиваясь от результата [3], предлагается метод, позволяющий строить полиномиальные по времени алгоритмы для замкнутого варианта задачи коммивояжера и для задачи Штейнера.

2 Определения

Дадим необходимые определения. Пусть $G(V, E)$ — неориентированный связный граф и $c : E \to R_+$ — функция стоимости, заданная на ребрах графа G. Стоимость цепи, цикла или дерева на графе определяется как сумма стоимостей ребер, входящих в эту цепь, цикл или дерево.

Цепь (цикл) на графе называется **гамильтоновой**, если она ровно по одному разу проходит через каждую вершину графа.

[1] Южный федеральный университет, madi@math.sfedu.ru

Незамкнутая задача коммивояжера: для заданного графа G и функции стоимости c найти гамильтонову *цепь* минимальной стоимости.

Замкнутая задача коммивояжера: для заданного графа G и функции стоимости c найти гамильтонов *цикл* минимальной стоимости.

Известно [1, 8], что как замкнутая, так и незамкнутая задачи коммивояжера являются NP-трудными, т. е. для этих задач в общем виде в настоящее время не существует полиномиальных по времени точных алгоритмов решения.

Задача коммивояжера остается NP-трудной даже при рассмотрении важного частного случая — метрической задачи коммивояжера, т. е. варианта, в котором функция стоимости ребер удовлетворяет требованиям к метрическим функциям:

– неотрицательность:

$$\forall x, y \in X \quad c(x, y) \geq 0\,;$$
$$c(x, y) = 0 \Leftrightarrow x = y\,;$$

– симметричность:

$$\forall x, y \in X \quad c(x, y) = c(y, x)\,;$$

– неравенство треугольника:

$$\forall x, y, z \in X \quad c(x, z) \leq c(x, y) + c(y, z)\,.$$

Иногда рассматривают более общий вариант задачи, допуская прохождение через каждую вершину более одного раза. В [9] показано, что в случае метрической задачи это обобщение не является существенным: оптимальный цикл (или цепь) проходит через каждую вершину ровно один раз даже в том случае, если разрешено проходить более одного раза.

Другим частным случаем задачи коммивояжера является задача с **ультраметрической** функцией стоимости. Метрическая функция называется ультраметрической, если помимо приведенных выше условий неотрицательности и симметричности она удовлетворяет также **усиленному неравенству треугольника**:

$$\forall x, y, z \in X \quad c(x, z) \leq \max\{c(x, y), c(y, z)\}\,.$$

В работе [3] приведен алгоритм, который за полиномиальное ($O(n^2)$, где n — число вершин графа) время строит на графе гамильтонову цепь (алгоритм НайтиГамильтоновуЦепь в данной статье). Таким образом, незамкнутая задача коммивояжера на графах с ультраметрической функцией стоимости является полиномиально разрешимой. В данной работе на основе результатов [3] показано, что полиномиально разрешимым является также и замкнутый вариант задачи коммивояжера с ультраметрической функцией стоимости ребер.

Обобщением задачи коммивояжера является задача нахождения минимального остова ограниченной степени [1, 10]: для заданного графа G, функции стоимости c и натурального числа k найти остовное дерево, у которого степени всех вершин не превосходят k и которое имеет минимальную стоимость среди деревьев такого вида. При $k = 2$ эта задача преобразуется в незамкнутую задачу коммивояжера и, таким образом, является NP-трудной в случае произвольной функции c. Из [3] немедленно следует, что эта задача также полиномиально разрешима для ультраметрических графов.

Еще одной известной NP-трудной (в общем случае) задачей является построение минимального дерева Штейнера [1]. Деревом Штейнера для заданного графа $G(V, E)$ и множества **терминальных** вершин $X \subseteq V$ называется подграф графа G, являющийся деревом и содержащий все терминальные вершины. Требуется построить дерево Штейнера, имеющее минимальную стоимость (т. е. минимальное дерево Штейнера). В данной работе показано, что в случае ультраметрической функции стоимости минимальное дерево Штейнера содержит только терминальные вершины и поэтому является минимальным остовным деревом на подграфе, порожденном множеством терминальных вершин. Из этого следует, что такое дерево может быть построено за полиномиальное время (например, алгоритмом Краскала или Прима).

3 Алгоритмы для ультраметрических графов

Пусть $G(V, E)$ — неориентированный связный граф и $c : E \to R_+$ — функция стоимости, заданная на ребрах графа G и удовлетворяющая требованиям к ультраметрическим функциям. В этом случае граф G без потери общности можно считать полным. Положим также $c(v, v) = 0$ для любой вершины v. Для упрощения формулировок всюду в данной статье **треугольником** (u, v, w) будем называть подграф графа G, порожденный множеством вершин $\{u, v, w\}$, т. е. состоящий из этих вершин и из ребер, соединяющих эти вершины между собой.

3.1 Построение минимальной гамильтоновой цепи

Полученные в данной работе результаты основаны на алгоритме Гасфилда [3]. Этот алгоритм приведен (с изменением обозначений на используемые в данной статье) ниже в виде процедуры НайтиГамильтоновуЦепь. Процедура получает на входе

граф $G(V, E)$ с n вершинами и за время $O(n^2)$ строит гамильтонову цепь P минимальной стоимости.

Процедура НайтиГамильтоновуЦепь(G)

1. $U := V$.
2. Положить P равным пустому пути.
3. Произвольно выбрать вершину $v \in V$.
4. Повторить $n - 1$ раз:
   (а) Удалить v из U.
   (б) Найти вершину $w \in U$ такую, что $c(v, w) \le c(v, u)$ для всех $u \in U$.
   (в) Добавить дугу (v, w) в P.
   (г) $v := w$.

Поскольку гамильтонова цепь является частным случаем остовного дерева и удовлетворяет ограничениям на степень вершины при любом $k > 1$, из результата [3] немедленно следует

Утверждение. Для неориентированного графа $G(V, E)$ с ультраметрической функцией стоимости и произвольного натурального числа $k > 1$ минимальное остовное дерево со степенями вершин, меньшими или равными k, может быть найдено за время $O(|V|^2)$.

3.2 Преобразование треугольника

Алгоритм НайтиГамильтоновуЦепь за полиномиальное время решает незамкнутую задачу коммивояжера. Однако он не адаптируется напрямую для решения других задач, рассматриваемых в данной статье. Их решение начнем с операции, которую будем называть «преобразование треугольника».

Пусть H — подграф графа G и для тройки вершин r, v и w ребра (r, v) и (r, w) принадлежат H, а ребро (v, w) не принадлежит. Преобразование треугольника (r, v, w) заключается в следующем:

Процедура ПреобразоватьТреугольник(H, r, v, w)

1. Из ребер (r, v) и (r, w) выбрать такое, стоимость которого больше или равна $c(v, w)$. Если этому требованию удовлетворяют оба ребра, то выбрать любое из них.
2. Удалить из H выбранное в п. 1 ребро.
3. Добавить к H ребро (v, w).

Рисунок 1 иллюстрирует преобразование треугольника для случая $c(r, v) \ge c(v, w)$. Пунктиром показаны ребра, не принадлежащие H.

Заметим, что на шаге 1 алгоритма ПреобразоватьТреугольник требуется выбрать ребро, удовлетворяющее определенному условию. Поэтому необходимо обосновать допустимость этого преобразования, т. е. показать, что такое ребро всегда найдется.

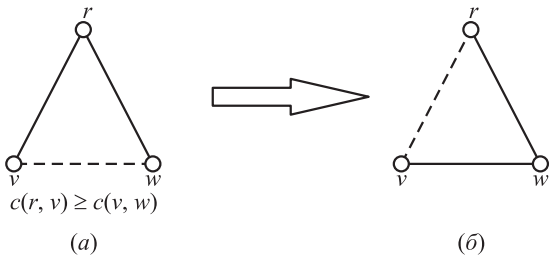


Рис. 1 Преобразование треугольника

Теорема 1. *Если функция стоимости является ультраметрической, то для любых (r, v, w) преобразование треугольника является допустимым и в результате его выполнения стоимость подграфа H не увеличивается.*

Доказательство. Для обоснования допустимости преобразования треугольника необходимо и достаточно показать, что на шаге 1 всегда найдется ребро, принадлежащее H (т. е. (r, v) или (r, w)), стоимость которого больше или равна стоимости ребра (v, w). Но это следует из усиленного неравенства треугольника

$$c(v, w) \le \max\{c(r, v), c(r, w)\}.$$

В силу правила выбора ребра на шаге 1 при замене этого ребра на ребро (v, w) общая стоимость подграфа H не увеличится. Теорема 1 доказана.

3.3 Задача коммивояжера

Алгоритм решения замкнутой задачи коммивояжера основан на применении полиномиального по времени алгоритма для построения гамильтоновой цепи — алгоритма НайтиГамильтоновуЦепь.

Для построения и обоснования алгоритма решения замкнутой задачи коммивояжера введем дополнительные обозначения и докажем два вспомогательных утверждения.

Пусть $G(V, E)$ — неориентированный граф, на ребрах которого задана ультраметрическая функция стоимости c. Обозначим:

$$c_{\max} = \max_{e \in E} c(e)\,;\ E_{\max} = \{e \in E : c(e) = c_{\max}\}\,.$$

Лемма 1. *Любая вершина $v \in V$ инцидентна ребру из $E_{\max}$.*

Доказательство. Выберем произвольную вершину $v \in V$ и произвольное ребро $e = (u, w) \in$

$\in E_{\max}$. Рассмотрим треугольник (u, v, w). В этом треугольнике ребро $e = (u, w)$ имеет стоимость $c_{\max}$. Поэтому, по усиленному неравенству треугольника, как минимум одно из ребер (v, u) или (v, w) также имеет стоимость $c_{\max}$. Лемма 1 доказана.

Лемма 2. *Множество V можно разбить на непересекающиеся подмножества (**кластеры**) $V_1, \ldots, V_k$ таким образом, что*

1. $\forall i \neq j, \forall u \in V_i, \forall v \in V_j$ *выполняется* $c(u, v) = c_{\max}$.
2. $\forall i$ и $\forall u, v \in V_i$ *выполняется* $c(u, v) < c_{\max}$.

Доказательство. Для любой вершины v через $R(v)$ обозначим множество вершин, соединенных с v ребрами со стоимостью, меньшей $c_{\max}$: $R(v) = \{u \in V : c(u, v) < c_{\max}\}$. Заметим, что каждое множество $R(v)$ не пусто, поскольку $v \in R(v)$. Для доказательства леммы достаточно показать, что для различных вершин $u, v \in V$ множества $R(u)$ и $R(v)$ либо не пересекаются, либо совпадают. Покажем это методом от противного.

Предположим, что существуют две вершины u и v ($u \neq v$), для которых $R(u)$ и $R(v)$ пересекаются, но не совпадают. Тогда существуют вершины x и y, не совпадающие ни с u, ни с v, такие что: $x \in R(u) \cap R(v)$ и $y \in R(u) \setminus R(v)$. Рассмотрим подграф, образованный вершинами v, x и y. Так как $y \notin R(v)$, то $c(v, y) = c_{\max}$. В соответствии с усиленным неравенством треугольника как минимум одно из ребер (v, x) или (x, y) также должно иметь стоимость $c_{\max}$. Но $x \in R(v)$, поэтому $c(v, x) < c_{\max}$. Следовательно, $c(x, y) = c_{\max}$. Но, с другой стороны, в подграфе, образованном вершинами u, x и y, стоимость каждого из ребер (u, x) и (u, y) меньше $c_{\max}$, поскольку $x, y \in R(u)$. Это в сочетании с $c(x, y) = c_{\max}$ противоречит усиленному неравенству треугольника.

Таким образом, $R(u)$ и $R(v)$ либо не пересекаются, либо совпадают. Поэтому в качестве кластеров можно взять различные множества вида $R(v)$, $v \in V$. Лемма 2 доказана.

Лемма 3. *Пусть $V = \bigcup_{i=1}^{k} V_i$ — разбиение множества вершин на кластеры и G_i $(i = 1, \ldots, k)$ — подграфы, порожденные этими кластерами. Тогда любой минимальный гамильтонов цикл Z^* на графе G может быть представлен в виде*

$$P_1, e_1, P_2, e_2, \ldots, e_{k-1}, P_k, e_k\,, \qquad (1)$$

где P_i $(i = 1, \ldots, k)$ — минимальные гамильтоновы цепи на G_i и ребра e_i принадлежат $E_{\max}$ (рис. 2). И наоборот, любой гамильтонов цикл, имеющий такой вид, является минимальным.

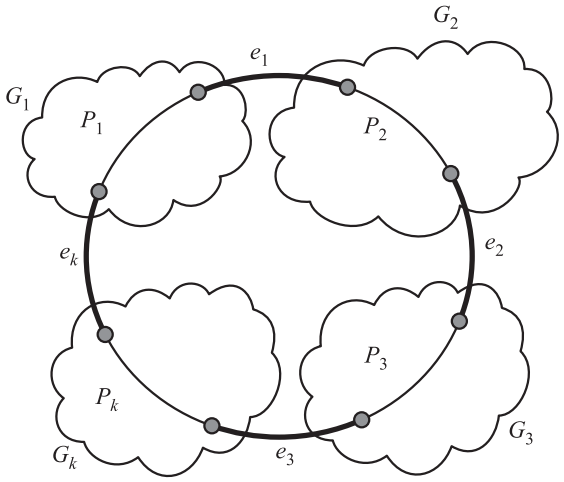


Рис. 2 Структура минимального гамильтонова цикла на ультраметрическом графе

Доказательство Для того чтобы доказать лемму, достаточно показать, что любой минимальный гамильтонов цикл Z^* имеет следующую структуру:

(i) заходя в какой-либо кластер G_i, выходит из него только после полного обхода всех вершин кластера. Иными словами, Z^* заходит и выходит из каждого кластера ровно по одному разу;

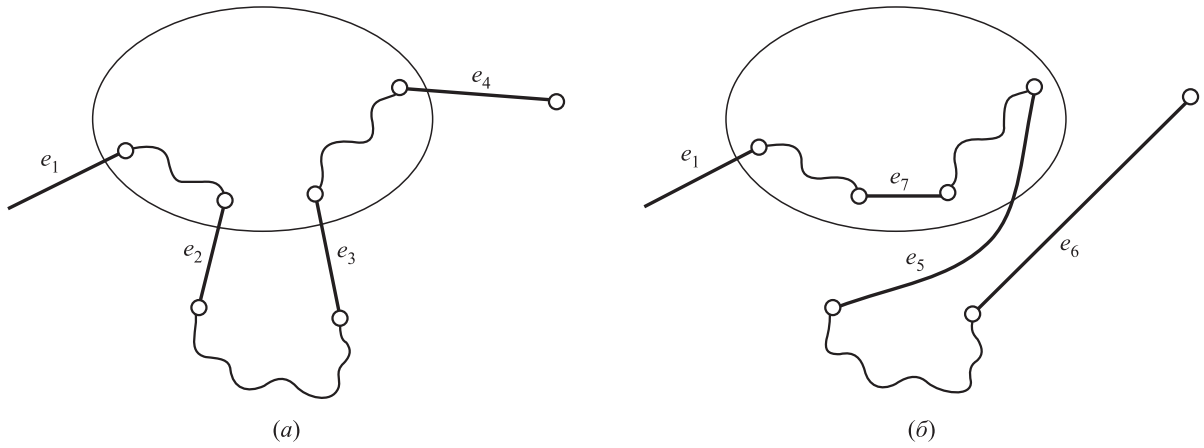
(ii) внутри кластера Z^* проходит по минимальной гамильтоновой цепи.

Рассмотрим гамильтонов цикл Z, не удовлетворяющий условию (i). Тогда Z имеет вид, изображенный на рис. 3, *а*, где ребра $e_1, \ldots, e_4$ имеют стоимость $c_{\max}$, так как соединяют вершины разных кластеров (кластер изображен овалом).

Преобразуем данный цикл так, как изображено на рис. 3, *б*. Здесь ребра e_1, e_5 и e_6 также имеют стоимость $c_{\max}$, а стоимость ребра e_7 меньше $c_{\max}$, так как оно находится внутри кластера. Допустимость преобразования следует из полноты графа G (т. е. ребро e_7 обязательно существует) и леммы 1 (существуют требуемые для преобразования ребра e_5 и e_6). В результате получим гамильтонов цикл меньшей стоимости. Это означает, что исходный цикл не был минимальным.

Если для гамильтонова цикла Z выполняется (i), но нарушается условие (ii), то Z не минимален, поскольку можно заменить его фрагмент внутри кластера на минимальную гамильтонову цепь и получить гамильтонов цикл меньшей стоимости.

Верно и обратное утверждение. Действительно, если гамильтонов цикл Z имеет вид (1), то он совпадает по стоимости с одним из минимальных гамильтоновых циклов и, следовательно, сам является минимальным. Лемма 3 доказана.



(*а*) (*б*)

Рис. 3 Преобразование цикла, не удовлетворяющего условию (i)

Теорема 2. *Для неориентированного графа $G(V, E)$ с ультраметрической функцией стоимости минимальный гамильтонов цикл может быть найден за время $O(|V|^2)$.*

Доказательство. Для нахождения минимального гамильтонова цикла применим алгоритм НайтиГамильтоновЦикл, приведенный ниже.

Процедура НайтиГамильтоновЦикл(G)

1. Построить кластеры $\{V_1, \ldots, V_k\}$.
   Пусть $G_i (i = 1, \ldots, k)$ — подграфы, порожденные кластерами V_i.
2. Для каждого i от 1 до k на графе G_i построить минимальную гамильтонову цепь P_i с помощью процедуры НайтиГамильтоновуЦепь.
3. Построить гамильтонов цикл Z, последовательно соединив для каждого i конец цепи P_i с началом цепи P_{i+1} ребром из $E_{\max}$ (конец P_k соединяется с началом P_1), т. е. цикл Z должен выглядеть как на рис. 2.

В силу леммы 2 всегда существует требуемое на шаге 1 разбиение множества вершин на кластеры. В силу леммы 1 на шаге 3 существует возможность соединить концы и начала построенных цепей требуемым образом. Из леммы 3 следует, что цикл Z совпадает по виду с минимальным гамильтоновым циклом и, следовательно, сам является минимальным.

Оценим временну́ю сложность алгоритма. Шаг 1 (разбиение на кластеры) можно выполнить, обходя граф поиском в глубину по ребрам из $E \setminus E_{\max}$. Каждая компонента связности, выделяемая при таком обходе, соответствует кластеру. Поэтому временна́я сложность шага 1 совпадает со сложностью поиска в глубину и не превышает $O(|V|^2)$.

На шаге 2 процедура НайтиГамильтоновуЦепь вызывается k раз. Пусть $n_i = |V_i|$ $(i = 1, \ldots, k)$. Тогда временна́я сложность шага 2 оценивается как $O\left(\sum\limits_{i=1}^{k} {n_i}^2\right)$, что не превышает $O\left(\left(\sum\limits_{i=1}^{k} n_i\right)^2\right)$, т. е. $O\left(|V|^2\right)$. Очевидно, что шаг 3 также может быть выполнен за время $O(|V|^2)$. Теорема 2 доказана.

3.4 Задача Штейнера

Покажем, что в случае ультраметрической функции стоимости минимальное дерево Штейнера совпадает с минимальным остовным деревом подграфа, порожденного множеством терминальных вершин (т. е. графа, состоящего из множества терминальных вершин и множества всех ребер, соединяющих терминальные вершины друг с другом). Таким образом, минимальное дерево Штейнера для графа с ультраметрической функцией стоимости можно строить за полиномиальное время уже известными алгоритмами (например, алгоритмом Прима).

Теорема 3. *Пусть $G(V, E)$ — связный неориентированный граф, $X \subseteq V$ — множество терминальных вершин, $c : E \to R_+$ — ультраметрическая функция стоимости. Тогда на G существует минимальное дерево Штейнера, состоящее только из терминальных вершин. Если стоимости всех ребер графа строго положительны, то любое минимальное дерево Штейнера на G содержит только терминальные вершины.*

Доказательство. Пусть $T(V_T, E_T)$ — минимальное дерево Штейнера для G, c и X.

Предположим, что T содержит нетерминальную вершину s. Если $\deg_T(s) = 1$, то эту вершину вместе с инцидентным ей ребром $(s, v) \in E_T$ можно удалить из T, получив штейнеровское дерево меньшей или равной (при $c(s, v) = 0$) стоимости.

Рассмотрим случай $\deg_T(s) \geq 2$. Пусть v и w — две вершины, инцидентные s на T. Применим процедуру ПреобразоватьТреугольник(T, s, v, w). В результате получится новое дерево Штейнера T', также являющееся минимальным (в силу теоремы 1). В T' степень вершины s на 1 меньше, чем в T.

Последовательным применением подобных преобразований получим дерево Штейнера T'', в котором степень s равна 1 и $c(T'') \leq c(T)$. Но, удалив из T'' вершину s вместе с инцидентным ей ребром (s, v), получим дерево Штейнера T^*, для которого $c(T^*) \leq c(T'') \leq c(T)$. Таким образом, T^* — минимальное дерево Штейнера для того же множества терминальных вершин и T^* содержит на одну нетерминальную вершину (вершина s) меньше, чем исходное дерево T. Очевидно, что, применив подобное преобразование несколько раз, можно получить минимальное дерево Штейнера, содержащее только терминальные вершины.

Рассмотрим случай, когда стоимости всех ребер на графе G строго положительны. Тогда после удаления вершины s и ребра (s, v) в соответствии с описанной выше процедурой получим дерево Штейнера T^* такое, что $c(T^*) < c(T)$. А это противоречит тому, что исходное дерево T является минимальным по стоимости деревом Штейнера. Таким образом, методом от противного доказано, что T не может содержать нетерминальные вершины. Теорема 3 доказана.

Если все вершины графа являются терминальными, то минимальное дерево Штейнера совпадает с минимальным остовным деревом. Поэтому справедливо следующее утверждение.

Следствие. Для графа с ультраметрической функцией стоимости минимальное дерево Штейнера является минимальным остовным деревом подграфа, порожденного множеством терминальных вершин.

4 Заключение

В данной работе проведен анализ ультраметрических версий нескольких задач, являющихся NP-трудными в общем случае. Для замкнутой задачи коммивояжера приведен полиномиальный алгоритм решения. Для задачи построения минимального дерева Штейнера показано, что в ультраметрическом случае решение совпадает с минимальным остовным деревом для подграфа и, следовательно, может быть построено за полиномиальное время одним из ранее известных алгоритмов.

Автор благодарит Б. Я. Штейнберга за ценные замечания и предложения.

Литература

1. *Гэри М., Джонсон Д.* Вычислительные машины и труднорешаемые задачи / Пер. с англ. — М.: Мир, 1982. 416 с. (*Garey M. R., Johnson D. S.* Computers and intractability: A guide to the theory of NP-completeness. — New York: W. H. Freeman & Co, 1979. 338 p.)
2. *Farach M., Kannan S., Warnow T.* A robust model for finding optimal evolutionary trees // Algorithmica. Special Issue on Computational Biology, 1995. Vol. 13. No. 1. P. 155–179.
3. *Гасфилд Д.* Строки, деревья и последовательности в алгоритмах. Информатика и вычислительная биология / Пер. с англ. — СПб.: Невский Диалект, БХВ-Петербург, 2003. 654 с. (*Gusfield D.* Algorithms on strings, trees and sequences: Computer science and computational biology. — Cambridge: Cambridge University Press, 1997. 556 p. http://www.cs.ucdavis.edu/~gusfield/ultraerrat/ultraerrat.html.)
4. *Moore N. C. A., Proseer P.* The ultrametric constraint and its application to phylogenetics // J. Artif. Intell. Res., 2008. Vol. 32. P. 901–938.
5. *Хаубольд Б., Вие Т.* Введение в вычислительную биологию. Эволюционный подход / Пер. с англ. — М.–Ижевск: НИЦ «Регулярная и хаотическая динамика», Ижевский институт компьютерных исследований, 2011. 424 с. (*Haubold B., Wiehe T.* Introduction to computational biology: An evolutionary approach. — 3rd ed. — Basel: Birkhäuser, 2006. 328 p.)
6. *Li X., Plaxton C. G., Tiwari M., Venkataramani A.* Online hierarchical cooperative caching // SPAA'04: 16th Annual ACM Symposium on Parallelism in Algorithms and Architectures Proceedings. — New York: ACM, 2004. P. 74–83.
7. *Кинцель В.* Спиновые стекла как модельные системы для нейронных сетей // Успехи физических наук, 1987. Т. 152. С. 123–131.
8. *Кристофидес Н.* Теория графов. Алгоритмический подход / Пер. с англ. — М.: Мир, 1978. 432 с. (*Christofides N.* Graph theory: An algorithmic approach (Computer science and applied mathematics). — New York: Academic Press, 1975. 400 p.)
9. *Майника Э.* Алгоритмы оптимизации на сетях и графах / Пер. с англ. — М.: Мир, 1981. 323 с. (*Meinika E.* Optimization algorithms on networks and graphs. — New York – Basel: Dekker, 1978. 356 p.)
10. *Bui T. N., Zrncic C. M.* An ant-based algorithm for finding degree-constrained minimum spanning tree // GECCO'06: 8th Annual Conference on Genetic and Evolutionary Computation Proceedings. — New York: ACM, 2006. P. 11–18.

Поступила в редакцию 30.07.13

ON POLYNOMIAL TIME COMPLEXITY OF ULTRAMETRIC VERSIONS OF CERTAIN NP-HARD PROBLEMS

M. G. Adigeev

Southern Federal University, 105/42 Bol'shaya Sadovaya Str., Rostov-on-Don 344006, Russian Federation

Abstract: The paper deals with important special cases of the travelling salesman problem and the Steiner tree problem. Both of these problems are NP-hard even in the metric case, i. e., for graphs whose edge cost function meets the triangle inequality. Even more severe restriction is imposed by the **strong** triangle inequality: $\forall x, y, z \in X \quad c(x, z) \leq \max\{c(x, y), c(y, z)\}$. The function which meets this inequality is called **ultrametric**. The analysis of graphs with an ultrametric edge cost function is presented. This analysis leads to an algorithm for building the minimal cost Hamiltonian cycle in time $O\left(n^2\right)$ where n is the number of vertices. For the Steiner tree problem, it is proven that in the case of an ultrametric edge function, the minimum Steiner tree includes only terminal vertices and thus may also be constructed in polynomial time, as a minimum spanning tree on a subgraph of the original graph.

Keywords: ultrametric function; strong triangle inequality; travelling salesman problem; Steiner tree; polynomial-time algorithms

DOI: 10.14375/19922264140207

Acknowledgments

The author thanks B. Ya. Shteinberg for his valuable comments and suggestions.

References

1. Garey, M. R., and D. S. Johnson. 1979. *Computers and intractability: A guide to the theory of NP-completeness*. New York: W. H. Freeman & Co. New York. 338 p.
2. Farach, M., S. Kannan, and T. Warnow. 1995. A robust model for finding optimal evolutionary trees. *Algorithmica*. Special Issue on Computational Biology. 13(1):155–179.
3. Gusfield, D. 1997. *Algorithms on strings, trees and sequences: Computer science and computational biology*. Cambridge: Cambridge University Press. 556 p. http://www.cs.ucdavis.edu/~gusfield/ultraerrat/ultraerrat.html.
4. Moore, N. C. A., and P. Proseer. 2008. The ultrametric constraint and its application to phylogenetics. *J. Artif. Intell. Res*. 32:901–938.
5. Haubold, B., and T. Wiehe. 2006. *Introduction to computational biology: An evolutionary approach*. 3rd ed. Basel: Birkhäuser. 328 p.
6. Li, X., C. G. Plaxton, M. Tiwari, and A. Venkataramani. 2004. Online hierarchical cooperative caching. *16th Annual ACM Symposium on Parallelism in Algorithms and Architectures (SPAA'04) Proceedings*. New York. 74–83.
7. Kintsel' V. 1987. Spinovye stekla kak model'nye sistemy dlya neyronnykh setey. [Spin glasses as model systems for neural networks]. *Uspekhi Fizicheskikh Nauk* [Advances in Physical Sciences] 152:123–131.
8. Christofides, N. 1975. *Graph theory: An algorithmic approach (Computer science and applied mathematics)*. New York: Academic Press. 400 p.
9. Meinika, E. 1978. *Optimization algorithms on networks and graphs*. New York – Basel: Dekker. 356 p.
10. Bui, T. N., and C. M. Zrncic. 2006. An ant-based algorithm for finding degree-constrained minimum spanning tree. *GECCO'06: 8th Annual Conference on Genetic and Evolutionary Computation Proceedings*. New York. 11–18.

Received July 30, 2013

Contributor

Adigeev Mikhail G. (b. 1973) — Candidate of Science (PhD) in technology, associate professor, Southern Federal University, 105/42 Bol'shaya Sadovaya Str., Rostov-on-Don 344006, Russian Federation; madi@math.sfedu.ru

ИНФОРМАТИКА И ЕЁ ПРИМЕНЕНИЯ, 2014. Т. 8. Вып. 2. С. 77–85

РЕШЕНИЕ ОБРАТНОЙ ЗАДАЧИ В МНОГОДИПОЛЬНОЙ МОДЕЛИ ИСТОЧНИКОВ МАГНИТОЭНЦЕФАЛОГРАММ МЕТОДОМ НЕЗАВИСИМЫХ КОМПОНЕНТ*

В. Е. Бенинг[1], М. А. Драницына[2], Т. В. Захарова[3], П. И. Карпов[4]

Аннотация: Настоящая работа посвящена изучению функциональных зон головного мозга человека. Функциональное картирование коры головного мозга является чрезвычайно сложной задачей, возникновение которой обусловлено современным уровнем развития методов неинвазивного исследования головного мозга. Магнитоэнцефалография (МЭГ), один из таких современных неинвазивных методов, — очень мощный инструмент, обладающий научным и прикладным медицинским потенциалом. Результатом проведения МЭГ являются большие массивы данных, несущие информацию о процессах, происходящих в головном мозге. В ходе обработки этих данных перед исследователем ставится некорректная обратная задача, заключающаяся в пространственной реконструкции источников МЭГ-сигналов в коре головного мозга человека с заданной точностью. На настоящий момент не существует универсальных инструментов для точного в достаточной степени решения такой обратной задачи при анализе МЭГ-сигналов. Одному и тому же распределению потенциалов на поверхности головы могут соответствовать различные зоны активности коры головного мозга. Однако при некоторых предположениях: источники потенциала дискретные, относятся к различным функциональным областям мозга, располагаются относительно неглубоко, — задача имеет однозначное решение. В данной работе предполагается, что МЭГ-сигнал представляет собой суперпозицию сигналов мультидиполей. Решение обратной задачи в таком случае называется многодипольным приближением. Нахождение источников активности проходит в два этапа: на первом методом независимых компонент производится декомпозиция исходных МЭГ-сигналов на конечное число независимых компонент; на втором по аналитической формуле рассчитываются координаты однодипольного источника активности для каждой отдельной независимой компоненты.

Ключевые слова: метод независимых компонент; нормальное распределение; токовый диполь; многодипольная модель; магнитоэнцефалограмма

DOI: 10.14375/19922264140208

1 Введение

Головной мозг человека — это орган центральной нервной системы. Он состоит из большого числа (до 200 млрд) нейронов, связанных между собой особыми связями, превращающими наш мозг во взаимосвязанную сеть. Взаимодействуя посредством этих связей, нейроны формируют электрические импульсы, которые управляют деятельностью всего организма. Ввиду высокой сложности организации мозга его работа до сих пор является недостаточно изученной областью.

Магнитоэнцефалография — это новый неинвазивный метод исследования активности головного мозга [1, 2]. Интерес к МЭГ в мире очень высок.

Начиная с 1992 г. финская компания Elekta Neuromag Oy занимается разработкой программного обеспечения в области МЭГ. В 2000 г. в США был образован Martinos Center for Biomedical Imaging при Массачусетском технологическом институте. В июле 2005 г. был запущен проект по компьютерному моделированию коры головного мозга человека BlueBrainProject. Над ним совместно работают компания IBM и Швейцарский федеральный технический институт Лозанны. С 2007 г. в Кембридже в одном из крупнейших центров по изучению психологии Medical Research Council Cognitionand Brain Sciences Unit (MRC) начинает действовать МЭГ-лаборатория. Исследовательская работа по изучению головного мозга ведется на медицинских фа-

* Работа выполнена при поддержке РНФ (проект 14-11-00364).

[1] Московский государственный университет им. М. В. Ломоносова, факультет вычислительной математики и кибернетики; Институт проблем информатики Российской академии наук, bening@yandex.ru

[2] Московский государственный университет им. М. В. Ломоносова, факультет вычислительной математики и кибернетики, margarita13april@mail.ru

[3] Московский государственный университет им. М. В. Ломоносова, факультет вычислительной математики и кибернетики, lsa@cs.msu.ru

[4] Национальный исследовательский технологический университет «МИСиС», karpov.petr@gmail.com

культетах: в старейшем университете Швейцарии в Базеле, в Швеции в Гетеборгском университете, в Хельсинкском технологическом университете.

В России впервые МЭГ-центр «Научно-образовательный центр нейрокогнитивных исследований» был создан в 2008 г. в Москве. И с 2011 г. на базе этого центра на факультете ВМК МГУ им. М. В. Ломоносова стали проводиться исследования по обработке МЭГ-сигналов. При этом одной из важнейших ставилась задача точной локализации областей активности нейронов [3].

Наиболее сложной является проблема повышения точности локализации первичной моторной коры (M1) и конкретно области представительства руки в зоне M1. Эта задача решалась при помощи метода вызванных потенциалов и построения ассоциативного фильтра [4, 5]. Был рассмотрен и иной статистический подход, основанный на различных способах кластеризации мозга [6, 7].

В данной работе решается обратная задача по локализации источников следующим образом. В случае однодипольного источника имеется аналитическое решение обратной задачи нахождения координат диполя. Авторами рассматривается многодипольная модель с конечным числом областей активности нейронов. Предлагается метод сведения многодипольной модели к решению обратной задачи для некоторого числа разных функциональных однодипольных источников. Это стало возможным в связи с применением метода независимых компонент (ICA — Independent Component Analysis) [8], который разделил смешанный МЭГ-сигнал на разные функциональные компоненты. Далее для каждой такой компоненты рассчитываются координаты источника мозговой активности.

2 Электромагнитное поле, создаваемое нейронной активностью

С физической точки зрения мозговая активность описывается с помощью классической электродинамики сплошных сред. Динамика электромагнитного поля определяется уравнениями Максвелла в среде [9], которые в системе СИ записываются следующим образом:

$$\left.\begin{aligned} \nabla \times \mathbf{H} &= \mathbf{j} + \frac{\partial D}{\partial t}\,; \\ \nabla \times \mathbf{E} &= -\frac{\partial B}{\partial t}\,; \\ \nabla \cdot \mathbf{B} &= 0\,; \\ \nabla \cdot \mathbf{D} &= \rho\,, \end{aligned}\right\} \tag{1}$$

где $\mathbf{H}$ — напряженность магнитного поля; $\mathbf{E}$ — напряженность электрического поля; $\mathbf{B}$ — магнитная индукция; $\mathbf{D}$ — электрическая индукция; а также материальными уравнениями, в которых заложены свойства среды:

$$\left.\begin{aligned} \mathbf{D} &= \epsilon\mathbf{E}\,; \\ \mathbf{B} &= \mu\mathbf{H}\,; \\ \mathbf{j} &= \sigma\mathbf{E}\,, \end{aligned}\right\} \tag{2}$$

где ϵ и μ — диэлектрическая и магнитная проницаемость среды соответственно; σ — проводимость среды; $\mathbf{j}$ — плотность электрического тока.

Для исследования мозговой активности используется два стандартных приближения [10]. Во-первых, считается, что магнитная проницаемость всех тканей головы совпадает с магнитной проницаемостью вакуума: $\mu = \mu_0$. Во-вторых, используется приближение квазистатического магнитного поля, при котором в уравнениях Максвелла (1) можно пренебречь всеми производными по времени, т. е. в любой момент времени электрическое и магнитное поля определяются мгновенным распределением всех зарядов и токов в системе, как если бы они были стационарными.

В квазистатическом приближении поле $\mathbf{E}$ оказывается безвихревым, поэтому можно ввести скалярный электрический потенциал

$$\mathbf{E} = -\nabla\varphi \tag{3}$$

и для расчета магнитного поля использовать закон Био–Савара–Лапласа:

$$\mathbf{B}(\mathbf{r}) = \frac{\mu_0}{4\pi}\int \frac{\mathbf{j}(\mathbf{r}')\times\mathbf{R}}{R^3}\,d^3r'\,, \tag{4}$$

где $\mathbf{r}$ — радиус-вектор точки, в которой вычисляется магнитное поле, интегрирование ведется по $\mathbf{r}'$ — всем точкам источника; $\mathbf{R} = \mathbf{r} - \mathbf{r}'$.

Плотность тока $\mathbf{j}$, создаваемого нейронной активностью, можно разделить на две компоненты: первичный ток и объемный (пассивный) ток:

$$\mathbf{j} = \mathbf{j}^p + \mathbf{j}^v\,. \tag{5}$$

Первичный ток — это ток, создаваемый непосредственно нейронной активностью (т. е. градиентами химического потенциала), и его распределение сильно локализовано в небольшой области головного мозга. Такое локализованное распределение плотности тока удобно моделировать с помощью понятия токового диполя, для которого плотность тока задается δ-функцией Дирака:

$$\mathbf{j}^p = \mathbf{Q}\delta(\mathbf{r} - \mathbf{r}_\mathbf{Q})\,, \tag{6}$$

где $\mathbf{Q}$ — дипольный момент токового диполя, расположенного в точке $\mathbf{r}_\mathbf{Q}$. В данной статье рассматриваются только токовые диполи (в отличие, например, от магнитных диполей), поэтому далее токовые диполи будут называться просто диполями. Заметим, что на клеточном уровне один диполь может порождаться большим количеством микроскопических первичных токов (называемых в этом случае возбужденными), которые вызываются десятками тысяч синхронно активируемых больших пирамидальных нейронов коры головного мозга [2]; несмотря на это, такие возбуждения хорошо аппроксимируются однодипольной моделью. Если одновременно проявляется активность в нескольких хорошо локализованных зонах головного мозга, то можно пользоваться многодипольным приближением [11]:

$$\mathbf{j}^p = \sum_{i=1}^{N} \mathbf{Q}_i \delta(\mathbf{r} - \mathbf{r}_{\mathbf{Q}_i}) \,,$$

где N — число диполей, каждый из которых имеет дипольный момент $\mathbf{Q}_i$ и расположен в точке $\mathbf{r}_{\mathbf{Q}_i}$.

Объемный ток создается макроскопическим электрическим полем и обеспечивает локальную электронейтральность, которую стремится нарушить первичный ток. Таким образом, объемный ток определяется материальным уравнением (2):

$$\mathbf{j}^v(\mathbf{r}) = \sigma(\mathbf{r})\mathbf{E}(\mathbf{r}) \,. \tag{7}$$

В отличие от первичного, объемный ток (порожденный некоторыми первичными токами) распределен по всему объему головы.

Применив разделение тока на первичный и вторичный (5) к закону Био–Савара–Лапласа (4), а также воспользовавшись уравнениями (3) и (7), получим

$$\mathbf{B}(\mathbf{r}) = \frac{\mu_0}{4\pi} \int (\mathbf{j}^p(\mathbf{r}') - \sigma(\mathbf{r}')\nabla\varphi(\mathbf{r}')) \times \frac{\mathbf{R}}{R^3}\, d^3r' \,,$$

что после преобразований можно записать следующим образом:

$$\mathbf{B}(\mathbf{r}) = \frac{\mu_0}{4\pi} \int (\mathbf{j}^p(\mathbf{r}') + \varphi(\mathbf{r}')\nabla'\sigma(\mathbf{r}')) \times \frac{\mathbf{R}}{R^3}\, d^3r' \,, \tag{8}$$

где оператор ∇' относится к переменной $\mathbf{r}'$.

В следующем разделе применяется общая теория электромагнитного поля, создаваемого нейронной активностью, для случая однодипольной сферической модели.

3 Обратная задача для однодипольной сферической модели головы

В данной работе рассматривается сферическая модель головы. Эта модель является самой простой и позволяет получить аналитические решения, но в то же время она улавливает большинство принципиальных эффектов. Эта модель дает не только качественные, но и даже неплохие количественные предсказания, если источник сигнала не расположен в непосредственной близости от центра аппроксимирующей сферы [12]. С небольшими уточнениями эту модель можно использовать для построения более реалистичных моделей головы [13].

В данной работе голова моделируется сферой радиуса R с центрально-симметричным распределением проводимости $\sigma(r)$. В такой системе магнитное поле вне головы можно вычислить, используя только первичные токи и не рассматривая вторичные [10].

Рассмотрим радиальную компоненту магнитного поля $B_r = \mathbf{B} \cdot \mathbf{e}_r$ ($\mathbf{e}_r = \mathbf{r}/r$ — единичный вектор, сонаправленный с радиус-вектором). Из формулы (8) следует, что вклад объемных токов в B_r равен нулю, так как векторы $\nabla'\sigma \sim \mathbf{r}'$, $\mathbf{R} = \mathbf{r} - \mathbf{r}'$ и $\mathbf{e}_r$ компланарны, поэтому их смешанное произведение равно нулю. Следовательно, в проекции на радиус-вектор в формулу (8) дает вклад только первичный ток:

$$\mathbf{B}(\mathbf{r}) = \frac{\mu_0}{4\pi} \int \mathbf{j}^p(\mathbf{r}') \times \frac{\mathbf{R}}{R^3}\, d^3r' \,. \tag{9}$$

Теперь рассмотрим однодипольную модель. Если источником является один диполь (6), то формула (9) превращается в

$$\mathbf{B}_\mathbf{r} = -\frac{\mu_0}{4\pi} \frac{\mathbf{Q} \times \mathbf{r}_\mathbf{Q}}{|\mathbf{r} - \mathbf{r}_\mathbf{Q}|^3}\, \mathbf{e}_r \,. \tag{10}$$

Так как радиальная компонента диполя не дает вклада в B_r вне головы, то, не теряя общности, можно считать, что диполь имеет ориентацию $\mathbf{Q} = Q\mathbf{e}_x$ и расположен в точке $\mathbf{r}_\mathbf{Q} = r_Q\mathbf{e}_z$ (рис. 1). Тогда формула (10) запишется следующим образом [13]:

$$B_r = -\frac{\mu_0}{4\pi} \frac{Qr_Q r \sin\theta\cos\phi}{(r^2 + r_Q^2 - 2rr_Q\cos\theta)^{3/2}} \,.$$

Отсюда видно, что $B_r = 0$ во всей плоскости $\phi = \pm\pi/2$. Найдем точки, в которых B_r достигает локальных экстремумов. В этих точках $\cos\phi = \pm 1$.

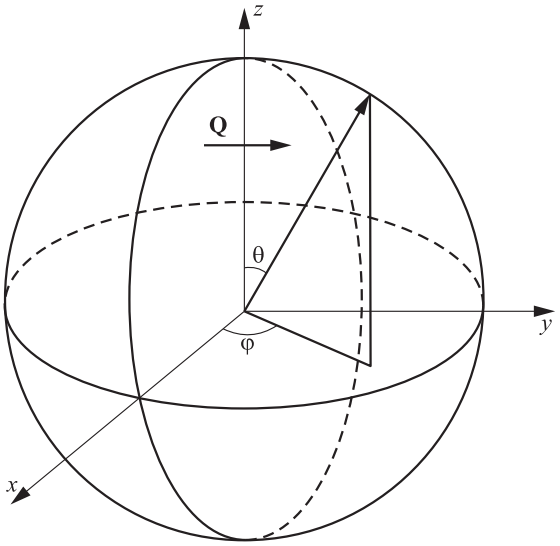


Рис. 1 Геометрия задачи

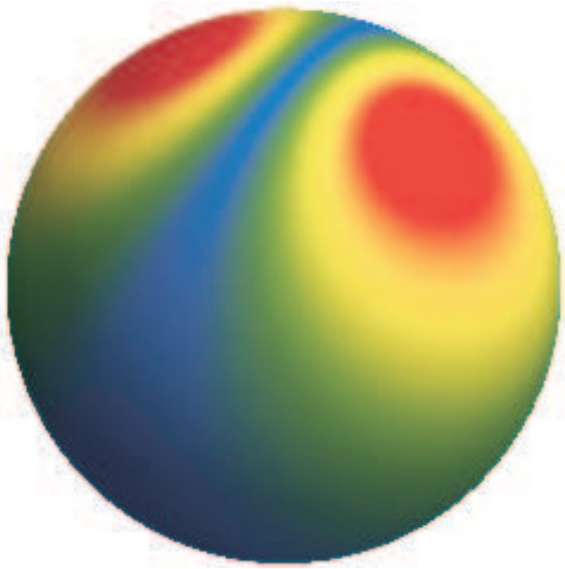


Рис. 2 Обратная задача в однодипольной сферической модели. Красным цветом показаны области максимума $|B_r|$. Диполь, создающий данное магнитное поле, должен лежать в плоскости симметрии системы

Находя локальный экстремум по углу θ, получим экстремальные значения угла:

$$\cos\theta = \frac{-(R^2 + r_Q^2) + \sqrt{(R^2 + r_Q^2)^2 + 12R^2 r_Q^2}}{2Rr_Q}. \quad (11)$$

Этот результат для прямой задачи дает возможность решить обратную задачу в предположении, что источником сигнала является один диполь. Если разрешить (11) относительно r_Q, получим

$$r_Q = R\frac{(3 - \cos^2\theta) - \sqrt{(3 - \cos^2\theta)^2 - 4\cos^2\theta}}{2\cos\theta}. \quad (12)$$

Таким образом, в однодипольной модели обратная задача решается следующим способом. Нужно найти на поверхности сферы две точки с максимальным значением радиальной компоненты магнитного поля $B_r = \pm B_{\max}$ (рис. 2). Тогда источник будет лежать в плоскости симметрии этих точек на расстоянии r_Q от центра сферы, которое задается формулой (12).

В следующих разделах полученный результат будет применен к многодипольной модели, которую методом ICA можно разделить на независимые однодипольные источники.

4 Метод независимых компонент

Метод независимых компонент является методом декомпозиции смеси случайных функций [8]. В рассматриваемой задаче метод ICA раскладывает регистрируемые МЭГ-сигналы в линейную комбинацию независимых случайных компонент.

4.1 Математическая модель метода независимых компонент

Пусть существует n случайных величин $x_1, \ldots, x_n$, каждая из которых представляет собой линейную комбинацию n случайных величин $s_1, \ldots, s_n$:

$$x_i = a_{i1}s_1 + a_{i2}s_2 + \cdots + a_{in}s_n, \;\; i = 1, \ldots, n\,,$$

где a_{ij} — некоторые действительные числа при $i, j = 1, \ldots, n$.

По определению s_i взаимно независимы (независимые компоненты). Модель ICA описывает, как данные наблюдений могут генерироваться в процессе смешивания независимых компонент s_i, и в этом смысле данная модель является порождающей. При этом, независимые компоненты s_i не могут наблюдаться непосредственно в ходе эксперимента, также неизвестными являются и смешивающие коэффициенты a_{ij}. Следовательно, для решения задачи есть только случайный вектор наблюдений, при этом, применяя ICA, требуется определить как независимые компоненты, так и смешивающие коэффициенты исходя из максимально общих предположений.

Зачастую удобнее использовать матричное представление модели. Пусть x — случайный вектор (вектор-столбец), $x = (x_1, \ldots, x_n)^{\mathrm{T}}$ и s — случайный вектор (вектор-столбец), $s = (s_1, \ldots, s_n)^{\mathrm{T}}$. Определим матрицу A с элементами $\{a_{ij}\}$. Тогда, используя векторные обозначения, можно записать модель ICA в виде

$$x = As\,.$$

4.2 Ограничения метода

Введем некоторые ограничения для обеспечения возможности оценки независимых компонент:

- статистическая независимость компонент;
- распределение независимых компонент обязано быть отличным от гауссовского.

Заметим, что в исходной модели не предполагается какое-либо известное распределение независимых компонент. Но в том случае, если это распределение известно, задача может быть значительно упрощена.

Для простоты изложения метода будем считать квадратной и обратимой смешивающую матрицу A.

Оценив матрицу A, можно вычислить обратную к ней разделяющую матрицу B и искомые независимые компоненты s:

$$s = Bx\,.$$

Без потери общности также будем полагать, что наблюдения x и независимые компоненты s имеют нулевое среднее.

4.3 Неопределенность метода

Отметим следующие особенности метода ICA:

- в представленной модели невозможно оценить дисперсии независимых компонент, причиной является одновременная неопределенность смешивающей матрицы и независимых компонент s (модель не изменится, если поделить и умножить на скалярную величину столбец a_i матрицы и s_i соответственно);
- в представленной модели невозможно оценить порядок независимых компонент, т. е. любую из найденных независимых компонент можно обозначить как первой, так и n-й.

4.4 Основная идея метода независимых компонент

Пусть имеется вектор наблюдений и согласно общей модели ICA он представляет собой линейную комбинацию независимых компонент:

$$x = As\,.$$

Будем считать, что независимые компоненты одинаково распределены. Для вычисления независимых компонент необходимо, с учетом обратимости смешивающей матрицы, разрешить уравнение:

$$s = A^{-1}x\,.$$

Определим векторы $y = b^{\mathrm{T}}x = \sum\limits_i b_i x_i = b^{\mathrm{T}}As$ (вектор b рассчитывается специальным образом и будет определен ниже) и $q = A^{\mathrm{T}}b$. Тогда можно записать

$$y = b^{\mathrm{T}}x = q^{\mathrm{T}}s = \sum_i q_i s_i\,.$$

Если вектор b^{T} будет совпадать с одной из строк обратной матрицы A^{-1} (допустим, k-й), тогда скалярное произведение $b^{\mathrm{T}}x$ совпадет с k-й независимой компонентой s_k. Понятно, что вектор q тогда будет иметь только одну ненулевую компоненту, k-ю и $q_k = 1$.

Но матрица A неизвестна, и поэтому точно рассчитать вектор b невозможно. Попробуем оценить вектор b, руководствуясь следующими рассуждениями.

Примем во внимание тот факт, что сумма независимых одинаково распределенных случайных величин имеет распределение, более близкое к гауссовскому, чем каждая из этих случайных величин сама по себе. Тогда случайная величина $y = b^{\mathrm{T}}x = q^{\mathrm{T}}s$ имеет распределение, максимально далекое от гауссовского в том случае, если случайная величина равна одной из независимых компонент s_i. Можно выбрать в качестве b вектор, который максимизирует негауссовость $y = b^{\mathrm{T}}x$. Этот вектор определяет вектор $q = A^{\mathrm{T}}b$ с единственной ненулевой компонентой, а вектор $y = b^{\mathrm{T}}x = q^{\mathrm{T}}s$ соответствует одной из независимых компонент. Таким образом, максимизация меры негауссовости $b^{\mathrm{T}}x$ позволяет получить одну из независимых компонент.

Решая задачу максимизации негауссовости по n-мерному вектору b, получают $2n$ локальных максимумов, по 2 максимума на каждую независимую компоненту: со знаком плюс и минус (s_i и $-s_i$).

4.5 Меры негауссовости

В качестве меры негауссовости наиболее часто при расчетах используют коэффициент эксцесса и негэнтропию [8]. Рассмотрим каждую из них в отдельности.

4.5.1 Эксцесс

Эксцесс случайной величины y с учетом нулевого среднего можно рассчитать по формуле:

$$\mathrm{kurt}(y) = \mathrm{E}\{y^4\} - 3(\mathrm{E}\{y^2\})^2.$$

Основные преимущества использования эксцесса как меры негауссовости — это простота вычислений и теоретических выкладок, однако коэффициент эксцесса не устойчив к выбросам.

Одним из методов нахождения локального максимума является градиентный метод. Это метод нахождения локального экстремума функции с помощью движения вдоль градиента. В данном случае

градиентный алгоритм с использованием эксцесса может быть записан в виде:

$$\triangle\, w \sim \mathrm{sign}(\mathrm{kurt}(w^{\mathrm{T}}z))\mathrm{E}\{z(w^{\mathrm{T}}z)^3\}\,. \qquad (13)$$

Для максимизации эксцесса $\mathrm{kurt}(y)$ с заданным вектором наблюдений z выбирается некоторый начальный вектор w, рассчитывается направление, по которому абсолютное значение эксцесса случайной величины $y = w^{\mathrm{T}}z$ растет наиболее быстро (по формуле (13)), и далее вектор w сдвигается в этом направлении. Решение состоит в построении последовательности векторов w, увеличивающих эксцесс $\mathrm{kurt}(y)$, т. е. меру негауссовости.

4.5.2 Негэнтропия

Рассмотрим еще одну меру негауссовости. Для этого введем следующее определение.

Отбеливание — это процесс получения случайного вектора с нулевым математическим ожиданием, компоненты которого не коррелированы и имеют единичные дисперсии.

Отбеливание часто используется в качестве предварительной обработки данных и состоит в линейном преобразовании случайного вектора x следующим образом: $z = Vx = VAs$, при этом $\mathrm{E}\{zz^{\mathrm{T}}\} = \mathrm{I}$. В данном случае также предполагаем, что наблюдаемый сигнал x прошел процедуру отбеливания.

Понятие негэнтропии (от *англ.* negative entropy) основывается на информационно-теоретических свойствах дифференциальной энтропии (далее — энтропии).

Одним из основных результатов теории информации является экстремальность гауссовской случайной величины в том смысле, что она имеет наибольшую энтропию среди всех случайных величин с одинаковыми дисперсиями. Отметим, что малой энтропией обладают случайные величины с островершинной плотностью распределения. Следовательно, энтропия может служить мерой негауссовости.

Наиболее подробно понятие и свойства энтропии изложены в [14].

Чтобы получить меру негауссовости, неотрицательную и равную нулю для гауссовских случайных величин, вводят понятие негэнтропии. Обозначим негэнтропию как J и определим ее по формуле

$$J(y) = H(y_{\mathrm{gauss}}) - H(y)\,,$$

где y_{gauss} — нормальный случайный вектор с такой же матрицей корреляции, как и у вектора y, а энтропия H случайного вектора y с плотностью $p_y(x)$ определяется как

$$H = -\int p_y(x)\log p_y(x)\,dx\,.$$

Использование негэнтропии как показателя негауссовости оправдано теоретическими соображениями, но, в отличие от эксцесса, негэнтропия обладает высокой сложностью вычислений.

Обычно в вычислениях используют некую аппроксимацию негэнтропии. Используя моменты высоких порядков, можно получить аппроксимацию вида

$$J(y) \approx \frac{1}{12}\left(\mathrm{E}\{y^3\}\right)^2 + \frac{1}{48}\left(\mathrm{kurt}(y)\right)^2\,.$$

Но такая оценка не устойчива к выбросам и характеризует в основном хвосты распределения, не отражая особенности распределения около его центра.

В случаях, когда известна некоторая информация о характере плотности распределения, аппроксимацию негэнтропии получают при помощи метода максимума энтропии.

Например, при помощи только одной неквадратичной функции G можно построить следующее приближение:

$$J(y) \approx \left[\mathrm{E}\{G(y)\} - \mathrm{E}\{G(z)\}\right]^2\,, \qquad (14)$$

где z — нормальная случайная величина с нулевым математическим ожиданием и единичной дисперсией. При этом, выбирая не слишком быстро возрастающую функцию G, можно получить надежную оценку негэнтропии.

На практике при выборе функции G руководствуются следующими требованиями.

1. Оценивание $\mathrm{E}G(X)$ не должно быть сложным статистически и оценка должна быть устойчивой к выборосам.
2. Функция $G(x)$ не должна расти быстрее, чем $|x|^2$.
3. Функция $G(x)$ должна отражать особенности распределения X.

Далее максимизируют негэнтропию, воспользовавшись ее аппроксимацией (14). Для этого применяется градиентный метод. В этом случае алгоритм может быть записан в виде:

$$\triangle\, w \sim \left(\mathrm{E}\{G(w^{\mathrm{T}}z)\} - \mathrm{E}\{G(v)\}\right)\mathrm{E}\{zg(w^{\mathrm{T}}z)\}\,;$$

$$w := \frac{w}{||w||}\,,$$

где v — случайная величина со стандартным нормальным распределением; функция g — производная функции G.

5 Заключение

Магнитоэнцефалография — это неинвазивный метод исследования функционирования головного мозга. Этот метод, при условии внедрения высокоточных математических методов обработки и интерпретации полученных сигналов, в перспективе может стать ключевым инструментом исследования в нейронауках. С помощью магнитоэнцефалографа на поверхности головы фиксируется магнитная активность нейронов, а затем на основе этих данных решается обратная задача по локализации самих источников активности.

Очевидна ценность метода как в научных исследованиях, так и в реальной клинической практике. Так, в ходе нейрохирургических вмешательств могут быть повреждены невосполнимые зоны головного мозга, что ведет к развитию необратимого нарушения различных функций (например, речевых, двигательных). Так как расположение функциональных зон в мозге человека индивидуально, для врача крайне важно иметь инструмент по локализации с высокой точностью этих областей.

Метод локализации, представленный в данной работе, разработан для простой модели: количество источников конечно, источники фиксированы, источники относятся к разным функциональным областям. В дальнейшем предполагается, усложняя модель исследования и привлекая суперкомпьютерную технику, приблизиться в смысле модели к реальному самому сложному органу центральной нервной системы – головному мозгу и решить задачу в максимально реальных условиях.

Литература

1. *Sarvas J.* Basic mathematical and electromagnetic concepts of the biomagnetic inverse problem // Physics in Medicine and Biology, 1987. Vol. 32. No. 1. P. 11–22.
2. *Baillet S., Mosher J. C., Leahy R. M.* Electromagnetic brain mapping // IEEE Signal Processing Magazine, 2001. Vol. 7. No. 2. P. 14–30.
3. *Friston K., Harrison L., Daunizeau J., Kiebel S., Phillips C., Trujillo-Barreto N., Henson R., Flandin G., Mattoutf J.* Multiple sparse priors for the M/EEG inverse problem // NeuroImage, 2008. Vol. 39. No. 4. P. 1104–1120.
4. *Захарова Т. В., Никифоров С. Ю., Гончаренко М. Б., Драницына М. А., Климов Г. А., Хазиахметов М. Ш., Чаянов Н. В.* Методы обработки сигналов для локализации невосполнимых областей головного мозга // Системы и средства информатики, 2012. Т. 22. Вып. 2. С. 157–176.
5. *Хазиахметов М. Ш., Захарова Т. В.* Об алгоритмах нахождения опорных точек миограммы для использования в локализации невосполнимых областей головного мозга // Статистические методы оценивания и проверки гипотез: Межвузовский сборник научных трудов, 2013. Т. 25. С. 56–63.
6. *Бенинг В. Е., Горшенин А. К., Королев В. Ю.* Асимптотически оптимальный критерий проверки гипотез о числе компонент смеси вероятностных распределений // Информатика и её применения, 2011. Т. 5. Вып. 3. С. 4–16.
7. *Захарова Т. В., Гончаренко М. Б., Никифоров С. Ю.* Метод решения обратной задачи магнитоэнцефалографии, основанный на кластеризации поверхности мозга // Статистические методы оценивания и проверки гипотез: Межвузовский сборник научных трудов, 2013. Т. 25. С. 120–125.
8. *Hyvärinen A., Karhunen J., Oja E.* Independent component analysis. — New York: John Wiley & Sons, 2001. 504 p.
9. *Landau L. D., Pitaevskii L. P., Lifshitz E. M.* Electrodynamics of continuous media. — New York: Pergamon, 1984. 432 p.
10. *Hämäläinen M., Hari R., Ilmoniemi R. J. Knuutila J., Lounasmaa O. V.* Magnetoencephalography — theory, instrumentation, and applications to noninvasive studies of the working human brain // Rev. Modern Phys., 1993. Vol. 65. No. 1. P. 413–497.
11. *Mosher J. C., Lewis P. S., Leahy R. M.* Multiple dipole modeling and localization from spatio-temporal MEG data // IEEE Trans. Biomedical Eng., 1992. Vol. 39. No. 6. P. 541.
12. *Uitert R., Weinstein D., Johnson C.* Can a spherical model substitute for a realistic head model in forward and inverse MEG simulations? // Biomag 2002: 13th Conference (International) on Biomagnetism Proceedings. — Berlin, Offenbach: VDE Verlag, 2002. P. 798–800.
13. *Ilmoniemi R. J., Hamalainen M. S., Knuutila J.* The forward and inverse problems in the spherical model // Biomagnetism: Applications and theory / Eds. H. Weinberg, G. Stroink, T. Katila. — New York: Pergamon, 1985. P. 278–282.
14. *Королев В. Ю., Бенинг В. Е., Шоргин С. Я.* Математические основы теории риска. — М.: Физматлит, 2011. 620 с.

Поступила в редакцию 3.05.14

INDEPENDENT COMPONENT ANALYSIS FOR THE INVERSE PROBLEM IN THE MULTIDIPOLE MODEL OF MAGNETOENCEPHALOGRAM'S SOURCES

V. E. Bening[1,2], M. A. Dranitsyna[1], T. V. Zakharova[1], and P. I. Karpov[3]

[1]Department of Mathematical Statistics, Faculty of Computational Mathematics and Cybernetics, M. V. Lomonosov Moscow State University, 1-52 Leninskiye Gory, GSP-1, Moscow 119991, Russian Federation
[2]Institute of Informatics Problems, Russian Academy of Sciences, 44-2 Vavilov Str., Moscow 119333, Russian Federation
[3]Department of Theoretical Physics and Quantum Technologies, College of New Materials and Nanotechnology, National University of Science and Technology "MISiS," 4 Leninskiy Prosp., Moscow, Russian Federation

Abstract: This paper is devoted to a challenging task of brain functional mapping which is posed due to the current techniques of noninvasive human brain investigation. One of such techniques is magnetoencephalography (MEG) which is very potent in the scientific and practical contexts. Large data retrieved from the MEG procedure comprise information about brain processes. Magnetoencephalography data processing sets a highly ill-posed problem consisting in spatial reconstruction of MEG-signal sources with a given accuracy. At the present moment, there is no universal tool for accurate solution of such inverse problem. The same distribution of potentials on the surface of a human head may be caused by activity of different areas within cerebral cortex. Nevertheless, under certain assumptions, this task can be solved unambiguously. The assumptions are the following: signal sources are discrete, belong to distinct functional areas of the brain, and have superficial location. The MEG-signal obtained is assumed to be a superposition of multidipole signals. In this case, the solution of the inverse problem is a multidipole approximation. The algorithm proposed assumes two main steps. The first step includes application of independent component analysis to primary/basic MEG-signals and obtaining independent components, the second step consists of treating these independent components separately and employing an analytical formula to them as for monodipole model to get the isolated signal source location for each component.

Keywords: independent component analysis; normal distribution; current dipole; multidipole model; magnetoencephalogram

DOI: 10.14375/19922264140208

Acknowledgments

The work was supported by the Russian Scientific Foundation (project 14-11-00364).

References

1. Sarvas, J. 1987. Basic mathematical and electromagnetic concepts of the biomagnetic inverse problem. *Phys. Medicine Biol.* 32(1):11–22.
2. Baillet, S., J. C. Mosher, R. M. Leahy. 2001. Electromagnetic brain mapping. *IEEE Signal Processing Magazine*. 7(2):14–30.
3. Friston, K., L. Harrison, J. Daunizeau, S. Kiebel, Ch. Phillips, N. Trujillo-Barreto, R. Henson, G. Flandin, and J. Mattoutf. 2008. Multiple sparse priors for the M/EEG inverse problem. *NeuroImage* 39:1104–1120.
4. Zakharova, T. V., S. Yu. Nikiforov, M. B. Goncharenko, M. A. Dranitsyna, G. A. Klimov, M. S. Khaziakhmetov, and N. V. Chayanov. 2012. Metody obrabotki signalov dlya lokalizatsii nevospolnimykh oblastey golovnogo mozga [Signal processing methods for the localization of non-renewable brain regions]. *Sistemy i Sredstva Informatiki — Systems and Means of Informatics* 22(2):157–176.
5. Khaziakhmetov, M. S., and T. V. Zakharova. 2013. Ob algoritmakh nakhozhdeniya opornykh tochek miogrammy dlya ispol'zovaniya v lokalizatsii nevospolnimykh oblastey golovnogo mozga [Algorithms for myogram reference points search with the aim of irrecoverable brain regions localization]. *Statisticheskie Metody Otsenivaniya i Proverki Gipotez. Mezhvuzovskiy Sbornik Nauchnykh Trudov* [Statistical methods for estimating and hypothesis testing. Interuniversity Collection of Research Papers]. Perm. 25:56–63.
6. Bening, V. Ye., A. K. Gorshenin, and V. Yu. Korolev. 2011. Asimptoticheski optimal'nyy kriteriy proverki gipotez o chisle komponent smesi veroyatnostnykh raspredeleniy

[Asymptotically optimum hypothesis test for number of components in mixture of probability distribution]. *Informatika i ee Primeneniya — Inform. Appl.* 5(3):4–16.

7. Zakharova, T. V., M. B. Goncharenko, and S. Yu. Nikiforov. 2013. Metod resheniya obratnoy zadachi magnitoentsefalografii, osnovannyy na klasterizatsii poverkhnosti mozga [Inverse problem solving method based on clustering of brain surface]. *Statisticheskie Metody Otsenivaniya i Proverki Gipotez. Mezhvuzovskiy Sbornik Nauchnykh Trudov* [Statistical methods for estimating and hypothesis testing. Interuniversity Collection of Research Papers]. Perm. 25:120–125.
8. Hyvärinen, A., J. Karhunen, and E. Oja. 2001. *Independent component analysis*. New-York: John Wiley & Sons. 504 p.
9. Landau, L. D., L. P. Pitaevskii, and E. M. Lifshitz. 1984. *Electrodynamics of continuous media*. New York: Pergamon. 432 p.
10. Hämäläinen, M., R. Hari, R. J. Ilmoniemi, J. Knuutila, and O. V. Lounasmaa. 1993. Magnetoencephalography — theory, instrumentation, and applications to noninvasive studies of the working human brain. *Rev. Modern Phys.* 65:413–497.
11. Mosher, J. C., P. S. Lewis, and R. M. Leahy. 1992. Multiple dipole modeling and localization from spatio-temporal MEG Data. *IEEE Trans. Biomedical Eng.* 39(6):541.
12. Uitert, R., D. Weinstein, and C. Johnson. 2002. Can a spherical model substitute for a realistic head model in forward and inverse MEG simulations? *13th Conference (International) on Biomagnetism Proceedings*. Jena, Germany. 798–800.
13. Ilmoniemi, R. J., M. S. Hamalainen, and J. Knuutila. 1985. The forward and inverse problems in the spherical model. *Biomagnetism: Applications and theory*. Eds. Weinberg, H., G. Stroink, and T. Katila. New York: Pergamon. 278–282.
14. Korolev, V. Yu., V. E. Bening, and S. Ya. Shorgin. 2011. Matematicheskie osnovy teorii riska [Mathematical basics of risk theory]. Moscow: Fizmatlit. 620 p.

Received May 3, 2014

Contributors

Bening Vladimir E. (b. 1954) — Doctor of Science in physics and mathematics; professor, Department of Mathematical Statistics, Faculty of Computational Mathematics and Cybernetics, M. V. Lomonosov Moscow State University, 1-52 Leninskiye Gory, GSP-1, Moscow 119991, Russian Federation; senior scientist, Institute of Informatics Problems, Russian Academy of Sciences, 44-2 Vavilov Str., Moscow 119333, Russian Federation; bening@yandex.ru

Dranitsyna Margarita A. (b. 1983) — PhD student, Department of Mathematical Statistics, Faculty of Computational Mathematics and Cybernetics, M. V. Lomonosov Moscow State University, 1-52 Leninskiye Gory, GSP-1, Moscow 119991, Russian Federation; margarita13april@mail.ru

Zakharova Tatyana V. (b. 1962) — Candidate of Science (PhD) in physics and mathematics, senior lecturer, Department of Mathematical Statistics, Faculty of Computational Mathematics and Cybernetics, M. V. Lomonosov Moscow State University, Moscow State University, 1-52 Leninskiye Gory, GSP-1, Moscow 119991, Russian Federation; lsa@cs.msu.ru

Karpov Peter I. (b. 1990) — PhD student, Department of Theoretical Physics and Quantum Technologies, College of New Materials and Nanotechnology, National University of Science and Technology "MISiS," 4 Leninskiy Prosp., Moscow, Russian Federation; karpov.petr@gmail.com

ИНФОРМАТИКА И ЕЁ ПРИМЕНЕНИЯ, 2014. Т. 8. Вып. 2. С. 86–97

ПОСТРОЕНИЕ АГРЕГИРОВАННЫХ ПРОГНОЗОВ ОБЪЕМОВ ЖЕЛЕЗНОДОРОЖНЫХ ГРУЗОПЕРЕВОЗОК С ИСПОЛЬЗОВАНИЕМ РАССТОЯНИЯ КУЛЬБАКА–ЛЕЙБЛЕРА*

А. П. Мотренко[1], В. В. Стрижов[2]

Аннотация: Данное исследование посвящено проблеме построения агрегированных прогнозов объемов железнодорожных грузоперевозок. Для получения агрегированных прогнозов требуется кластеризовать временны́е ряды таким образом, чтобы распределения временны́х рядов внутри кластера совпадали. При решении задачи кластеризации требуется оценить близость между временны́ми рядами, исходя из их эмпирических распределений. Вводится критерий принадлежности временны́х рядов одному распределению, основанный на расстоянии Кульбака–Лейблера между гистограммами временны́х рядов. Приводится теоретическое и практическое исследование предложенного критерия. Решается задача кластеризации временны́х рядов на основе матрицы парных расстояний между ними.

Ключевые слова: эмпирическая функция распределения; расстояние между гистограммами; расстояние Кульбака–Лейблера; задача двух выборок

DOI: 10.14375/19922264140209

1 Введение

Особенностью задачи прогнозирования объема погрузок по историческим данным о загруженности железнодорожной сети различными группами грузов является необходимость определить оптимальный уровень детализации [1, 2]: по видам перевозимых грузов, по наборам станций, по кодам вагонов. Требуется получить прогноз как для объема погрузок в целом, так и для отдельных групп грузов. При этом спрогнозированный объем погрузок в целом может не совпадать с суммой прогнозов по отдельным группам. Для повышения согласованности полученных прогнозов предлагается [3] вместо прогноза «в целом» объединять ряды только в том случае, если их распределения совпадают, чтобы агрегированные данные имели тот же статистический смысл, что и исходные ряды. Для решения задачи агрегации временны́х рядов необходимо определить расстояние между временны́ми рядами таким образом, чтобы оно отражало близость эмпирических распределений между рядами.

В литературе по математической статистике вводится множество коэффициентов, показывающих, что некоторые два распределения P и Q близки друг к другу. Такие коэффициенты в различных источниках называются расстоянием между распределениями [4], мерами разделяющей информации [5], мерами статистического расстояния [6]. В работе [7] описан метод порождения коэффициентов $d(P, Q)$ «непохожести» двух распределений, обладающих некоторыми стандартными свойствами, например:

- коэффициент $d(P, Q)$ должен быть определен на всех парах распределений с одним носителем;
- значение $d(P, Q)$ должно быть минимально при $P = Q$;
- при любом измеримом преобразовании носителя распределений P и Q расстояние между ними не увеличивается.

Идея метода [7] заключается в том, чтобы рассмотреть различные выпуклые функции случайной величины Q/P. С точки зрения распределения P математическое ожидание Q/P независимо от Q, а дисперсия стремится к нулю при $Q \to P$. Также в [7] показано, что многие известные функции расстояния могут быть получены этим методом. В частности, им могут быть порождены все f-дивергенции [8] и в том числе расстояние Кульбака–Лейблера [4]. В работе [9] приведено сравнение многих известных расстояний с точки зрения скорости сходимости эмпирического распределения к истинному, а также качественного поведения функции расстояния при сходимости. При решении задачи кластеризации в обработке изображений были введены меры [10, 11], основанные на метрике Вассерштейна.

* Работа выполнена при поддержке РФФИ (грант 13-07-13139).

[1] Московский физико-технический институт, anastasia.motrenko@gmail.com

[2] Вычислительный центр Российской академии наук им. А. А. Дородницына, strijov@ccas.com

В работе [1] для оценки близости распределений используется расстояние Кульбака–Лейблера между гистограммами, построенными по временны́м рядам. В данной работе показано, что распределение расстояние Кульбака–Лейблера между гистограммами из одного распределения в пределе ограничено сверху распределением χ^2.

Предложен критерий для решения задачи двух выборок, основанный на расстоянии Кульбака–Лейблера между гистограммами временны́х рядов. Продемонстрировано применение критерия к решению задачи двух выборок для различных пар распределений и показана его состоятельность. Для набора временны́х рядов о железнодорожных грузоперевозках решается задача кластеризации с помощью алгоритма кратчайшего незамкнутого пути [12] на основе матрицы парных расстояний [13, 14] между рядами. При решении задачи кластеризации ряды группируются по типу груза.

2 Постановка задачи

Задан набор временны́х рядов $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_S\}$, где каждый ряд $\mathbf{x}_j = \{x_j(i) \in \mathbb{R}\}_{i=1}^{m_j}$ — это последовательность реализаций некоторого стационарного случайного процесса. Требуется кластеризовать набор

$$\mathbf{X} = \bigsqcup_{k=1}^{K} \mathbf{X}_k\,, \quad \mathbf{X}_k = \{\mathbf{x}_j,\, j \in \mathcal{A}_k\}\,,$$
$$\{1, \ldots, S\} == \bigsqcup_{k=1}^{K} \mathcal{A}_k\,, \quad (1)$$

разбив набор $\mathbf{X}$ на K наборов $\mathbf{X}_k$ временны́х рядов таких, что все ряды в $\mathbf{X}_k$ принадлежат одному и тому же распределению. Здесь $\mathcal{A}_k$ — множество индексов временны́х рядов k-го кластера.

В силу стационарности случайного процесса $\mathbf{x}$ пренебрежем последовательностью значений ряда $\mathbf{x}$. Представим временной ряд $\mathbf{x}$ как выборку X реализаций некоторой случайной величины с распределением P:

$$X = \{x \in \mathbb{R} |\text{ для некоторого } i \in \{1, \ldots, m\} : x(i) = x\}\,. \quad (2)$$

Для решения задачи об агрегировании временны́х рядов $\mathbf{x}$ и $\mathbf{x}'$ будем сравнивать гистограммы, построенные по выборкам X и X', сопоставленным каждому из рядов в соответствии с (2). Опишем подробнее процедуру построения гистограммы.

Пусть объемы выборок X и X' равны m и m' соответственно. Разобьем область значений случайной величины из P на N промежутков $[a_i, a_{i+1}]$ и обозначим $p_i = P(a_i < x \le a_{i+1})$ вероятность случайной величине с распределением P принять значение из i-го промежутка; n_i и n'_i — количество объектов выборок X и X', попавших в i-й промежуток. Обозначим $\hat{P}_m$ гистограмму, построенную по выборке X объема m из распределения P. Гистограмма $\hat{P}_m$ задается набором оценок

$$\hat{P}_m(a_i < x \le a_{i+1}) = \frac{n_i}{m} = \hat{p}_i\,, \quad i = 1, \ldots, N-1\,, \quad (3)$$

вероятности p_i.

Для решения задачи кластеризации (1) воспользуемся алгоритмом нахождения кратчайшего незамкнутого пути между временными рядами. Результатом применения алгоритма является минимальное остовное дерево: граф с $n-1$ ребрами, покрывающий все n вершин, ребра которого обладают минимальной суммарной длиной. Удалив из минимального остовного дерева $K-1$ самых длинных ребер, получим кластеризацию вершин графа на K кластеров $\mathbf{X}_k$. Вершинами графа являются исследуемые временны́е ряды $\mathbf{x}_j$; длина ребра, соединяющего две вершины, равна расстоянию между соответствующими временны́ми рядами. Найдя расстояние между всеми парами рядов, получим матрицу парных расстояний D. В качестве расстояний между рядами $\mathbf{x}_r$ и $\mathbf{x}_s$ будем использовать симметризованное расстояние Кульбака–Лейблера между гистограммами $\hat{P}_{m_r}$ и $\hat{P}_{m_s}$, построенными по временны́м рядам:

$$D(r,s) = \frac{2m_r m_s}{m_r + m_s}\Big(D_{\mathrm{KL}}(\hat{P}_{m_r}||\hat{P}_{m_s}) + D_{\mathrm{KL}}(\hat{P}_{m_s}||\hat{P}_{m_r})\Big). \quad (4)$$

Первый сомножитель в правой части снимает зависимость от объема выборки. Необходимость его введения будет объяснена в следующем разделе.

Конечной целью кластеризации временны́х рядов с учетом расстояний между ними является повышение согласованности агрегированных прогнозов. Для оценки качества кластеризации будем рассматривать несогласованность

$$\delta(i) = \left| \sum_{k}^{K} \hat{\mathbf{X}}_k(i) - \sum_{j}^{n} \hat{\mathbf{x}}_j(i) \right| \quad (5)$$

при прогнозировании по наборам $\mathbf{X}_k$ временны́х рядов и отдельным временны́м рядам $\mathbf{x}_j$. Здесь $\hat{\mathbf{X}}_k(i)$ — прогноз агрегированного ряда в момент i; $\hat{x}_j(i)$ — прогноз j-го ряда в момент времени i. Чем

меньше несогласованность, тем качественнее выполнена кластеризация. Очевидно, что наименьшее значение $\delta_{\min} = 0$ выражения (5) достигается при $K = S$, поэтому предлагается ограничить число K или ввести в (5) штраф $h(K)$ за его повышение:

$$K = \arg\min_K \left(\sum_{i=1}^{m} \delta(i) + h(K) \right).$$

В данной работе ограничимся рассмотрением предложенного критерия принадлежности временны́х рядов к одному распределению и кластеризации временны́х рядов на основе расстояния между ними.

3 Статистическая значимость расстояния Кульбака–Лейблера

Чтобы показать, что результаты кластеризации временны́х рядов на основе расстояния Кульбака–Лейблера между ними статистически значимы, необходимо исследовать распределение расстояния Кульбака–Лейблера между гистограммами, построенными по выборкам X и X' из одного распределения P. В данном разделе будет показано, что, хотя расстояние Кульбака–Лейблера не имеет предельного распределения, для него можно получить предельные оценки сверху.

Пусть пока выборки X и X' имеют одинаковый объем m. Рассмотрим расстояние $D_{\mathrm{KL}}(\hat{P}_m||\hat{P}'_m)$ между гистограммами $\hat{P}_m$ и $\hat{P}'_m$. По определению расстояние Кульбака–Лейблера $D_{\mathrm{KL}}(Q||P)$ между распределениями Q и P равно

$$D_{\mathrm{KL}}(Q||P) = \int P f\left(\frac{Q}{P}\right), \tag{6}$$

где $f(t) = t \ln t$. Функция f строго выпукла и дважды дифференцируема в единице, и, повторяя рассуждения из [8], разложим подынтегральное выражение из правой части (6) по f в окрестности единицы:

$$P(x) f\left(\frac{Q(x)}{P(x)}\right) = f(1) + f'(1)(Q(x) - P(x)) + \\ + \frac{f''(1)}{2} \frac{(Q(x) - P(x))^2}{P(x)} + P(x) o\left(\left(\frac{Q(x)}{P(x)} - 1\right)^3\right),$$

где $f(1) = 0$; $f''(1) = 1$. Подставив вместо Q распределение $\hat{P}_m$, определяемое (3), и просуммировав по i, получим соотношение:

$$D_{\mathrm{KL}}(\hat{P}_m||P) = \sum_{i=1}^{N} p_i f\left(\frac{\hat{p}_i}{p_i}\right) = \\ = \frac{1}{2} \sum_{i}^{N} \frac{(\hat{p}_i - p_i)^2}{p_i} + \sum_{i=1}^{N} p_i \cdot \varepsilon \left(\left(\frac{\hat{p}_i}{p_i} - 1\right)^3\right) \sim \\ \sim \frac{1}{2m} \sum_{i}^{N} \frac{(n_i - m p_i)^2}{m p_i}$$

и следующий предельный переход:

$$2m D_{\mathrm{KL}}(\hat{P}_m||Q) \sim m \sum_{i=1}^{N} \frac{(\hat{p}_i - p_i)^2}{p_i} = \\ = \sum_{i=1}^{N} \frac{(n_i - m p_i)^2}{m p_i} \to \chi^2_{\mathcal{N}} \text{ при } m \to \infty. \tag{7}$$

Докажем следующую теорему:

Теорема 1. *Случайная величина* $2m D_{\mathrm{KL}}(Q||\hat{P}_m) \to \chi^2_{\mathcal{N}}$ *по распределению при* $m \to \infty$.

Доказательство. Аналогично доказательству предельного перехода (7) разложим $D_{\mathrm{KL}}(Q||\hat{P}_m)$ по степеням $f(t)$ вблизи единицы и получим:

$$D_{\mathrm{KL}}(Q||\hat{P}_m) \sim \frac{1}{2} \sum_{i=1}^{N} \frac{(\hat{P}_m(\xi_i) - Q(\xi_i))^2}{\hat{P}_m(\xi)} = \\ = \frac{1}{2m} \sum_{i=1}^{N} \frac{(n_i - m p_i)^2}{n_i}.$$

Пусть $G_m(x)$ — функция распределения случайной величины $\sum\limits_{i=1}^{N} (n_i - m p_i)^2 / (m p_i)$; $F_m(x)$ — случайной величины $\sum\limits_{i=1}^{N} (n_i - m p_i)^2 / n_i$. Так как $G_m(x)$ сходится поточечно к $F_{\chi^2_{N-1}}$ при $m \to \infty$, имеем:

$$\left| G_m(x) - F_{\chi^2_{N-1}} \right| < \frac{\epsilon}{2} \quad \forall m > m'.$$

Докажем, что $|G_m(x) - F_m(x)| \to 0$ при $m \to \infty$. Для этого покажем, что $\forall \varepsilon > 0$ найдется объем выборки m_0 такой, что для всех $m > m_0$ выполняется

$$\mathrm{P}\left(\left| \frac{(n_i - m p_i)^2}{n_i} - \frac{(n_i - m p_i)^2}{m p_i} \right| \le \frac{\varepsilon}{N} \right) > 1 - \varepsilon. \tag{8}$$

Согласно центральной предельной теореме,

$$\frac{n_i - m p_i}{p_i (1 - p_i) \sqrt{m}} \to \mathcal{N}(0, 1)$$

по распределению при $m \to \infty$, причем для скорости сходимости имеет место неравенство Берри–Эссеена:

$$|Q_m(x) - \Phi(x)| \le \frac{A}{\sqrt{m}},$$

где $Q_m(x)$ — функция распределения величины $(n_i - mp_i)/(p_i(1-p_i)\sqrt{m})$; $\Phi(x)$ — функция стандартного нормального распределения; A — некоторая константа. Тогда вероятность

$$\mathrm{P}\left(\left|\frac{n_i - mp_i}{p_i(1-p_i)\sqrt{m}}\right| < C\right) = Q_m(C) - Q_m(-C) \geq \geq 2\Phi(C) - 1 - \frac{2A}{\sqrt{m}}. \quad (9)$$

Пусть, кроме того, выполняется $0 < 1 - p \leq p_i \leq \leq p < 1$. Тогда с вероятностью $\mathrm{P}_C \geq 2\Phi(C) - 1$ выполняется

$$\left|\frac{(n_i - mp_i)^2}{n_i} - \frac{(n_i - mp_i)^2}{mp_i}\right| = \frac{|n_i - mp_i|^3}{mn_ip_i} \leq \leq \frac{C^3(1-p_i)^3p_i^2}{n_i}\sqrt{m} \leq \frac{C^3(1-p)^3p}{\sqrt{m} - C(1-p)}.$$

Обозначим $m_1 = [4C^2(1-p)^2]$, тогда при $m > m_1$ имеет место $\sqrt{m} - C(1-p_i) > (1/2)\sqrt{m}$ и

$$\left|\frac{(n_i - mp_i)^2}{n_i} - \frac{(n_i - mp_i)^2}{mp_i}\right| \leq \frac{2C^3(1-p_i)^3p_i}{\sqrt{m}}.$$

Тогда для фиксированного ε определим

$$C_\varepsilon = \frac{\epsilon^{1/3}m^{1/6}}{(1-p_i)(2p_iN)^{1/3}};$$

$$\mathrm{P}_m(\varepsilon) = 2\Phi(C_\varepsilon) - 1 - \frac{2A}{\sqrt{m}}.$$

При заданном ε вероятность $\mathrm{P}_m(\varepsilon) \to 1$ при $m \to \to \infty$, поэтому найдется m_2 такое, что для любого $m > m_2$ выполнено $\mathrm{P}_m(\varepsilon) > 1 - \varepsilon$. Выбрав $m_0 = \max(m_1, m_2)$, получим утверждение (8). Тогда

$$\left|\sum_{i=1}^{N}\frac{(n_i - mp)^2}{n_i} - \frac{(n_i - mp)^2}{mp}\right| \leq$$

$$\leq \sum_{i=1}^{N}\left|\frac{(n_i - mp)^2}{n_i} - \frac{(n_i - mp)^2}{mp}\right| < \varepsilon \text{ при } m > m_0.$$

Из только что доказанного следует, что $|F_m(x) - G_m(x)| \to 0$ при $m \to 0$. Тогда $\forall\epsilon > 0\ \exists m'' :$ при $m > m''$ выполняется

$$\left|F_m(x) - F_{\chi^2_{N-1}}\right| < |F_m(x) - G_m(x)| + + \left|G_m(x) - F_{\chi^2_{N-1}}\right| < \frac{\epsilon}{2} + \frac{\epsilon}{2}.$$

Доказанная теорема и утверждение (7) позволяют получить оценки распределения случайных величин $2mD_{\mathrm{KL}}(\hat{P}_m||Q)$ и $2mD_{\mathrm{KL}}(Q||\hat{P}_m)$ при больших m. Для решения задачи кластеризации (1) потребуется также исследовать поведение расстояния Кульбака–Лейблера $D_{\mathrm{KL}}(\hat{P}_m||\hat{P}_l)$ между гистограммами, построенными по выборкам X и X' различных длин m и l. Воспользовавшись неравенством треугольника

$$D_{\mathrm{KL}}(\hat{P}_m||\hat{P}_l) \leq D_{\mathrm{KL}}(\hat{P}_m||Q) + D_{\mathrm{KL}}(Q||\hat{P}_l),$$

получим следствия из теоремы 1.

Следствие 1. $2mD_{\mathrm{KL}}(\hat{P}_m||\hat{P}'_m) \leq \chi^2_{2\mathcal{N}}$ в пределе при $m \to \infty$.

Следствие 2. Пусть выборки X, X' растут таким образом, что $m/l \to \rho$, $0 < \rho < \infty$. Тогда

$$2\frac{ml}{m+l}D_{\mathrm{KL}}(\hat{P}_m||\hat{P}_l) \leq \chi^2_{2\mathcal{N}}$$

в пределе при $m, l \to \infty$.

Доказательство. Действительно, при выполнении условия $m/l \to \rho$, $0 < \rho < \infty$, имеем

$$\frac{l}{m+l} \to \frac{1}{1+\rho}; \quad \frac{m}{m+l} \to \frac{\rho}{1+\rho};$$

$$2\frac{ml}{m+l}D_{\mathrm{KL}}(\hat{P}_m||\hat{P}_l) \leq \frac{l}{m+l}2mD_{\mathrm{KL}}(\hat{P}_m||Q) + + \frac{m}{m+l}2lD_{\mathrm{KL}}(Q||\hat{P}_l) \to \chi^2_{2\mathcal{N}}.$$

Обозначим величину $(2ml/(m+l))D_{\mathrm{KL}}(\hat{P}_m||\hat{P}'_m)$ через $\xi_{m,l}$. Следствие 2 дает верхнюю оценку поведения случайной величины $\xi_{m,l}$ при больших m и l, а именно: пусть $\eta \sim \chi^2_{2\mathcal{N}}$, тогда при достаточно больших m и l для любого элементарного исхода w из вероятностного пространства Ω выполнено $\xi_{m,l}(w) < \eta(w)$. Следовательно, для любого $x \in \mathbb{R}$ верно

$$\mathrm{P}(\xi_{m,l} < x) \geq \mathrm{P}(\eta < x). \quad (10)$$

В следующем разделе покажем, как этот факт будет использоваться для проверки принадлежности временны́х рядов к одному распределению.

4 Проверка принадлежности временны́х рядов к одному распределению

Для решения задачи агрегирования временны́х рядов $\mathbf{x}$ и $\mathbf{x}'$ необходимо уметь принимать решение о принадлежности временны́х рядов к одному распределению. Опишем процедуру проверки гипотезы о принадлежности выборок X и X', составленных (2) из временны́х рядов $\mathbf{x}$ и $\mathbf{x}'$. Пусть нулевая

гипотеза H_0 состоит в принадлежности выборок X и X' к одному распределению:

$$H_0 : P(x) = P'(x)\,.$$

Сформулируем критерий проверки гипотезы H_0 при альтернативе $H_1 : P(x) \neq P'(x)$. Для этого определим критическую область $U(\alpha)$ для статистики $t_{m,l}$ с уровнем значимости α:

$$U(\alpha) = \{t : \bar{t}_{1-\alpha} > t \text{ или } t > \bar{t}_\alpha\}\ ,$$

где критическое значение $\bar{t}_\alpha$ определяется соотношением

$$\mathrm{P}(t > \bar{t}_\alpha | H_0) = \alpha\,. \tag{11}$$

Так как предельное распределение величины $\xi_{m,l}$ неизвестно, будем использовать критическую область, задаваемую распределением $\chi^2_{2\mathcal{N}}$. Будем говорить, что данные отвергают гипотезу H_0 в случае, если статистика $t_{m,l}$ принадлежит критической области

$$U^{\chi^2}(\alpha) = \left\{t : \bar{t}^{\chi^2}_{1-\alpha} > t \text{ или } t > \bar{t}^{\chi^2}_\alpha\right\}, \tag{12}$$

где $\bar{t}^{\chi^2}$ — критическое значение величины $\chi^2_{2\mathcal{N}}$:

$$\mathrm{P}(t > \bar{t}^{\chi^2}_\alpha | t \sim \chi^2_{2\mathcal{N}}) = \alpha\,.$$

Из неравенства (10) и определения (11) критических значений следует, что критические области U и U^{χ^2} не сравнимы, т. е.

$$\bar{t}_{1-\alpha} < \bar{t}^{\chi^2}_{1-\alpha}\,; \quad \bar{t}_\alpha < \bar{t}^{\chi^2}_\alpha\,.$$

Это означает, что возможны следующие ситуации.

1. Случай $\bar{t}^{\chi^2}_{1-\alpha} < t_{m,l} < \bar{t}_\alpha$, когда статистика $t_{m,l}$ одновременно принадлежит истинной, но неизвестной критической области U и вычислимой критической области U^{χ^2}.
2. Случай $\bar{t}_{1-\alpha} < t_{m,l} < \bar{t}^{\chi^2}_{1-\alpha}$, когда статистика $t_{m,l}$ принадлежит истинной, но неизвестной критической области U и не принадлежит U^{χ^2}. Так как $t_{m,l} \in U$, то с высокой вероятностью гипотеза H_0 неверна и есть риск принять неверное решение об истинности гипотезы H_0. Таким образом, зазор между $\bar{t}_{1-\alpha}$ и $\bar{t}^{\chi^2}_{1-\alpha}$ повышает вероятность ошибки второго рода.
3. Случай $\bar{t}_\alpha < t_{m,l} < \bar{t}^{\chi^2}_\alpha$, когда статистика $t_{m,l}$ попадает в U^{χ^2}, хотя на самом деле $t_{m,l}$ не принадлежит U. В этом случае велика вероятность, что H_0 верна, но решение будет принято в пользу H_1. Таким образом, зазор между $\bar{t}_\alpha$ и $\bar{t}^{\chi^2}_\alpha$ повышает вероятность ошибки первого рода.

Второй случай разрешается следующим образом: использование симметризованного расстояния позволяет перейти от двусторонних критериев U и U^{χ^2} вида (12) к односторонним критериям:

$$U_1(\alpha) = \{t : t > \bar{t}_\alpha\}\ ; \quad U^{\chi^2}_1(\alpha) = \left\{t : t > \bar{t}^{\chi^2}_\alpha\right\}.$$

В этом случае $U^{\chi^2}_1 \subseteq U_1$ и справедливо следствие:

$$t_{m,l} \in U^{\chi^2}_1 \Rightarrow t_{m,l} \in U_1\,.$$

Кроме того, далее будет показано (теорема 2), что при увеличении объема выборки m вероятность отклонить гипотезу H_0 с помощью критерия (12) в случае, если гипотеза H_0 неверна, стремится к единице. Влияние третьего случая на возможность применения критерия (12) для принятия нулевой гипотезы исследуется экспериментально. Эксперименты, приведенные ниже и в разд. 5, показывают, что при истинности нулевой гипотезы области U и U^{χ^2} достаточно близки для принятия верного решения.

Пример применения критерия (12) при истинности H_0. На рис. 1 изображены гистограммы для двух выборок из стандартного нормального распределения (рис. 1, *а* и 1, *б*) и стандартного нормального распределения с шумом $\varepsilon \sim 0{,}1R[0,1]$ (рис. 1, *в* и 1, *г*), а также зависимость расстояния Кульбака–Лейблера $D_{\mathrm{KL}}(\hat{P}_m || \hat{P}'_m)$ между дискретными распределениями, задаваемыми гистограммами $\hat{P}_m$ и $\hat{P}'_m$, между выборками одинакового объема m от объема выборки m и область допустимых значений с точки зрения критерия (12) (рис. 1, *д* и 1, *е*). Здесь вместо критических значений $\bar{t}^{\chi^2}_{1-\alpha}$, $\bar{t}^{\chi^2}_\alpha$ и $t_m = 2mD_{\mathrm{KL}}$ отложены величины $\bar{t}^{\chi^2}_{1-\alpha}/(2m)$, $\bar{t}^{\chi^2}_\alpha/(2m)$ и $t_m/(2m)$, чтобы продемонстрировать масштаб расстояния Кульбака–Лейблера и наличие сходимости. Рисунки показывают, что в данном случае использование распределения $\chi^2_{2\mathcal{N}}$ в качестве оценки предельного распределения статистики t_m позволяет принять верное решение о принадлежности рядов к одному распределению. Кривые *1* и *3* показывают границу области, в которую вошло $1 - \alpha = 90\%$ выборки, и задают оценку критической области для $D_{\mathrm{KL}}(\hat{P}_m || \hat{P}'_m)$. Более подробно результаты описаны в разд. 5.

Покажем теперь, что критерий (12) также можно использовать для отвержения гипотезы H_0.

Теорема 2. *Критерий* (12) *состоятелен:*

$$\lim_{m\to\infty} \mathrm{P}(t_m \in U | H_1) = 1\,,$$

т. е. вероятность отвергнуть гипотезу H_0, если распределения временны́х рядов X и X' не совпадают, с увеличением выборки стремится к единице.

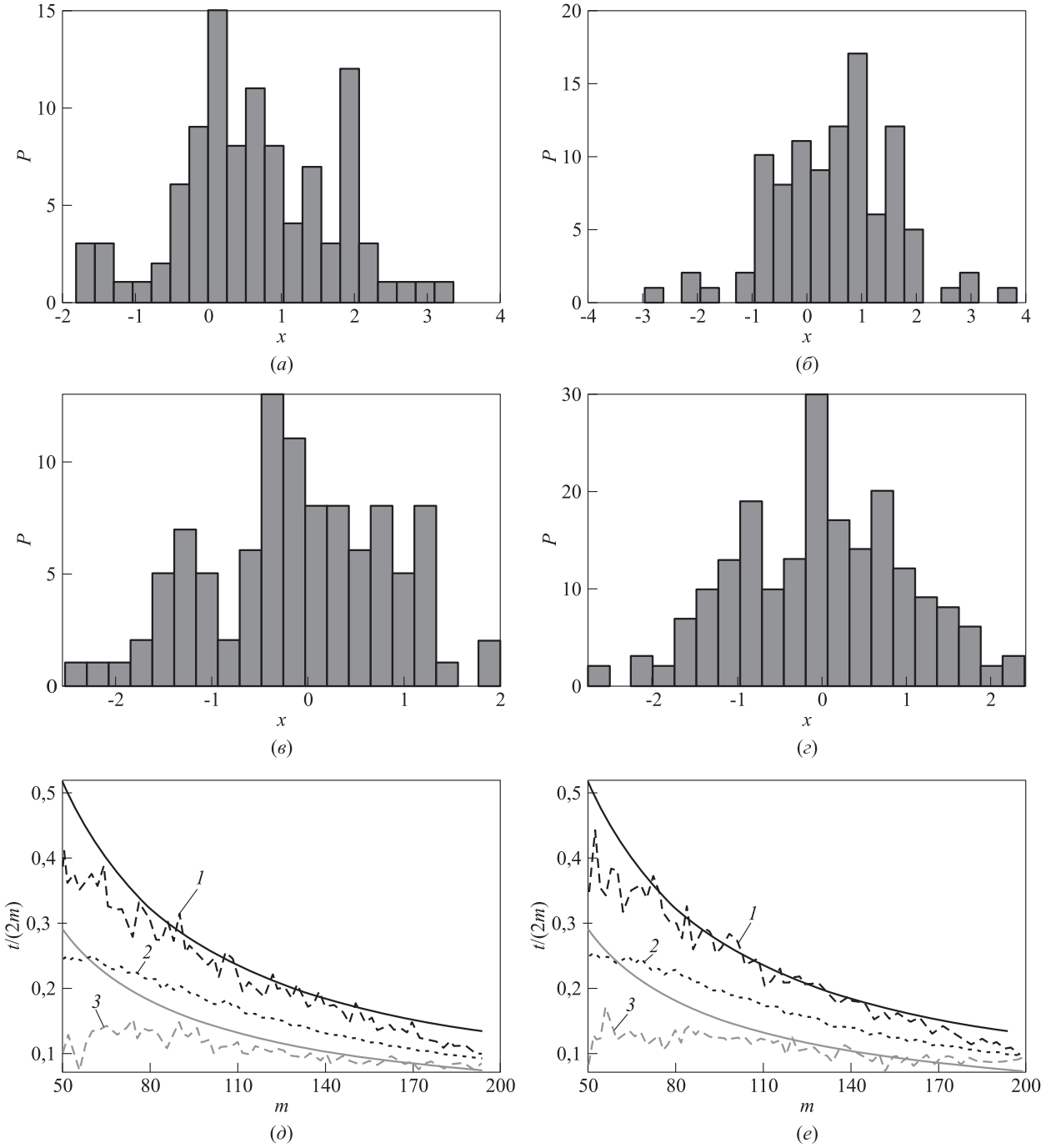


Рис. 1 Гистограммы, построенные по двум выборкам из нормального распределения (*а*, *б*) и зашумленного (*в*, *г*) и зависимость статистики t_m от объема выборки (*д*, *е*) (сплошными кривыми отмечены границы доверительного интервала для t_m при $\alpha = 0{,}1$): *1* — $\overline{t}^{\chi^2}_{1-\alpha}/(2m)$; *2* — $\overline{t}^{\chi^2}_{\alpha}/(2m)$; *3* — $t_m/(2m)$

Доказательство. Пусть функции распределения P и P' временны́х рядов не совпадают. Тогда найдется $x^* \in \mathbb{R}$, при котором значения этих функций различны: $P(x^*) \neq P'(x^*)$. Следовательно, найдется такой способ разбиения пространства $\mathbb{R}$, что для некоторого i вероятность попадания в i-й промежуток не одинакова для рассматриваемых случайных величин:

$$P(a_i < x \leq a_{i+1}) = p_i \neq p'_i = P'(a_i < x \leq a_{i+1})\,.$$

Пусть $p_i > p'_i$. Согласно (9) при больших m с вероятностью $\mathrm{P} > (2\Phi(C_1) - 1)(2\Phi(C_2) - 1)$ выполнено

$$|n_i - mp_i| < C_1\sqrt{m}\,; \quad |n'_i - mp'_i| < C_2\sqrt{m}\,.$$

Для любого $\varepsilon > 0$ найдется константа C_ε : $\mathrm{P} > (2\Phi(C_\varepsilon) - 1)^2 > 1 - \varepsilon$. Выберем $C_1 = C_2 = C_\varepsilon$. Тогда $(n_i - n'_i) > m(p_i - p'_i) + O(\sqrt{m})$ и

$$\frac{(n_i - n'_i)^2}{n_i} > m\frac{(p_i - p'_i)^2}{p_i} + O(\sqrt{m}) > Cm\,. \quad (13)$$

Следовательно, для любого $\alpha \in (0, 1)$ при достаточно больших m

$$t_m = 2mD_{\mathrm{KL}}(\hat{P}^1_m||\hat{P}^2_m) \sim \sum_{i=1} \frac{(n_i - n'_i)^2}{n_i} > Cm > \bar{t}_\alpha$$

с вероятностью $\mathrm{P} > 1 - \epsilon$, т. е. вероятность $\mathrm{P}(t_m > \bar{t}_\alpha) \to 1$ при $m \to \infty$.

5 Вычислительный эксперимент

Работа критерия была рассмотрена на различных парах распределений. Для выбранной пары распределений повторялась следующая процедура:

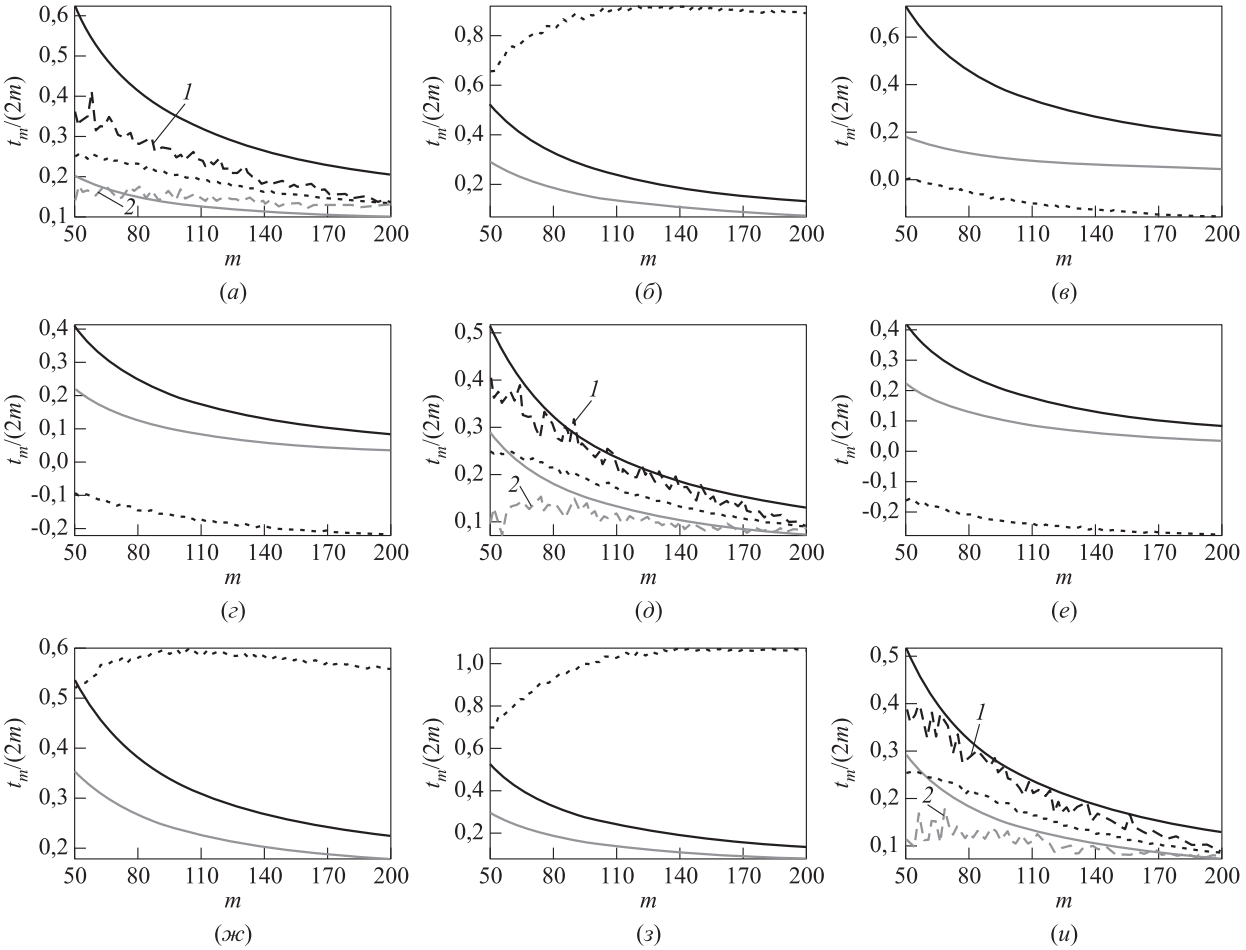


Рис. 2 Зависимость статистики t_m от объема выборки для различных пар распределений; сплошными кривыми отмечены границы доверительного интервала для t_m при $\alpha = 0{,}1$: (*а*) $\mathrm{Exp}\,(1)||\mathrm{Exp}\,(1)$; (*б*) $\mathrm{Exp}\,(1)||\mathcal{N}(0, 1)$; (*в*) $\mathrm{Exp}\,(1)||R(0, 1)$; (*г*) $\mathcal{N}(0, 1)||\mathrm{Exp}\,(1)$; (*д*) $\mathcal{N}(0, 1)||\mathcal{N}(0, 1)$; (*е*) $\mathcal{N}(0, 1)||R(0, 1)$; (*ж*) $R(0, 1)||\mathrm{Exp}\,(1)$; (*з*) $R(0, 1)||\mathcal{N}(0, 1)$; (*и*) $R(0, 1)||R(0, 1)$

(1) генерировались выборки X и X' одинакового объема m;

(2) по выборкам строились гистограммы $\hat{P}_m$ и $\hat{P}'_m$ с фиксированным числом разбиений $N = 20$ и вычислялись расстояния Кульбака–Лейблера $D_{\mathrm{KL}}(\hat{P}_m||\hat{P}'_m)$;

(3) расстояния усреднялись по 1000 генерациям выборок;

(4) объем m выборки увеличивался.

На каждом из графиков на рис. 2 отложены расстояния $D_{\mathrm{KL}}(\hat{P}_m|\hat{P}'_m)$ в зависимости от объема выборки и критические значения $\bar{t}^{\chi^2}/(2m)$ при заданном уровне значимости α. Заметим, что в случае различных распределений расстояние $D_{\mathrm{KL}}(\hat{P}_m|\hat{P}'_m)$ быстро попадает в критическую область и характер его зависимости от m согласуется с оценкой (13). Отрицательные значения, не характерные для расстояния Кульбака–Лейблера, возникают при численном приближении интеграла (6), когда распределение в знаменателе под знаком логарифма имеет большую область определения. Именно из-за отрицательных значений был использован двусторонний критерий. В дальнейших экспериментах было использовано симметризованное расстояние Кульбака–Лейблера (4), что позволило использовать односторонний критерий. На графиках, иллюстрирующих применение критерия к выборкам из одного распределения (рис. 2, *а*, *д*, *и*), также показаны штриховыми кривыми *1* и *2* значения $\bar{t}/(2m)$, где $\bar{t}$ — оценки критических значений, полученные экспериментально:

$$\bar{t}_\alpha = \min\left\{\bar{t} : \frac{1}{M}\sum_{t\in T}[t > \bar{t}] < \alpha\right\}.$$

Здесь суммирование индикаторной функции $[t > \bar{t}]$ ведется по всей выборке T статистик t, полученных по M генерациям пар выборок X и X' (в данном эксперименте $M = 1000$). Видно, что, хотя критические области U^{χ^2} и U не совпадают, при истинности гипотезы H_0 статистика t_m не попадает ни в U^{χ^2}, ни в U.

Оценка фактического значения α. Так как распределение статистики t_m не совпадает с распределением χ^2, уровень значимости α, при котором определяется критическая область U^{χ^2}, не соответствует реальному уровню значимости критерия. Чтобы оценить реальный уровень значимости решения о принятии или отвержении гипотезы H_0, необходимо подсчитать долю объектов выборки T, попавших в U^{χ^2} при заданном α:

$$\hat{\alpha} = \frac{1}{M}\sum_{t\in T}\left[t > \bar{t}^{\chi^2}_\alpha\right].$$

Результаты отражены на рис. 3. Из рисунков следует, что для достижения уровня значимости $\alpha = 0{,}1$ нужно использовать в качестве оценки U критическую область U^{χ^2} с уровнем значимости $\alpha = 0{,}001$.

Кластеризация временны́х рядов, отражающих железнодорожные грузоперевозки. Продемонстрировав таким образом статистическую значимость расстояния Кульбака–Лейблера, построим кластеризацию набора временны́х рядов Российских железных дорог. Данные о перевозках включают даты отправления и прибытия, станции внутри железнодорожной ветки и группы грузов. Для исследования были выбраны временны́е ряды с весами вагонов, нагруженных различными группами грузов, агрегированные по станциям. Вначале из рассмотрения были исключены ряды, содержащие менее 50 отсчетов времени. Матрицы D симметризованных расстояний Кульбака–Лейблера (4) для набора исследуемых временны́х рядов изображены на рис. 4. Решая задачу кластеризации (1), необходимо стремиться привести матрицу D к блочному виду. В левом столбце таблицы показано, как именно перевозимые группы грузов были разбиты на кластеры для случая $K = 5$.

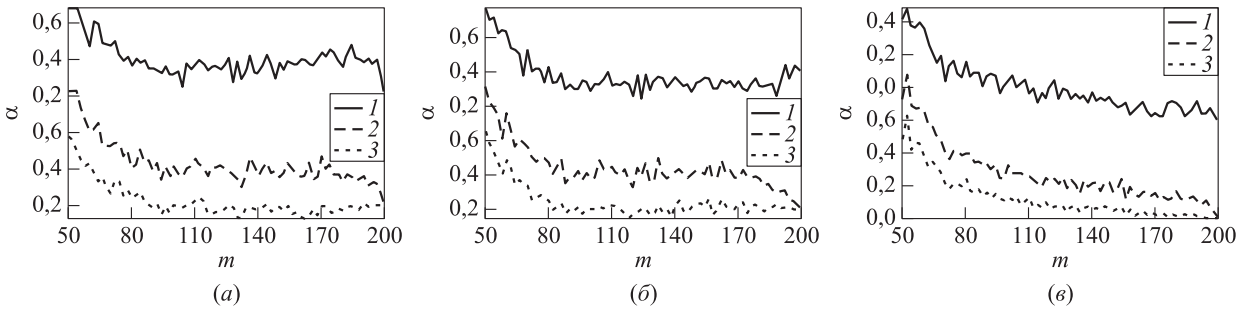


Рис. 3 Зависимость фактического уровня значимости от объема выборки при различных уровнях значимости критерия $\chi^2_{2\mathcal{N}}$ (*1* — $\alpha = 10^{-1}$; *2* — 10^{-2}; *3* — $\alpha = 10^{-3}$): (*а*) Exp (1)||Exp (1); (*б*) $\mathcal{N}(0,1)||\mathcal{N}(0,1)$; (*в*) $R(0,1)||R(0,1)$

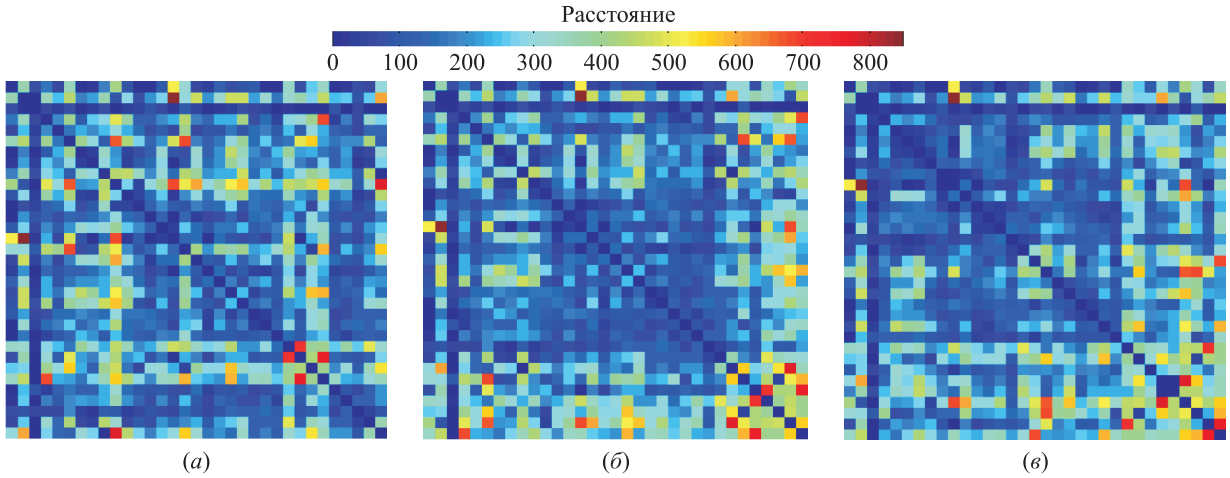


Рис. 4 Симметризованные матрицы попарных расстояний Кульбака–Лейблера между временны́ми рядами для различных групп грузов: до кластеризации (*а*) и после нее для $K = 5$ (*б*) и 10 кластеров (*в*)

Группы грузов

1 — Каменный уголь; 2 — Кокс; 3 — Нефть и нефтепродукты; 7 — Руда железная и марганцевая; 8 — Руда цветная и серное сырье; 9 — Черные металлы; 10 — Метизы и оборудование; 11 — Металлические конструкции; 14 — Сельскохозяйственные машины; 15 — Автомобили; 16 — Цветные металлы; изделия из них и лом цветных металлов; 17 — Химические и минеральные удобрения; 18 — Химикаты и сода; 20 — Промышленное сырье и формовочные материалы; 22 — Огнеупоры; 23 — Цемент; 25 — Сахар; 26 — Мясо и масло животное; 27 — Рыба; 28 — Картофель, овощи и фрукты; 36 — Комбикорма; 38 — Жмыхи; 39 — Бумага; 42 — Грузы в контейнерах	1 — Каменный уголь; 2 — Кокс; 3 — Нефть и нефтепродукты; **6 — Флюсы**; **21 — Шлаки гранулированные**; 7 — Руда железная и марганцевая; 8 — Руда цветная и серное сырье; 10 — Метизы и оборудование; 11 — Металлические конструкции; 12 — Метизы; 14 — Сельскохозяйственные машины; 15 — Автомобили; 16 — Цветные металлы, изделия из них и лом цветных металлов; 17 — Химические и минеральные удобрения; 18 — Химикаты и сода; 20 — Промышленное сырье и формовочные материалы; 22 — Огнеупоры; 23 — Цемент; 25 — Сахар; 26 — Мясо и масло животное; 27 — Рыба; 28 — Картофель, овощи и фрукты; 36 — Комбикорма; 38 — Жмыхи; 39 — Бумага; 42 — Грузы в контейнерах
43 — Остальные и сборные грузы; 19 — Строительные грузы	9 — Черные металлы; 12 — Метизы
31 — Промышленные товары народного потребления; 34 — Зерно	43 — Остальные и сборные грузы; 19 — Строительные грузы; 35 — Продукты перемола
13 — Лом черных металлов	31 — Промышленные товары народного потребления; **29 — Поваренная соль**; 34 — Зерно
35 — Продукты перемола	13 — Лом черных металлов

Ряды, содержащие менее 50 отсчетов, последовательно присоединялись к каждому из рядов, содержавших более 50 отсчетов. Затем для объединенного ряда и исходного ряда длиной более 50 отсчетов проверялась гипотеза о принадлежности рядов к одному распределению. Результаты проверки гипотезы и значения статистики $t_{m,l}$ изображены на рис. 5, *а*, *б*. Строки соответствуют временны́м рядам длиной менее 50 отсчетов, столбцы — длиной более 50 отсчетов. Результат проверки гипотезы H_0 отмечен синим цветом, если гипотеза отвергается, красным — если гипотеза принимается. Было выполнено слияние рядов «Поваренная соль» и «Продукты промышленного потребления», а также «Флюсы» и «Шлаки гранулированные» с временны́м рядом «Нефть и нефтепродукты». Затем снова была выполнена кластеризация, результаты которой занесены в правый столбец таблицы. Вид матрицы парных расстояний после слияния временны́х рядов и их кластеризации представлен на рис. 5, *в*.

6 Заключение

Расстояние Кульбака–Лейблера широко применяется для сравнения распределений, однако счи-

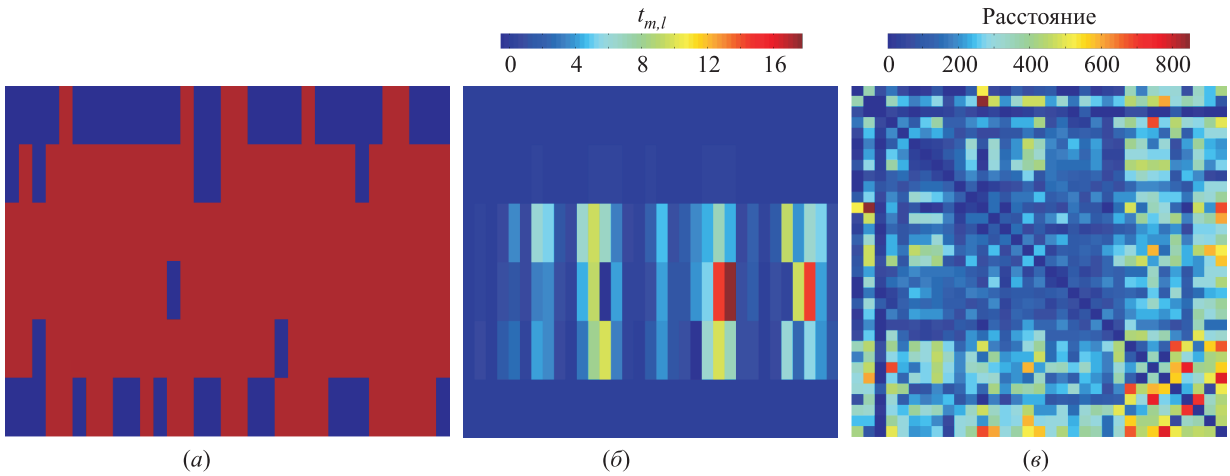


Рис. 5 Результат проверки гипотезы о принадлежности исходного временно́го ряда и временно́го ряда, получаемого присоединением к нему некоторого более короткого ряда, к одному распределению (*а*); значения статистики $t_{m,l}$ при проверке гипотезы H_0 (*б*); матрица парных расстояний после объединения временны́х рядов и их кластеризации (*в*)

тается непригодным для использования в статистических целях из-за того, что не имеет предельного распределения. В данной работе показано, что распределение расстояния Кульбака–Лейблера в пределе ограничено сверху хи-квадратом. Это дает, пусть и ограниченную, возможность использования расстояния Кульбака–Лейблера между распределениями в качестве статистики для проверки гипотезы о принадлежности двух выборок одному распределению и позволяет говорить о статистической значимости расстояния Кульбака–Лейблера. Код, позволяющий выполнить процедуру проверки нулевой гипотезы о принадлежности выборок к одному распределению на основе расстояния Кульбака–Лейблера между их гистограммами, находится в свободном доступе [15]. В работе продемонстрировано использование предлагаемого критерия, а также приведен пример кластеризации временны́х рядов на основе расстояния Кульбака–Лейблера.

Литература

1. *Вальков А. С., Кожанов Е. М., Медведникова М. М., Хусаинов Ф. И.* Непараметрическое прогнозирование загруженности системы железнодорожных узлов по историческим данным // Машинное обучение и анализ данных, 2012. Т. 1. № 4. С. 448–465.
2. *Вальков А. С., Кожанов Е. М., Мотренко А. П., Хусаинов Ф. И.* Построение кросс-корреляционных зависимостей при прогнозе загруженности железнодорожного узла // Машинное обучение и анализ данных, 2013. Т. 1. № 5. С. 505–518.
3. *Медведникова М. М.* Согласование агрегированных непараметрических прогнозов временны́х рядов // Машинное обучение и анализ данных, 2014 (в печати). Т. 1. № 8.
4. *Kullback S.* Information theory and statistics. — New York: Wiley, 1959. 395 p.
5. *Chernoff H.* A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations // Ann. Math. Stat., 1952. Vol. 4. No. 23. P. 493–655.
6. *Kolmogorov A. N.* On the approximation of distributions of sums of independent summands by infinitely divisible distributions // Contributions to statistics. — Oxford: Pergamon Press, 1965. P. 158–174.
7. *Ali S. M., Silvey S. D.* A general class of coefficients of divergence of a distribution from another // J. R. Stat. Soc. Ser. B (Methodoligical), 1966. Vol. 1. No. 28. P. 131–142.
8. *Csiszar I., Shields P.* Information theory and statistics: A tutorial // Foundations Trend Comm. Inform. Theory, 2004. No. 4. P. 417–528.
9. *Gibbs A. L., Su F. E.* On choosing and bounding probability metrics // Intern. Stat. Rev., 2002. Vol. 3. No. 70. P. 419–435.
10. *Mallows C.* A note on asymptotic joint normality // Ann. Math. Stat., 1972. Vol. 42. No. 2. P. 508–515.
11. *Irpino A., Verde R., Lechevallier Y.* Dynamic clustering of histograms using Wasserstein metric // Advances in computational statistics. — Heidelberg: Physica-Verlag, 2006. P. 869–876.
12. *Двоенко С. Д.* Неиерархический дивизимный алгоритм кластеризации // Автоматика и телемеханика, 1999. № 4. С. 117–123.
13. *Стрижов В. В., Кузнецов М. П., Рудаков К. В.* Метрическая кластеризация последовательностей аминокислотных остатков в ранговых шкалах // Математическая биология и биоинформатика, 2012. Т. 7. № 1. С. 345–359.
14. *Двоенко С. Д., Пшеничный Д. О.* О метрической коррекции матриц парных сравнений // Машинное обучение и анализ данных, 2013. Т. 1. № 5. С. 606–620.

15. *Мотренко А. П.* Статистический тест для проверки гипотезы о принадлежности двух выборок одному распределению на основе расстония Кульбака–Лейблера // Algorithms of machine learning. — Sourceforge, 2014. http://sourceforge.net/p/mlalgorithms/code/HEAD/tree/Group874/Motrenko2014KL/code/KLtest.m.

Поступила в редакцию 12.13.14

OBTAINING AN AGGREGATED FORECAST OF RAILWAY FREIGHT TRANSPORTATION USING KULLBACK–LEIBLER DISTANCE

A. P. Motrenko[1] and V. V. Strijov[2]

[1]Moscow Institute of Physics and Technology, 9 Institutskiy per., Dolgoprudny, Moscow Region 141700, Russian Federation

[2]Dorodnicyn Computing Centre, Russian Academy of Sciences, 40 Vavilov Str., Moscow 119333, Russian Federation

Abstract: This study addresses the problem of obtaining an aggregated forecast of railway freight transportation. To improve the quality of aggregated forecast, the time series clusterization problem is solved in such a way that the time series in each cluster belong to the same distribution. To solve the clusterization problem, it is necessary to estimate the distance between empirical distributions of the time series. A two-sample test based on the Kullback–Leibler distance between histograms of the time series is introduced. Theoretical and experimental research of the suggested test is provided. Also, as a demonstration, the clusterization of a set of railway time series based on the Kullback–Leibler distance between time series is obtained.

Keywords: empirical distribution function; distance between histograms; Kullback–Leibler distance; two-sample problem

DOI: 10.14375/19922264140209

Acknowledgments

The work was supported by the Russian Foundation for Basic Research (grant No. 13-07-13139).

References

1. Val'kov, A. S., E. M. Kozhanov, M. M. Medvednikova, and F. I. Khusainov. 2012. Neparametricheskoe prognozirovanie zagruzhennosti sistemy zheleznodorozhnykh uzlov po istoricheskim dannym [Nonparametric forecasting of railroad stations occupancy according to historical data]. *Mashinnoe Obuchenie i Analiz Dannykh* [J. Machine Learning and Data Analysis] 4(1):448–465.

2. Val'kov, A. S., E. M. Kozhanov, A. P. Motrenko, and F. I. Khusainov. 2013. Postroenie kross-korrelyatsionnykh zavisimostey pri prognoze zagruzhennosti zheleznodorozhnogo uzla [Constructing a cross-correlation model to forecast the utilization of a railway junction station]. *Mashinnoe Obuchenie i Analiz Dannykh* [J. Machine Learning and Data Analysis]. 5(1):505–518.

3. Medvednikova, M. M. 2014 (in press). Soglasovanie agregirovannykh neparametricheskikh prognozov vremennykh ryadov [Matching of aggeragared nonparametric forecasts of time series]. *Mashinnoe Obuchenie i Analiz Dannykh* [J. Machine Learning and Data Analysis] 8(1). [In Russian.]

4. Kullback, S. 1959. *Information theory and statistics*. New York: Wiley. 395 p.

5. Chernoff, H. 1952. A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. *Ann. Math. Stat.* 4(23):493–655

6. Kolmogorov, A. N. 1965. On the approximation of distributions of sums of independent summands by infinitely divisible distributions. *Contributions to statistics*. Oxford: Pergamon Press. P. 158–174.

7. Ali, S. M., and S. D. Silvey. 1966. A general class of coefficients of divergence of a distribution from another. *J. R. Stat. Soc. Series B (Methodoligical)* 1(28):131–142.

8. Csiszar, I., and P. Shields. 2004. Information theory and statistics: A tutorial. *Foundations Trend Comm. Inform. Theory* 4:417–528.

9. Gibbs, A. L., and F. E. Su. 2002. On choosing and bounding probability metrics. *Intern. Stat. Rev.* 3(70):419–435.

10. Mallows, C. 1972. A note on asymptotic joint normality. *Ann. Math. Stat.* 42(2):508–515.

11. Irpino, A., R. Verde, and Y. Lechevallier. 2006. Dynamic clustering of histograms using Wasserstein metric. *COMPSTAT*. 869–876.

12. Dvoenko, S. D. 1999. Neierarkhicheskiy divizimnyy algoritm klasterizatsii [Nonhierarchical divisible clasterization algorithm]. *Avtomatika i Telemekhanika* [Automation and Remote Control] 4:117–123.

13. Strizhov, V. V., M. P. Kuznetsov, and K. V. Rudakov. 2012. Metricheskaya klasterizatsiya posledovatel'nostey aminokislotnykh ostatkov v rangovykh shkalakh [Metric clustering of sequences of amino acid residues in rank scales]. *Matematicheskaya Biologiya i Bioinformatika* [Mathematical Biology and Bioinformatics] 7(1):345–359.

14. Dvoenko, S. D., and D. O Pshenichnyy. 2013. O metricheskoy korrektsii matrits parnykh sravneniy [On metric correction of matrices of pairwise comparisons]. *Mashinnoe Obuchenie i Analiz Dannykh* [J. Machine Learning and Data Analysis] 5(1):606–620.

15. Motrenko, A. P. 2014. Statisticheskiy test dlya proverki gipotezy o prinadlezhnosti dvukh vyborok odnomu raspredeleniyu na osnove rasstoniya Kul'baka–Leyblera [A statistical test for the two-sampe problem based on the Kullback–Leibler distance]. http://sourceforge.net/p/mlalgorithms/code/HEAD/tree/Group874/Motrenko2014KL/code/KLtest.m.

Received March 12, 2014

Contributors

Motrenko Anastasia P. (b. 1992) — student, Moscow Institute of Physics and Technology, 9 Institutskiy per., Dolgoprudny, Moscow Region 141700, Russian Federation; anastasia.motrenko@gmail.com

Strijov Vadim V. (b. 1967) — Candidate of science (PhD) in physics and mathematics, associate professor, scientist, Dorodnicyn Computing Center, Russian Academy of Sciences, 40 Vavilov Str., Moscow 119333, Russian Federation; strijov@ccas.com

ИНФОРМАТИКА И ЕЁ ПРИМЕНЕНИЯ, 2014. Т. 8. Вып. 2. С. 98–110

ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ КОРПУСНЫХ ИССЛЕДОВАНИЙ: ПРИНЦИПЫ ПОСТРОЕНИЯ КРОССЛИНГВИСТИЧЕСКИХ БАЗ ДАННЫХ*

Н. В. Бунтман[1], Анна А. Зализняк[2], И. М. Зацман[3], М. Г. Кружков[4], Е. Ю. Лощилова[5], Д. В. Сичинава[6]

Аннотация: Рассматривается информационная технология создания кросслингвистических баз данных текстов на русском языке и их переводов на французский язык, называемых параллельными текстами. Разработанные принципы построения этой базы данных обеспечивают реализацию уникального сочетания трех видов двуязычного поиска: лексического, грамматического и лексико-грамматического. Отличительной чертой рассматриваемой технологии является одновременное формирование русско-французского параллельного подкорпуса Национального корпуса русского языка (НКРЯ) и кросслингвистической базы данных глагольных лексико-грамматических форм русского языка и их функциональных эквивалентов во французских переводах. Подкорпус и база данных обладают разной глубиной выравнивания: в первом случае оно выполняется на уровне предложений, а во втором — на уровне конструкций. Теоретическое значение создания этой базы данных заключается в обеспечении исследований как в области двуязычной контрастивной грамматики, так и в направлении создания грамматики русского языка, опирающейся на современную эмпирическую базу и информационные технологии корпусной лингвистики. Ее основное прикладное назначение заключается в повышении качества машинного перевода.

Ключевые слова: параллельный корпус; информационная технология; кросслингвистические базы данных; двуязычный лексико-грамматический поиск; корпусная лингвистика; контрастивная грамматика
DOI: 10.14357/19922264140210

1 Введение

Возникновение параллельных электронных корпусов ознаменовало начало новой эры контрастивных лингвистических исследований; пионерскими в этой области стали работы 1990-х гг. Стига Йоханссона с англо-норвежским корпусом. Сочетание методов современной компьютерной лингвистики с возможностями сопоставления текстов на двух и более языках, предоставляемыми параллельными корпусами, обеспечило возможность осуществления контрастивного лингвистического анализа на принципиально новом уровне точности (ср. [1]). Благодаря таким корпусам за прошедшие два десятилетия в этой области были достигнуты значительные успехи как в плане разработки методик анализа, так и в плане создания оригинальных лексикографических описаний. О перспективах контрастивных грамматических исследований на базе параллельных корпусов см. работы [1–7].

В работе [8] была описана технология формирования русско-французского параллельного подкорпуса НКРЯ, содержащего литературные произведения на русском языке и их переводы на французский язык.

В настоящее время русско-французский подкорпус НКРЯ содержит тексты произведений совокупным объемом в 2 млн словоупотреблений. При этом часть параллельных текстов представлена в *поливариантном формате*, т. е. одно произведение на русском языке выровнено в корпусе по предложениям с *несколькими вариантами его перевода* на французский язык. Совокупный объем поливариантных текстов — 700 тыс. словоупотреблений, т. е. больше трети всего параллельного русско-французско-русского подкорпуса НКРЯ.

Параллельные корпуса стали включаться в состав НКРЯ с 2005 г. [2, 4, 9, 10]. Сейчас он включает восемь двуязычных параллельных подкорпусов

*Работа выполнена в ИПИ РАН при поддержке фонда «Династия» (грант NG13-036) и РФФИ (грант № 13-06-00403).

[1]Московский государственный университет им. М. В. Ломоносова, факультет иностранных языков и регионоведения, nabunt@hotmail.com

[2]Институт языкознания Российской академии наук; Институт проблем информатики Российской академии наук, anna.zalizniak@gmail.com

[3]Институт проблем информатики Российской академии наук, izatsman@yandex.ru

[4]Институт проблем информатики Российской академии наук, magnit75@yandex.ru

[5]Институт проблем информатики Российской академии наук, lena0911@mail.ru

[6]Институт русского языка Российской академии наук, mitrius@gmail.com

с русским языком оригинала или перевода (английский, немецкий, французский, испанский, итальянский, польский, украинский и белорусский) и один многоязычный параллельный подкорпус. Подкорпус параллельных текстов на русском и французском языках появился в составе НКРЯ в декабре 2012 г. Технологию, используемую для формирования этого подкорпуса и описанную ранее в работе [8], обозначим как Parallel Corpus technology или ParCor-технология.

В 2013 г. ParCor-технология была дополнена новыми операциями: была создана база данных глагольных форм русского языка и вариантов их перевода на французский язык (далее — БД) и сформирован поливариантный подкорпус параллельных текстов на русском и французском языках (далее — подкорпус). Новая технология дала возможность формировать БД одновременно с пополнением подкорпуса и реализовать три вида двуязычного поиска глагольных форм и их переводов: лексического, грамматического и лексико-грамматического. Например, в этой БД можно задать и выполнить запрос на поиск параллельных выровненных текстовых фрагментов, в которых в русском оригинале употреблена глагольная форма прошедшего времени несовершенного вида, а в параллельных французских фрагментах — passé composé. Технологию, ориентированную на одновременное формирование подкорпуса и БД с двуязычным поиском параллельных глагольных форм в оригинальном и переведенных текстах обозначим как Database Parallel Corpus technology или DBParCor-технология.

Цель статьи состоит в том, чтобы описать назначение, задачи и принципы построения БД, формируемой на основе параллельных текстов НКРЯ, а также функции двуязычного поиска, реализованные в этой БД.

2 Назначение базы данных и принципы ее построения

Принципы построения БД во многом диктовались ее назначением. Она создавалась как инструмент описания русской грамматической семантики «в зеркале французского языка», а также с целью уточнения положений русско-французской контрастивной грамматики. При выработке принципов построения БД использовались работы Гака [11, 12], Кузнецовой [13], Гиро-Вебер [14] и др. Эти работы, однако, появились в докорпусную эпоху; теперь, когда созданы и регулярно пополняются русско-французские корпуса, стали доступны параллельные тексты в цифровой форме, их сопоставление и анализ дает возможность уточнить описание русско-французской контрастивной грамматики.

В ходе разработки БД учитывалось то, что объектом анализа являются соответствия глагольных категорий русского и французского языка в параллельных текстах. Было определено несколько новых терминов, которые отражают существо принципов построения БД.

Ключевыми являются понятия «лексико-грамматическая форма», или ЛГФ, и «базовый вид ЛГФ», определения которых даны ниже.

Определение 1. Под *лексико-грамматической формой* (ЛГФ) понимается совокупность элементов конкретного предложения, обладающая набором признаков, задаваемых базовым видом ЛГФ.

Определение 2. Под *базовым видом ЛГФ* понимается определенная комбинация значений следующих параметров:

- категориальная принадлежность языковой единицы (в данной статье рассматриваются только глагольные ЛГФ, т. е. значение этого параметра фиксировано);
- набор значений грамматических категорий, релевантных для выбранного класса единиц;
- (факультативно) определенные элементы структуры предложения, задающие «конструкцию»; например: «PastPF + *если бы*».

Другими словами, базовый вид ЛГФ представляет собой некоторую комбинацию значений глагольных категорий, в совокупности с определенными элементами структуры предложения задающую некоторую «конструкцию»[1].

В процессе формирования БД было выделено 15 базовых видов ЛГФ русского языка; это так называемое множество-источник (табл. 1)[2]. Количество базовых видов ЛГФ французского языка (множество-цель) не фиксировано, поскольку оно возрастает по мере пополнения БД; в текущем варианте БД оно составляет 25 единиц (табл. 2).

Помимо базовых видов ЛГФ для каждого из двух языков сформировано множество дополнительных признаков, которые позволяют специфицировать тип конструкции. А именно: дополнительные признаки характеризуют либо состав глагольной груп-

[1] В значении, которое придается этому понятию в Грамматике конструкций [15–17].

[2] В текущую версию БД включены только те ЛГФ русского языка, которые содержат глагол в финитной форме (т. е. исключались безличные глаголы, слова категории состояния, причастия, деепричастия, а также перифразы с глаголом *быть*). В дальнейшем состав рассматриваемых видов глагольных форм будет расширяться.

Таблица 1 Множество-источник базовых видов глагольных ЛГФ русского языка

	Полное название базового вида ЛГФ	Сокращенное обозначение базового вида ЛГФ
1.	Настоящее	Pres
2.	Прошедшее НСВ	Past-IPF
3.	Прошедшее СВ	Past-PF
4.	Простое будущее	Fut-PF
5.	Сложное будущее	Fut-IPF
6.	Императив СВ	Imperat-PF
7.	Императив НСВ	Imperat-IPF
8.	Форма с *бы* СВ	Past-PF+*бы*
9.	Форма с *бы* НСВ	Past-IPF+*бы*
10.	Форма с *если бы* СВ	Past-PF+*если бы*
11.	Форма с *если бы* НСВ	Past-IPF+*если бы*
12.	Форма с *чтобы* СВ	Past-PF+*чтобы*
13.	Форма с *чтобы* НСВ	Past-IPF+*чтобы*
14.	Форма с *было* СВ	Past-PF+*было*
15.	Форма с *было* НСВ	Past-IPF+*было*

Таблица 2 Множество-цель базовых видов ЛГФ французского языка

	Полное название базового вида ЛГФ	Сокращенное обозначение базового вида ЛГФ
1.	Présent	Pr
2.	Passé composé	PasCom
3.	Passé simple	PasSim
4.	Imparfait	Imparf
5.	Plus-que-parfait	PqParf
6.	Passé antérieur	PasAnt
7.	Passé immédiat	PasIm
8.	Futur simple	Fut
9.	Futur antérieur	FutAnt
10.	Futur immédiat	FutIm
11.	Impératif	Imperat
12.	Subjonctif présent	SubjPres
13.	Subjonctif passé	SubjPas
14.	Subjonctif imparfait	SubjImparf
15.	Subjonctif plus-que-parfait	SubjPqParf
16.	Conditionnel présent	CondPr
17.	Conditionnel passé	CondPas
18.	Participe présent	PartPr
19.	Participe passé	PartPas
20.	Participe passé composé	PartPasComp
21.	Gérondif	en PartPr
22.	Infinitif	Inf
23.	Préposition+infinitif	Prep+Inf
24.	Préposition+infinitif passé	Prep+InfPas
25.	Substantif	Subst

пы (например, наличие при глаголе подчиненного инфинитива, модального детерминанта или отрицания), либо тип предложения, в котором употреблена данная ЛГФ (например, придаточное, вопросительное предложение, диалогическая реплика) (табл. 3 и 4). Каждый признак приложим или ко всем, или к некоторым из базовых видов ЛГФ. На всех рисунках статьи дополнительные признаки указаны в квадратных скобках после базового вида ЛГФ.

Определение 3. Комбинацию базового вида ЛГФ с одним или несколькими из дополнительных признаков назовем *видом ЛГФ*.

Принципы установления соответствия в параллельных выровненных текстах между русскими и французскими ЛГФ состоят в следующем. Сначала из фразы русского оригинала вычленяется фрагмент, включающий ЛГФ, базовый вид которой принадлежит множеству-источнику (см. табл. 1). Далее ищется ее «функционально эквивалентный фрагмент» (ФЭФ)[1] во французском переводе, из которого извлекается ЛГФ, базовый вид которой принадлежит множеству-цели (см. табл. 2).

Лексико-грамматическая форма русского языка и соответствующая ей ЛГФ французского языка образуют *моноэквиваленцию* (см. определение 4 и табл. 5). Если в процессе анализа ФЭФ оказывается, что нужный базовый вид французской ЛГФ в табл. 2 отсутствует, то множество-цель может быть пополнено. Случаи, когда для русской ЛГФ французский эквивалент не найден, отмечаются в БД специальной пометой (Nondetermined), и в процессе обработки данных они пока не учитываются[2].

Поиск ФЭФ и выявление содержащейся в нем ЛГФ французского языка являются первой задачей, решение которой обеспечивается разработанным вариантом БД. Для описания других задач БД, рассмотренных в следующем разделе, определим еще пять терминов: «моноэквиваленция», «тип моноэквиваленции», «полиэквиваленция», «тип полиэквиваленции» и «гиперэквиваленция».

Определение 4. *Моноэквиваленция* (МЭ) — это двухместный кортеж вида $\langle R_n(i); F_m(j)\rangle$, где первую позицию занимает i-е вхождение ЛГФ базового вида R_n русского языка (см. табл. 1) в оригинальном тексте. Вторую позицию занимает j-е вхождение ЛГФ базового вида F_m французского языка (см. табл. 2) в одном из вариантов перевода i-го вхождения русской ЛГФ. Все МЭ, входящие в БД, имеют идентификационный номер.

[1] Термин «функционально эквивалентный фрагмент» введен в работе [2], см. также [9].

[2] Речь идет о таких случаях, когда семантическое содержание, заключенное в выбранной ЛГФ оригинала, передано в переводе столь существенно иными лексическими средствами, что установление соответствия между ЛГФ при помощи того аппарата, который имеется на сегодня, оказывается невозможно. Например: *ты [. . .] так теребишь за носы, что* ***еле держатся*** *— tu tirais tellement sur leur nez [. . .] que* ***tu as failli le leur arracher***.

Таблица 3 Дополнительные признаки для базовых видов ЛГФ русского языка

Полное название дополнительного признака	Сокращенное обозначение дополнительного признака
Подчиненный инфинитив СВ	[SubInf-PF]
Подчиненный инфинитив НСВ	[SubInf-IPF]
Модальный детерминант	[ModDet]
Отрицание	[Neg]
Вопросительное предложение	[Interrog]
Восклицательное предложение	[Exclam]
Глагол, вводящий прямую речь	[VerbDirSp]
Глагол в составе диалогической реплики	[DialRepl]
Глагол в придаточном предложении	[Sub]
Глагол в изъяснительном придаточном	[SubCompl]
Глагол в определительном придаточном	[SubAttr]

Таблица 4 Дополнительные признаки для базовых видов ЛГФ французского языка

Полное название дополнительного признака	Сокращенное обозначение дополнительного признака
Подчиненный инфинитив	[SubInf]
Подчиненный инфинитив прошедшего времени	[SubInfPas]
Добавление подчиняющего предиката	[+SuperPred]
Модальный детерминант	[ModDet]
Отрицание	[Neg]
Вопросительное предложение	[Interrog]
Восклицательное предложение	[Exclam]
Глагол, вводящий прямую речь	[VerbDirSp]
Глагол в составе диалогической реплики	[DialRepl]
Глагол в придаточном предложении	[Sub]
Глагол в изъяснительном придаточном	[SubCompl]
Глагол в определительном придаточном	[SubAttr]
Глагол в условном придаточном	[SubCond]
Accusativus cum infinitivo	[Acc.c.Inf]
Faire + Infinitif	[*faire* + Inf]
Laisser + Infinitif	[*laisser* + Inf]
Sembler + Infinitif	[*sembler* + Inf]
Paraître + Infinitif	[$paraître$ + Inf]

Таблица 5 Моноэквиваленция, зарегистрированная в БД под номером 4711

№ МЭ	ЛГФ русского языка	Вид ЛГФ русского языка	ЛГФ перевода	Вид ЛГФ перевода
4711	потом [. . .] плотно **запер** все двери	Past-PF [ModDet]	après **avoir** bien **fermé** toutes les portes	Prep + InfPas [Sub]

Определение 5. *Типом моноэквиваленции* называется кортеж базовых видов ЛГФ русского и французского языка $\langle R_n; F_m\rangle$, например ⟨Past-PF; Prep + InfPas⟩ (см. 3-й и 5-й столбцы в табл. 5).

Определение 6. *Полиэквиваленция* — это двухместный кортеж вида $\langle R_n(i); \{F_m(j), F_k(r), \ldots\}\rangle$, представляющий собой объединение нескольких моноэквиваленций с идентичной первой позицией ($\langle R_n(i); F_m(j)\rangle$, $\langle R_n(i); F_k(r)\rangle$ и т. д.), отражающих разные варианты перевода одного и того же i-го вхождения ЛГФ базового вида R_n в русском оригинальном тексте: $F_m(j)$ — это ЛГФ французского языка, идентифицированная в первом переводе и соответствующая i-му вхождению русской ЛГФ, $F_k(r)$ — во втором переводе и т. д. (табл. 6).

Определение 7. *Типом полиэквиваленции* называется кортеж базовых видов ЛГФ русского и французского языка $\langle R_n; \{F_m, F_k, \ldots\}\rangle$, например ⟨Pres-IPF; {Pr,Pr}⟩ (см. 2-й и 5-й столбцы в табл. 6).

Таблица 6 Две моноэквиваленции (№№ 596, 5927), составляющие полиэквиваленцию*

ЛГФ русского языка	Вид ЛГФ русского языка	ЛГФ в текстах французских переводов и их виды		
		Номер моноэквиваленции	ЛГФ в текстах французских переводов	Вид ЛГФ французского языка
Я иногда в театр **хожу**	Pres-IPF [ModDet] [DialRepl]	596	**Il m'arrive d'aller** au théâtre,	Pr [SubInf] [+SuperPred] [DialRepl]
		5927	Non, je **vais parfois** au théâtre, et en visite.	Pr [ModDet] [DialRepl]

*Французские ЛГФ, входящие в данную полиэквиваленцию, имеют одинаковый базовый вид, но различаются на уровне дополнительных признаков, указанных в квадратных скобках.

Определение 8. *Гиперэквиваленция* — это двухместный кортеж вида $\langle R_n; \{F\}\rangle$, репрезентирующий соответствие между базовым видом ЛГФ русского языка R_n и множеством базовых видов эквивалентных ЛГФ французского языка, входящих во вторую позицию моноэквиваленций БД с ЛГФ базового вида R_n.

Другими словами, каждая гиперэквиваленция включает один базовый вид ЛГФ русского языка R_n и список базовых видов ЛГФ французского языка — при условии, что хотя бы одна ЛГФ базового вида из этого списка образовала в БД моноэквиваленцию с русской ЛГФ базового вида R_n.

Используя определенные выше термины, перечислим те задачи, для решения которых предназначена спроектированная БД:

- построение моно-, поли- и гиперэквиваленций;
- двуязычный лексический, грамматический и лексико-грамматический поиск моно- и полиэквиваленций;
- вычисление частотности для каждого типа моно- или полиэквиваленций.

Для решения этих задач был разработан веб-интерфейс, который позволяет пользователям-лингвистам взаимодействовать с БД в онлайн-режиме с помощью распространенных веб-браузеров (Internet Explorer, Mozilla Firefox, Google Chrome). Для создания и ведения БД используется СУБД Microsoft SQL Server.

Функции БД можно разделить на две основные группы:

(1) первая группа функций служит для построения и редактирования моноэквиваленций (см. рис. 1 для функции редактирования);

(2) вторая группа функций — для поиска уже построенных моно- и полиэквиваленций (см. рис. 2 с интерфейсом поиска и просмотра полиэквиваленций).

Группа функций построения и редактирования моноэквиваленций в БД позволяет отфильтровывать выровненные фрагменты оригинального и переводных текстов по названию книги, автору перевода и присутствующих в этих фрагментах видам ЛГФ. Используя эти функции, пользователь-лингвист может просматривать выровненные фрагменты параллельных текстов с целью формирования моноэквиваленций.

На начало 2014 г. построено 10 527 моноэквиваленций и на их основе автоматически было сгенерировано 4128 полиэквиваленций (т. е. объединений моноэквиваленций из разных переводов одного оригинального текста с одной и той же ЛГФ русского языка в первой позиции кортежа).

3 Двуязычный поиск

На странице поиска и просмотра полиэквиваленций пользователи БД могут видеть подборки полиэквиваленций (см. рис. 2), которые генерируются в соответствии с поисковым запросом. Пользователи БД могут осуществлять поиск моно- и полиэквиваленций, используя следующие поисковые признаки: название русского произведения, французский перевод, базовые виды и признаки ЛГФ русского и французского языка, лексемы оригинала и переводов, искомые тексты как последовательности знаков, включая знаки препинания (ср. опцию «поиск точных форм» в НКРЯ).

Поисковые признаки можно задавать как по отдельности, так и в сочетании. В результате выполнения поискового запроса можно узнать число найденных полиэквиваленций, удовлетворяющих заданным поисковым признакам, и посмотреть их.

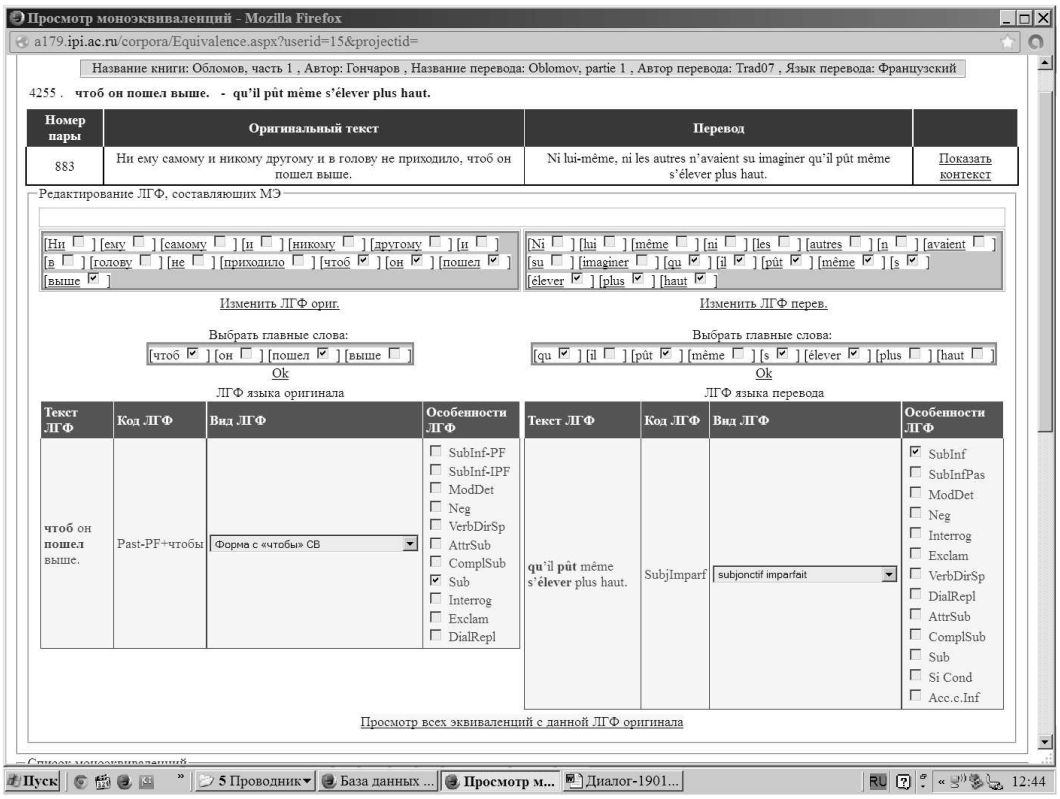


Рис. 1 Интерфейс для редактирования моноэквиваленций

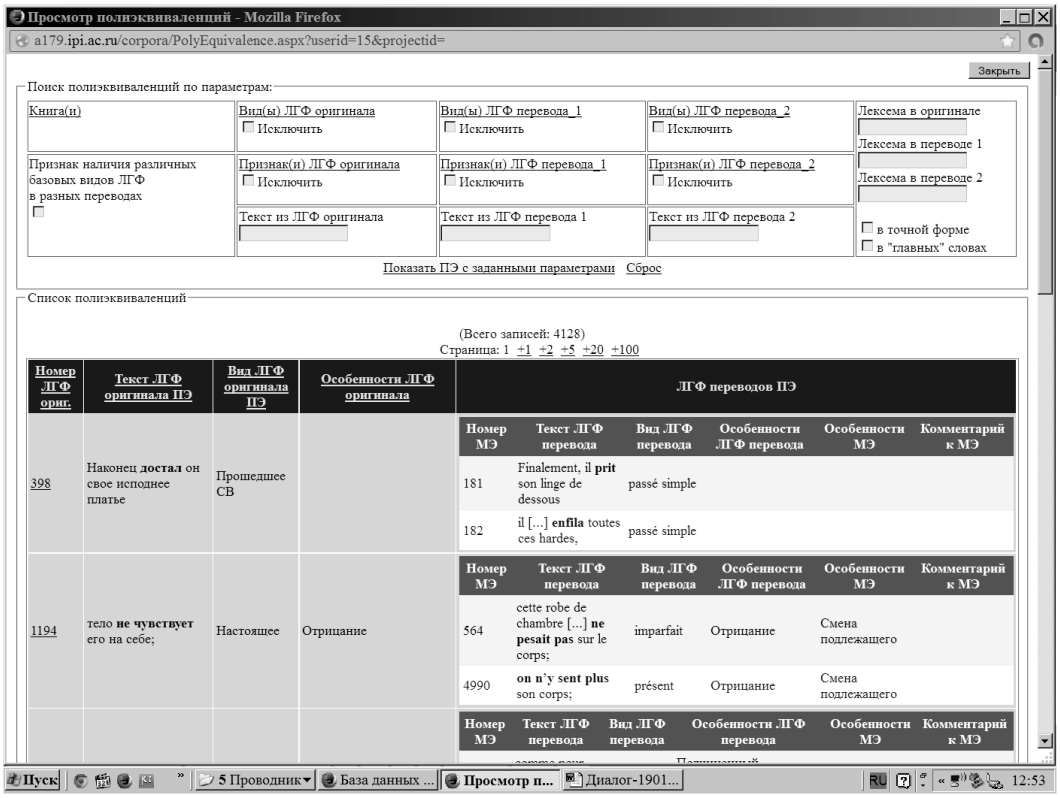


Рис. 2 Интерфейс для поиска и просмотра полиэквиваленций

Таблица 7 Две полиэквиваленции, найденные в БД по поливариантному двуязычному грамматическому запросу

Полиэквиваленция		Номер МЭ	ЛГФ перевода	Базовый вид ЛГФ перевода
он **решил оставить** [. . .] липовые и дубовые деревья	Past-PF [SubInf]	2931	alors qu'**il garderait** les [. . .] tilleuls et chênes,	CondPr
		8011	Il **décida de laisser** tels quels les [. . .] tilleuls et les chênes,	PasSim [SubInf]
он **решил** [. . .] яблони и груши **уничтожить**	Past-PF [SubInf]	2932	il **se débarrasserait** des pommiers et des poiriers	CondPr
		8013	Il **décida** [. . .] **de supprimer** les pommiers et les poitiers	PasSim [SubInf]

Принципиально новой является функция двуязычного грамматического поиска, который применим как к одному, так и одновременно к нескольким переводам (поливариантный двуязычный запрос). Например, если задать один базовый вид ЛГФ русского языка «Past-PF» и соответстующие в двух вариантах перевода два базовых вида ЛГФ французского языка CondPr и PasSim, то в БД будут найдены две полиэквиваленции с заданными поисковыми грамматическими признаками, отраженные в табл. 7.

Кроме базового вида ЛГФ в поисковом запросе могут задаваться также дополнительные признаки для ЛГФ русского языка (из табл. 3) и для ЛГФ французского языка (из табл. 4).

Например, если задать поисковый запрос «Pres [SubInf-PF]» в русской фразе и «CondPr [SubInf]» хотя бы в одном из двух ее переводов, то в БД будут найдены три полиэквиваленции с заданными видами ЛГФ (при этом в найденных полиэквиваленциях могут присутствовать и другие дополнительные признаки) (табл. 8).

Приведенные примеры двуязычного грамматического поиска говорят о том, что разработанная БД является на сегодняшний день уникальным лингвистическим ресурсом, который может быть использован для исследования не только глагольных форм, но и более широкого спектра языковых единиц (см. разд. 4).

4 Исследование лингвоспецифичной лексики с помощью базы данных

В настоящее время DBParCor-технология и БД адаптируются к исследованию лингвоспецифичных единиц (ЛСЕ) русского языка «в зеркале ино-

Таблица 8 Три полиэквиваленции, найденные в БД по видам ЛГФ

Полиэквиваленция		Номер МЭ	ЛГФ переводов	Вид ЛГФ переводов
Не может постараться для барина!	Pres [SubInf-PF] [Neg]	661	**Tu pourrais** tout de même **faire un effort** pour ton maître!	CondPr [SubInf] [Exclam]
		5897	Il **ne peut** même **pas faire** un petit effort pour son maître!	Présent [SubInf] [Exclam]
теперь **можете отдать**	Pres [SubInf-PF]	945	maintenant vous **pouvez** me **rembourser**.	Présent [SubInf]
		7584	alors vous **pourriez** peut-être **me rembourser**?	CondPr [SubInf] [ModDet] [Interrog]
Разве я **могу** все это [. . .] **перенести**?	Pres [SubInf-PF] [Interrog]	8939	Est-ce que je **puis** [. . .] le **supporter?**	Présent [SubInf] [Interrog]
		8940	Je **pourrais** [. . .] **supporter** tout ça?	CondPr [SubInf] [Interrog]

странных языков», включая французский. Для решения задач проекта «Контрастивное корпусное исследование специфических черт семантической системы русского языка», финансируемого по гранту РФФИ, была разработана оригинальная методология контрастивного корпусного анализа, осуществляемого с помощью БД, которая опирается на концептуальный аппарат контрастивного анализа русских лексико-грамматических форм, описанный выше. Данная методология предусматривает:

- статистическое и/или экспертное обоснование гипотез лингвоспецифичности лексических единиц русского языка на основе анализа текстов двуязычных корпусов, БД и других текстовых источников;
- статистическое обоснование с помощью БД гипотез лингвоспецифичности лексических единиц русского языка, сформулированных в ходе предшествующих исследований на основании семантического анализа;
- статистическую и экспертную верификацию гипотез с использованием БД.

Если для статистического обоснования гипотез могут использоваться разные информационные ресурсы (книги, корпуса или БД), то для верификации гипотез используется только БД, так как ключевым этапом верификации является построение моно- и полиэквиваленций и вычисление частотности их типов. Построение моно- и полиэквиваленций позволяет документировать процесс статистической верификации гипотез, а также согласовывать и документировать результаты верификации, выполненной лингвистами-экспертами (экспертная верификация гипотез).

Проведенные эксперименты по формированию, обоснованию, статистической и экспертной верификации гипотез с помощью БД показали, что для их верификации потребуется увеличить ее объем.

Разработанная методика построения статистической и экспертной верификации гипотез основана на использовании количественного статистического и качественного экспертного методов. Суть статистического метода заключается в следующем. Для каждой языковой единицы из списка потенциально ЛСЕ русского языка определяется число ее переводных эквивалентов в тексте переводов, имеющихся в параллельном корпусе, вычисляются частотности переводных эквивалентов и их разброс. Для определения числа переводных эквивалентов могут использоваться книжные источники, корпуса и БД: лингвист-эксперт анализирует причины разброса частотности переводных эквивалентов и отбрасывает те случаи, когда разброс не связан с лингвоспецифичностью (в частности, он может быть обусловлен различием в способе лексикализации, например русскому *плавать* в английском языке соответствует три разных глагола, обозначающих три различных вида плавания: *swim*, *sail*, *float*). Оставшиеся статистически выявленные лексические единицы считаются гипотетически лингвоспецифичными.

Разработанная методика включает стадию уточнения степени лингвоспецифичности языковой единицы. На этой стадии, в частности, проводится анализ условий появления рассматриваемой единицы русского языка в обратных переводах (т. е. множество «стимулов» перевода на русский язык). Чем больше таких стимулов, тем больше вероятность лингвоспецифичности рассматриваемой единицы.

Категория лингвоспецифичных слов находится в отношении пересечения с категорией безэквивалентной лексики, т. е. имеется множество языковых единиц, относящихся к обеим категориям, но есть и непересекающиеся области. Статистические методы применимы только к тем лингвоспецифичным словам, которые относятся также к категории безэквивалентной лексики. Если слово не принадлежит к этой категории, то применяется качественный экспертный метод построения гипотезы, который включает детальный сопоставительный семантический анализ рассматриваемой лексической единицы и его переводного эквивалента и выявление возможных расхождений в составе компонентов, формирующих их семантическую структуру.

Верификация гипотез лингвоспецифичности языковых единиц выполняется лингвистами-экспертами с использованием БД, сформированной на основе параллельных текстов двуязычного корпуса. Каждая построенная в БД моноэквиваленция включает гипотетически лингвоспецифичную лексическую единицу русского языка и один из ее переводных эквивалентов, найденных в параллельных текстах. (На данном этапе ограничим исследуемый материал, с одной стороны, переводами на французский язык и с другой — лингвоспецифичными личными глагольными формами; в дальнейшем исследовательская база будет расширена в обоих направлениях: по числу языков и по спектру конструкций).

Если сопоставительный статистический и семантический анализ гипотетически лингвоспецифичной лексической единицы и ее эквивалентов позволяет лингвисту-эксперту выявить специфический смысловой компонент, присутствующий в русском слове и отсутствующий в переводе, то лингвоспецифичность этой единицы признается им верифицированной. Особую ценность для нужд

семантического анализа представляют полиэквиваленции, которые предоставляют в распоряжение эксперта данные о границах вариативности перевода интересующей его единицы в контексте, зафиксированном в полиэквиваленции.

Так, в двух переводах фразы *Давно собирался к тебе* из БД (рис. 3), оба французских глагола *s'apprêter* и *se préparer* (буквально 'готовиться') не содержат специфического смыслового компонента неконтролируемости, заключенного в русском глаголе *собираться* (см. [18]) и, наоборот, усиливают, по сравнению с оригиналом, семы приготовления и прилагаемых усилий.

База данных предоставляет в распоряжение пользователя практически полный спектр семантических компонентов, составляющих ту сложную концептуальную конфигурацию, которая заключена, например, в русском глаголе *успеть* (ср. [18–20]), что позволяет уточнить проведенный ранее анализ этого лингвоспецифичного слова (рис. 4). Сопоставление с двумя французскими переводами, где использовано выражение со словом 'время', особенно ясно выявляет эти дополнительные семы, определяющие лингвоспецифичный характер данного русского глагола. В русском оригинале речь идет не столько о возможной нехватке времени, сколько о наклонности и способности, с одной стороны, и о случайности и удаче — с другой; кроме того, в русском глаголе имеется отсутствующая во французском переводном эквиваленте оценочная сема (ср. существительное *успех*).

К каждой единице предварительного списка ЛСЕ русского языка применяется процедура анализа, включающая следующие шаги:

- формирование и выполнение поискового запроса по данной лексеме, который позволяет выявить в БД все включающие ее моно- и полиэквиваленции;
- анализ грамматической составляющей полученных моно- и полиэквиваленций, в том числе статистическое распределение реально встречающихся в узусе грамматических форм;
- анализ лексической составляющей полученных моно- и полиэквиваленций, в том числе статистические параметры выбора переводного эквивалента;
- интерпретация полученных результатов с точки зрения семантического анализа исходной единицы русского языка и оценки степени ее лингвоспецифичности.

Разработанная методология контрастивного корпусного анализа специфических черт семантической системы русского языка предполагает использование уже имеющейся БД глагольных форм, а также построение лексических моноэквиваленций, что планируется осуществить в дальнейшем. При этом базовые виды лексических моноэквиваленций будут определяться включенными в них ЛСЕ.

5 Заключение

Сформированная БД позволила уточнить ряд положений русско-французской контрастивной грамматики. В частности, список соответствий, описанных в работах [11, 12] и частично суммированных в работе [13]:

- инвертирован (в работах Гака и Кузнецовой материал рассматривается в направлении от французского к русскому, так как конечной целью там является интерпретация значения и функции форм французского языка);
- существенно расширен, т. е. установлены новые типы переводных соответствий;
- подвергнут статистической оценке.

Особый интерес представляют полученные результаты частотного анализа переводных соответствий. В частности, корреляция между оппозициями «совершенный *vs.* несовершенный вид» в русском языке и «passe composé/passe simple *vs.* imparfait» во французском может быть уточнена на основе количественных показателей: базовому виду русской ЛГФ Past-IPF лишь в 49,4% случаев соответствует базовый вид французской ЛГФ Imparf и в 21% случаев — PasCom/PasSim; особенно значимой представляется последняя цифра, отражающая широту семантического диапазона русского несовершенного вида.

Разработанная методология, DBParCor-технология и созданная БД, сформированная на основе выровненных текстов поливариантного параллельного корпуса, позволили также уточнить семантику русских глагольных форм: варианты перевода на французский язык, обладающий более детализированной сеткой грамматических противопоставлений в области темпорально-модальных значений, выявляют определенные семантические компоненты, заключенные в значении русских глагольных форм.

В заключение отметим, что разработанная DBParCor-технология может быть адаптирована для использования в других кросслингвистических проектах, целью которых является приведение в соответствие знаний о русском языке современному состоянию лингвистической теории и эмпирической базе, представленной современными электронными корпусами, с одной стороны, и, с

Давно **собирался** к тебе, —	Depuis longtemps je **m'apprêtais** à te rendre visite.
	Il y a déjà longtemps que je **me préparais** à venir te voir,

Рис. 3 Лингвоспецифичная единица *собираться*

Когда это он **успел** опять лечь-то	Quand est-ce qu'il **a trouvé le temps** de se recoucher
	Mais, comment **a**-t-il **eu le temps** de se recoucher?

Рис. 4 Глагол *успеть*

другой стороны, потребностям современной системы образования, а также требованиям, предъявляемым новыми информационными технологиями машинного перевода. Необходимость использования кросслингвистических моделей для разработки технологий машинного перевода была обоснована в работах [21–23].

DBParCor-технология может быть использована в проектах, посвященных изучению на базе параллельных выровненных текстов лексико-грамматических форм других категорий без изменения структуры БД или с небольшими ее изменениями. Для адаптации DBParCor-технологии нужно сформировать перечень используемых языков, определить списки базовых видов ЛГФ и их дополнительных признаков для языков оригинала и перевода в соответствии с целями конкретных проектов.

Литература

1. *Aijmer K., Altenberg B.* Advances in corpus-based contrastive linguistics. Studies in honour of Stig Johansson. — Amsterdam: John Benjamins, 2013. 295 p.
2. *Добровольский Д. О., Кретов А. А., Шаров С. А.* Корпус параллельных текстов // Научная и техническая информация. Сер. 2. Информационные процессы и системы, 2005. № 6. С. 16–27.
3. Корпусные исследования по русской грамматике / Под ред. К. Л. Киселевой, Е. В. Рахилиной, В. А. Плунгяна, С. Г. Татевосова. — М.: Пробел-2000, 2009. 516 с.
4. *Добровольский Д. О.* Корпус параллельных текстов в исследовании культурно-специфичной лексики // Национальный корпус русского языка: 2006–2008. Новые результаты и перспективы. — СПб.: Нестор-История, 2009. С. 383–401.
5. *Сичинава Д. В., Шведова М. А.* Параллельные корпуса в составе Национального корпуса русского языка: технологии и решаемые задачи // Компьютерная лингвистика: научное направление и учебная дисциплина. — Гомель: ГГУ им. Ф. Скорины, 2010. С. 30–34.
6. *Сичинава Д. В.* Комплексное исследование одноязычного и параллельного корпусов в грамматических исследованиях // Корпусная лингвистика-2011: Труды Междунар. конф. — СПб.: СПбГУ, 2011. С. 316–322.
7. *Сичинава Д. В., Архангельский Т. А.* Параллельные белорусско-русский и русско-белорусский корпуса: совместный проект Национального корпуса русского языка // Труды школы-семинара TEL-2012. — Казань: КФУ, 2012. С. 54–60.
8. *Loiseau S., Sitchinava D. V., Zalizniak A. A., Zatsman I. M.* Information technologies for creating the database of equivalent verbal forms in the Russian-French multivariant parallel corpus // Информатика и её применения, 2013. Т. 7. Вып. 2. С. 100–109.
9. *Добровольский Д. О., Кретов А. А., Шаров С. А.* Корпус параллельных текстов: архитектура и возможности использования // Национальный корпус русского языка: 2003–2005. — М.: Индрик, 2005. С. 263–296.
10. *Андреева Е. Г., Касевич В. Б.* Грамматика и лексика (на материале англо-русского корпуса параллельных текстов) // Национальный корпус русского языка: 2003–2005. — М.: Индрик, 2005. С. 297–307.
11. *Гак В. Г.* Русский язык в сопоставлении с французским. — М.: УРСС, 2006. 264 с.
12. *Гак В. Г.* Сравнительная типология французского и русского языков. — М.: УРСС, 2009. 288 с.
13. *Kouznetsova I. N.* Grammaire contrastive du francais et du russe. — M.: Nestor Academic Publs., 2009. 272 p.
14. *Guiraud-Weber M.* Essais de syntaxe russe et contrastive. — Aix: Université de Provence, 2011. 337 p.
15. *Goldberg A.* Constructions: A Construction Grammar approach to argument structure. — Chicago: Univ. of Chicago Press, 1995. 265 p.
16. *Goldberg A.* Constructions at work. The nature of generalization in grammar. — Oxford: Oxford Univ. Press, 2006. 290 p.
17. Лингвистика конструкций / Под ред. Е. В. Рахилиной. — М.: Азбуковник, 2010. 584 с.
18. *Зализняк Анна А., Левонтина И. Б.* Отражение «национального характера» в лексике русского языка (размышления по поводу книги: *Wierzbicka Anna*. Semantics, culture, and cognition. Universal human concepts in culture-specific configurations. — N.Y., Oxford: Oxford Univ. Press, 1992) // Russian Linguistics, 1996. Vol. 20. No. 2/3. P. 237–264.
19. *Виноградов В. В.* История слов. — М.: Толк, 1994. 1138 с.
20. *Плунгян В. А.* Конструкция с *успеть* и *не успеть* в русском языке XIX–XX вв.: корпусное исследование // Русский язык XIX века: Проблемы изучения и лек-

сикографического описания. — СПб.: Наука, 2004. С. 112–115.

21. *Kozerenko E. B.* Cognitive approach to language structure segmentation for machine translation algorithms // MLMTA'03: Conference (International) on Machine Learning; Models, Technologies and Applications Proceedings. — Las Vegas: CSREA Press, 2003. P. 49–55.
22. *Козеренко Е. Б.* Лингвистические фильтры в статистических моделях машинного перевода // Информатика и её применения, 2010. Т. 4. Вып. 2. С. 83–92.
23. *Kozerenko E. B.* Syntactic transformations modelling for hybrid machine translation // ICAI'11, WORLDCOMP'11 Proceedings. — Las Vegas: CSREA Press, 2011. P. 875–881.

Поступила в редакцию 29.03.14

INFORMATION TECHNOLOGIES FOR CORPUS STUDIES: UNDERPINNINGS FOR CROSS-LINGUISTIC DATABASE CREATION

N. V. Buntman[1], Anna A. Zaliznyak[2,3], I. M. Zatsman[3], M. G. Kruzhkov[3], E. Yu. Loshchilova[3], and D. V. Sitchinava[4]

[1]Faculty of Foreign Languages and Area Studies, M. V. Lomonosov Moscow State University, 31-a Lomonosov Str., Moscow 119192, Russian Federation
[2]Institute of Linguistics, Russian Academy of Sciences, 1-1 Bolshyi Kislovskiy pereulok, Moscow 125009, Russian Federation
[3]Institute of Informatics Problems, Russian Academy of Sciences, 44-2 Vavilov Str., Moscow 119333, Russian Federation
[4]Institute of Russian Language, Russian Academy of Sciences, 18/2 Volkhonka Str., Moscow 119019, Russian Federation

Abstract: Information technology for creation of cross-linguistic databases of Russian texts with French translations (also known as parallel texts) is considered. The underlying principles of the developed database provide a unique combination of three types of bilingual search: lexical, grammatical, and lexico-grammatical. A distinctive feature of the considered technology is simultaneous creation of Russian-French parallel subcorpus within the National Russian Corpus and of the cross-linguistic database of Russian verbal lexico-grammatical forms and their French functional equivalents. The subcorpus and the database have different levels of alignment: the former is aligned at the level of sentences, and the later at the level of constructions. The academic relevance of the developed database is due to its support of bilingual contrastive grammar development, as well as to its role in creation of Russian grammar based on the modern empirical base and information technologies of corpus linguistics. The main practical application of the database consists in improvement of quality of machine translation.

Keywords: parallel corpus; information technology; cross-linguistic databases; bilingual lexical grammar search; corpus linguistics; contrastive grammar

DOI: 10.14357/19922264140210

Acknowledgments

The work was performed in the Institute of Informatics Problems of the Russian Academy of Sciences with financial support of Foundation "Dynasty" (grant NG13-036) and Russian Foundation for Basic Research (grant No. 13-06-00403).

References

1. Aijmer, K., and B. Altenberg. 2013. *Advances in corpus-based contrastive linguistics. Studies in honour of Stig Johansson*. Amsterdam: John Benjamins. 295 p.
2. Dobrovolsky, D. O., A. A. Kretov, and S. A. Sharoff. 2005. Korpus parallel'nykh tekstov [Corpus of parallel texts]. *Nauchnaya i Tekhnicheskaya Informatsiya. Ser. 2. Informatsionnye protsessy i sistemy* [Scientific and technical information. Ser. 2: Information processes and systems] 6:16–27.
3. Kiseleva, K. L., E. V. Rahilina, V. A. Plungjan, and S. G. Tatevosov, eds. 2009. *Korpusnye issledovaniya po russkoy grammatike* [Corpus studies on Russian grammar]. Moscow: Probel-2000. 516 p.
4. Dobrovolsky, D. O. 2009. Korpus parallel'nykh tekstov v issledovanii kul'turno-spetsifichnoy leksiki [A corpus

of parallel texts and studying culture-specific lexicon]. *Natsional'nyy korpus russkogo yazyka: 2006–2008. Novye rezul'taty i perspektivy* [Russian National Corpus: 2006–2008. New results and prospects]. St. Petersburg: Nestor-Istoriya. 383–401.

5. Sitchinava, D. V., and M. A. Shvedova. 2010. Parallel'nye korpusa v sostave Natsional'nogo korpusa russkogo yazyka: Tekhnologii i reshaemye zadachi [Parallel corpora of the Russian National Corpus: Technologies and problems]. *Komp'yuternaya lingvistika: Nauchnoe napravlenie i uchebnaya distsiplina* [Computational linguistics: Scientific field and academic discipline]. Gomel': Gomel' University. 30–34.
6. Sitchinava, D. V. 2011. Kompleksnoe issledovanie odnoyazychnogo i parallel'nogo korpusov v grammaticheskikh issledovaniyakh [Comprehensive study of monolingual and parallel corpora in grammatical studies]. *Korpusnaya Lingvistika-2011: Trudy Konferentsii* [Corpus-Based Linguistics-2011 Proceedings]. St. Petersburg. 316–322.
7. Sitchinava, D. V., and T. A. Arhangel'skij. 2012. Parallel'nye belorussko-russkiy i russko-belorusskiy korpusa: Sovmestnyy proekt Natsional'nogo korpusa russkogo yazyka [Parallel Belarusian-Russian and Russian-Belarusian corpora: Joint project of the Russian National Corpus]. School-Seminar TEL-2012 Proceedings. Kazan': Kazan' University. 54–60.
8. Loiseau, S., D. V. Sitchinava, A. A. Zalizniak, and I. M. Zatsman. 2013. Information technologies for creating the database of equivalent verbal forms in the Russian-French multivariant parallel corpus. *Informatika i ee Primeneniya — Inform. Appl.* 7(2):100–109.
9. Dobrovolsky, D. O., A. A. Kretov, and S. A. Sharoff. 2005. Korpus parallel'nykh tekstov: Arkhitektura i vozmozhnosti ispol'zovaniya [Corpus of parallel texts: Architecture and usage]. *Natsional'nyy korpus russkogo yazyka: 2003–2005* [Russian National Corpus 2003–2005]. Moscow: Indrik. 263–296.
10. Andreeva, E. G., and V. B. Kasevich. 2005. Grammatika i leksika (na materiale anglo-russkogo korpusa parallel'nykh tekstov) [Grammar and lexicon in the English-Russian corpus of parallel texts]. *Natsional'nyy korpus russkogo yazyka: 2003–2005* [Russian National Corpus 2003–2005]. Moscow: Indrik. 297–307.
11. Gak, V. G. 2006. *Russkiy yazyk v sopostavlenii s frantsuzskim* [Russian language compared to French]. Moscow: URSS. 264 p.
12. Gak, V. G. 2009. *Sravnitel'naya tipologiya frantsuzskogo i russkogo yazykov* [Comparative typology of French and Russian]. Moscow: URSS. 288 p.
13. Kouznetsova, I. N. 2009. *Grammaire contrastive du francais et du russe*. Moscow: Nestor Academic Publs. 272 p.
14. Guiraud-Weber, M. 2011. *Essais de syntaxe russe et contrastive*. Aix: Université de Provence. 337 p.
15. Goldberg, A. 1995. *Constructions: A construction grammar approach to argument structure*. Chicago: Univ. of Chicago Press. 265 p.
16. Goldberg, A. 2006. *Constructions at work. The nature of generalization in language*. Oxford: Oxford Univ. Press. 290 p.
17. Rakhilina, E. V., ed. 2010. *Lingvistika konstruktsiy* [Construction linguistics]. Moscow: Azbukovnik. 584 p.
18. Zaliznjak, Anna A., and I. B. Levontina. 1996. Otrazhenie "natsional'nogo kharaktera" v leksike russkogo yazyka (razmyshleniya po povodu knigi: Anna Wierzbicka. 1992. *Semantics, culture, and cognition. Universal human concepts in culture-specific configurations*. — New York, Oxford: Oxford Univ. Press) [Representation of "national character" in the Russian lexicon (reflections on the book: Anna Wierzbicka. 1992. *Semantics, culture, and cognition. Universal human concepts in culture-specific configurations*. New York, Oxford: Oxford Univ. Press]. *Russian Linguistics* 20:237–264.
19. Vinogradov, V. V. 1994. *Istoriya slov* [History of words]. Moscow: Tolk. 1138 p.
20. Plungjan, V. A. 2004. Konstruktsiya s uspet' i ne uspet' v russkom yazyke XIX–XX vv.: Korpusnoe issledovanie [Constructions with "uspet"' and "ne uspet"' in Russian language in XIX–XX centuries: Corpus-based studies]. *Russkiy yazyk XIX veka: Problemy izucheniya i leksikograficheskogo opisaniya* [Russian language in XIX century: Studies and lexicographical description]. St. Petersburg: Nauka. 112–115.
21. Kozerenko, E. B. 2003. Cognitive approach to language structure segmentation for machine translation algorithms. *MLMTA'03: Conference (International) on Machine Learning; Models, Technologies and Applications Proceedings*. Las Vegas. 49–55.
22. Kozerenko, E. B. 2010. Lingvisticheskie fil'try v statisticheskikh modelyakh mashinnogo perevoda [Linguistic filters for statistical machine translation models]. *Informatika i ee Primeneniya — Inform. Appl.*] 4(2):83–92.
23. Kozerenko, E. B. 2011. Syntactic transformations modelling for hybrid machine translation. *ICAI'11, WORLDCOMP'11 Proceedings*. Las Vegas. 875–881.

Received March 29, 2014

Contributors

Buntman Nadezhda V. (b. 1957) — Candidate of Science (PhD) in philology, associated professor, Faculty of Foreign Languages and Area Studies, M. V. Lomonosov Moscow State University, 31-a Lomonosov Str., Moscow 119192, Russian Federation; nabunt@hotmail.com

Zalizniak Anna A. (b. 1959) — Doctor of Science in philology, leading scientist, Institute of Linguistics, Russian Academy of Sciences, 1-1 Bolshoy Kislovskiy pereulok, Moscow 125009, Russian Federation; Institute of Informatics Problems, Russian Academy of Sciences, 44-2 Vavilov Str., Moscow 119333, Russian Federation; anna.zalizniak@gmail.com

Zatsman Igor M. (b. 1952) — Doctor of Science in technology, Head of Department, Institute of Informatics Problems, Russian Academy of Sciences, 44-2 Vavilov Str., Moscow 119333, Russian Federation; izatsman@yandex.ru

Kruzhkov Mikhail G. (b. 1975) — leading programmer, Institute of Informatics Problems, Russian Academy of Sciences, 44-2 Vavilov Str., Moscow 119333, Russian Federation; magnit75@yandex.ru

Loshchilova Elena J. (b. 1960) — scientist, Institute of Informatics Problems, Russian Academy of Sciences, 44-2 Vavilov Str., Moscow 119333, Russian Federation; lena0911@mail.ru

Sitchinava Dmitri V. (b. 1980) — Candidate of Science (PhD) in philology, senior scientist, Institute of the Russian Language, Russian Academy of Sciences, 18/2 Volkhonka Str., Moscow 119019, Russian Federation; mitrius@gmail.com

ИНФОРМАТИКА И ЕЁ ПРИМЕНЕНИЯ, 2014. Т. 8. Вып. 2. С. 111–121

ПОСТРОЕНИЕ МОДЕЛЕЙ СИСТЕМНОЙ ДИНАМИКИ В УСЛОВИЯХ ОГРАНИЧЕННОЙ ЭКСПЕРТНОЙ ИНФОРМАЦИИ*

О. Г. Кантор[1], С. И. Спивак[2]

Аннотация: Системная динамика — методология изучения сложных динамических систем, ориентированная на проведение компьютерного эксперимента. Построение моделей системной динамики во многом зависит от имеющейся экспериментальной информации и квалификации экспертов. Имитационный эксперимент с «плохими» моделями может привести к существенному или даже полному искажению свойств изучаемой системы. В настоящей работе приводится описание метода построения моделей системной динамики, представляющего собой комплекс математических моделей, в основу которых положены идеи подхода Л. В. Канторовича к математической обработке экспериментальных данных, и вычислительных процедур. Важным преимуществом при этом является возможность включения в модель значимых с точки зрения исследователя условий, влияющих на ее адекватность. Апробация разработанного метода осуществлялась на примере моделирования численности населения Российской Федерации.

Ключевые слова: модели системной динамики; точечные и интервальные оценки параметров моделей; подход Л. В. Канторовича; предельно допустимые погрешности измерений

DOI: 10.14357/19922264140211

1 Введение

Основу концепции системной динамики составляет представление функционирования изучаемой системы в виде совокупности потоков ресурсов, определяющих ее. При этом подразумевается достаточно высокая степень агрегирования, в результате чего рассмотрению подлежат лишь наиболее значимые факторы [1–4].

Основой для построения уравнений системной динамики служат дифференциальные модели. При построении математических моделей системной динамики используются переменные двух типов: системные уровни и темпы. Системные уровни полностью описывают состояние системы в произвольный момент времени. Изменение системных уровней вызвано соответствующими темпами, которые, в свою очередь, зависят от одного или нескольких системных уровней (но не от других темпов). В некоторых случаях в целях более детального отражения процессов, протекающих в изучаемой системе, и/или более удобной формы записи уравнений модели системной динамики могут использоваться вспомогательные переменные.

В моделях системной динамики для всех системных уровней пишутся уравнения одного и того же типа [1, 3]:

$$\frac{d\bar{x}}{dt} = f\left(\bar{x}, \bar{a}\right) = \bar{x}^{+} - \bar{x}^{-} , \qquad (1)$$

где $(\bar{x}, \bar{a})$— вектор-функция, зависящая от переменных $\bar{x}$ и параметров $\bar{a}$ модели; $\bar{x}^{+}$ и $\bar{x}^{-}$ — положительный и отрицательный темпы скорости системных уровней $\bar{x}$, каждый из которых включает в себя все факторы, вызывающие соответственно рост и убывание $\bar{x}$.

Аналитическое решение систем дифференциальных уравнений (1) с учетом размерности реальных задач представляет собой практически неразрешимую задачу. Поэтому традиционным является переход от дифференциальных уравнений (1) к их разностным аналогам:

$$\Delta\bar{x} = f\left(\bar{x}, \bar{a}\right) , \qquad (2)$$

для численного интегрирования которых разработано большое количество методов. (Далее для определенности будем говорить только о модели (2), подразумевая, что если ее параметры известны, то и модель (1) также определена.)

Следует заметить, что темпы в случае использования модели (1) показывают закон изменения соответствующих системных уровней, а в случае их разностных аналогов (2) — каким образом изменяются соответствующие системные уровни за временной интервал, равный шагу моделирования, выбор которого во многом зависит от имеющейся экспериментальной информации, получаемой из фактических наблюдений за исследуемой системой.

*Работа выполнена при финансовой поддержке РФФИ (проект № 13-01-00749).

[1]Институт социально-экономических исследований Уфимского научного центра Российской академии наук, o_kantor@mail.ru

[2]Башкирский государственный университет, semen.spivak@mail.ru

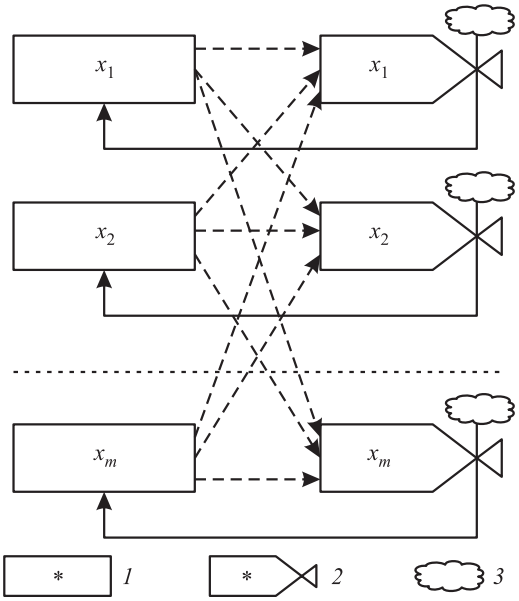


Рис. 1 Потоковая диаграмма модели системной динамики общего вида: *1* — системные уровни; *2* — системные темпы; *3* — неучтенные факторы

В некоторых случаях уравнения системной динамики составляются на основе очевидных логических связей между системными уровнями и темпами. Ключевая роль при этом отводится экспертам, опыт и знания которых позволяют полагаться на их мнения и оценки. В более сложных ситуациях, например, когда ни исследователю, ни экспертам до конца не ясно, каким образом выбранные для анализа системные уровни и темпы взаимодействуют друг с другом, определение точного вида зависимостей (2) представляет собой самостоятельную задачу.

Одним из способов наглядного отображения процессов, протекающих в изучаемой системе, являются потоковые диаграммы (рис. 1), которые обеспечивают целостное представление структуры уравнений (2), включая отображение причинно-следственных связей и петель обратных связей. Потоковые диаграммы, по сути, представляют собой инструмент системного анализа проблемы, применение которого способствует ее детальному пониманию и в ряде случаев позволяет осуществить декомпозицию задачи на несколько самостоятельных задач меньшей размерности, тем самым снижая общую трудоемкость процесса получения решения.

Отсутствие априори известных закономерностей является характерной особенностью задач, связанных с моделированием социально-экономических систем, при исследовании которых необходимо учитывать ряд особенностей, наиболее существенные из которых — ограниченность имеющейся информации и ее неточность. Ограниченность информации обусловливается изменчивостью самих социально-экономических систем, в процессе развития которых может возникать или исчезать необходимость сбора той или иной статистической информации, при этом сам по себе процесс сбора и обработки информации является достаточно длительным и трудоемким. Неточность исходной информации объясняется многостадийностью процесса ее сбора, существенной долей субъективизма, а в некоторых случаях и умышленным искажением информации с целью «приукрасить» реальное положение дел. При изучении социально-экономических систем не представляется возможным оценить точность полученных наблюдений с помощью применения стандартных подходов, основанных на сравнении данных с эталонными значениями, известными изначально или полученными на основе многократного проведения наблюдений в одних и тех же условиях, по причине их отсутствия или невозможности организации соответствующего эксперимента. В таких условиях едва ли разумно полностью полагаться на экспертные оценки, если даже таковые и будут получены.

Аналогичные проблемы могут возникать и не только при изучении социально-экономических систем. Полагаясь на мнение эксперта, исследователь всегда рискует, в силу того что даже самый квалифицированный эксперт может оказаться неправым и все усилия исследователя по построению математической модели и дальнейшие эксперименты с ней окажутся напрасными.

2 Постановка задачи

В условиях ограниченности экспертных представлений об исследуемой системе исследователь может обладать информацией лишь самого общего характера, что не позволяет ему идентифицировать зависимости (2). Возникающую на этапе построения уравнений системной динамики неопределенность целесообразно подразделять на два типа:

(1) *неопределенность первого типа*, характеризующуюся отсутствием информации о значениях параметров зависимостей заданной функциональной структуры, что соответствует ситуации, когда известны закономерности, связывающие системные уровни и системные темпы, но параметры, фигурирующие в них, подлежат определению;

(2) *неопределенность второго типа*, при которой неизвестна сама функциональная структура связи между системными уровнями и темпами.

Неопределенность второго типа характерна для ситуаций, когда исследователь либо не доверяет мнению экспертов, либо в принципе не прибегает к их опросу, а его собственных знаний об изучаемом объекте недостаточно для того, чтобы самостоятельно определить функциональные связи и соответствующие им параметры.

Очевидно, что «раскрывать» неопределенность второго типа можно посредством последовательного рассмотрения приемлемых с точки зрения исследователя структур для уравнений (2). При этом каждый такой отдельный случай порождает неопределенность первого типа, что и обусловливает актуальность методов определения значений параметров моделей системной динамики в соответствующих условиях.

Любая задача определения значений параметров моделей заданной функциональной структуры, по сути, сводится к поиску такого набора параметров $\bar{a}$, которые обеспечивали бы приемлемые с точки зрения исследователя качественные характеристики. Выбор конкретного инструментария для определения совокупности параметров модели зависит лишь от объема знаний и предпочтений самого исследователя. Большое значение при этом имеет цель, которую ставит перед собой исследователь. Так, если целью является максимально точное соответствие имеющимся наблюдениям, то, скорее всего, определяющим критерием будет выступать близость расчетных и экспериментальных данных; если же целью исследования ставится адекватность модели каким-либо свойствам реального объекта, то логично в качестве определяющего критерия выбрать степень соответствия модели именно таким свойствам. Следует особо отметить, что критерий, обеспечивающий достижение лучшего значения по одной из качественных характеристик, вовсе не обязательно будет обеспечивать хорошее значение по другой характеристике.

Определение параметров модели (2) на основании имеющейся экспериментальной информации сопряжено с двумя существенными проблемами:

(1) число наблюдений n в практических задачах превышает (а чаще — существенно превышает) число параметров, подлежащих определению, поэтому решаемые задачи априори являются *переопределенными*;

(2) такого рода задачи изначально являются *некорректными*, так как приближенный характер исходной информации влечет невыполнение требований, предъявляемых к корректно поставленным задачам (существование точного решения, его единственность и устойчивость к малым изменениям исходных данных) [5].

Перечисленные проблемы ограничивают, а иногда и вовсе исключают возможность применения классических методов, таких как, например, методы статистического анализа. Если же исследователь будет обладать информацией о диапазонах вариации каждого из параметров $\bar{a}$, то определение непосредственного вида модели (2) может быть осуществлено посредством реализации специального численного эксперимента, в ходе которого из множества допустимых значений параметров $\bar{a}$ будет выбран единственный набор $\bar{a}^*$, доставляющий оптимальное значение некоторому критерию, отражающему, по мнению исследователя, соответствие расчетных и экспериментальных данных.

В некоторых случаях диапазоны значений параметров модели могут быть заданы исходя из их смысловой нагрузки, однако в общем случае их определение является самостоятельной задачей. Решение этой задачи «на глазок» может привести либо к слишком большим диапазонам вариации параметров модели, что потребует при организации численного эксперимента по определению параметров $\bar{a}^*$ существенных временны́х затрат и повышенных требований к производительности вычислительной техники, либо к ситуации, когда в заданном множестве по результатам численного эксперимента оптимальный набор параметров $\bar{a}^0$ попросту не будет существовать.

Исследование в настоящей работе посвящено разработке метода определения значений параметров моделей системной динамики в условиях ограниченности информации относительно значений параметров искомых зависимостей.

3 Описание метода определения диапазона вариации параметров моделей системной динамики

Для решения задачи определения диапазона вариации параметров моделей системной динамики авторы разработали метод, базирующийся на подходе, основоположником которого является Л. В. Канторович, впервые высказавший идеи получения точных двусторонних границ для параметров моделей и областей расположения искомых и наблюдаемых величин [6]. (Дальнейшее описание предлагаемого метода осуществляется на примере построения зависимости для одного отдельно взятого системного уровня, для которого необходимо определить точный вид функциональной зависимости $dx/dt = f(x, \bar{a}) = x^+ - x^-$ и, соответственно, ее разностного аналога $\Delta x = f(x, \bar{a})$.)

Традиционно проверка соответствия расчетных и экспериментальных данных осуществляется по-

средством введения в рассмотрение величин отклонений:

$$\eta_j = \Delta x^{\text{расч}}|_j - \Delta x^{\text{эксп}}|_j = \\ = f\left(x^{\text{расч}}|_j, \bar{a}\right) - \Delta x^{\text{эксп}}|_j\,, \quad j = \overline{1,n}\,,$$

где $\Delta x^{\text{эксп}}|_j$ — известное из наблюдений изменение переменной модели в j-й момент времени; $x^{\text{расч}}|_j$ — рассчитанное согласно модели значение переменной x в j-й момент времени; n — общее число имеющихся наблюдений.

Стандартный путь решения задачи определения значений параметров модели (2) заключается в минимизации отклонений $\{\eta_j, j = \overline{1,n}\}$ в смысле некоторого введенного критерия. В рамках математической статистики дается обоснование вида такого критерия в случае известного закона распределения погрешности измерений. Так, если погрешности измерений подчиняются нормальному закону распределения или распределению Лапласа, то критерий соответственно принимает вид

$$\sum_{j=1}^{n} \frac{1}{\sigma_j^2}\,\eta_j^2\,, \tag{3}$$

или

$$\sum_{j=1}^{n} \frac{1}{\sigma_j}\,|\eta_j|\,,$$

где σ_j^2 — дисперсия измерений, $j = \overline{1,n}$.

В реальных системах, как правило, информация о законе распределения погрешности измерений отсутствует, в то время как доступной является информация о предельно допустимой погрешности измерений (именно этот факт и был взят за основу Канторовичем в работе [6]). Условие того, что модель описывает наблюдаемые величины, приводит к системе неравенств

$$|\eta_j| = \left|f\left(x^{\text{расч}}|_j, \bar{a}\right) - \Delta x^{\text{эксп}}|_j\right| \le \varepsilon_j\,,\ j = \overline{1,n}\,, \tag{4}$$

где ε_j — погрешность j-го измерения, численное решение которой предполагает использование в качестве начального приближения хотя бы одной точки, обеспечивающей справедливость всех соотношений (4).

Основной принцип интервального оценивания в свете подхода Канторовича состоит в том, что величины отклонений η_j должны находиться в пределах погрешностей измерений ε_j. Описывая предлагаемый им подход к обработке наблюдений [6], Канторович считал, что исследователь должен располагать информацией о величине предельно допустимой погрешности ε_j. Однако далеко не всегда это является возможным (например, при исследовании социально-экономических систем в силу упомянутых выше особенностей). Более того, даже в тех случаях, когда известны величины погрешностей измерений ε_j, система (4) может оказаться несовместной.

В этой связи, по мнению авторов, целесообразно величины ε_j рассматривать как неизвестные, что позволит осуществлять поиск точки, гарантирующей справедливость всех соотношений (4), исходя из условия обеспечения оптимума любого критерия, характеризующего соответствие расчетных и экспериментальных данных. В качестве такого критерия авторы применили функцию

$$\max_{1\le j\le n} |\eta_j|\,, \tag{5}$$

используемую в методе выравнивания по Чебышёву. Основная идея метода выравнивания по Чебышёву заключается в приближении экспериментальных данных таким способом, чтобы обеспечивалась равномерная точность описания во всей исследуемой области, а из всей экспериментальной информации фактически используется $k+1$ точка, где k — число искомых параметров. Очевидно, что оптимальные параметры модели (2) должны минимизировать норму (5), а это эквивалентно решению задачи

$$\min_{\Omega}\ \max_{1\le j\le n} |\eta_j|\,, \tag{6}$$

где Ω — множество допустимых значений искомых параметров $\bar{a}$, в качестве которого в первом приближении может быть взят многомерный параллелепипед произвольного размера.

Задача (6) сводится [5] к решению оптимизационной задачи с помощью введения дополнительного параметра λ, такого что

$$|\eta_j| \le \lambda\,, \quad j = \overline{1,n}\,.$$

В результате задача (6) может быть формализована следующим образом:

$$\left.\begin{array}{l} \lambda \to \min\limits_{\Omega,\lambda}\,; \\ -\lambda \le f\left(x^{\text{расч}}|_j, \bar{a}\right) - \Delta x^{\text{эксп}}|_j \le \lambda\,,\ j = \overline{1,n}\,; \\ \lambda \ge 0\,. \end{array}\right\} \tag{7}$$

Данная задача, в отличие от задачи (3), является совместной всегда. Ее решением будет набор точечных оценок параметров модели (2) $\bar{a}^*$ и значение λ^*.

Определение диапазонов вариации параметров $\bar{a}$ может быть осуществлено посредством решения оптимизационных задач вида

$$\left.\begin{array}{l} a_i \to \min\limits_{\Omega}(\max\limits_{\Omega})\,; \\ \left| f\left(x^{\text{расч}}|_j\,, \bar{a}\right) - \Delta x^{\text{эксп}}|_j \right| \le \lambda^*\,, \;\; j = \overline{1,n}\,, \end{array}\right\} \quad (8)$$

для каждого отдельно взятого компонента a_i вектора $\bar{a}$. При этом в целях получения большей информации о диапазонах значений параметров модели (2) правые части системы ограничений могут варьироваться, т. е. вместо модели (8) могут рассматриваться модели вида

$$a_i \to \min_{\Omega}(\max_{\Omega})\,;$$
$$\left| f\left(x^{\text{расч}}|_j\,, \bar{a}\right) - \Delta x^{\text{эксп}}|_j \right| \le \lambda^*(1+\delta)\,, \;\; j = \overline{1,n}\,,$$

где δ — числовой параметр, $\delta \ge 0$.

4 Упрощения и допущения

Основная сложность практической реализации описанного подхода применительно к моделям системной динамики вытекает из спецификации правых частей соотношений (2) и заключается, во-первых, в большой размерности задач (7) и (8) и, во-вторых, в нелинейности вектор-функции $f(\bar{x}, \bar{a})$, что существенно усложняет процесс решении. Для решения перечисленных проблем авторы использовали идеи линеаризации уравнений системной динамики (2) по параметрам $\bar{a}$ [7]. Благодаря этому задача определения диапазонов вариации параметров модели (2) (т. е. решения задач (7) и (8)) сводится к классическим задачам линейного программирования, для решения которых разработано множество эффективных алгоритмов. Такое упрощение приводит к определенной потере точности искомых интервалов, однако названный недостаток можно считать компенсированным за счет экономии временны́х и программных ресурсов.

Важным преимуществом разработанного метода является возможность включения в модель на этапе численного интегрирования системы (2) различных дополнительных условий, соблюдение которых продиктовано очевидными соображениями, что позволяет повысить степень адекватности модели, но не осуществимо в рамках классических методов. В качестве примера такого рода условий могут быть названы ограничения на будущие значения переменных модели (2) или их предполагаемые приращения. Совокупность всех таких условий может быть формализована в виде ограничений $G(\bar{x}, \bar{a}) \subset S^0$, подлежащих включению в модели (7) и (8).

К определению параметров моделей системной динамики авторы подошли с позиций учета следующих аспектов: необходимо добиваться, во-первых, близости расчетных и экспериментальных данных, во-вторых, — минимально возможной области предельно допустимых погрешностей аппроксимации; в-третьих, — минимального уровня вариации оцениваемых параметров. (Под минимально возможной областью предельно допустимых погрешностей аппроксимации подразумевается область значений величин $\{\eta_j\}$ с минимальным диаметром.) Соблюдение третьего принципа позволяет снизить неопределенность, обусловленную неединственностью решения поставленной задачи.

С этих позиций в рамках предлагаемого подхода формализация показателей качественных характеристик модели (2), а именно точности, адекватности и пр., может быть обеспечена посредством задания целевой функции и ограничений, отражающих каждая в отдельности одну из качественных характеристик. Способ их непосредственного задания должен определяться исследователем на основе анализа специфики самой задачи и цели моделирования. Так, возможным вариантом критерия близости расчетных и экспериментальных данных на стадии численного интегрирования системы (2) может служить средняя ошибка аппроксимации, что позволяет реализовать все названные принципы [8]. (Средняя ошибка аппроксимации рассчитывается по формуле

$$\bar{A} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i^{\text{расч}} - y_i^{\text{эксп}}}{y_i^{\text{эксп}}}\right| \cdot 100\%\,,$$

где $y_i^{\text{расч}}$ и $y_i^{\text{эксп}}$ — соответственно рассчитанные согласно полученной модели и известные из наблюдений значения исследуемого фактора в i-й момент времени; n — общее число наблюдений.)

Заметим, что формализация критерия, характеризующего точность по совокупности уравнений модели (2), представляет собой самостоятельную задачу, решение которой находится в компетенции исследователя [9].

5 Апробация метода

Апробация разработанного метода осуществлялась при моделировании численности населения Российской Федерации. Общий вид исследованной авторами модели системной динамики в терминах разностных уравнений следующий:

$$\left.\begin{array}{l} \Delta N = a_1 N^{\alpha_1} D^{\beta_1} I^{\gamma_1} - a_2 N^{\alpha_2} D^{\beta_2} I^{\gamma_2}\,; \\ \Delta D = a_3 N^{\alpha_3} D^{\beta_3} I^{\gamma_3} - a_4 N^{\alpha_4} D^{\beta_4} I^{\gamma_4}\,; \\ \Delta I = a_5 N^{\alpha_5} D^{\beta_5} I^{\gamma_5} - a_6 N^{\alpha_6} D^{\beta_6} I^{\gamma_6}\,, \end{array}\right\} \quad (9)$$

Таблица 1 Исходные данные для модели (9)

Год	Численность населения РФ N, чел.	Душевые доходы D, руб./чел. в год	Индекс потребительских цен I, доли ед.
1998	147 802 133	12 122,4	1,844
1999	147 539 426	19 906,8	1,365
2000	146 890 128	27 373,2	1,202
2001	146 303 611	36 744,0	1,186
2002	145 649 334	47 366,4	1,151
2003	144 963 650	62 044,8	1,120
2004	144 168 205	76 923,6	1,117
2005	143 474 219	97 342,8	1,109
2006	142 753 551	122 352,0	1,090
2007	142 220 968	151 232,4	1,119
2008	142 008 800	179 287,2	1,133
2009	141 904 000	202 282,8	1,088

где N — численность населения РФ; D — душевые доходы за год; I — индекс потребительских цен. Информационную базу исследования составили данные официальной статистической отчетности за период с 1998 по 2009 гг. (табл. 1).

В целях сокращения проблем вычислительного характера была использована двухэтапная процедура, основанная на применении разработанного метода к линеаризованным правым частям модели (9). Целесообразность двухэтапной процедуры обусловлена необходимостью задания центра разложения в процедуре линеаризации и спецификой уравнений (9). Используя разложение их правых частей в ряд Маклорена по переменным $\{a_i, \alpha_i, \beta_i, \gamma_i\}$, $i = \overline{1,6}$, получим систему:

$$\left.\begin{aligned} \Delta N &\approx a_1 - a_2\,; \\ \Delta D &\approx a_3 - a_4\,; \\ \Delta I &\approx a_5 - a_6\,, \end{aligned}\right\} \qquad (10)$$

в которой отсутствуют параметры, характеризующие показатели степеней всех переменных модели. Именно по этой причине на первом этапе с помощью разложения в ряд Маклорена (10) определялись точечные и интервальные оценки величин $\{a_{2i-1} - a_{2i}\}$, $\left\{(a_{2i-1} - a_{2i})^0\right\}$ и $\left\{\left[(a_{2i-1} - a_{2i})^-\,; (a_{2i-1} - a_{2i})^+\right]\right\}$, $i = \overline{1,3}$.

Разложение правых частей соотношений модели (9) в ряд Тейлора с центром в точке $\left\{a_i^0, \alpha_i = 0, \beta_i = 0, \gamma_i = 0\right\}$, $i = \overline{1,6}$, имеют следующий вид (на примере первого уравнения):

$$\Delta N \approx a_1 + a_1^0 \ln N \cdot \alpha_1 + a_1^0 \ln D \cdot \beta_1 + a_1^0 \ln I \cdot \gamma_1 - \\ - a_2 - a_2^0 \ln N \cdot \alpha_2 - a_2^0 \ln D \cdot \beta_2 - a_2^0 \ln I \cdot \gamma_2\,. \qquad (11)$$

Таким образом, используя разложение в ряд Тейлора (11), можно осуществлять идентификацию всех параметров первого уравнения модели (9). (Строго говоря, вместо точечных и интервальных оценок параметров a_1 и $_2$ будут найдены точечные и интервальные оценки для выражения $(a_1 - a_2)$. Однако это обстоятельство несущественно усложнит процесс численного эксперимента по определению оптимального набора параметров системы (9).) Выбор центров разложения в окрестности нулевых значений искомых параметров был продиктован стремлением получить для них как можно меньшие значения, что, в свою очередь, объясняется смыслом каждого из слагаемых исследуемых уравнений системной динамики: рост и уменьшение каждой из введенных в рассмотрение переменных должны обладать умеренной интенсивностью. Следует заметить, что использование только разложения в ряд Тейлора требует обоснованного подхода к выбору точки, служащей центром разложения.

Разработанный авторами метод (см. разд. 3) применялся для каждого уравнения модели (9) в отдельности в рамках каждого из двух описанных выше этапов процедуры, основанной на линеаризации исследуемых зависимостей. Далее для определенности все рассуждения приводятся на примере первого уравнения системы (9). Для остальных уравнений этой системы все рассуждения аналогичны.

Модель (7) для первого уравнения системы (10) имеет вид:

$$\left.\begin{aligned} \lambda &\to \min_{a_1, a_2, \lambda}\,; \\ -\lambda \le a_1 - a_2 - \Delta N^{\text{эксп}}\big|_j &\le \lambda\,,\ \ j = \overline{1,11}\,; \\ a_1 &\ge 0\,; \\ a_2 &\ge 0\,; \\ \lambda &\ge 0\,, \end{aligned}\right\} \qquad (12)$$

где $\Delta N^{\text{эксп}}\big|_j$ — годовые приращения величины N. Требования неотрицательности параметров a_1 и a_2 следуют из смысла слагаемых уравнений системной динамики. Результатом решения задачи (12) являются величины $(a_1 - a_2)^0$ и λ^0.

Для определения диапазона вариации величины $a_1 - a_2$ решалась задача вида (8), которая для первого уравнения системы (10) имеет вид:

$$\left.\begin{aligned} a_1 - a_2 &\to \min_{a_1, a_2}(\max_{a_1, a_2})\,; \\ \left|a_1 - a_2 - N^{\text{эксп}}\big|_j\right| &\le \lambda^0\,,\ \ j = \overline{1,11}\,; \\ a_1 &\ge 0\,; \\ a_2 &\ge 0\,. \end{aligned}\right\} \qquad (13)$$

Результаты численной реализации задач (12) и (13) (табл. 2) позволяют осуществлять обоснованный выбор центра разложения на втором этапе

Таблица 2 Оптимальные решения задач (12) и (13)

Параметр	Точечные оценки $(a_1-a_2)^0$	Минимальное значение $\left((a_1-a_2)^0\right)^{\min}$	Максимальное значение $\left((a_1-a_2)^0\right)^{\max}$	λ^0
a_1-a_2	−392457,5	−795445,0	10530,0	402987,5
a_3-a_4	1514,45	622,20	2406,70	892,25
a_5-a_6	0,227	0,088	0,365	0,139

процедуры линеаризации системы (9), при этом полученные интервальные оценки в случае необходимости предоставляют исследователю дополнительную степень свободы при определении его координат.

Далее на основании полученных результатов (см. табл. 2) и разложения (11) решались задачи определения точечных:

$$\left.\begin{aligned}
&\lambda \to \min_{\left\{a_i,\alpha_i,\beta_i,\gamma_i \mid j=\overline{1,2}\right\},\lambda};\\
&-\lambda \le a_1^0 + a_1 + a_1^0 \ln N\cdot\alpha_1 + a_1^0 \ln D\cdot\beta_1 +\\
&\quad + a_1^0 \ln I\cdot\gamma_1 - a_2^0 - a_2 - a_2^0 \ln N\cdot\alpha_2 -\\
&\quad - a_2^0 \ln D\cdot\beta_2 - a_2^0 \ln I\cdot\gamma_2 -\\
&\quad - \Delta N^{\text{эксп}}\big|_j \le \lambda;\ \ j=\overline{1,11};\\
&a_1 \ge 0;\\
&a_2 \ge 0;\\
&\lambda \ge 0
\end{aligned}\right\} \quad (14)$$

и интервальных оценок параметров модели (9):

$$\left.\begin{aligned}
&a_1-a_2 \to \min_{\left\{a_i,\alpha_i,\beta_i,\gamma_i \mid j=\overline{1,2}\right\}};\\
&a_1^0 + a_1 + a_1^0 \ln N\cdot\alpha_1 + a_1^0 \ln D\cdot\beta_1 +\\
&+ a_1^0 \ln I\cdot\gamma_1 - a_2^0 - a_2 - a_2^0 \ln N\cdot\alpha_2 -\\
&- a_2^0 \ln D\cdot\beta_2 - a_2^0 \ln I\cdot\gamma_2 - \Delta N^{\text{эксп}}\Big|_j \le \lambda^*,\\
&\qquad\qquad j=\overline{1,11};\\
&a_1 \ge 0;\\
&a_2 \ge 0.
\end{aligned}\right\} \quad (15)$$

Результатом решения задачи (14) являются точечные оценки $(a_1-a_2)^*$, $\alpha^*_{1,2}$, $\beta^*_{1,2}$, $\gamma^*_{1,2}$, λ^*.

В (15) приведен вид оптимизационной задачи для определения интервалов значений величины (a_1-a_2). Для остальных параметров модели системной динамики составлялись аналогичные задачи, при этом на диапазон их значений накладывались ограничения (столбец 2 табл. 3), которые в сочетании с условиями на неотрицательность параметров $\left\{a_i, i=\overline{1,6}\right\}$ формировали упомянутое выше множество Ω (15).

Результаты численной реализации моделей (14) и (15) (см. табл. 3) были использованы для организации специальной вычислительной процедуры по определению оптимального набора параметров модели (9) (рис. 2).

Как отмечалось ранее, существенным преимуществом разработанного метода является возможность учета априорных ограничений на значения параметров искомых зависимостей, известных из очевидных соображений, что позволяет значительно сократить неопределенность решаемых задач. В качестве таких ограничений были использованы условия на приращения переменных:

$$\left|\Delta N|_j\right| \le 0{,}006N, \qquad j=\overline{1,12}; \quad (16)$$

$$\left|\Delta D|_j\right| \le 0{,}7D, \qquad j=\overline{1,12}; \quad (17)$$

$$\left|\Delta I|_j\right| \le 0{,}7I, \qquad j=\overline{1,12}, \quad (18)$$

и их будущие значения:

$$|N_{13}^{\text{расч}} - N_{12}^{\text{эксп}}| \le 100\,000; \quad (19)$$

$$|D_{13}^{\text{расч}} - D_{12}^{\text{эксп}}| \le 120\,000. \quad (20)$$

Условие (16) обусловлено максимальным за весь период 1998–2009 гг. изменением показателя численности населения: в 2004 г. численность населения РФ сократилась на 0,6% (что соответствует примерно 800 тыс. чел.). Условия (17) и (18) ограничивают рост переменных D (душевых доходов за год) и I (индекса потребительских цен) величиной в 70%. Условия (19) и (20) отражают тенденцию изменения соответствующих переменных, сложившуюся к 2010 г. ($N_{13}^{\text{расч}}$ и $D_{13}^{\text{расч}}$ — рассчитанные согласно модели (9) прогнозные значения переменных N и D в момент времени $j=13$, т. е. для 2010 г.).

В качестве числовых критериев, характеризующих точность каждого уравнения модели (9), рассматривались средние ошибки аппроксимации, для которых приемлемым считался уровень, не превы-

Таблица 3 Оценки параметров модели (9)

Параметр	Ограничения на вариацию	Точечные оценки	Минимальное значение параметра	Максимальное значение параметра
$a_1 - a_2$	—	22,03	−14 151 439,74	23,4
α_1	[0; 5]	5,00	0,00	5,00
β_1	[0; 5]	1,02	1,02	5,00
γ_1	[−5; 5]	5,00	−5,00	5,00
α_2	[0; 5]	1,41	0,11	1,412
β_2	[0; 5]	0,00	0,00	1,03
γ_2	[−5; 5]	4,06	1,47	4,06
$a_3 - a_4$	—	−7173,5	−7173,5	−3459,3
α_3	[0; 2]	0,13	0,00	0,13
β_3	[0; 2]	0,32	0,32	0,33
γ_3	[−2; 2]	−1,18	−1,21	−1,14
α_4	[0; 2]	2,00	0,00	2,00
β_4	[0; 2]	0,00	0,00	2,00
γ_4	[−2; 2]	1,99	−2,00	2,00
$a_5 - a_6$	–	−6,59	−9,60	7,59
α_5	[0; 3]	0,00	0,00	2,99
β_5	[0; 3]	0,32	0,32	0,33
γ_5	[−2; 3]	1,70	−1,99	3,00
α_6	[0; 3]	3,00	0,00	3,00
β_6	[0; 3]	0,01	0,01	3,00
γ_6	[−2; 3]	1,70	−2,00	3,00

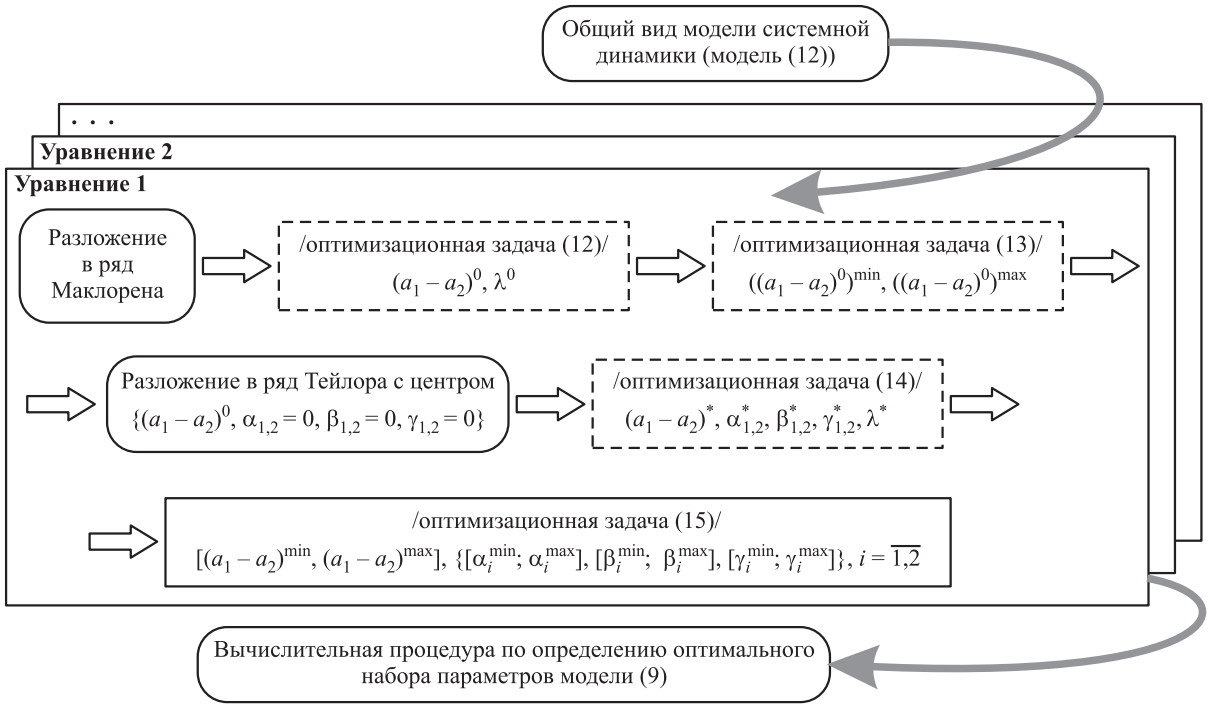
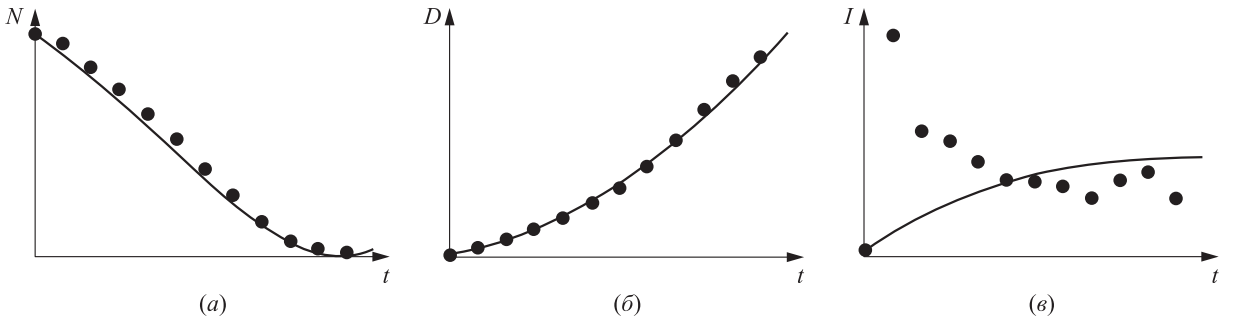


Рис. 2 Схема разработанного метода определения оптимального набора параметров модели системной динамики



(*а*) (*б*) (*в*)

Рис. 3 Графическая иллюстрация результатов численного интегрирования системы (21) методом Рунге–Кутты (кривые): (*а*) $\bar{A}_N = 0{,}13\%$; (*б*) $\bar{A}_D = 3{,}33\%$; (*в*) $\bar{A}_I = 5{,}86\%$. Значки — экспериментальные данные

шающий 10% [10] (т. е. $\bar{A}_N \le 10\%$, $\bar{A}_D \le 10\%$, $\bar{A}_I \le 10\%$).

Критерий оптимальности набора значений параметров модели (9) определялся на основе модуля вектора средних ошибок аппроксимации, компонентами которого являлись рассчитываемые средние ошибки аппроксимации по каждому уравнению модели (2):

$$\sqrt{\bar{A}_N^2 + \bar{A}_D^2 + \bar{A}_I^2} \to \min .$$

В качестве метода численного интегрирования системы (9) авторы выбрали метод Рунге–Кутты 4-го порядка ввиду его высокой точности и меньшей склонности к возникновению неустойчивости решения. По результатам вычислительной процедуры был определен оптимальный с позиций заданных условий и критерия набор параметров исследуемой модели численности населения Российской Федерации:

$$\left.\begin{aligned} &\frac{dN}{dt} = \\ &= 8{,}139 \cdot 10^{-22} \frac{N^{2,05} D^2}{I^2} - 64{,}1 \frac{N^{0,33} D^{0,3}}{I^{0,3}} ; \\ &\frac{dD}{dt} = 560 D^{0,35} - 9900\, I ; \\ &\frac{dI}{dt} = 0{,}131 I^{-0,4} - 0{,}0072 \frac{N^{0,092} D^{0,092}}{I^{0,092}} . \end{aligned}\right\} \quad (21)$$

Модель (21) характеризуется хорошими показателями точности (рис. 3) и адекватности, что позволило, в частности, эффективно решать задачи получения прогнозных оценок. Так, полученные прогнозные оценки согласно модели (21) в дальнейшем были подтверждены фактическими наблюдениями (табл. 4).

6 Заключение

Для успешной реализации метода системной динамики необходимо использовать математические модели, обладающие «хорошими» качественными характеристиками. Предложенный в работе подход к построению уравнений системной динамики позволяет определять оптимальный с точки зрения исследователя вид модели в условиях отсутствия четких представлений о функциональных связях между переменными. Полученные в ходе апробации результаты свидетельствуют о том, что описанный в настоящей работе метод определения точечных и интервальных оценок параметров моделей системной динамики, основанный на использовании идей подхода Л. В. Канторовича к обработке наблюдений, является эффективным инструментом для подготовки и организации численного эксперимента по определению оптимального набора параметров моделей заданной структуры. Существенным преимуществом разработанного подхода является возможность соблюдения ряда дополнительных условий, значимых на взгляд исследователя, учесть которые классическими статистическими методами невозможно.

Таблица 4 Сравнение прогнозных и фактических значений численности населения РФ, тыс. чел.

Источник	На 1 января 2010 г.	На 1 января 2011 г.	На 1 января 2012 г.
По данным Федеральной службы государственной статистики	142 833,0	142 865,0	143 056,0
Согласно модели системной динамики (21)	142 025,2	142 649,1	143 793,4
Погрешность	807,8 (0,57%)	215,9 (0,15%)	737,4 (0,52%)

Литература

1. *Форрестер Дж.* Мировая динамика / Пер. с англ. — М.: Наука, 1978. (*Forrester J.* World dynamics. — Wright-Allen Press, 1971. 144 p.)
2. *Wolctenholme E.* System enquiry — a system dynamic approach. — Chichester, England: John Wiley and Sons, 1990. 238 p.
3. *Белолипецкий В. М., Шокин Ю. И.* Математическое моделирование в задачах окружающей среды. — Новосибирск: Инфолио-пресс, 1997.
4. *Sterman J. D.* Business dynamics systems thinking and modeling for a complex world. — Irwin: McGrow-Hill, 2000. 1008 p.
5. *Спивак С. И.* Информативность кинетических измерений // Вестник Башкирского ун-та, 2009. Т. 14. № 3. С. 1056–1059.
6. *Канторович Л. В.* О некоторых новых подходах к вычислительным методам и обработке наблюдений // Сибирский математический журнал, 1962. Т. 3. № 5. С. 701–709.
7. *Моисеев Н. Н.* Математические задачи системного анализа. — М.: Наука, 1981. 488 с.
8. *Спивак С. И., Кантор О. Г.* Качество моделей математической обработки наблюдений социально-экономических систем // Системы управления и информационные технологии, 2012. № 2(48). С. 44–49.
9. *Гайнанов Д. А., Кантор О. Г., Казаков В. В.* Оценка уровня социально-экономического развития территориальных систем на основе метрического анализа // Вестник Томского гос. ун-та, 2009. № 322. С. 138–144.
10. Эконометрика / Под ред. Елисеевой И. И. — М.: Финансы и статистика, 2008. 576 с.

Поступила в редакцию 25.10.13

CONSTRUCTION OF SYSTEM DYNAMICS MODELS IN CONDITIONS OF LIMITED EXPERT INFORMATION

O. G. Kantor[1], S. I. Spivak[2]

[1]Institute of Social and Economic Researches of Ufa Scientific Centre RAS; 71 Av. Oktyabrya, Ufa 450054, Russian Federation
[2]Bashkir State University, 32 Validy Str., Ufa 450076, Russian Federation

Abstract: System dynamics is a methodology for studying of complex dynamic systems focused on conducting computer simulations. Construction of system dynamics models is largely dependent on the available experimental information and expert judgments. A simulation experiment with "bad" models can lead to significant or even total distortion of the system properties. This paper describes a method of constructing the system dynamics models, which is a set of mathematical models, based on the idea of the Kantorovich approach to the mathematical treatment of experimental data, and computational procedures. An important advantage is the possibility of including in the model significant conditions which are important to researcher and affect model adequacy. The developed method was tested on the example of the Russian population.

Keywords: system dynamics models; point and interval estimation of model parameters; the Kantorovich approach; the maximum permissible error of measurement

DOI: 10.14357/19922264140211

Acknowledgments

The work was supported by the Russian Foundation for Basic Research (project No. 13-01-00749).

References

1. Forrester, J. 1971. *World dynamics*. Wright-Allen Press. 144 p.
2. Wolctenholme, E. 1990. *System enquiry — a system dynamic approach*. Chichester, England: John Wiley and Sons. 238 p.
3. Belolipeckij, V. M., and Ju. I. Shokin. 1997. *Matematicheskoe modelirovanie v zadachakh okruzhayushchey sredy* [Mathematical modeling in environmental problems]. Novosibirsk: Infolio-Press. 240 p.
4. Sterman, J. D. 2000. *Business dynamics systems thinking and modeling for a complex world*. Irwin: McGrow-Hill. 1008 p.
5. Spivak, S. I. 2009. Informativnost' kineticheskikh izmereniy [Informative kinetic measurements]. *Vestnik Bashkirskogo Un-ta* [Bashkir University Bulletin] 3(14):1056–1059.

6. Kantorovich, L. V. 1962. O nekotorykh novykh podkhodakh k vychislitel'nym metodam i obrabotke nablyudeniy [Some new approaches to computational methods and treatment of observations]. *Sibirskiy Matematicheskiy Zhurnal* [Siberian Mathematical J.] 5(3):701–709.
7. Moiseev, N. N. 1981. *Matematicheskie zadachi sistemnogo analiza* [Mathematical problems of system analysis]. Moscow: Nauka. 488 p.
8. Spivak, S. I., and O. G. Kantor. 2012. Kachestvo modeley matematicheskoy obrabotki sotsial'no-ekonomicheskikh sistem [Quality of mathematical processing observations models of socio-economic systems]. *Sistemy Upravleniya i Informatsionnye Tekhnologii* [Management and Information Technology] 2(48):44–49.
9. Gajnanov, D. A., O. G. Kantor, and V. V. Kazakov. 2009. Otsenka urovnya sotsial'no-ekonomicheskogo razvitiya territorial'nykh sistem na osnove metricheskogo analiza [Estimation of level of social and economic development of territorial systems based on metric analysis]. *Vestnik Tomskogo Gos. Un-ta* [Bulletin of Tomsk State University] 322:138–144.
10. Eliseeva, I. I., ed. [Ed. I. I. Eliseeva.]. 2008. *Ekonometrika* [Econometrics]. Moscow: Finance and Statistics. 576 p.

Received October 25, 2013

Contributors

Kantor Olga G. (b. 1971) — Candidate of Science (PhD) in physics and mathematics, associate professor, senior scientist, Institute of Social and Economic Researches, Ufa Scientific Centre of the Russian Academy of Sciences, 71 Av. Oktyabrya, Ufa 450054, Russian Federation; o_kantor@mail.ru

Spivak Semen I. (b. 1945) — Doctor of Science in physics and mathematics, professor, Head of Department, Bashkir State University, 32 Validy Str., Ufa 450076, Russian Federation; semen.spivak@mail.ru

ИНФОРМАТИКА И ЕЁ ПРИМЕНЕНИЯ, 2014. Т. 8. Вып. 2. С. 122–129

ДЕКЛАРАТИВНЫЕ СТРУКТУРЫ ЗНАНИЙ В ПРОБЛЕМНО-ОРИЕНТИРОВАННЫХ СИСТЕМАХ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

А. Г. Мацкевич[1]

Аннотация: Описаны методы и средства представления знаний в виде декларативных структур на основе расширенных семантических сетей (РСС) для формирования баз знаний (БЗ) систем искусственного интеллекта, а также приведены примеры реализованных интеллектуальных систем обработки знаний для различных предметных областей. Для обработки декларативных структур знаний, представленных в виде РСС, разработан специализированный язык логического программирования ДЕКЛ. На языке ДЕКЛ были реализованы лингвистические процессоры (ЛП), осуществляющие перевод предложений естественного языка (ЕЯ) (русского и английского) в структуры БЗ, а также обратный перевод из внутренних (глубинных) представлений в поверхностные формы русского или английского языка.

Ключевые слова: интеллектуальные системы; представление знаний; обработка естественного языка; семантические сети; логическое программирование

DOI: 10.14357/19922264140212

1 Введение

Стремительно развивается тенденция к глобализации информации: она проявляется практически во всех сферах жизни общества. В то же время тексты на ЕЯ, в том числе в оцифрованном виде, являются основным способом хранения и передачи знаний. Во многих случаях человек не в силах прочитать и осмыслить даже малую часть того, что ему предлагается. В настоящее время решением этой глобальной проблемы занимаются, в частности, системы искусственного интеллекта. Основными компонентами таких систем являются БЗ, предназначенные для хранения информации в форме, удобной для ее обработки, а также средства, необходимые для преобразования текстов на ЕЯ в такую форму. Кроме этого, в состав таких систем входят компоненты, с помощью которых обеспечивается эффективный поиск и анализ информации и решаются другие насущные задачи.

В ИПИ РАН накоплен большой опыт разработки интеллектуальных систем обработки текстовых знаний. В коллективе, научным руководителем которого на протяжении многих лет являлся доктор технических наук, профессор Игорь Петрович Кузнецов, создана линия систем искусственного интеллекта, построенных на аппарате представления знаний в виде РСС. Еще в 1980-е гг. в своих монографиях [1, 2] и в докторской диссертации И. П. Кузнецов предложил использовать РСС для описания декларативных знаний.

Семантическая сеть в этом варианте состоит из множества вершин, представляющих объекты. Из вершин составляются элементарные фрагменты (ЭФ), каждый из которых представляет k-местное отношение. Во фрагмент вводятся две дополнительные вершины: одна соответствует отношению, а другая — всей совокупности упомянутых объектов с учетом их отношения. Эти вершины, как и любые другие вершины, могут стоять на местах объектов в других фрагментах, что обеспечивает высокие изобразительные возможности и гибкость: представление отношений между отношениями, между совокупностями связанных объектов и т. п. Из таких фрагментов и составляются сети, названные расширенными семантическими сетями. Как показали исследования [1–3], подобные сети оказываются удобными для представления различных языковых конструкций, в которых отглагольные существительные, представляющие определенное действие, сами могут связываться в рамках глагольных форм. Такие сети могут служить основой для решения многих логико-аналитических задач.

2 Языки представления и обработки знаний

Для обработки декларативных структур знаний, представленных в виде РСС, разработан специализированный язык логического программирования ДЕКЛ. Он обеспечивает представление декларатив-

[1]Институт проблем информатики Российской академии наук; Московский технический университет связи и информатики (МТУСИ), xmag@mail.ru

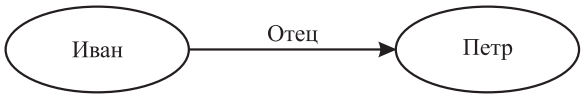


Рис. 1 Обычная семантическая сеть из вершин и дуги

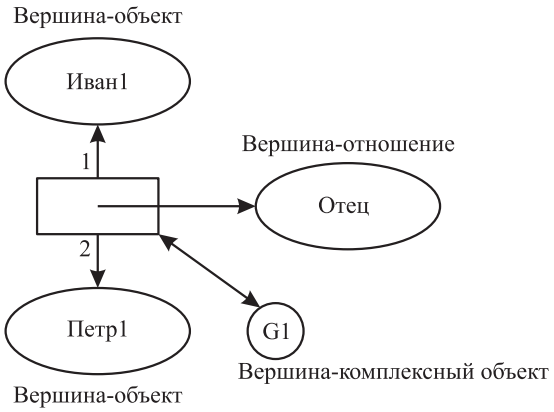


Рис. 2 Расширенная семантическая сеть из вершин-объектов, вершины-отношения и вершины-комплексного объекта

ных структур знаний в виде РСС и имеет средства их обработки [4, 5].

Обычные семантические сети состоят из вершин (они соответствуют объектам) и связывающих их дуг, которые соответствуют отношениям. Например, «*Иван отец Петра*» представляется в виде графа, представленного на рис. 1.

В РСС используются не слова ЕЯ, например русского ОТЕЦ, ИВАН, ПЁТР, а конкретные объекты базы знаний — ИВАН1, ИВАН2, . . . , ПЕТР1, соответствующие референтам, т. е. людям с именами Иван, Петр. База знаний может содержать много подобных объектов.

Такое различие необходимо для более точного представления информации. Далее, вместо дуги используется специальная вершина связи и ЭФ.

На рис. 2 представлено, что *ИВАН1 отец ПЕТР1*. Цифры 1, 2 возле стрелок указывают, что ИВАН1 — это первый объект отношений, а ПЕТР1 — 2-й. Если поменять их местами, то уже будет *ПЕТР1 отец ИВАН1*. При этом вершина-отношение (ОТЕЦ) выделена как самостоятельная. Она тоже может быть связана отношением. Далее, выделена *комплексная вершина* G1, которая соответствует объектам ИВАН1, ПЕТР1 с их отношением. Все это образует *часть семьи* — более сложный объект, которому сопоставлена вершина G1. На языке ДЕКЛ такой ЭФ записывается в виде:

ОТЕЦ(ИВАН1,ПЕТР1/G1).

Это предикатная форма записи ЭФ. Если вершина-отношение (ОТЕЦ) и комплексная вершина (G1)

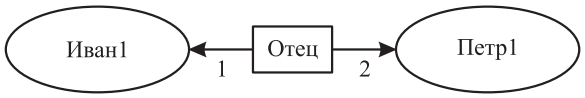


Рис. 3 Упрощенная РСС из вершин-объектов и вершины-связи

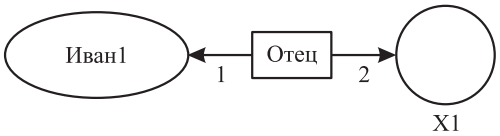


Рис. 4 Упрощенная РСС с вершиной-переменной

не связаны с какими-либо другими вершинами (не входят в другие ЭФ в качестве вершин-объектов), то будем использовать более простую запись ЭФ (рис. 3).

Запись сети рис. 3 в предикатном виде: ОТЕЦ(ИВАН1,ПЕТР1).

В РСС различают 2 типа вершин.

1. Вершины, соответствующие *определенным объектам*, отношениям, классам объектов (ИВАН1, ОТЕЦ, . . .).
2. Вершины, соответствующие *неопределенным объектам* (X1, X2, . . . , Xn). Они называются ***x-вершинами*** или ***вершинами-переменными***. Их обозначения выносятся за рамки вершин, указывая таким образом, что они не означены.

Сеть рис. 4 представляет: *ИВАН1 отец неизвестно кому* — ОТЕЦ(ИВАН1, X1)

С помощью сети рис. 5 представлено, что *ИВАН1 и ПЕТР1 как-то связаны между собой*, но неизвестно, каким отношением — X2(ИВАН1,ПЕТР1).

В дальнейшем вводится понятие «отношение в широком смысле» и допускается множество объектов отношений (более двух). В этом случае ЭФ в общем виде будет выглядеть, как на рис. 6. Здесь представлено N-*арное отношение* R1 *между объектами* A1, A2, . . . , AN, которые образуют комплексный объект G2. Такой ЭФ записывается в предикатном виде как $R1(A1, A2, \ldots, AN/G2)$.

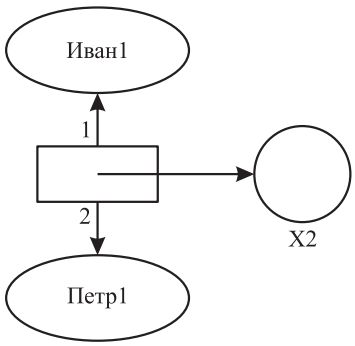


Рис. 5 Расширенная семантическая сеть с вершиной-переменной

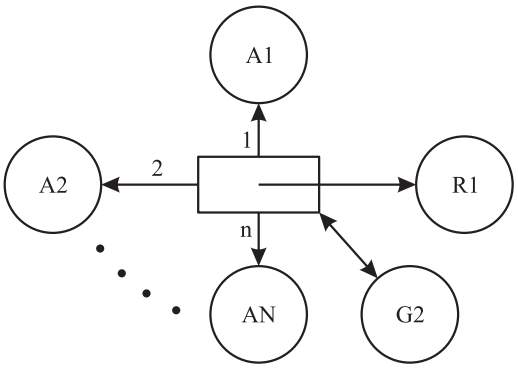


Рис. 6 Расширенная семантическая сеть с N-арным отношением R1 между объектами A1, . . . , AN, образующими комплексный объект G2

Посредством РСС вводятся представления отношений, описываются родовидовые деревья и другие конструкции для комплексных объектов. В РСС используется принцип наследования свойств: свойства каждой вершины более высокого уровня справедливы для всех ее вершин более низкого уровня. Это позволяет значительно сократить объем знаний, используя свойства (отношения) классов объектов и автоматически перенося их на конкретные объекты.

Такое разнообразие «изобразительных средств» делает РСС очень гибким и практически универсальным средством описания декларативных знаний.

Для обработки структур знаний, представленных в виде РСС, разработана инструментальная среда ДЕКЛ, включающая в себя язык ДЕКЛ, а также собственную базу данных — для хранения РСС и текстов. Последние версии этой среды написаны на языке Object Pascal в среде программирования Delphi. Язык ДЕКЛ основан на применении правил ЕСЛИ. . .ТО. . ., которые содержат правую и левую часть. Для левой части ищутся сопоставимые структуры в БЗ. Если в БЗ оказалась сеть (РСС), сопоставимая с левой частью правила, тогда правило делается применимым и выполняются действия, записанные в правой части этого правила. В частности, могут быть вызваны другие правила.

В языке есть специальные конструкции, обеспечивающие интерфейсы пользователя, например работу с файлами, построение окон, выдачу на экран, обработчики событий (например, нажатия клавиш, действия с мышкой) и др.

Далее приведен простой пример представления декларативных знаний в виде БЗ и программы, написанной на языке логического программирования ДЕКЛ, выдающей студентов группы ГР_1.

```
{== База знаний ==}
СТУДЕНТ(ГР_1,STUD_1) ФАМ(IVANOV,STUD_1)
    ИМЯ(IVAN,STUD_1)
СТУДЕНТ(ГР_1,STUD_2) ФАМ(PETROV,STUD_2)
    ИМЯ(PETER,STUD_2)
СТУДЕНТ(ГР_1,STUD_3) ФАМ(SIDOROV,STUD_3)
    ИМЯ(IVAN,STUD_3)
СТУДЕНТ(ГР_1,STUD_4) ФАМ(SMIRNOV,STUD_4)
    ИМЯ(PETER,STUD_4)
СТУДЕНТ(ГР_1,STUD_5) ФАМ(IVANOV,STUD_5)
    ИМЯ(ANDREY,STUD_5)
СТУДЕНТ(ГР_2,STUD_6) ФАМ(PETROV,STUD_6)
    ИМЯ(OLEG,STUD_6)
{== Программа ==}
BEG(/91+) {= начало программы – код 91+ =}
{= с продукции START начинается применение программы =}
START:IF THEN
{B:PAR(1,9) = Включение трассировки.
    B:PAR(1,0) – выключение =}
T!:STUD_OUT(ГР_1)
B:IN() {= ДЕКЛ встает и ожидает нажатия клавиши =}
B:HALT(); {= Выход из ДЕКЛ =}
{== Поиск по группе X1 ее студентов X2
    с выдачей ФИО ==}
STUD_OUT(X1):IF СТУДЕНТ(X1,X2) ФАМ(X10,X2)
    ИМЯ(X20,X2) THEN
B:BK() {= Переход на новую строку =}
B:A(« Студент гр. »,X1) {= Выдача на экран =}
B:A(«: »,X10,« »,X20);
END(/92+) {= Конец программы – код 92+ =}
@BL(91-,92-) {= указывает на зону, к которой не должны
    применяться продукции =}
```

Если сохранить БЗ и программу в файле TEST_1.Z и написать в командной строке ДЕКЛ_WIN.EXE TEST_1.Z, то на экран будет выдано следующее:

Студент гр. ГР_1: IVANOV IVAN
Студент гр. ГР_1: PETROV PETER
Студент гр. ГР_1: SIDOROV IVAN
Студент гр. ГР_1: SMIRNOV PETER
Студент гр. ГР_1: IVANOV ANDREY

3 Лингвистические процессоры на основе расширенных семантических сетей

На языке ДЕКЛ были реализованы средства, строящие РСС по текстам ЕЯ [6–10], называемые лингвистическими процессорами. С использованием ЛП созданы такие объектно-ориентированные системы, как «Криминал», «Аналитик», «LINGVO-MASTER», «Антитеррор».

На рис. 7 представлена схема, которую можно считать общей для этих систем.

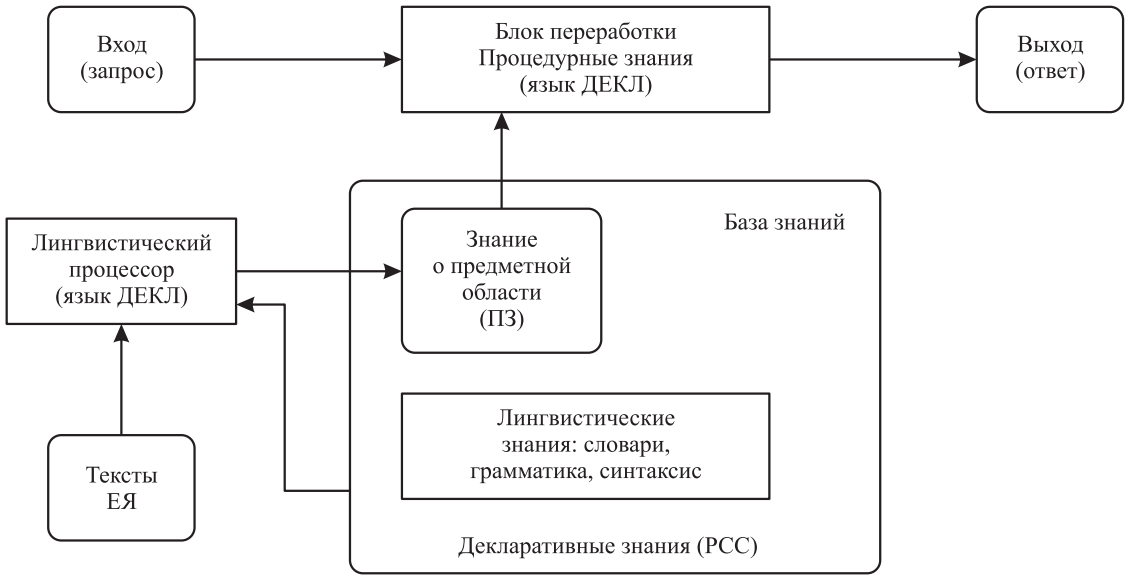


Рис. 7 Схема функционирования объектно-ориентированных систем, использующих ЛП

В данной схеме с помощью ЛП накапливаются *предметные знания* (ПЗ), которые определяют ответы, вырабатываемые системой на тот или иной запрос. Сам ЛП настраивается на работу с входными текстами с помощью *лингвистических знаний* (ЛЗ). Так как преобразование текстов и их обработка имеют много общих задач, то перспективным является использование единого инструментария: представление предметных и лингвистических знаний с помощью РСС и их обработка программами на языке ДЕКЛ. Таким образом формируются декларативные знания в виде базы предметных и базы лингвистических знаний.

Более того, ЛП может быть использован для поддержания режима ответа на запросы, выраженные на ЕЯ. Запросы представляются в виде РСС, и поиск ответа идет на уровне обработки структур знаний. На это и ориентирован язык ДЕКЛ.

4 Реализованные логико-аналитические системы

Уникальной по своим возможностям системой, использующей РСС и написанной на языке ДЕКЛ, является логико-аналитическая система «Криминал» [6, 8]. Эта система нашла свое применение в аналитических отделах ГУВД г. Москвы и МВД.

Система «Криминал» базируется на документах, поступающих из различных источников. Это сводки, объяснительные, служебные записки служб органов внутренних дел; записные книжки фигурантов; отчеты, документы общего назначения; газетные публикации; словесные портреты фигурантов и т. д. Эти документы обрабатываются ЛП в автоматическом режиме, по каждому из них строится РСС. Далее следует постлингвистическая обработка. Она заключается в логическом анализе и выделении наиболее значимых характеристик документа с точки зрения сотрудников правоохранительных органов: орудий преступления, способа его совершения, способа проникновения и др. Осуществляется дополнение документа атрибутами в соответствии с принятыми классификаторами. Вся выделенная информация образует РСС, которая называется содержательным портретом документа, где представлены значимые элементы текста и их связи. Такие портреты (РСС) образуют БЗ, которая является основой для логико-аналитической обработки.

На основе содержательных портретов строятся предметные каталоги. Это списки фигурантов, адресов и других объектов (которые были выявлены из документов), упорядоченные по алфавиту. Такие списки делают поиск направленным. Пользователь может выбрать из них любой объект для последующего анализа.

На основе декларативных знаний, полученных в ходе анализа документов, в системе «Криминал» реализованы следующие задачи логико-аналитической обработки:

- поиск похожих происшествий и фигурантов по информации, извлеченной автоматически из имеющихся источников;
- контекстный поиск документов;
- поиск фигурантов по словесному портрету;

– поиск информации по запросам на естественном (русском) языке;
– объяснение результатов поиска;
– анализ и отображение связей между фигурантами;
– оценка степени причастности фигуранта к происшествию;
– упорядочение фигурантов по степени их криминальной и преступной активности;
– выявление организованных преступных формирований (на основе связей фигурантов);
– статистическая обработка информации, выдача усредненных и оценочных данных, характеризующих динамику изменения криминогенных процессов во времени.

Приведем типовой пример документа и его содержательного портрета (РСС).

1.05.2013г. в 7.10 Иванова Ивана Ивановича 1953г.р прож. ул. Ивановская 25-1-273, работает АОЗТ «ХДУ», зам. директора, о том, что 1-05-98г. неизвестные от д.22 кор.3 по ул.Сенная, похитили а/м ГАЗ 31029, черная, 1995 г/в, дв.402-0019476, кузов 0285927 . . .

Его содержательный портрет имеет следующий вид:

```
ДОК_(221,′TEXT_98.TXT′,′S_CRI.NL′)
ДАТА_(#1.5.2013,2013,МАЙ,~1,7.1/4+) 4–(221,ДАТА_)
FIO(ИВАНОВ,ИВАН,ИВАНОВИЧ,1953/5+)
    5–(221,FIO)
АДР_(УЛ.,ИВАНОВСКАЯ,25,1,273/6+) 6–(221,АДР_)
ПРОЖ.(5–,6–/7+)
ОРГ_(АОЗТ,ХДУ/8+) 8–(221,ОРГ_)
РАБ_(5–,8–,ЗАМ.,ДИРЕКТОР/9+) 10–(221,РАБ_)
FIO(" "," "," "," ",НЕСКОЛЬКО/10+) 10–(221,FIO)
НЕИЗВЕСТНЫЙ(10–)
АВТО_(АВТОМАШИНА,ГАЗ,31029,ЧЕРНЫЙ,
    1995,Г/В,ДВ.,402,19476,  КУЗОВ,285927,УЧЕТ/11+)
11–(221,АВТО_)
УГНАТЬ(10–,11–/12+)
ДАТА_(#1.5.2013,2013,МАЙ,~1/14+) 4–(221,ДАТА_)
КОГДА(12–,14–)
АДР_(УЛ.,СЕННАЯ,ДОМ,22,КОРП.,3/15+)
    15–(221,АДР_) ГДЕ(12–,15–)
ПРЕДЛ_(221,4–,5–,6–,8–,9–,О,ТОМ,12–,14–,15–)
```

Первый фрагмент ДОК_(221,′TEXT_98.TXT′, ′S_CRI.NL′) указывает, что содержательный портрет построен на основе документа 221 из файла ′TEXT_98.TXT′. При этом были использованы лингвистические знания ′S_CRI.NL′. Второй фрагмент представляет дату. Здесь используются внутрисистемные коды, которые соответствуют определенным вершинам РСС без мнемонических обозначений. Причем 4+ значит, что создается новая вершина без имени, а 4– значит, что вершина без имени используется. Эти фрагменты необходимы для быстрого поиска нужных фрагментов, когда в оперативной памяти (БЗ) находится множество содержательных портретов.

Далее следует фрагмент FIO(ИВАНОВ,ИВАН, ИВАНОВИЧ,1953/5+), представляющий фигуранта, затем следует фрагмент АДР_(УЛ., ИВАНОВСКАЯ, 25,1,273/6+), соответствующий выделенному адресу, а за ним фрагмент ПРОЖ. (5–,6–/7+), в котором с использованием внутренних кодов записана информация о том, что 5–, т. е. ИВАНОВ ИВАН ИВАНОВИЧ, проживает по адресу 6, т. е. УЛ.,ИВАНОВСКАЯ,25,1,273. Этой выделенной информации соответствует вершина 7+.

Фрагменты ПРЕДЛ_ представляют предложения с аргументами: кодами фрагментов, которые представляют объекты и действия, и словами, которые никуда не вошли. За счет фрагментов ПРЕДЛ_ текст может быть восстановлен по содержательному портрету документа.

Содержательные портреты запоминаются в БЗ системы. На основе этой БЗ и построенных предметных каталогов реализованы различные виды семантического поиска: поиск по признакам и связям, поиск связанных объектов на различных уровнях, поиск похожих фигурантов и происшествий, поиск по приметам.

Таким образом, логико-аналитическая система «Криминал» стала важным этапом в развитии использования механизмов РСС для представления декларативных структур знаний.

Система LINGVO-MASTER обеспечивает автоматическую формализацию различного рода справок и сообщений (автобиографических данных, заявок на работу, резюме), представляющих собой тексты на ЕЯ в достаточно произвольной форме [10]. Результатами формализации являются выделенные значимые компоненты (информационные объекты) и их отображение в поля некоторой базы данных и соответствующей формы (анкеты) с фиксированными полями. Тогда становится возможным использование типовых средств для решения пользовательских задач, например поиска работы по запросу пользователя или поиска специалиста, удовлетворяющего требованиям работодателя. Общая схема такой системы отличается от схемы рис. 7, хотя имеет много общих частей. Методики извлечения семантической информации, разработанные в системе «Криминал», используются и здесь. Лингвистический процессор системы LINGVO-MASTER в своей основе имеет тот же ЛП. Так же строится содержательный портрет документа. Конечно, в БЗ (см. рис. 7) изменяются ЛЗ, а знания о предметной области (ПЗ), представлен-

ные в виде РСС, являются исходными данными для обратного ЛП. Обратный ЛП служит для преобразования содержательных портретов (РСС) в компоненты ЕЯ и для их отображения в поля формы или сайта. Этот процессор имеет свои ЛЗ, с помощью которых задается последовательность выдачи рубрик (полей) и то, какими объектами они должны заполняться. Для выделения таких объектов служат их имена, а также связи, заданные в РСС. Для каждого выделенного объекта строится его описание — из входящих в него нормализованных слов. Далее ищется сегмент предложения, соответствующий объекту. Этот сегмент и выдается в качестве результата.

Таким образом, в системе LINGVO-MASTER, во-первых, обработка идет на уровне структур знаний (РСС) с использованием созданного для этого инструментария (языка ДЕКЛ). Отсюда возможность вовлечения в процесс анализа семантических категорий и различного рода связей. И, во-вторых, основные процессоры в данной системе сделаны как оболочки, которые легко подстраивать под предметную область и особенности текстов за счет лингвистических и экспертных знаний. Это очень важно, когда требуется обработка реальных текстов. На стадии проектирования удается учесть лишь малую часть их особенностей. Дальнейшее совершенствование (и качество системы) определяется удобством и возможностями средств подстройки.

Система отлаживалась на материалах кадрового портала HeadHunter и обеспечивала следующие результаты: коэффициент шумов в компонентах (лишних слов в объектах) — не более 1%–2% и потерь (отсутствие нужных слов) — не более 3%.

5 Заключение

Тот факт, что основные процессоры в реализованных системах сделаны как оболочки, которые легко подстраивать под предметную область и особенности текстов за счет лингвистических и экспертных знаний, открывает большие возможности использования РСС и языка ДЕКЛ для создания проблемно-ориентированных систем искусственного интеллекта. Благодаря этому достоинству была создана англоязычная версия системы обработки автобиографических данных, а также система «Антитеррор», ориентированная на выявление информации о террористической деятельности из материалов СМИ [11, 12].

Литература

1. *Кузнецов И. П.* Механизмы обработки семантической информации. — М.: Наука, 1978. 175 с.
2. *Кузнецов И. П.* Семантические представления. — М.: Наука, 1986. 296 с.
3. *Кузнецов И. П.* Расширяющиеся системы активного диалога. — М.: Наука, 1982. 309 с.
4. *Кузнецов И. П. Пузанов В. В., Шарнин М. М.* Система обработки декларативных структур знаний ДЕКЛАР-2. — М.: ИПИАН, 1989. 106 с.
5. *Кузнецов И. П. Шарнин М. М.* Интеллектуальный редактор знаний на основе расширенных семантических сетей // Системы и средства информатики, 1993. Вып. 5. С. 14–21.
6. *Кузнецов И. П.* Методы обработки сводок с выявлением особенностей фигурантов и происшествий // Диалог-98: Труды Междунар. семинара по компьютерной лингвистике и ее приложениям. — Казань: Хэтер, 1998. Т. 2. С. 691–700.
7. *Кузнецов И. П., Козеренко Е. Б., Шарнин М. М.* Семантико-ориентированная система фактографического поиска со входом на русском и английском языках // Диалог-98: Труды Междунар. семинара по компьютерной лингвистике и ее приложениям. — Казань: Хэтер, 1998. Т. 2. С. 821–830.
8. *Кузнецов И. П., Мацкевич А. Г.* Лингвистический процессор для автоматического выявления из текстов значимой информации с ее компоновкой в рамках указанных шаблонов // Диалог 2001: Труды Междунар. семинара по компьютерной лингвистике и ее приложениям. — М.: Наука, 2001. Т. 2. С. 134–137.
9. *Kuznetsov I., Kozerenko E.* The system for extracting semantic information from natural language texts // MLMTA-03: Conference (International) on Machine Learning Proceedings. — Las Vegas: CSREA, 2003. P. 75–80.
10. *Кузнецов И. П., Мацкевич А. Г.* Семантико-ориентированный лингвистический процессор для автоматической формализации автобиографических данных // Компьютерная лингвистика и интеллектуальные технологии: Труды Междунар. конференции «Диалог'2006». — М.: РГГУ, 2006. С. 317–322.
11. *Кузнецов И. П., Мацкевич А. Г.* Англоязычная версия системы автоматического выявления значимой информации из текстов естественного языка // Компьютерная лингвистика и интеллектуальные технологии: Труды Междунар. конф. «Диалог'2005». — М.: Наука, 2005. С. 303–311.
12. *Кузнецов И. П., Козеренко Е. Б., Мацкевич А. Г.* Принципы организации объектно-ориентированных систем обработки неформализованной информации // Искусственный интеллект: Журнал НАН Украины, 2010. Вып. 3. С. 227–237.

Поступила в редакцию 7.05.14

DECLARATIVE KNOWLEDGE STRUCTURES IN PROBLEM-ORIENTED SYSTEMS OF ARTIFICIAL INTELLIGENCE

A. G. Matskevich[1,2]

[1]Institute of Informatics Problems, Russian Academy of Sciences, 44-2 Vavilov Str., Moscow 119333, Russian Federation

[2]Moscow Technical University of Communications and Informatics (MTUCI), 8a Aviamotornaya Str., Moscow 111024, Russian Federation

Abstract: The article describes the techniques and tools of knowledge representation in the form of declarative structures based on the Extended Semantic Networks (ESN) formalism to produce intelligent systems knowledge bases and it provides the examples of intellectual knowledge processing systems for different domains. For processing the declarative knowledge structures presented in the form of ESN, a specialized logical programming language DECL has been developed. In DECL, the linguistic processors have been implemented executing translation of natural language (Russian and English) sentences into structures of a knowledge base, as well as reverse translation of inner representations into the surface forms of Russian or English.

Keywords: intelligent systems; knowledge representation; natural language processing; semantic networks; logical programming

DOI: 10.14357/19922264140212

References

1. Kuznetsov, I. P. 1978. *Mekhanizmy obrabotki semanticheskoy informatsii* [Semantic information processing mechanisms]. Moscow: Nauka. 175 p.
2. Kuznetsov, I. P. 1986. *Semanticheskie predstavleniya* [Semantic representations]. Moscow: Nauka. 296 p.
3. Kuznetsov, I. P. 1982. *Rasshiryayushchiesya sistemy aktivnogo dialoga* [Extending systems of active dialogue]. Moscow: Nauka. 309 p.
4. Kuznetsov, I. P., V. V. Puzanov, and M. M. Sharnin. 1989. *Sistema obrabotki deklarativnykh struktur znaniy DEKLAR-2* [The system of declarative knowledge structures processing]. Moscow: IPIAN. 106 p.
5. Kuznetsov, I. P., and M. M. Sharnin. 1993. Intellektual'nyy redaktor znaniy na osnove rasshirennykh semanticheskikh setey [Intelligent knowledge editor based on the extended semantic networks]. *Sistemy i Sredstva Informatiki — Systems and Means of Informatics* 5:14–21.
6. Kuznetsov, I. P. 1998. Metody obrabotki svodok s vyyavleniem osobennostey figurantov i proisshestviy [The methods of reports processing with the identification of criminal acts and participants characteristics]. *Dialog-98: Trudy Mezhdunar. Seminara po Komp'yuternoy Lingvistike i ee Prilozheniyam* [Dialogue-98: Seminar (International) in Computational Linguistics and Its Applications Proceedings]. Kazan': Kheter. 2:691–700.
7. Kuznetsov, I. P., E. B. Kozerenko, and M. M. Sharnin. 1998. Semantiko-orientirovannaya sistema faktograficheskogo poiska so vkhodom na russkom i angliyskom yazykakh [The semantic-oriented system of factual search with the input in Russian and in English]. *Dialog-98: Trudy Mezhdunar. Seminara po Komp'yuternoy Lingvistike i ee Prilozheniyam* [Dialogue-98: Seminar (International) in Computational Linguistics and Its Applications Proceedings]. Kazan': Kheter. 2:821–830.
8. Kuznetsov, I. P., and A. G. Matskevich. 2001. Lingvisticheskiy protsessor dlya avtomaticheskogo vyyavleniya iz tekstov znachimoy informatsii s ee komponovkoy v ramkakh ukazannykh shablonov [The linguistic processor for automatic identification of the meaningful information from texts and its arrangement within the framework of language templates]. *Dialog 2001: Trudy Mezhdunar. Seminara po Komp'yuternoy Lingvistike i ee Prilozheniyam* [Dialogue-2001: Seminar (International) in Computational Linguistics and Its Applications Proceedings]. Moscow: Nauka. 2: 134–137.
9. Kuznetsov, I., and E. Kozerenko. 2003. The system for extracting semantic information from natural language texts. *MLMTA-03: Conference (International) on Machine Learning Proceedings*. Las Vegas: CSREA. 75–80.
10. Kuznetsov, I. P., and A. G. Matskevich. 2006. Semantiko-orientirovannyy lingvisticheskiy protsessor dlya avtomaticheskoy formalizatsii avtobiograficheskikh dannykh [The semantic-oriented linguistic processor for automatic formalization of autobiogrphy data]. *Komp'yuternaya Lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunar. Konf. "Dialog'2006"* [Computational Linguistics and Intelligent Technologies: Conference (International) "Dialogue'2006" Proceedings]. Moscow: RGGU. 317–322.
11. Kuznetsov, I. P., and A. G. Matskevich. 2005. Anglo-yazychnaya versiya sistemy avtomaticheskogo vyyavleniya

znachimoy informatsii iz tekstov estestvennogo yazyka [The English version of the system for automatic extraction of the meaningful information from natural language texts]. *Komp'yuternaya Lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunar. Konf. "Dialog'2005"* [Computational Linguistics and Intelligent Technologies: Conference (International) "Dialogue'2005" Proceedings]. Moscow: Nauka. 303–311.

12. Kuznetsov, I. P., E. B. Kozerenko, and A. G. Matskevich. 2010. Printsipy organizatsii ob"ektno-orientirovannykh sistem obrabotki neformalizovannoy informatsii [The principles of organization of the object-oriented systems of unstructured information processing]. *Iskusstvennyy Intellekt: Zhurnal NAN Ukrainy* [Artificial Intelligence: The Ukraine National Academy of Sciences J.] 3:227–237.

Received May 7, 2014

Contributor

Matskevich Andrey G. (b. 1953) — senior scientist, Institute of Informatics Problems, Russian Academy of Sciences, 44-2 Vavilov Str., Moscow 119333, Russian Federation; associate professor, Moscow Technical University of Communications and Informatics (MTUCI), 8a Aviamotornaya Str., Moscow 111024, Russian Federation; xmag@mail.ru

ИНФОРМАТИКА И ЕЁ ПРИМЕНЕНИЯ, 2014. Т. 8. Вып. 2. С. 130–144

УНИВЕРСАЛЬНАЯ ТЕХНОЛОГИЯ ОЦЕНКИ БЛИЗОСТИ ИНФОРМАЦИОННЫХ ОБЪЕКТОВ

Л. А. Кузнецов[1]

Аннотация: Изложена технология определения степени подобия информационных объектов, которые представлены текстами или графическими изображениями. Объекты формализуются вероятностными моделями. Структура модели задается алгеброй на минимальном наборе изобразительных компонентов объекта. Количественными характеристиками структуры объектов являются распределения вероятностей на заданной алгебре. Количество информации в объектах оценивается энтропией. На энтропиях задается мера информационного подобия сопоставляемых объектов. Показана методика формирования оценки для текстовых и графических объектов. Приведены примеры реализации алгоритмов оценки и показана более высокая эффективность разработанных методов по сравнению с методами, описанными в литературе. Технология формирования образов информационных объектов и сравнения их семантического содержания является универсальной. Показаны возможности адаптации разработанной технологии к содержательным характеристикам исследуемых объектов.

Ключевые слова: информационный объект; текст; изображение; вероятностная модель; семантическое подобие; энтропия; мера подобия

DOI: 10.14375/19922264140213

1 Введение

Оценка близости информационных объектов является наиболее распространенным компонентом информационных технологий. На компоненте оценки информационного подобия объектов базируются технологии поиска информации, сравнения проектов, оценки оригинальности научных результатов и т. п. На этом компоненте в дальнейшем будут разработаны автоматические процедуры дифференцированной оценки знаний, включающие оценку уровня близости текстов на естественных и формальных языках, структурированных и неструктурированных графических объектов, т. е. всех компонентов представления знаний, к определению уровня усвоения которых сводится проверка качества обучения. Совокупность процедур оценки близости информационных объектов, представленных на разных «языках», позволит создавать автоматизированные системы проверки качества на всех уровнях обучения. Такие системы позволят исключить тестовый самообман, обеспечить объективность полноценной оценки качества подготовки, устранить возможность коррупции и разнообразных подтасовок в сфере образования.

Ниже излагается универсальная технология формальной оценки уровня подобия информационных объектов, имеющих однотипное представление на естественном языке (тексты), на формальном языке (формулы математические, химические и др.), графическое представление (схемы, чертежи, картины).

Информационные объекты обычно представляют собой композицию количественных данных и неформализованных сведений, которые могут быть представлены текстовой и графической информацией. В зависимости от содержания и предназначения объекта доля текстовой и графической информационных составляющих может иметь определяющее значение. Поэтому разработка моделей формального описания и оценки подобия объектов, представленных текстовой и графической информацией, является важной задачей, решению которой посвящена работа.

В информационно-поисковых системах при классификации текстов, при проверке текстов на плагиат [1] применяются статистические подходы на основе векторно-пространственной модели текста, предложенный Солтоном с соавторами в 1975 г. [2]. В ней текст представляется вектором частот входящих в него слов, а оценка близости текстов равна косинусу угла между векторами текстов.

Более эффективные инструменты оценки близости информационных объектов могут быть синтезированы на основе их представления в виде однотипных вероятностных моделей, допускающих

[1]Российская академия народного хозяйства и государственной службы при Президенте Российской Федерации (Липецкий филиал), Kuznetsov.Leonid48@gmail.com

структуризацию и количественное сопоставление содержащейся в них информации. Исследования показали, что вероятностная модель позволяет формально и с произвольной глубиной детализации описывать объекты, представленные текстами на естественных и формальных языках [3] или структурированными [4] и неструктурированными графическими изображениями [5]. Вероятностная модель позволяет формализовать объекты в виде системы классов различной информационной значимости. Система классов может адаптироваться к содержательной специфике сопоставляемых объектов и алфавитам их представления. Оценка уровня близости объектов производится на системе классов с учетом их информационно-семантической значимости. Процедуры формирования системы классов на заданном алфавите и уровня их значимости могут быть реализованы в виде самонастраивающихся по принципу обратной связи систем.

2 Вероятностная модель

Сопоставляемые информационные объекты формализуются в виде вероятностных моделей. Абстрактная вероятностная модель эксперимента с конечным числом исходов, или просто вероятностная модель, вводится в теории вероятностей для формального представления результатов произвольного эксперимента. Модель представляет собой совокупность трех составляющих [6]:

$$M = \{\Omega, \aleph, P(A_i)\}, \qquad (1)$$

где $\Omega = \{\omega_1, \omega_2, \ldots, \omega_n\}$ — множество элементарных событий (исходов или реализаций исследуемой случайной величины); $\aleph = \{A_1, A_2, \ldots, A_m\}$ — множество (система) случайных событий (алгебра); $P(A_i)$, $i = 1, 2, \ldots, m$, — вероятности случайных событий.

В контексте оценки близости информационных объектов под случайными величинами могут пониматься величины или элементы, совокупностью которых представляется объект. Случайной величиной или элементом информационного объекта, представленного текстом на естественном языке, является слово. Для структурированного графического объекта, представленного, например, электрической схемой, случайной величиной является стандартизованное обозначение элементов электрических схем. Для неструктурированного графического объекта, представленного, например, картиной, случайными величинами являются цвет и координаты пикселов, отражающих ее на мониторе.

В соответствии с содержанием информационных объектов случайные величины, используемые для их представления, принимают свои значения — элементарные события — из конечных множеств, соответствующих содержанию. Под элементарными событиями понимаются неделимые при сравнении элементы ω, из которых формируются объекты. Случайная величина «слово» (для русского текста) может принимать значения всех слов, имеющихся в словарях русского языка. Случайная величина «элемент электрических схем» может принимать значения элементов из соответствующего ГОСТа [7]. Случайная величина «пиксел» принимает значения из палитры цветов, представимых на мониторе, и возможных координат его положения.

Любые объекты представляют собой множества реализаций, или, по терминологии теории вероятностей, элементарных исходов соответствующих случайных величин. Текст представляется множеством различных слов, которые являются реализациями случайной величины «слово», электрическая схема представляется набором значений величины «элемент электрических схем», картина — множеством реализаций величины «пиксел». Современные возможности информационных технологий позволяют достаточно просто проверить полную идентичность информационных объектов, представленных в электронном виде. Но в большинстве задач обработки информации требуется не установка идентичности, а оценка степени близости информационных объектов, отраженная некоторой количественной мерой.

Такая задача может быть решена с использованием вероятностной модели информационных объектов, позволяющей структурировать множество реализаций случайной величины $\Omega = \{\omega_1, \omega_2, \ldots, \omega_n\}$ введением системы классов $\aleph = \{A_1, A_2, \ldots, A_m\}$. Классы A_j, $j = 1, 2, \ldots, m$, позволяют отразить разнообразные содержательные особенности и информационную значимость отдельных совокупностей реализаций случайной величины. Классами может быть формально представлена семантика информационных объектов, являющаяся определяющей при оценке их подобия.

Случайные события A_i конструируются на множестве реализаций случайной величины $\Omega = \{\omega_1, \omega_2, \ldots, \omega_n\}$ с помощью операций $\cup$ — сложения случайных событий (объединения множеств), $\cap$ — произведения случайных событий (пересечения множеств), $\bar{}$ — отрицания событий (дополнения множеств). Случайные события A_i представляют подмножества множества Ω. Случайные события, получаемые из A_i с помощью перечисленных операций, также принадлежат алгебре $\aleph$. В ал-

гебру $\aleph$ входят невозможное событие (пустое множество $\emptyset$) и достоверное событие (множество Ω).

Целью практического применения вероятностной модели (1) является структуризация информации, содержащейся в множестве Ω, и выявление некоторых неслучайных ее характеристик, скрытых во множестве случайных реализаций $\{\omega_1, \omega_2, \ldots, \omega_n\}$. Именно эти содержательные характеристики исследуемой случайной величины и отражаются в модели (1) случайными событиями A_i, составляющими алгебру $\aleph$.

Введение алгебры случайных событий $\aleph = \{A_1, A_2, \ldots, A_m\}$ задает систему содержательных (качественных) характеристик, позволяющих разделить множество элементарных событий $\Omega = \{\omega_1, \omega_2, \ldots, \omega_n\}$ на классы $A_1, A_2, \ldots, A_m$. Принципиальным в данном контексте свойством алгебры является возможность конструирования новых случайных событий из множества уже имеющихся, потому что новые случайные события, полученные объединением, пересечением и отрицанием принадлежащих алгебре событий, также принадлежат алгебре, или точнее:

$$\text{из } A_i,\ A_j \in \aleph \text{ следует } A_i \cap A_j \in \aleph,$$
$$A_i \cup a_j \in \aleph \text{ и } \overline{A}_i \in \aleph,\ i, j \in [1, m]. \quad (2)$$

На основании (2) алгебра на каждом новом этапе формирования (эволюции) модели (1) может неограниченно расширяться и изменяться применением операций сложения, перемножения и отрицания к системе случайных событий предшествующего этапа.

Количественной мерой, определяющей соотношение случайных событий, образующих алгебру, является третий элемент модели (1) — вероятности случайных событий $P(A_i)$, которые находятся по вероятностям элементарных событий:

$$P(A_i) = \sum_{\omega_j \in A_i} p(\omega_j), \quad (3)$$

где $p(\omega_j)$ — вероятность элементарного события ω_j.

Можно видеть, что вероятностная модель (1) позволяет отразить всю возможную информацию о случайной величине, которая может быть формально извлечена из множества ее реализаций Ω. Система случайных событий $\aleph$ обеспечивает возможность выявления неоднородности элементов множества Ω, а вероятности случайных событий $A_i \in \aleph$, $i = 1, 2, \ldots, m$, позволяют количественно оценить степень неоднородности. Возможность модификации алгебры (2) позволяет осуществлять адаптацию структуры вероятностной модели (1), направление и результаты которой могут определяться некоторыми функционалами, заданными на распределении вероятностей (3).

3 Вероятностная модель текстовых информационных объектов

Оценка близости информационных объектов может осуществляться сопоставлением их вероятностных моделей. Современные информационные технологии позволяют представить в электронном виде информацию, полностью характеризующую объекты самого разного содержания. При этом информация на электронных носителях обычно представляет собой комбинацию текстов на естественном языке и графических изображений. Поэтому для формального представления содержательных информационных объектов в виде вероятностных моделей (1)–(3) необходимы инструменты трансформации текстов и графических изображений в такие абстрактные модели.

Пусть информационный объект представлен текстом на естественном языке. Будем иметь в виду структурированные языки, наделенные морфологией и синтаксисом. Рассмотрим погружение информационного объекта в вероятностную модель (1), или, что одно и то же, формирование по тексту, представленному на естественном языке, вероятностной модели (1). Вероятностная модель позволяет осуществить на необходимом уровне детализации разложение моделируемых текстов на морфологические и синтаксические компоненты, которые в обобщенном виде отражают однотипные семантические представления.

Уровень детализации структуры текстов определяется из условия конечного предназначения вероятностной модели — оценки степени семантического подобия объектов (текстов). При разработке технологии можно исходить из общепринятой практики неавтоматизированной экспертной оценки семантической близости информационных объектов, представленных текстами на естественном языке. Эксперт в содержании текстов выделяет и сопоставляет определяющие семантические аспекты: (1) объекты, т. е. о чем или о ком сообщается в текстах; (2) действия объекта или с объектом; (3) образ и условия действия (как, при каких условиях, где, когда действует объект или осуществляются действия с ним); (4) результат действий и т. п.

Отражению всех содержательных аспектов в структурированных языках соответствуют определенные синтаксические конструкции, опирающиеся на морфологию. В результате разложения сопоставляемых текстов по алгебрам единой структуры тексты трансформируются в обобщенные предложения, в которых роль отдельных членов играют

введенные компоненты алгебры — случайные события A_i, $i = 1, 2, \ldots, m$, представляющие обобщенные подлежащие, сказуемые, дополнения, обстоятельства и т. п. Оценка близости текстов на множестве компонентов алгебры, с учетом семантической ценности слов и конструкций, позволяет принципиально изменить качество оценки подобия текстов по сравнению с отмеченным методом простого подсчета одинаковых слов в текстах (пересечения списков слов).

Для иллюстрации формирования вероятностной модели текста [8] ниже используется простой объект, представленный на естественном языке нижеследующим абзацем текста статьи, выделенным курсивом.

Адаптация абстрактной модели (1) *к описанию информационных объектов, представленных текстовой и графической информацией, достигается конкретизацией случайных величин и алгебры к содержательной специфике объекта.*

Содержание этого предложения может быть выражено бесконечным множеством семантически равноценных фраз, в которых могут использоваться наборы других слов. Из этого следует, что с информационно-семантических позиций слово может трактоваться случайной величиной, а конкретные слова, использованные в этом (и любом другом) предложении — реализациями, элементарными исходами случайной величины «слово». Поэтому выделенное предложение может трактоваться как результат эксперимента, в котором случайная величина «слово» получила следующие элементарные исходы (множество реализаций):

$$\begin{gathered}\Omega_1 = \{\omega_1 = \text{адаптация},\ \omega_2 = \text{абстрактной},\\ \omega_3 = \text{модели},\ \omega_4 = 1,\ \omega_5 = \text{к},\ \omega_6 = \text{описанию},\\ \omega_7 = \text{информационных},\ \omega_8 = \text{объектов},\\ \omega_9 = \text{представленных},\ \omega_{10} = \text{текстовой},\\ \omega_{11} = \text{и},\ \omega_{12} = \text{графической},\ \omega_{13} = \text{информацией},\\ \omega_{14} = \text{достигается},\ \omega_{15} = \text{конкретизацией},\\ \omega_{16} = \text{случайных},\ \omega_{17} = \text{величин},\ \omega_{18} = \text{и},\\ \omega_{19} = \text{алгебры},\ \omega_{20} = \text{к},\ \omega_{21} = \text{содержательной},\\ \omega_{22} = \text{специфике},\ \omega_{23} = \text{объекта}\}. \quad (4)\end{gathered}$$

Называя элементарные исходы, как оговорено выше, просто элементами, можно сказать, что исследуемый текст состоит из элементов (4). Различные слова, образующие Ω_1, несут существенно отличающуюся семантическую нагрузку, которая связана с их морфологической и синтаксической принадлежностью. Поэтому дифференциация слов — элементарных исходов Ω_1 с учетом их семантической значимости позволяет значительно повысить уровень адекватности формальной модели, отражающей информационные объекты. При дифференциации слов целесообразно использовать естественную структуру языка, регламентируемую морфологией и синтаксисом.

В структурированных языках морфология определяет принадлежность слов к частям речи, их грамматические категории и формы. Семантический вес слов определяется их принадлежностью к определенным частям речи. Синтаксис регламентирует строй языка, место и роль в предложении отдельных слов, которые отражают их семантическую значимость. Роль и место слов в предложении, регулируемые синтаксисом, тесно связаны с их морфологией, определяющей принадлежность слов к частям речи и форму представления. Морфология и синтаксис дополняют друг друга и позволяют синтезировать алгебру, достаточно полно отражающую семантическую нагрузку слов в тексте.

Алгебра может синтезироваться на морфологической основе. При этом система случайных событий, по которым распределяются слова, синтезируется на частях речи. Для большинства структурированных языков это существительные, прилагательные, числительные, местоимения и глаголы.

Синтаксис также может быть принят за основу системы случайных событий, по которым распределяются слова, составляющие текст. В этом случае система событий будет представлена наборами подлежащих, сказуемых, определений, дополнений и других членов предложения. Система случайных событий, по которым раскладываются тексты, может конструироваться на комплексной основе морфологии и синтаксиса.

Используя для иллюстрации простейшую морфологическую алгебру, отражающую знаменательные части речи: существительные, прилагательные, числительные, местоимения, наречия и глаголы (предлоги и союзы игнорируются), получаем следующую систему случайных событий:

$$\begin{gathered}\aleph_1 = \{A_1 — \text{существительное},\\ A_2 — \text{прилагательное},\ A_3 — \text{числительное},\\ A_4 — \text{глагол}\}, \quad (5)\end{gathered}$$

где

$$\begin{gathered}A_1 = \{\omega_1 = \text{адаптация},\ \omega_3 = \text{модели},\\ \omega_6 = \text{описанию},\ \omega_8 = \text{объектов},\\ \omega_{13} = \text{информацией},\ \omega_{15} = \text{конкретизацией},\\ \omega_{17} = \text{величин},\ \omega_{19} = \text{алгебры},\\ \omega_{22} = \text{специфике},\ \omega_{23} = \text{объекта}\};\\ A_2 = \{\omega_2 = \text{абстрактной},\\ \omega_7 = \text{информационных},\\ \omega_9 = \text{представленных},\ \omega_{10} = \text{текстовой},\end{gathered}$$

ω_{12} = графической, ω_{16} = случайных,
ω_{21} = содержательной$\}$;

$$A_3 = \{\omega_4 = 1\};$$

$$A_4 = \{\omega_{14} = \text{достигается}\}.$$

Четыре слова (к, и, и, к), содержащиеся в Ω_1 (4), исключены из дальнейшего рассмотрения, так что общее число сохраненных элементарных событий — количество исходов — 19. Вероятность каждого из них $p(\omega_j) = 1/19$. Вероятности случайных событий будут равны:

$$P(A_1) = \frac{10}{19};\ \ P(A_2) = \frac{7}{19};$$

$$P(A_3) = \frac{1}{19};\ \ P(A_4) = \frac{1}{19}.$$

При этом выполняется условие нормированности вероятности:

$$P(A_1) + P(A_2) + P(A_3) + P(A_4) = 1.$$

Формальное представление выделенного текста в виде вероятностной модели (1) имеет вид:

$$M_1 = \{\Omega_1, \aleph_1, P(A_i)\}, \qquad (6)$$

где Ω_1 — множество (4) элементарных исходов — слов, принадлежащих знаменательным частям речи; $\aleph_1$ — система случайных событий (5), ассоциируемая (в примере) со знаменательными частями речи; $P(A_i)$, $i = 1, 2, 3, 4$, — вероятности случайных событий.

Индекс 1 в модели (6) подчеркивает, что это модель конкретного информационного объекта, определенного множеством элементарных исходов Ω_1. При сопоставлении информационных объектов, представленных текстами, они формализуются моделями M_1 и M_2 вида (6), которые отражают их вероятностную структуру на введенной алгебре. Уровень подобия объектов оценивается по близости распределений вероятностей по единой для обоих объектов алгебре. Количественной мерой подобия является взаимная информация в объектах, определение которой рассматривается ниже.

4 Вероятностная модель графических информационных объектов

Графические информационные объекты, оценку содержательной близости которых требуется производить при решении различных практических задач, формализуются в виде вероятностной модели (1). Построение формального образа (1) графического объекта зависит от наличия или отсутствия некоторого конечного «алфавита» в его исходном представлении. Под «алфавитом» в данном случае понимается конечный набор определенных графических структур — элементов, из которых компонуются сопоставляемые в процессе оценки близости графические информационные объекты.

Наличие «алфавитов» характерно для изображений большинства искусственно создаваемых объектов: сооружений и схем различного содержания и предназначения. Такие объекты можно назвать структурированными. Отличными от них являются неструктурированные изображения, например объекты искусства или визуальные изображения некоторых фаз, картин протекания технологических процессов и состояний в них некоторого континуума.

Построение формального образа (1) структурированного объекта может осуществляться по изложенной в предыдущем пункте методике. Наличие конечного алфавита или конечного набора элементов позволяет все элементы моделируемого объекта однозначно связать с конечным множеством $\aleph$ классов A_j элементов. Классы A_j могут конструироваться из элементов ω_i, $i = 1, 2, \ldots, n$, в соответствии с (2) и отражать любые сложные композиции из элементов и классов, соответствующие содержательным представлениям в предметной области. По близости однотипных классов оценивается степень отличия или подобия сопоставляемых объектов. Эти классы, подобно частям речи в предыдущем примере, объявляются случайными событиями A_k, $k = 1, 2, \ldots, K$, множество которых образует алгебру $\aleph$. В результате сопоставляемые графические объекты, отражающие проекты здания или электрические схемы, формализуются в виде вероятностных моделей (1), структура которых детализируется по существенным для оценки степени их подобия компонентам. Количественная мера близости объектов определяется по распределению вероятностей на единой для них алгебре $\aleph$.

Неструктурированные графические объекты представляют собой некоторые непрерывные изображения, степень близости которых требуется оценить. Близкая задача решается в биометрических системах, обеспечивающих оценку сходства контролируемого изображения с заданным его эталоном на определенном уровне доверительной вероятности [9]. При этом системы настраиваются на определенный класс объектов (эталонов). В рассматриваемом подходе предполагается, что объекты могут иметь произвольную палитру и очертания. В этом случае для представления объектов в виде

вероятностных моделей (1) их необходимо структурировать.

Любое изображение в электронном виде представляется множеством точек — пикселов. Каждый пиксел содержит в себе закодированную информацию о цвете, формат представления которой зависит от используемой цветовой модели. Цветовая модель с помощью современных графических редакторов может быть трансформирована в любую другую цветовую модель, которая по тем или иным причинам более удобна для исследователя.

Наиболее распространенной в современных графических форматах является цветовая модель RGB (см., например, [10, 11]). Модель RGB является аддитивной, в ней новые цвета образуются путем добавления основных цветов к базовому черному цвету. Как следует из названия самой модели, основными цветами являются красный, зеленый и синий. Каждая точка несет в себе информацию об интенсивностях этих трех основных цветов, из которых образуется все множество видимых человеком цветовых оттенков. На представление интенсивности каждого цвета отводится 8 бит (или один октет), так что цвет может иметь 256 уровней интенсивности (от 0 до 255). Сочетание цветовых интенсивностей $(0, 0, 0)$ соответствует черному цвету, $(255, 255, 255)$ — белому цвету. Суммарное количество цветов, которое можно представить данной моделью, равно $256 \times 256 \times 256 = 16\,777\,216$. Цветовое пространство RGB можно представить в виде трехмерного куба с осями R, G и B. Сторона куба имеет длину 256 единиц с координатами начала отрезка 0 и конца отрезка 255.

При переходе к вероятностной модели (1) случайными величинами считаются интенсивности цветов в точке. В существующих алгоритмах сравнения изображений (например, в методе цветовых гистограмм [10, 11]) RGB-пространство делится на несколько непересекающихся подпространств, или областей. При обработке изображений подсчитывается количество пикселов, попадающих в каждую из выделенных областей пространства RGB. В результате получается гистограмма распределения частот по подпространствам. Сравнением гистограммы исследуемого объекта с гистограммой эталона оценивается степень их близости. При этом вследствие предопределенности эталона исключается необходимость отражения геометрии объекта.

Оценка близости произвольных графических объектов требует сопоставления не только цветовой палитры, но и привязки геометрии объектов к системе координат для сопоставления конфигурации. Разработана модификация метода цветовых гистограмм, позволяющая отразить конфигурацию сопоставляемых объектов [12]. Модификация состоит в добавлении к традиционному RGB-пространству координатных осей X и Y. Комбинированное с координатными осями $RGBXY$-пространство позволяет хранить информацию о цветовой интенсивности и пространственном расположении точек изображения.

Пятимерное $RGBXY$-пространство может быть разбито на систему непересекающихся подпространств, которые позволяют определить систему случайных событий A_i, $i = 1, 2, \ldots, m$, образующих алгебру $\aleph$. Сетка разбиения необязательно равномерная, что позволяет детализировать образы изображений. Распределение точек изображения по подпространствам формирует распределение вероятностей для образов сопоставляемых объектов. В результате не структурированные исходно графические объекты формализуются в виде вероятностных моделей (1).

5 Технология оценки близости объектов

Технология оценки степени близости сопоставляемых объектов, представленных их образами в виде вероятностных моделей (1), может базироваться на представлениях теории информации. В теории информации вводится мера количества информации, содержащегося в случайной величине, которая позволяет определить количественные меры соотношения информации в случайных объектах, характеризующие уровень близости объектов.

Информационные объекты исходно представлены различными наборами элементарных событий $\Omega_1 = (\omega_1^1, \omega_2^1, \ldots, \omega_{N_1}^1)$, $\Omega_2 = (\omega_1^2, \omega_2^2, \ldots \ldots, \omega_{N_2}^2)$ — реализаций случайных величин. В данном контексте под случайными величинами понимаются минимальные неделимые (атомарные) элементы, наборами которых представляются информационные объекты: слова, элементы изображения формул, схем, пикселы и т. п. Элементарными событиями — значениями случайных величин (реализациями) будут конкретные слова исследуемого текста, элементы изображения конкретных формул, схем, $RGBXY$-значения пикселов и т. п.

Для представления объектов в виде вероятностных моделей (образов) вводится единая алгебра $\aleph = \{A_1, A_2, \ldots, A_m\}$, структура которой должна максимально полно отражать принципиальные компоненты информационной сущности сопоставляемых объектов. Содержание информационных объектов $\Omega_1 = (\omega_1^1, \omega_2^1, \ldots, \omega_{N_1}^1)$, $\Omega_2 = (\omega_1^2, \omega_2^2, \ldots \ldots, \omega_{N_2}^2\}$ раскладывается по системе событий $\aleph = = \{A_1, A_2, \ldots, A_m\}$. В результате информационные

объекты представляются в виде систем случайных событий:

$$\left.\begin{aligned}\Omega_1 = (\omega_1^1, \omega_2^1, \ldots, \omega_{N_1}^1) \Rightarrow \\ \Rightarrow \aleph = \{A_1^1, A_2^1, \ldots, A_m^1\}\,; \\ \Omega_2 = (\omega_1^2, \omega_2^2, \ldots, \omega_{N_2}^2) \Rightarrow \\ \Rightarrow \aleph = \{A_1^2, A_2^2, \ldots, A_m^2\}\,.\end{aligned}\right\} \quad (7)$$

На основании свойства (2) из случайных событий с помощью операций над множествами могут быть синтезированы новые случайные события, принадлежащие алгебре $\aleph = \{A_1, A_2, \ldots, A_m\}$. Количественной характеристикой распределения реализаций $\Omega_1 = (\omega_1^1, \omega_2^1, \ldots, \omega_{N_1}^1)$, $\Omega_2 = (\omega_1^2, \omega_2^2, \ldots, \omega_{N_2}^2)$, составляющих объекты, по системе случайных событий $\aleph = \{A_1, A_2, \ldots, A_m\}$ служат, в соответствии с (3), эмпирические вероятности

$$P(A_j^1) = \sum_{w_i^1 \in A_j^1} p(\omega_i^1)\,;\;\; P(A_j^2) = \sum_{\omega_i^2 \in A_j^2} p(\omega_i^2)\,. \quad (8)$$

В результате информационные объекты M_1 и M_2, заданные множествами реализаций $\Omega_1 = (\omega_1^1, \omega_2^1, \ldots, \omega_{N_1}^1)$, $\Omega_2 = (\omega_1^2, \omega_2^2, \ldots, \omega_{N_2}^2)$, представляются в виде вероятностных моделей

$$\left.\begin{aligned}M_1 = \{\Omega_1, \aleph, P(A_j^1)\}\,; \\ M_2 = \{\Omega_2, \aleph, P(A_j^2)\}\,.\end{aligned}\right\} \quad (9)$$

Назовем формализованное представление информационных объектов (9) в виде вероятностных моделей (1) образами объектов. Задача состоит в оценке подобия объектов, имеющих образы (9).

Формальная оценка степени подобия объектов может базироваться на количественных характеристиках их образов. Адекватной основой для синтеза таких характеристик представляется теория информации, основоположником которой является К. Шеннон [13]. В теории информации количество информации, содержащееся в случайной величине, может оцениваться энтропией, которая определяется по распределению вероятностей случайной величины. Энтропия вероятностной модели, отражающая количество информации в ней, может определяться в виде

$$H = -\sum_{A_j \in \aleph} P(A_j) \ln P(A_j)\,, \quad (10)$$

где $P(A_j)$ — вероятности случайных событий (8).

На основании энтропии могут быть синтезированы различные меры близости информационных объектов. Учитывая аналогию терминов и представлений теории вероятностей и теории множеств, воспользуемся графической иллюстрацией,

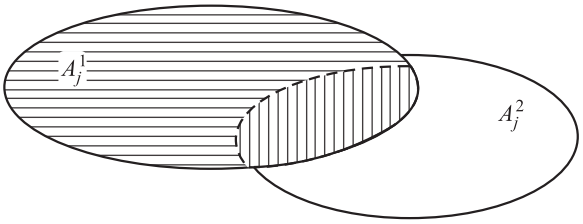


Рис. 1 Геометрическая интерпретация операций с количествами информации, содержащимися в одноименных случайных событиях системы $\aleph$ двух объектов

представленной на рис. 1, для пояснения существа предлагаемых мер близости информационных объектов.

Площадь эллипсов A_j^1 и A_j^2 на рис. 1 условно отображает вероятности случайных событий A_j^1 и A_j^2 (множества реализаций ω, принадлежащих этим событиям). Для отдельных реализаций и сформированных из них случайных событий вероятности могут быть определены по (8). Но для оценки близости объектов необходимо рассматривать объект, представляющий собой объединение исходных. Множество его реализаций получается объединением множеств (7) $\Omega = (\Omega_1 + \Omega_2)$, и количество элементов в объединенном объекте равно сумме $N = N_1 + N_2$. Поэтому вероятности комбинированных случайных событий, показанных на рис. 1, определяются следующим образом:

$$p_i(\omega_i^l) = \frac{n(\omega_i^l)}{N_1 + N_2}\,;$$
$$P_j^l(A_j^l) = \sum_{\omega_i^l \in A_j^l} p_i(\omega_i^l)\,;\;\; l = 1, 2\,,$$

где $n(\omega_i^l)$ — количество реализаций ω_i^l в множестве реализаций Ω^l.

Для оценки близости могут быть использованы комбинированные случайные события, определенные для объединенного объекта $\Omega = (\Omega_1 + \Omega_2)$. Они получаются (см. рис. 1) композицией случайных событий первого A_j^1, $j = 1, 2, \ldots, m$, и второго A_j^2, $j = 1, 2, \ldots, m$, объектов с помощью отмеченных выше операций для всех компонентов системы $\aleph$. Это следующие комбинированные случайные события:

(1) сумма (объединение) случайных событий A_j^1 и A_j^2, которая на рис. 1 представляется площадью обоих эллипсов и определяется в виде

$$A_j^{1+2} = A_j^1 + A_j^2 = \{\omega \in \Omega | \omega \in [A_j^1 + A_j^2]\}\,; \quad (11a)$$

(2) произведение (пересечение) случайных событий A_j^1 и A_j^2, включающее реализации, при-

надлежащие одновременно A_j^1 и A_j^2, и определяемое в виде

$$A_j^{1\&2} = A_j^1 \& A_j^2 = \{\omega \in \Omega | \omega \in A_j^1 \,\&\, \omega \in A_j^2\}\,, \quad (11б)$$

на рис. 1 ему соответствует площадь с вертикальной штриховкой;

(3) разность случайных событий A_j^1 и A_j^2 (событие, включающее элементы $\omega \in \Omega$, принадлежащие A_j^1 и не принадлежащие A_j^2), которая определяется в виде

$$A_j^{1-2} = A_j^1 - A_j^2 = \{\omega \in \Omega | \omega \in A_j^1 \,\&\, \omega \notin A_j^2\}\,, \quad (11в)$$

на рис. 1 ей соответствует площадь с горизонтальной штриховкой;

(4) разность случайных событий A_j^2 и A_j^1, отражаемая на рис. 1 областью без штриховки и определяемая в виде

$$A_j^{2-1} = A_j^2 - A_j^1 = \{\omega \in \Omega | \omega \in A_j^2 \,\&\, \omega \notin A_j^1\}\,. \quad (11г)$$

События (11) позволяют отразить уровень совпадения или различия множеств реализаций, входящих в однотипные случайные события системы $\aleph = \{A_1, A_2, \ldots, A_m\}$, по которой раскладываются образы сопоставляемых объектов. В зависимости от сущности объекта его реализации представляют собой слова естественного или искусственного (формального) языка, компоненты структурированных или неструктурированных графических объектов.

Нетрудно видеть, что пересечение (11б) отражает множество общих для обоих объектов реализаций j-го типа, которое и определяет степень близости объектов: чем больше значение (11б), тем больше степень их подобия. Наоборот, разности (11в), (11г), независимо от знака, отражают отличие объектов, которое возрастает с увеличением разности.

Сумма (11а) представляет объединение реализаций (элементарных событий) обоих объектов. Сопоставляемые объекты известны в виде множеств Ω_1 и Ω_2 элементарных событий, так что их объединение представляет общее множество элементарных событий

$$\Omega = (\Omega_1 + \Omega_2) = \\ = (\omega_1^1, \omega_2^1, \ldots, \omega_{N_1}^1, \omega_1^2, \omega_2^2, \ldots, \omega_{N_2}^2)\,. \quad (12)$$

На множестве Ω могут быть определены вероятности всех типов комбинированных случайных событий (11) на всех компонентах алгебры $\aleph = \{A_1, A_2, \ldots, A_m\}$ в виде суммы вероятностей $p(\omega_i)$ реализаций ω_i^1 и ω_i^2:

$$P(A_j^V) = \sum_{\omega_i \in A_j^V} p(\omega_i) = \sum_{\omega_i \in A_j^V} \frac{n(\omega_i)}{N_1 + N_2}\,, \\ j = 1, 2, \ldots, m\,,$$

где $n(\omega_i)$ — количество исходов ω_i в множестве Ω, соответствующих указанной верхним индексом V операции из набора (11).

Количество информации в случайных событиях определяется энтропией (10), которая может быть вычислена по распределению вероятностей событий любого типа из (11) по системе $\aleph$. Собственно энтропия является абстрактной величиной и не может характеризовать степень близости объектов. Но если использовать отношение энтропий, характеризующих количество информации в комбинированных случайных событиях, определенных на множестве (12), то можно получить меры близости, достаточно адекватные задаче оценки подобия объектов.

Например, система случайных событий (11б) отражает объем общей для сопоставляемых объектов информации, который количественно может оцениваться энтропией, вычисляемой по вероятностям системы (11б):

$$H(1\&2) = -\sum_{j=1}^{m} P(A_j^{1\&2}) \ln P(A_j^{1\&2})\,. \quad (13)$$

Информация, отличающая объекты 1 и 2, отражается случайными событиями (11в). Количественно объем этой информации определяется энтропией, вычисляемой по вероятностям событий (11в):

$$H(1-2) = -\sum_{j=1}^{m} P(A_j^{1-2}) \ln P(A_j^{1-2})\,. \quad (14)$$

Меры степени отличия объектов могут содержательно соответствовать интуитивным представлениям о близости как о расстоянии между объектами. Отношение количества различающей объекты информации (14) к количеству информации, общей для обоих объектов (13), представляет одну из таких безразмерных величин

$$\rho 1 = \frac{H(1-2)}{H(1\&2)}\,. \quad (15)$$

Для оценки диапазона изменения величины (15) можно рассмотреть два предельных варианта: (1) объекты идентичны и (2) объекты не имеют общих элементов. Значение (15) в первом варианте получается из условий идентичности множеств элементарных событий, образующих объекты: $\Omega_1 = \Omega_2 \to N_2 = N_1$ и $(\omega_1, \omega_2, \ldots, \omega_{N_1}) = (\omega_1, \omega_2, \ldots, \omega_{N_2})$. Отсюда следует совпадение определенных на Ω_1 и Ω_2 случайных событий A_j^1, A_j^2, $j = 1, 2, \ldots, m$, и равенство распределений ве-

роятностей $P_1(A_j^1) = P_2(A_j^2)$. Поэтому разность вероятностей равна нулю: $P(A_j^{1-2}) = P(A_j^1 - A_j^2) = = P(0)$, $j = 1, 2, \ldots, m$. В теории информации принято соглашение $P(0)\ln P(0) = 0$; следовательно, $H(1{-}2) = 0$.

Сопоставляемые объекты предполагаются независимыми. Пересечение множеств $A_j^1 = A_j^2$ равно $A_j^{1\&2} = A_j^1 \& A_j^2 = 2A_j^1 = 2A_j^2$ и $P(A_j^{1\&2}) = = P(A_j^1) \times P(A_j^2)$, $j = 1, 2, \ldots, m$; следовательно, $H(1\&2)_j = H(A_j^1) + H(A_j^2)$ и общая энтропия (13) будет отражать количество информации в обоих объектах $H(1\&2)$, отличное от нуля. Поэтому в первом варианте идентичности объектов значение (15) $\rho 1 = 0/H(1\&2) = 0$.

Второй вариант, когда объекты не имеют общих элементов, дает разность, равную содержанию уменьшаемого, а пересечение — равное нулю. Поэтому получается $\rho 1 = \infty$. Так что мера (15) изменяется от нуля для совпадающих объектов до бесконечности для объектов, не имеющих общих элементов, что соответствует представлениям о расстоянии, как мере близости.

Могут быть использованы и другие, близкие по смыслу $\rho 1$, меры. Например, в числителе (15) энтропию, отражающую вероятности событий (11в), можно заменить энтропией, отражающей общее количество информации в объектах, вычисляемой по вероятностям событий (11а). В этом случае мера получается в виде

$$\rho 2 = \frac{H(1+2)}{H(1\&2)}\,, \qquad (16)$$

где $H(1+2)$ вычисляется по (14) заменой событий (11в) событиями (11а).

Мера (16) также может интерпретироваться расстоянием между объектами, которое для идентичных объектов будет равно единице, а для абсолютно разных — бесконечности. Для придания содержательного соответствия технологии оценки близости объектов существу объектов и решаемым задачам меры подобия типа (15) или (16) могут градуироваться в соответствующих единицах. Разработанная технология позволяет дифференцировать информационные объекты по системе случайных событий — алгебре $\aleph$, отражающей их семантическую значимость.

При разработке своего подхода к оценке близости информационных объектов автор сознательно сделал упор на использование табличного их представления по следующей причине. Информационные объекты разного типа являются композициями некоторых атомарных (неделимых при исследовании) элементов. На этих атомарных элементах конструируется алгебра $\aleph$ вероятностной модели (1). После определения алгебры информационный объект раскладывается по системе случайных событий и представляется таблицей, столбцы которой именуются случайными событиями.

В реляционных базах данных таблица именуется отношением, а столбцы — атрибутами. Дело, конечно, не в названиях, а в том, что табличное представление информационного объекта превращает его в обычное отношение структуры данных. Для работы с отношением (в данном контексте — с представлением вероятностного образа объекта) могут быть использованы операции реляционной алгебры, которые позволяют из исходных отношений конструировать и автоматически формировать новые отношения, наращивая уровень их сложности композицией предшествующих атрибутов. Поэтому табличное представление образов информационных объектов открывает возможности использовать реляционную алгебру для автоматизации процедур их исследования и детализации представления объектов.

Например, синтаксис определяет построение из слов — атомарных компонентов языка — различных конструкций, несущих определенную семантическую нагрузку. Поэтому, начиная с алгебры, представленной атомарными компонентами, могут вводиться композиционные конструкции атомарных компонентов с постепенным наращиванием сложности из компонентов предшествующих уровней. Выполняться все эти операции могут стандартными средствами реляционных баз данных.

На основе математического аппарата реляционных баз данных может быть реализована автоматизированная рекурсивная процедура формирования структуры вероятностных образов (алгебры) сравниваемых объектов. Начальное приближение $\aleph(0)$ системы случайных событий (7) задается на уровне атомарных событий, формируются вероятностные модели $M_1(0)$ и $M_2(0)$ (9), определяется мера их близости. Затем на основе $\aleph(0)$ начального приближения, свойств (2) алгебры, синтезируются новые случайные события и их атрибуты — композиции атрибутов из $\aleph(0)$. В результате получается алгебра $\aleph(1)$, по атрибутам которой формируются новые отношения (таблицы) $M_1(1)$ и $M_2(1)$, определяется мера близости объектов и сравнивается с полученной на $M_1(0)$ и $M_2(0)$. На основании сравнения значений меры выбирается продолжение рекурсии или прекращение процесса уточнения оценки близости объектов.

Усиление дифференциации семантической значимости отдельных компонентов алгебры $\aleph$ в конкретных задачах достигается введением весовых коэффициентов. Для настройки технологии на конкретные объекты и задачи может применяться

детализация алгебры, адаптация весов и градуировки меры подобия объектов на основании эмпирической информации. Эти возможности позволяют создавать инструменты эффективной оценки близости содержательной сущности информационных объектов.

6 Иллюстрация применения технологии

Тестирование разработанной технологии оценки близости информационных объектов было осуществлено на текстовых и графических объектах. Результаты ее применения изложены в ряде работ. В [8] приведены результаты экспериментальной проверки возможности применения разработанной технологии для автоматизированной оценки знаний студентов. В штатном режиме контроля знаний группа из 22 студентов написала изложение на английском языке. Изложения были проверены и оценены по 100-балльной шкале преподавателем по стандартной методике. Затем тексты работ студентов и исходный текст, прочитанный преподавателем, были введены в систему, в которой работы студентов представлялись в виде образов-копий $M_1(S)$, $S = 1, 2, \ldots, 22$, а исходный текст принимался за эталон M_2.

Близость ответов эталону оценивалась по изложенной выше методике. Для представления объектов в эксперименте использовалась следующая система событий (7): $A_1 =$ существительное; $A_2 =$ глагол; $A_3 =$ прилагательное; $A_4 =$ наречие; $A_5 =$ числительное; $A_6 =$ неопределенное слово.

Количество взаимной информации $H^{\text{Э\&О}}$ было вычислено по (13) для всех ответов. Для придания абстрактной энтропии содержательного смысла она с использованием оценок, выставленных преподавателем, была проградуирована в единицах 100-балльной системы оценок, принятой в вузе. Параметры модели градуировки определялись методом наименьших квадратов в двух вариантах: первый в виде $y = a + bH^{\text{Э\&О}}$ и второй в виде $y_1 = a_0 + \sum\limits_i a_i H_i^{\text{Э\&О}}$, где $H_i^{\text{Э\&О}}$ — энтропии случайных событий $A_i^{\text{Э\&О}}$, $i = 1, 2, \ldots, 6$. Фактически во втором варианте параметры $a_i, i = 1, 2, \ldots, 6$, отражают различный «вклад» семантических компонентов в соответствие между эталоном и ответом и иллюстрируют возможности адаптации градуировки к содержательным особенностям проверяемых дисциплин, к методикам оценки и т. п. Более подробно детали структуризации текстов, определения взаимной информации и методики ее градуировки изложены в [8].

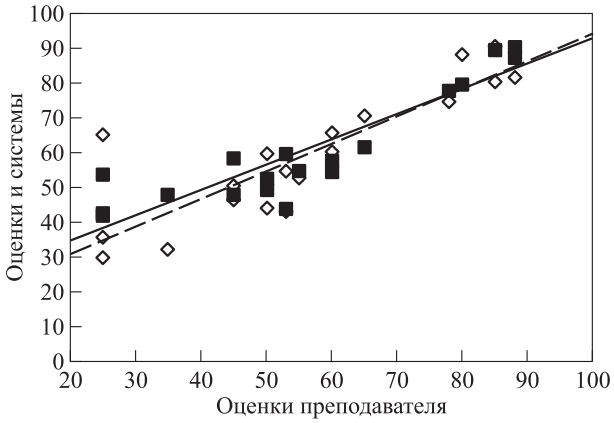


Рис. 2 Сопоставление оценок преподавателя и оценок системы: *1* — при градуировке без взвешивания частей речи; *2* — при взвешивании частей речи

На рис. 2 приводится графическая иллюстрация результата из [8], отражающая уравнения градуировки и отличие оценки, определяемой автоматически по близости ответа эталону, от оценки, выставленной преподавателем.

Среднеквадратичное отклонение оценок, выставленных преподавателем, от оценок по количеству информации y при градуировке по первому варианту составляет 11,04 балла, коэффициент корреляции равен 0,847, при градуировке по второму варианту уклонение составило 6,324 балла, а множественный коэффициент корреляции достиг 0,955.

При детальном анализе выяснилось, что две оценки, выставленные преподавателем и существенно выпадавшие из общего ряда, были не в полной мере адекватны содержанию изложений. Исключение двух этих точек и расчет по оставшимся 20 точкам принципиально изменяет результат: среднеквадратичная ошибка оценки y становится равной 1,593 балла, а оценки y_1 — 1,248 балла.

Регрессия, полученная для этого эмпирического материала, следуя векторно-пространственной модели текста, предложенной Г. Солтоном [2], для всех 22 изложений дает среднеквадратичное отклонение прогноза оценки от оценки преподавателя 13,27 балла. После удаления из массива двух выпадавших точек ошибка составила 2,155 балла, т. е. почти в два раза выше, чем при использовании информационной модели 1, 248 балла.

Разработанная технология формального сопоставления информационного содержания текстов и синтеза количественной меры семантической близости слов языка позволяет автоматизировать исследования проблем в области языкознания. Например, семантические отношения между словами (синонимы, антонимы, паронимы, гипонимы

Таблица 1 Алгебра событий и характеристики сопоставления схем

Случайные события	Перечень	Количество	Вероятность
A_1 — элементы	R4, L1, . . .	n_1	n_1/N
A_2 —ветви	L1C1, L2C2,. . .	n_2	n_2/N
A_3 — узлы	T1BC1R4,. . .	N_3	N_3/N

и т. п.) определяются в настоящее время субъективно представителями разных школ. Субъективные представления не формализуются, не могут быть упорядочены или сопоставлены.

Между тем понятно, что в основе определения семантических отношений лежат различные композиции множеств оттенков значений сопоставляемых слов (11), которые представлены в существующих словарях. Изложенный подход позволяет формализовать и, следовательно, автоматизировать процедуры количественной оценки меры «информационного расстояния» между словами. На этой основе могут быть введены объективные, количественные меры синонимичности, антонимичности и т. п. Интересные результаты в этой области могут быть получены с использованием нечетких отношений вместо четких, показанных на рис. 1.

В работе [14] изложена технология разработки универсального метрического тезауруса языка на примере русского языка. Технология базируется на формировании для каждого слова языка его содержательного образа в виде вероятностной модели. Существуют специализированные словари, ориентированные на определенные сферы деятельности и знаний. Обычно эти словари составляются экспертами в области языкознания на основании их субъективных представлений о семантике слов.

Технология представления текстов вероятностными моделями позволяет автоматизировать создание тезауруса языка в виде обобщенных содержательных образов отдельных слов языка. Обобщенный образ формируется в виде структурированной на выбранной алгебре $\aleph$ суммы отражающих смысл слова словарных статей из всех доступных словарей. Тезаурус формируется автоматически по введенным электронным версиям имеющихся словарей и представляет словарь, в котором с каждым словом связан его максимально полный содержательный образ.

Разработанный универсальный метрический тезаурус позволяет формально, опираясь на всю имеющуюся информацию о значении слов, и в этом смысле объективно, решать проблему синонимов при оценке близости текстов. Метрическая оценка семантической близости слов производится автоматически по изложенной технологии сопоставлением их образов. Мерой близости является расстояние, определяемое в виде (15) или (16) по обобщенным образам слов. При появлении новых версий электронных словарей они могут вводиться в систему и автоматически ассимилироваться ею. Технология создания тезауруса с использованием образов слов в виде вероятностных моделей (1) инвариантна по отношению к структурированным языкам и может быть использована в любом из них. Универсальный электронный тезаурус [14] является мощным инструментом для исследований в сфере филологии и языкознания.

В работе [4] показана возможность применения информационной технологии и вероятностной модели (1) для оценки близости схем. Схемы используются в процессе изучения многих дисциплин технического направления и поэтому широко представлены в обучающих системах. Для примера использовались электрические схемы, множества элементарных событий при представлении которых формируются элементами, регламентированными стандартом [7].

В табл. 1 показана система случайных событий, составляющих алгебру в этом случае, и количественные характеристики схем, при этом n_i, $i = 1, 2, 3$, — количество компонентов (реализаций) i-го типа, $N = n_1 + n_2 + n_3$ — общее количество компонентов в схеме.

Проверка осуществлялась на схеме, содержащей компоненты $n_1 = 41, n_2 = 34, n_3 = 18, N = 93$. Ошибки экзаменуемых моделировались устранением компонентов из схем-ответов. На рис. 3 показаны результаты увеличения информационного расстояния между эталоном и ответами по мере нарастания их ошибочности.

Полученные результаты показали [4] возможность синтеза автоматизированных процедур для оценки уровня соответствия схем их эталонному образу. Такие процедуры позволяют, в частности, разрабатывать в обучающих системах модули проверки ответов обучаемых в этой области.

Применение разработанной технологии для оценки близости неструктурированных графических объектов [5] показало возможность на ее

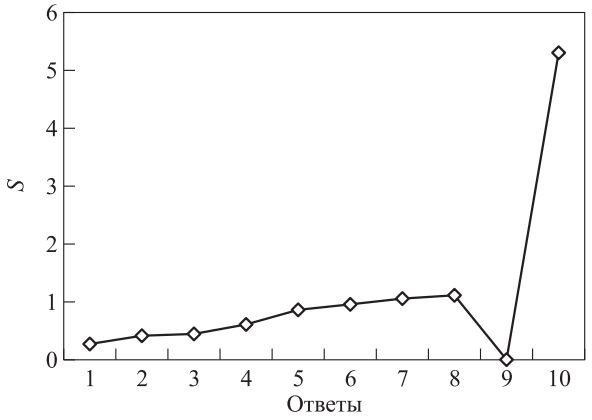


Рис. 3 График роста информационного расстояния при последовательном увеличении отклонений схем-ответов от эталона

основе существенного повышения уровня достоверности оценки. Этот вывод следует непосредственно из данных воспроизводимой табл. 2 из [5], в которой показаны сопоставляемые объекты и меры их близости, полученные по разным технологиям. Объекты выбраны так, чтобы их подобие и отличие были очевидны: уровень подобия объектов 1 и 3 намного выше, чем объектов 1 и 2, или 2 и 3.

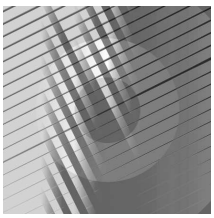
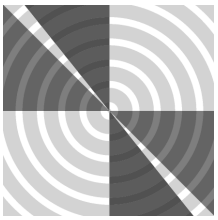
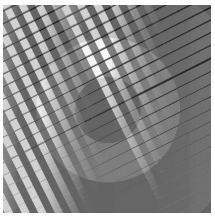
В работе [5] приведены результаты исследования различных мер для оценки степени близости объектов. Для исследования были реализованы стандартные, используемые на практике методы и оригинальные, опирающиеся на сопоставление количеств информации в объектах.

Для стандартного метода цветовых гистограмм [9–11], в основу которого положена RGB-модель, были реализованы алгоритмы, использующие модификации разбиения RGB-пространства на подпространства и RGB-осей — на интервалы. Цветовые модели объектов, как это принято в биометрических системах, представлялись в виде цветовых гистограмм. Оценка близости объектов производилась по корреляции их цветовых гистограмм.

Затем были разработаны алгоритмы, реализующие оригинальную технологию представления объектов в виде вероятностных моделей, синтезируемых разбиением пятимерного $RGBXY$-пространства на систему непересекающихся подпространств, образующих алгебру $\aleph = \{A_1, A_2, \ldots, A_m\}$.

На алгебре $\aleph = \{A_1, A_2, \ldots, A_m\}$ определялись комбинированные события (11а) и (11б) и соответствующие им распределения вероятностей для объектов 1, 2, 3, показанных в табл. 2. На распределениях вероятностей на системе введенных классов $\aleph$ вычислялись соответствующие энтропии. В исследовании [5] близость объектов в информационном масштабе оценивалась энтропийным расстоянием

Таблица 2 Результаты оценки близости объектов разными методами

Тип	Метод	Характеристики для образов			
		Образ 1	Образ 2	Образ 3	Совместная
	Графическое представление				—
Классические	Разбиение RGB-осей	—	—	—	$R_{12} = 0{,}897$ $R_{13} = 0{,}735$
	Разбиение RGB-пространства	—	—	—	$R_{12} = 0{,}7$ $R_{13} = 0{,}746$
Разработанные	Модифицированный метод	—	—	—	$R_{12} = 0{,}324$ $R_{13} = 0{,}913$
	Интегральный метод	$E_1 = 461{,}811$	$E_2 = 463{,}793$	$E_3 = 462{,}152$	$R_{12} = 0{,}055$ $R_{13} = 0{,}922$
	Дифференциальный метод	$E_1 = 1134{,}189$	$E_2 = 1280{,}136$	$E_3 = 1234{,}701$	$R_{12} = 0{,}069$ $R_{13} = 0{,}938$
	Энтропийное расстояние	$H_{12}(\cup) = 971{,}602$ $H_{13}(\cup) = 1227{,}317$	$H_{12}(\cap) = 110{,}171$	$H_{13}(\cap) = 740{,}361$	$\rho 2^{12} = 4{,}41$ $\rho 2^{13} = 0{,}829$

между ними, определяемым в виде (16), которое изменяется от 1 для совпадающих объектов до ∞ для объектов, не имеющих общих элементов.

Оценка близости графических объектов, показанных в табл. 2, была произведена по существующим и разработанным технологиям. При использовании стандартного метода цветовых гистограмм оценка осуществлялась коэффициентом корреляции цветовых гистограмм объектов, которые обозначены R_{ij}, где i, j — номера объектов в первой строке табл. 2.

В алгоритмах биометрической идентификации [9] пороговое значение степени соответствия устанавливается обычно на уровне 0,65. Из табл. 2 следует, что значения коэффициента подобия, полученные по классическим алгоритмам разбиения RGB-осей на интервалы и RGB-пространств на прямоугольные параллелепипеды, превышают пороговое значение для пар 1, 2 и 1, 3. Это свидетельствует об идентичности всех трех объектов. Оценки близости всех трех объектов практически совпадают, хотя отличие объекта 2 от 1 и 3 очевидно.

Введение в оценку близости объектов системы координат (модифицированный метод $RGBXY$) существенно изменяет результат. Коэффициент подобия образов 1 и 2 при использовании модифицированного алгоритма принимает значение $R = 0{,}324$, которое более чем в 2 раза ниже коэффициентов 0,897 и 0,7, полученных с использованием классических алгоритмов. Одновременно коэффициент подобия объектов 1 и 3 увеличился. Сопоставление «модифицированных пространственно-цветовых гистограмм» показывает, что объекты 1 и 3 примерно в 3 раза «ближе», чем объекты 1 и 2.

Информационный метод дает адекватную сопоставляемым изображениям меру соответствия в виде расстояния между ними. Расстояние $\rho 2$ пропорционально различию объектов и тем больше, чем меньше степень их совпадения. Расстояние $\rho 2_{1,2} = 4{,}41$ между объектами 1 и 2 более чем в пять раз превышает расстояние $\rho 2_{1,3} = 0{,}829$ между объектами 1 и 3. Можно видеть, что реально следующая из рисунков, включенных в табл. 2, близость объектов 1 и 3 отражается близостью значения $\rho 2_{1,3} = 0{,}829$ к минимально возможному, равному 1, для совпадающих объектов.

Градуировка шкалы меры $\rho 2$ (и иных, получаемых информационным методом) может быть осуществлена на основании реального эмпирического материала. Технология вычисления оценки инвариантна по отношению к разбиению объектов на подмножества. Разбиение объектов может производиться с переменным шагом по осям координат, что позволяет акцентировать внимание на областях с высокой плотностью информации использованием мелкого шага. Процедура оценки близости объектов может быть итеративной с использованием на каждом следующем шаге более детальных образов объектов, получаемых детализацией алгебры $\aleph$, использованной на предыдущем этапе оценки. И сетка разбиения, и веса областей могут автоматически адаптироваться в процессе функционирования системы.

7 Заключение

Формальная оценка семантического подобия информационных объектов может базироваться на количественной мере сопоставления их содержания. Случайность элементарных компонентов описания информационных объектов и произвольность формы представления объектов в информационных технологиях выдвигает актуальную задачу разработки формальной технологии и меры оценки уровня их семантического подобия. Объекты представляются множествами случайных реализаций набора элементарных изобразительных компонентов.

Элементарные компоненты различны для разных форм представления, но являются общими для сопоставляемых объектов. Единый подход к содержанию информационных объектов как множеству реализаций случайных величин позволяет разработать единую технологию оценки их близости. Различия представления объектов разного типа сводятся лишь к разным наборам элементарных изобразительных компонентов.

Информационные объекты разного типа формализуются вероятностными моделями, в которых реализации группируются в случайные события в соответствии с их информационной ценностью. Количественными характеристиками разложения информационных объектов по системам случайных событий являются распределения вероятностей. Распределения вероятностей определяют энтропии объектов. Энтропии отражают количества информации в сопоставляемых информационных объектах и позволяют синтезировать количественную меру их близости, подобную метрической.

Технология позволяет разработать универсальные эффективные процедуры оценки подобия информационных объектов, представленных графически и текстами на естественном или искусственном языке. Процедуры могут использоваться в системах поиска информации, оценки близости текстовых и графических объектов, автоматизированной проверки уровня усвоения знаний.

Литература

1. *Manning Ch. D., Raghavan P., Schütz H.* An introduction to information retrieval. — Cambridge: University Press, 2009. 569 p.
2. *Salton G., Wong A., Yang C. S.* A vector space model for automatic indexing // Comm. ACM, 1975. Vol. 18. No. 11. P. 613–620.
3. *Кузнецов Л. А.* Вероятностно-статистическая оценка адекватности информационных объектов // Информатика и её применения, 2011. Т. 5. Вып. 4. С. 39–50.
4. *Кузнецов Л. А., Кузнецова В. Ф., Антонов Д. И.* Оценка близости графических объектов на примере электрических схем с помощью информационного критерия // Открытое и дистанционное образование, 2013. № 2(50). С. 35–43.
5. *Кузнецов Л. А., Бугаков Д. А.* Разработка меры оценки информационного расстояния между графическими объектами // Информационно-управляющие системы, 2013. № 1. С. 74–79.
6. *Гнеденко Б. В.* Курс теории вероятностей. — 9-е изд., испр. — М.: ЛКИ, 2007. 448 с.
7. ГОСТ 2.743-91 ЕСКД. Обозначения условные графические в схемах. Элементы цифровой техники. — М.: Госстандарт, 1991. 75 с.
8. *Кузнецов Л. А., Кузнецова В. Ф.* Оценка семантической адекватности текстов информационным методом // Информатика и её применения, 2013. Т. 7. Вып. 1. С. 19–29.
9. *Гаспарян А. В., Киракосян А. А.* Система сравнения отпечатков пальцев по локальным признакам // Вестник РАУ. Сер. Физико-математические и естественные науки, 2006. № 2. С. 85–91.
10. *Swain M. J., Ballard D. H.* Color indexing // Int. J. Computer Vision, 1991. Vol. 7. No. 1. P. 11–32.
11. *Sticker M., Orengo M.* Similarity of color images // SPIE Conference Proceedings, 1995. Vol. 2420. P. 381–392.
12. *Кузнецов Л. А., Бугаков Д. А.* Развитие метода сравнения и классификации графических объектов // Вестник компьютерных и информационных технологий, 2013. № 2(104). С. 11–16.
13. *Шеннон К.* Работы по теории информации и кибернетике. — М.: Изд-во ИЛ, 1963. 833 с.
14. *Кузнецов Л. А., Кузнецова В. Ф., Капнин А. В.* Универсальный метрический тезаурус русского языка // Информатика и её применения, 2013. Т. 7. Вып. 3. С. 27–35.

Поступила в редакцию 10.12.13

UNIVERSAL TECHNOLOGY OF INFORMATION OBJECTS PROXIMITY ASSESSMENT

L. A. Kuznetsov

Russian Presidential Academy of National Economy and Public Administration (Lipetsk Branch), 3 Internatsional'naya Str., Lipetskaya oblast, Lipetsk 398050, Russian Federation

Abstract: The paper outlines the technology used to determine the degree of similarity of information objects, which are represented by text or graphic images. Objects are formalized by probabilistic models. The structure of the model is set by an algebra on a minimum set of graphic components of an object. Quantitative characteristics of the structure of objects are the probability distributions on the algebra. The amount of information in objects is estimated by entropy. The similarity measure of information objects is based on entropy. The paper describes the method of estimating the proximity of text and graphic objects. The paper provides several examples of estimation algorithms implementation. It is shown that the developed method is more efficient compared to the methods described in the literature. The technology used to form images of information objects and to compare their semantic content is universal. It is possible to adapt the technology to the meaningful characteristics of objects being analyzed.

Keywords: information object; text; image; probabilistic model; semantic similarity; entropy; measure of similarity

DOI: 10.14375/19922264140213

References

1. Manning, Ch. D., P. Raghavan, and H. Schütz. 2009. *An introduction to information retrieval*. Cambridge: University Press. 569 p.
2. Salton, G., A. Wong, and C. S. Yang. 1975. A vector space model for automatic indexing. *Comm. ACM*. 11:613–620.
3. Kuznetsov, L. A. 2011. Veroyatnostno-statisticheskaya otsenka adekvatnosti informatsionnykh ob"ektov [Probabilistic and statistical evaluation of the adequacy of information objects]. *Informatika i ee Primeneniya — Inform. Appl.* 5(4):39–50.
4. Kuznetsov, L. A., V. F. Kuznetsova, and D. I. Antonov. 2013. Otsenka blizosti graficheskikh ob"ektov na primere

elektricheskikh skhem s pomoshch'yu informatsionnogo kriteriya [Estimation of the distance graphical objects on the example of electrical circuits using information criterion]. *Otkrytoe i Distantsionnoe Obrazovanie* [Open and Distance Education] 2:35–43.

5. Kuznetsov, L. A., and D. A. Bugakov. 2013. Razrabotka mery otsenki informatsionnogo rasstoyaniya mezhdu graficheskimi ob"ektami [Development of measures assessing the information distance between graphic objects]. *Informatsionno-Upravlyayushchie Sistemy* [Information and Control Systems] 1:74–79.
6. Gnedenko, B. V. 2007. *Kurs teorii veroyatnostey* [Course of probability theory]. Moscow: LKI Publs. 448 p.
7. GOST 2.743-91 ESKD. Oboznacheniya uslovnye graficheskie v skhemakh. Elementy tsifrovoy tekhniki [State Standard 2.743-91 ESKD. Graphic symbols in schemes. Elements of digital technology]. M.: Gosstandart, 1991. 75 p.
8. Kuznetsov, L. A., and V. F. Kuznetsova. 2013. Otsenka semanticheskoy adekvatnosti tekstov informatsionnym metodom [Evaluation of the semantic adequacy of texts by information method]. *Informatika i ee Primeneniya — Inform. Appl.* 7(1):19–29.
9. Gasparyan, A. V., and A. A. Kirakosyan. 2006. Sistema sravneniya otpechatkov pal'tsev po lokal'nym priznakam [Fingerprint comparisons on local characteristics]. *Vestnik RAU. Ser. Fiziko-Matematicheskie i Estestvennye Nauki* [Herald of RAU. Physics, Mathematics, and Natural Sciences ser.] 2:85–91.
10. Swain, M. J., and D. H. Ballard. 1991. Color indexing. *Int. J. Computer Vision* 7(1):11–32.
11. Sticker, M., and M. Orengo. 1995. Similarity of color images. *SPIE Conference Proceedings* 2420:381–392.
12. Kuznetsov, L. A., and D. A. Bugakov. 2013. Razvitie metoda sravneniya i klassifikatsii graficheskikh ob"ektov [Development of the method of comparison and classification]. *Vestnik Komp'yuternykh i Informatsionnykh Tekhnologiy* [Computer and Information Bulletin Technology] 2(104):11–16.
13. Shennon, K. 1948. A mathematical theory of communication. Pt. I, II. *Bell. Syst. Techn. J.* 27(3):379–423; 27(4):623–656.
14. Kuznetsov, L. A., V. F. Kuznetsova, and A. V. Kapnin. 2013. Universal'nyy metricheskiy tezaurus russkogo yazyka [Universal Russian language thesaurus metric]. *Informatika i ee Primeneniya — Inform. Appl.* 7(3):27–35.

Received December 10, 2013

Contributor

Kuznetsov Leonid A. (b. 1942) — Doctor of Science in technology, professor, Honored Scientist of Russian Federation, Head of Department, Russian Presidential Academy of National Economy and Public Administration (Lipetsk Branch), 3 Internatsional'naya Str., Lipetskaya oblast, Lipetsk 398050, Russian Federation; Kuznetsov.Leonid48@gmail.com

ОБ АВТОРАХ

Адигеев Михаил Георгиевич (р. 1973) — кандидат технических наук, доцент Южного федерального университета

Бенинг Владимир Евгеньевич (р. 1954) — доктор физико-математических наук, профессор кафедры математической статистики факультета вычислительной математики и кибернетики Московского государственного университета им. М. В. Ломоносова; старший научный сотрудник Института проблем информатики Российской академии наук

Босов Алексей Вячеславович (р. 1969) — доктор технических наук, заведующий сектором Института проблем информатики Российской академии наук

Бунтман Надежда Валентиновна (р. 1957) — кандидат филологических наук, доцент факультета иностранных языков и регионоведения Московского государственного университета им. М. В. Ломоносова

Драницына Маргарита Александровна (р. 1983) — аспирант кафедры математической статистики факультета вычислительной математики и кибернетики Московского государственного университета им. М. В. Ломоносова

Жаворонкова Юлия Вадимовна (р. 1990) — программист-разработчик, ООО «КМ Медиа»

Зализняк Анна Андреевна (р. 1959) — доктор филологических наук, ведущий научный сотрудник Института языкознания Российской академии наук и Института проблем информатики Российской академии наук

Захарова Татьяна Валерьевна (р. 1962) — кандидат физико-математических наук, старший преподаватель кафедры математической статистики факультета вычислительной математики и кибернетики Московского государственного университета им. М. В. Ломоносова

Зацман Игорь Моисеевич (р. 1952) — доктор технических наук, заведующий отделом Института проблем информатики Российской академии наук

Кантор Ольга Геннадиевна (р. 1971) — кандидат физико-математических наук, доцент, старший научный сотрудник Института социально-экономических исследований Уфимского научного центра Российской академии наук (ИСЭИ УНЦ РАН)

Карпов Петр Игоревич (р. 1990) — аспирант кафедры теоретической физики и квантовых технологий Института новых материалов нанотехнологий Национального исследовательского технологического университета «МИСиС»

Кружков Михаил Григорьевич (р. 1975) — ведущий программист Института проблем информатики Российской академии наук

Кудрявцев Алексей Андреевич (р. 1978) — кандидат физико-математических наук, доцент кафедры математической статистики факультета вычислительной математики и кибернетики Московского государственного университета им. М. В. Ломоносова

Кузнецов Леонид Александрович (р. 1942) — доктор технических наук, профессор, заслуженный деятель науки РФ, заведующий кафедрой Российской академии народного хозяйства и государственной службы при Президенте РФ (Липецкий филиал)

Лощилова Елена Юрьевна (р. 1960) — научный сотрудник Института проблем информатики Российской академии наук

Лукашенко Олег Викторович (р. 1986) — кандидат физико-математических наук, младший научный сотрудник Института прикладных математических исследований Карельского научного центра Российской академии наук; преподаватель Петрозаводского государственного университета

Мацкевич Андрей Георгиевич (р. 1953) — старший научный сотрудник Института проблем информатики Российской академии наук; доцент Московского технического университета связи и информатики (МТУСИ)

Миронов Андрей Михайлович (р. 1966) — кандидат физико-математических наук, старший науч-

ный сотрудник Института проблем информатики Российской академии наук

Морозов Евсей Викторович (р. 1947) — доктор физико-математических наук, профессор, ведущий научный сотрудник Института прикладных математических исследований Карельского научного центра Российской академии наук; профессор Петрозаводского государственного университета

Мотренко Анастасия Петровна (р. 1992) — студентка Московского физико-технического института

Пагано Микеле (р. 1968) — доктор наук (PhD) по электронике, доцент Университета г. Пиза, Италия

Печинкин Александр Владимирович (р. 1946) — доктор физико-математических наук, профессор, главный научный сотрудник Института проблем информатики Российской академии наук, профессор Российского университета дружбы народов

Разумчик Ростислав Валерьевич (р. 1984) — кандидат физико-математических наук, старший научный сотрудник Института проблем информатики Российской академии наук

Синицын Игорь Николаевич (р. 1940) — доктор технических наук, профессор, заслуженный деятель науки РФ, заведующий отделом Института проблем информатики Российской академии наук

Сичинава Дмитрий Владимирович (р. 1980) — кандидат филологических наук, старший научный сотрудник Института русского языка Российской академии наук

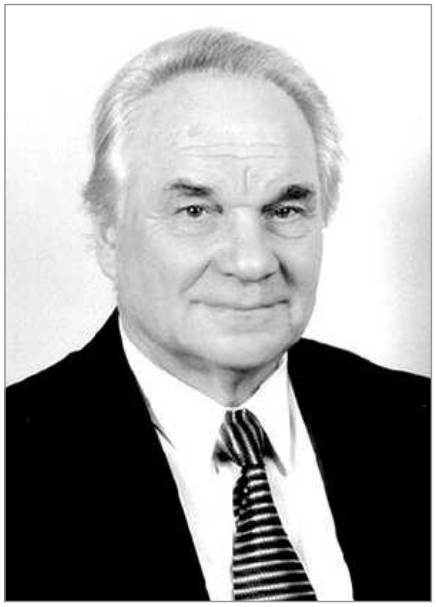
Спивак Семен Израилевич (р. 1945) — доктор физико-математических наук, профессор, заведующий кафедрой Башкирского государственного университета (БашГУ)

Стрижов Вадим Викторович (р. 1967) — кандидат физико-математических наук, доцент, научный сотрудник Вычислительного центра им. А. А. Дородницына Российской академии наук

Шоргин Сергей Яковлевич (р. 1952) — доктор физико-математических наук, профессор, заместитель директора Института проблем информатики Российской академии наук

ЮБИЛЕИ

К 85-летию научного руководителя Федерального государственного бюджетного учреждения науки Института системного анализа Российской академии наук, главного редактора журнала «Информатика и её применения» академика Российской академии наук С. В. Емельянова



18 мая 2014 года исполнилось 85 лет академику РАН С. В. Емельянову.

Станислав Васильевич Емельянов родился 18 мая 1929 г. в г. Воронеже. Учился в Московском авиационном институте на факультете приборостроения и систем управления летательных аппаратов (1947–1952), затем (1953–1957) — в аспирантуре (без отрыва от производства) при Институте автоматики и телемеханики АН СССР (ныне Институт проблем управления РАН).

В 1952 г. С. В. Емельянов поступил на работу в Институт автоматики и телемеханики, где прошел путь от инженера до заместителя директора по науке. С 1976 г. работает в Институте системных исследований АН СССР (в настоящее время Институт системного анализа РАН — ИСА РАН). С 1993 по 2003 гг. С. В. Емельянов — директор ИСА РАН; в настоящее время — научный руководитель ИСА РАН.

Член-корреспондент АН СССР (1970), действительный член РАН (1984), академик-секретарь Отделения информатики, вычислительной техники и автоматизации РАН (1992–2002). В настоящее время — заместитель академика-секретаря Отделения нанотехнологий и информационных технологий (ОНИТ РАН), руководитель секции информационных технологий и автоматизации.

Основные научные результаты С. В. Емельянова относятся к теории систем переменной структуры; теории бинарного управления и новых типов обратной связи; глобальной управляемости и стабилизации нелинейных систем; технологии системного моделирования и системного проектирования средств автоматизации; геометрическим методам анализа нелинейных систем; робастной устойчивости и стабилизации неопределенных систем.

С. В. Емельянов — основатель известной научной школы. Он подготовил более 30 докторов и 70 кандидатов наук; среди его учеников — академики и члены-корреспонденты РАН, члены других академий, руководители институтов, фирм.

Является автором 25 книг и свыше 278 статей. Получил 72 патента на изобретения.

Заведующий кафедрой нелинейных динамических систем и процессов управления факультета вычислительной математики и кибернетики МГУ (с 1989 г.). Почетный профессор МГУ (1998), заслуженный профессор МГУ (1999).

Лауреат Ленинской премии (1972), Государственной премии СССР (1980), премии Совета министров СССР (1981), Государственной премии Российской Федерации (1994), Премии Президиума РАН им. акад. А. А. Андронова (2000), лауреат Ломоносовской премии МГУ по науке I степени (2002), Премии Правительства РФ в области науки и технологий (2009), Премии Правительства РФ в области образования (2012).

С. В. Емельянов награжден орденами Октябрьской Революции (1974), Дружбы народов (1979), «За заслуги перед Отечеством» III степени (1999, 2004), Почета (2010), а также орденами Кирилла и Мефодия (Болгария), «За заслуги» (Польша).

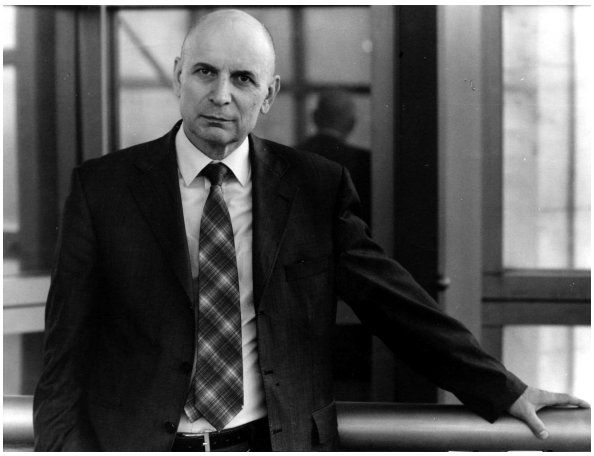
С. В. Емельянов является главным редактором журнала РАН «Информатика и ее применения», осуществляя общее руководство выработкой редакционной политики и процессом издания нашего журнала.

Он является также главным редактором журналов РАН «Информационные технологии и вычислительные системы», «Искусственный интеллект и принятие решений» и членом редакционных коллегий журналов РАН «Автоматика и телемеханика», «Дифференциальные уравнения», «Доклады РАН».

С. В. Емельянов — Председатель Совета по математике при Министерстве образования РФ; является членом ученых и специализированных советов в МГУ, ИСА РАН и ИПИ РАН.

Редакционный совет и Редакционная коллегия журнала «Информатика и ее применения» сердечно поздравляют Станислава Васильевича Емельянова с юбилеем и желают ему крепкого здоровья и новых научных достижений.

К 60-летию директора Федерального государственного бюджетного учреждения науки Института проблем информатики Российской академии наук, заместителя главного редактора журнала «Информатика и её применения» академика Российской академии наук И. А. Соколова



27 марта 2014 года исполнилось 60 лет академику РАН И. А. Соколову.

Игорь Анатольевич Соколов — известный ученый в области теоретической и прикладной информатики, основатель научной школы в области информационных технологий для распределенных автоматизированных информационно-управляющих систем.

И. А. Соколов окончил Московский государственный университет им. М. В. Ломоносова (факультет вычислительной математики и кибернетики) в 1976 г., аспирантуру там же — в 1979 г., работал в НИИ систем связи и управления ЦНПО «Каскад», с 1992 г. работает в Институте проблем информатики Российской академии наук (ИПИ РАН), с 1999 г. — директор ИПИ РАН.

В 2003 г. избран членом-корреспондентом РАН, в 2008 г. — академиком РАН.

В июне 2013 г. избран главным ученым секретарем Президиума РАН.

И. А. Соколов опубликовал более 150 научных трудов, в том числе 7 монографий, он является автором 23 зарегистрированных изобретений и программ.

Основные научные результаты И. А. Соколова связаны с разработкой инструментальных комплексов программных средств анализа и расчета вероятностно-временны́х характеристик систем в рамках моделей с дискретным и непрерывным временем, обоснованием и разработкой принципов построения и системотехнических решений по архитектуре крупномасштабных информационных систем двойного применения, базовым информационным и телекоммуникационным технологиям, обеспечению информационной безопасности.

Научные результаты И. А. Соколова позволили разработать, под его руководством и при его участии, специализированные информационные технологии, аппаратные и программные средства, комплексы, на основе которых создан ряд информационных систем национального масштаба.

В качестве Генерального конструктора руководит разработкой и развитием системы информационного обеспечения управления государством, автоматизированной системы управления и информационного обеспечения принятия управлен-

ческих решений органов безопасности и системы распределенных ситуационных центров, работающих по единому регламенту взаимодействия. Член Научного совета при Совете Безопасности РФ, член президиума Научно-технического совета ВПК, председатель Совета РАН по исследованиям в области обороны. Председатель диссертационных советов в ИПИ РАН и в НИИ АА. Научные достижения И. А. Соколова в области создания систем информационного обеспечения безопасности мегаполиса отмечены Премией правительства РФ (2004 г.). Награжден ведомственными наградами Совета Безопасности РФ и ГУСП Президента РФ.

Является заведующим кафедрой информационной безопасности факультета Вычислительной математики и кибернетики МГУ им. М. В. Ломоносова и кафедрой проблем информатики МИРЭА.

И. А. Соколов уделяет большое внимание организационной работе по редактированию и изданию научных журналов. Он является заместителем главного редактора журнала РАН «Информатика и её применения» и на этом посту выполняет основную текущую работу по отбору статей в журнал, организации их редактирования и публикации. И. А. Соколов является также главным редактором журнала РАН «Системы и средства информатики», членом редакционной коллегии журналов «Информационные технологии и вычислительные системы», «Системы высокой доступности», «Право и Кибербезопасность», членом редакционного совета журнала «Проблемы информатики».

Редакционный совет и Редакционная коллегия журнала «Информатика и её применения» сердечно поздравляют Игоря Анатольевича Соколова с 60-летием и желают крепкого здоровья и новых научных достижений.

Правила подготовки рукописей для публикации в журнале «Информатика и её применения»

Журнал «Информатика и её применения» публикует теоретические, обзорные и дискуссионные статьи, посвященные научным исследованиям и разработкам в области информатики и ее приложений.

Журнал издается на русском языке. По специальному решению редколлегии отдельные статьи могут печататься на английском языке.

Тематика журнала охватывает следующие направления:

- теоретические основы информатики;
- математические методы исследования сложных систем и процессов;
- информационные системы и сети;
- информационные технологии;
- архитектура и программное обеспечение вычислительных комплексов и сетей.

1. В журнале печатаются статьи, содержащие результаты, ранее не опубликованные и не предназначенные к одновременной публикации в других изданиях.

   Публикация не должна нарушать закон об авторских правах.

   Направляя рукопись в редакцию, авторы сохраняют все права собственников данной рукописи и при этом передают учредителям и редколлегии неисключительные права на издание статьи на русском языке (или на языке статьи, если он отличен от русского) и на ее распространение в России и за рубежом. Авторы должны представить в редакцию письмо в следующей форме:

   Соглашение о передаче права на публикацию:

   «*Мы, нижеподписавшиеся, авторы рукописи «. . . », передаем учредителям и редколлегии журнала «Информатика и её применения» неисключительное право опубликовать данную рукопись статьи на русском языке как в печатной, так и в электронной версиях журнала. Мы подтверждаем, что данная публикация не нарушает авторского права других лиц или организаций.*

   Подписи авторов: (ф. и. о., дата, адрес)».

   Это соглашение может быть представлено в бумажном виде или в виде отсканированной копии (с подписями авторов).

   Редколлегия вправе запросить у авторов экспертное заключение о возможности публикации представленной статьи в открытой печати.

2. К статье прилагаются данные автора (авторов) (см. п. 8). При наличии нескольких авторов указывается фамилия автора, ответственного за переписку с редакцией.

3. Редакция журнала осуществляет экспертизу присланных статей в соответствии с принятой в журнале процедурой рецензирования.

   Возвращение рукописи на доработку не означает ее принятия к печати.

   Доработанный вариант с ответом на замечания рецензента необходимо прислать в редакцию.

4. Решение редколлегии о публикации статьи или ее отклонении сообщается авторам.

   Редколлегия может также направить авторам текст рецензии на их статью. Дискуссия по поводу отклоненных статей не ведется.

5. Редактура статей высылается авторам для просмотра. Замечания к редактуре должны быть присланы авторами в кратчайшие сроки.

6. Рукопись предоставляется в электронном виде в форматах MS WORD (.doc или .docx) или LaTeX (.tex), дополнительно — в формате .pdf, на дискете, лазерном диске или электронной почтой. Предоставление бумажной рукописи необязательно.

7. При подготовке рукописи в MS Word рекомендуется использовать следующие настройки.

   Параметры страницы: формат — A4; ориентация — книжная; поля (см): внутри — 2,5, снаружи — 1,5, сверху — 2, снизу — 2, от края до нижнего колонтитула — 1,3.

   Основной текст: стиль — «Обычный», шрифт — Times New Roman, размер — 14 пунктов, абзацный отступ — 0,5 см, 1,5 интервала, выравнивание — по ширине.

   Рекомендуемый объем рукописи — не свыше 20 страниц указанного формата.

Сокращения слов, помимо стандартных, не допускаются. Допускается минимальное количество аббревиатур.

Все страницы рукописи нумеруются.

Шаблоны примеров оформления представлены в Интернете: http://www.ipiran.ru/journal/template.doc

8. Статья должна содержать следующую информацию на ***русском и английском языках***:

   – название статьи;
   – Ф.И.О. авторов, на английском можно только имя и фамилию;
   – место работы, с указанием почтового адреса организации и электронного адреса каждого автора;
   – сведения об авторах, в соответствии с форматом, образцы которого представлены на страницах: http://www.ipiran.ru/journal/issues/2013_07_01_rus/authors.asp и http://www.ipiran.ru/journal/issues/2013_07_01_eng/authors.asp;
   – аннотация (не менее 100 слов на каждом из языков). Аннотация — это краткое резюме работы, которое может публиковаться отдельно. Она является основным источником информации в информационных системах и базах данных. Английская аннотация должна быть оригинальной, может не быть дословным переводом русского текста и должна быть написана хорошим английским языком. В аннотации не должно быть ссылок на литературу и, по возможности, формул;
   – ключевые слова — желательно из принятых в мировой научно-технической литературе тематических тезаурусов. Предложения не могут быть ключевыми словами.

9. Требования к спискам литературы.

   Ссылки на литературу в тексте статьи нумеруются (в квадратных скобках) и располагаются в каждом из списков литературы в порядке первых упоминаний.

   Списки литературы представляются в двух вариантах:

   (1) **Список литературы к русскоязычной части**. Русские и английские работы — на языке и в алфавите оригинала;
   (2) **References**. Русские работы и работы на других языках — в латинской транслитерации с переводом на английский язык; английские работы и работы на других языках — на языке оригинала.

   Необходимо для составления списка “References” пользоваться размещенной на сайте http://translit.ru/ бесплатной программой транслитерации русского текста в латиницу, при этом в закладке «варианты. . . » следует выбрать опцию BGN.

   Список литературы “References” приводится полностью отдельным блоком, повторяя все позиции из списка литературы к русскоязычной части, независимо от того, имеются или нет в нем иностранные источники. Если в списке литературы к русскоязычной части есть ссылки на иностранные публикации, набранные латиницей, они полностью повторяются в списке “References”.

   Ниже приведены примеры ссылок на различные виды публикаций в списке “References”.

   Описание статьи из журнала:

   Zagurenko, A. G., V. A. Korotovskikh, A. A. Kolesnikov, A. V. Timonov, and D. V. Kardymon. 2008. Tekhniko-ekonomicheskaya optimizatsiya dizayna gidrorazryva plasta [Technical and economic optimization of the design of hydraulic fracturing]. *Neftyanoe hozyaystvo* [*Oil Industry*] 11:54–57.

   Zhang, Z., and D. Zhu. 2008. Experimental research on the localized electrochemical micromachining. *Rus. J. Electrochem.* 44(8):926–930. doi:10.1134/S1023193508080077.

   Описание статьи из электронного журнала:

   Swaminathan, V., E. Lepkoswka-White, and B. P. Rao. 1999. Browsers or buyers in cyberspace? An investigation of electronic factors influencing electronic exchange. *JCMC* 5(2). Available at: http://www.ascusc.org/jcmc/vol5/issue2/ (accessed April 28, 2011).

   Описание статьи из продолжающегося издания (сборника трудов):

   Astakhov, M. V., and T. V. Tagantsev. 2006. Eksperimental’noe issledovanie prochnosti soedineniy “stal’–kompozit” [Experimental study of the strength of joints “steel–composite”]. *Trudy MGTU “Matematicheskoe modelirovanie slozhnykh tekhnicheskikh sistem”* [*Bauman MSTU “Mathematical Modeling of Complex Technical Systems” Proceedings*]. 593:125–130.

Описание материалов конференций:

Usmanov, T. S., A. A. Gusmanov, I. Z. Mullagalin, R. Ju. Muhametshina, A. N. Chervyakova, and A. V. Sveshnikov. 2007. Osobennosti proektirovaniya razrabotki mestorozhdeniy s primeneniem gidrorazryva plasta [Features of the design of field development with the use of hydraulic fracturing]. *Trudy 6-go Mezhdunarodnogo Simpoziuma "Novye resursosberegayushchie tekhnologii nedropol'zovaniya i povysheniya neftegazootdachi"* [*6th Symposium (International) "New Energy Saving Subsoil Technologies and the Increasing of the Oil and Gas Impact" Proceedings*]. Moscow. 267–272.

Описание книги (монографии, сборники):

Lindorf, L. S., and L. G. Mamikoniants, eds. 1972. *Ekspluatatsiya turbogeneratorov s neposredstvennym okhlazhdeniem* [*Operation of turbine generators with direct cooling*]. Moscow: Energy Publs. 352 p.

Latyshev, V. N. 2009. *Tribologiya rezaniya. Kn. 1: Friktsionnye protsessy pri rezanii metallov* [*Tribology of cutting. Vol. 1: Frictional processes in metal cutting*]. Ivanovo: Ivanovskii State Univ. 108 p.

Описание переводной книги (в списке литературы к русскоязычной части необходимо указать: / Пер. с англ. — после названия книги, а в конце ссылки указать оригинал книги в круглых скобках):

1. В русскоязычной части:

   Тимошенко С. П., Янг Д. Х., Уивер У. Колебания в инженерном деле / Пер. с англ. — М.: Машиностроение, 1985. 472 с. (*Timoshenko S. P., Young D. H., Weaver W.* Vibration problems in engineering. — 4th ed. — N.Y.: Wiley, 1974. 521 p.)

2. В англоязычной части:

   Timoshenko, S. P., D. H. Young, and W. Weaver. 1974. *Vibration problems in engineering*. 4th ed. N.Y.: Wiley. 521 p.

Описание неопубликованного документа:

Latypov, A. R., M. M. Khasanov, and V. A. Baikov. 2004. Geology and production (NGT GiD). Certificate on official registration of the computer program No. 2004611198. (In Russian, unpubl.)

Описание интернет-ресурса:

Pravila tsitirovaniya istochnikov [Rules for the citing of sources]. Available at: http://www.scribd.com/doc/1034528/ (accessed February 7, 2011).

Описание диссертации или автореферата диссертации:

Semenov, V. I. 2003. Matematicheskoe modelirovanie plazmy v sisteme kompaktnyy tor [Mathematical modeling of the plasma in the compact torus]. D.Sc. Diss. Moscow. 272 p.

Kozhunova, O. S. 2009. Tekhnologiya razrabotki semanticheskogo slovarya informatsionnogo monitoringa [Technology of development of semantic dictionary of information monitoring system]. PhD Thesis. Moscow: IPI RAN. 23 p.

Описание ГОСТа:

GOST 8.586.5-2005. 2007. Metodika vypolneniya izmereniy. Izmerenie raskhoda i kolichestva zhidkostey i gazov s pomoshch'yu standartnykh suzhayushchikh ustroystv [Method of measurement. Measurement of flow rate and volume of liquids and gases by means of orifice devices]. Moscow: Standardinform Publs. 10 p.

Описание патента:

Bolshakov, M. V., A. V. Kulakov, A. N. Lavrenov, and M. V. Palkin. 2006. Sposob orientirovaniya po krenu letatel'nogo apparata s opticheskoy golovkoy samonavedeniya [The way to orient on the roll of aircraft with optical homing head]. Patent RF No. 2280590.

10. Присланные в редакцию материалы авторам не возвращаются.
11. При отправке файлов по электронной почте просим придерживаться следующих правил:
    - указывать в поле subject (тема) название журнала и фамилию автора;
    - использовать attach (присоединение);
    - в состав электронной версии статьи должны входить: файл, содержащий текст статьи, и файл(ы), содержащий(е) иллюстрации.
12. Журнал «Информатика и её применения» является некоммерческим изданием. Плата за публикацию не взимается, гонорар авторам не выплачивается.

Адрес редакции журнала «Информатика и её применения»:
Москва 119333, ул. Вавилова, д. 44, корп. 2, ИПИ РАН
Тел.: +7 (499) 135-86-92 Факс: +7 (495) 930-45-05
e-mail: rust@ipiran.ru (Сейфуль-Мулюков Рустем Бадриевич)
http://www.ipiran.ru/journal/issues/

Requirements for manuscripts submitted to Journal "Informatics and Applications"

Journal "Informatics and Applications" (Inform. Appl.) publishes theoretical, review, and discussion articles on the research and development in the field of informatics and its applications.

The journal is published in Russian. By a special decision of the editorial board, some articles can be published in English.

The topics covered include the following areas:

- theoretical fundamentals of informatics;
- mathematical methods for studying complex systems and processes;
- information systems and networks;
- information technologies; and
- architecture and software of computational complexes and networks.

1. The Journal publishes original articles which have not been published before and are not intended for publication in other editions. An article submitted to the Journal must not violate the Copyright law. Sending the manuscript to the Editorial Board, the authors retain all rights of the owners of the manuscript and transfer the nonexclusive rights to publish the article in Russian (or the language of the article, if not Russian) and its distribution in Russia and abroad to the Founders and the Editorial Board. Authors should submit a letter to the Editorial Board in the following form:

   Agreement on the transfer of rights to publish:

   "*We, the undersigned authors of the manuscript ". . . ", pass to the Founder and the Editorial Board of the Journal "Informatics and Applications" the nonexclusive right to publish the manuscript of the article in Russian (or in English) in both print and electronic versions of the Journal. We affirm that this publication does not violate the Copyright of other persons or organizations.*

   Author(s) signature(s): (name(s), address(es), date).

   This agreement should be submitted in paper form or in the form of a scanned copy (signed by the authors).
2. A submitted article should be attached with **the data on the author(s)** (see item 8). If there are several authors, the contact person should be indicated who is responsible for correspondence with the Editorial Board and other authors about revisions and final approval of the proofs.
3. The Editorial Board of the Journal examines the article according to the established reviewing procedure. If the authors receive their article for correction after reviewing, it does not mean that the article is approved for publication. The corrected article should be sent to the Editorial Board for the subsequent review and approval.
4. The decision on the article publication or its rejection is communicated to the authors. The Editorial Board may also send the reviews on the submitted articles to the authors. Any discussion upon the rejected articles is not possible.
5. The edited articles will be sent to the authors for proofread. The comments of the authors to the edited text of the article should be sent to the Editorial Board as soon as possible.
6. The manuscript of the article should be presented electronically in the MS WORD (.doc or .docx) or LaTeX (.tex) formats, and additionally in the .pdf format. All documents may be sent by e-mail or provided on a CD or diskette. A hard copy submission is not necessary.
7. The recommended typesetting instructions for manuscript.

   Pages parameters: format A4, portrait orientation, document margins (cm): left — 2.5, right — 1.5, above — 2.0, below — 2.0, footer 1.3.

   Text: font —Times New Roman, font size — 14, paragraph indent — 0.5, line spacing — 1.5, justified alignment.

   The recommended manuscript size: not more than 20 pages of the specified format.

   Use only standard abbreviations. Avoid abbreviations in the title and abstract. The full term for which an abbreviation stands should precede its first use in the text unless it is a standard unit of measurement.

   All pages of the manuscript should be numbered.

   The templates for the manuscript typesetting are presented on site: http://www.ipiran.ru/journal/template.doc.
8. The articles should enclose data both in **Russian and English**:

   - title;
   - author's name and surname;
   - affiliation — organization, its address with ZIP code, city, country, and official e-mail address;
   - data on authors according to the format: (see site)
     http://www.ipiran.ru/journal/issues/2013_07_01/authors.asp and
     http://www.ipiran.ru/journal/issues/2013_07_01_eng/authors.asp;
   - abstract (not less than 100 words) both in Russian and in English. Abstract is a short summary of the article that can be published separately. The abstract is the main source of information on the article and it could be included in leading information systems and data bases. The abstract in English has to be an original text and should not be an exact translation of the Russian one. Good English is required. In abstracts, avoid references and formulae;

– indexing is performed on the basis of keywords. The use of keywords from the internationally accepted thematic Thesauri is recommended.

Important! Keywords must not be sentences.

9. References. Russian references have to be presented both in English translation and Latin transliteration (refer http://www.translit.ru, option BGN).

Please take into account the following examples of Russian references appearance:

Article in journal:

Zhang, Z., and D. Zhu. 2008. Experimental research on the localized electrochemical micromachining. *Rus. J. Electrochem.* 44(8):926–930. doi:10.1134/S1023193508080077.

Journal article in electronic format:

Swaminathan, V., E. Lepkoswka-White, and B. P. Rao. 1999. Browsers or buyers in cyberspace? An investigation of electronic factors influencing electronic exchange. *JCMC* 5(2). Available at: http://www.ascusc.org/jcmc/vol5/issue2/ (accessed April 28, 2011).

Article from the continuing publication (collection of works, proceedings):

Astakhov, M. V., and T. V. Tagantsev. 2006. Eksperimental'noe issledovanie prochnosti soedineniy "stal'–kompozit" [Experimental study of the strength of joints "steel–composite"]. *Trudy MGTU "Matematicheskoe modelirovanie slozhnykh tekhnicheskikh sistem"* [*Bauman MSTU "Mathematical Modeling of Complex Technical Systems" Proceedings*]. 593:125–130.

Conference proceedings:

Usmanov, T. S., A. A. Gusmanov, I. Z. Mullagalin, R. Ju. Muhametshina, A. N. Chervyakova, and A. V. Sveshnikov. 2007. Osobennosti proektirovaniya razrabotki mestorozhdeniy s primeneniem gidrorazryva plasta [Features of the design of field development with the use of hydraulic fracturing]. *Trudy 6-go Mezhdunarodnogo Simpoziuma "Novye resursosberegayushchie tekhnologii nedropol'zovaniya i povysheniya neftegazootdachi"* [*6th Symposium (International) "New Energy Saving Subsoil Technologies and the Increasing of the Oil and Gas Impact" Proceedings*]. Moscow. 267–272.

Books and other monographs:

Lindorf, L. S., and L. G. Mamikoniants, eds. 1972. *Ekspluatatsiya turbogeneratorov s neposredstvennym okhlazhdeniem* [*Operation of turbine generators with direct cooling*]. Moscow: Energy Publs. 352 p.

Dissertation and Thesis:

Kozhunova, O. S. 2009. Tekhnologiya razrabotki semanticheskogo slovarya informatsionnogo monitoringa [Technology of development of semantic dictionary of information monitoring system]. PhD Thesis. Moscow: IPI RAN. 23 p.

State standards and patents:

GOST 8.586.5-2005. 2007. Metodika vypolneniya izmereniy. Izmerenie raskhoda i kolichestva zhidkostey i gazov s pomoshch'yu standartnykh suzhayushchikh ustroystv [Method of measurement. Measurement of flow rate and volume of liquids and gases by means of orifice devices]. M.: Standardinform Publs. 10 p.

Bolshakov, M. V., A. V. Kulakov, A. N. Lavrenov, and M. V. Palkin. 2006. Sposob orientirovaniya po krenu letatel'nogo apparata s opticheskoy golovkoy samonavedeniya [The way to orient on the roll of aircraft with optical homing head]. Patent RF No. 2280590.

References in Latin transcription are presented in the original language.

References in the text are numbered according to the order of their first appearance; the number is placed in square brackets. All items from the reference list should be cited.

10. Manuscripts and additional materials are not returned to Authors by the Editorial Board.
11. Submissions of files by e-mail must include:
    – the journal title and author's name in the "Subject" field;
    – an article and additional materials have to be attached using the "attach" function;
    – an electronic version of the article should contain the file with the text and a separate file with figures.
12. "Informatics and Applications" journal is not a profit publication. There are no charges for the authors as well as there are no royalties.

Editorial Board address:
IPI RAN, Vavilova Str., 44, block 2, Moscow 119333, Russia
Ph.: +7 (499) 135 86 92, Fax: +7 (495) 930 45 05
e-mail: rust@ipiran.ru (to Prof. Rustem Seyful-Mulyukov)
http://www.ipiran.ru/english/journal.asp