Информатика и её применения

Том 8 Выпуск 4 Год 2014

СОДЕРЖАНИЕ

Информатика и её применения

Том 8 Выпуск 4 Год 2014

СОДЕРЖАНИЕ

Технический редактор *Л. Кокушкина* Художественный редактор *М. Седакова*
Сдано в набор 01.11.14. Подписано в печать 11.12.14. Формат 60 x 84 / 8
Бумага офсетная. Печать цифровая. Усл.-печ. л. 19,0. Уч.-изд. л. 16,0. Тираж 100 экз.
Заказ № 2330
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. 4. С. 3–10

СОВМЕСТНОЕ СТАЦИОНАРНОЕ РАСПРЕДЕЛЕНИЕ ЧИСЛА ЗАЯВОК В НАКОПИТЕЛЕ И В БУНКЕРЕ ПЕРЕУПОРЯДОЧЕНИЯ В МНОГОКАНАЛЬНОЙ СИСТЕМЕ ОБСЛУЖИВАНИЯ С ПЕРЕУПОРЯДОЧЕНИЕМ ЗАЯВОК*

А. В.Печинкин[1], Р. В. Разумчик[2]

Аннотация: Рассматривается функционирующая в непрерывном времени многоканальная система обслуживания с накопителем бесконечной емкости и переупорядочением заявок. В систему поступает пуассоновский поток заявок, время обслуживания каждым прибором распределено по экспоненциальному закону с одним и тем же параметром. При поступлении в систему всем заявкам присваивается порядковый номер. На выходе из системы сохраняется порядок между заявками, установленный при входе в нее. Заявки, завершившие обслуживание и нарушившие установленный порядок, накапливаются на выходе системы в бункере переупорядочения (БП), который также имеет неограниченную емкость. Найдено совместное стационарное распределение числа заявок в накопителе и суммарного числа заявок в БП в терминах вычислительных алгоритмов и производящих функций (ПФ). Приведены примеры расчетов по полученным соотношениям.

Ключевые слова: многолинейная система массового обслуживания; переупорядочение; стационарное распределение числа заявок

DOI: 10.14357/19922264140401

1 Введение

Для функционирования ряда информационно-телекоммуникационных систем и для предоставления на их основе услуг необходимо соблюдение требования сохранения порядка в потоке передаваемых сообщений. Различные действия, необходимые для этого, можно объединить в одно понятие — переупорядочение. Для изучения влияния переупорядочения на качество функционирования информационно-телекоммуникационных систем к настоящему времени предложено множество моделей, которые в своей основе используют методы и модели теории массового обслуживания. Исследуемая система обычно представляется в виде системы или сети массового обслуживания с одним или несколькими входящими потоками сообщений. Эффект переупорядочения часто моделируется с помощью дополнительной очереди (БП), в которую попадают сообщения, обработанные в системе, и ожидают там до тех пор, пока порядок следования сообщений нельзя будет восстановить. Некоторый обзор работ в этом направлении можно найти в [1, 2], а некоторые последние результаты — в [3–8].

Настоящая работа является развитием [9], в которой рассматривается система массового обслуживания (СМО) с переупорядочением в виде марковской многоканальной системы обслуживания неограниченной емкости и бункером переупорядочения, также имеющим неограниченную емкость. В [9] была получена система уравнений равновесия для совместного стационарного распределения числа заявок в системе и бункере переупорядочения и приведены некоторые результаты численных расчетов. Однако несомненный интерес представляют две задачи, не освещенные в [9], которые и являются предметом данной статьи, а именно: разработка рекуррентного алгоритма расчета вышеупомянутого совместного стационарного распределения и нахождение этого распределения в терминах ПФ.

Статья организована таким образом. В разд. 2 приводится подробное описание системы. В разд. 3 дается рекуррентный алгоритм расчета совместного стационарного распределения, а в разд. 4 показано, как совместное стационарное распределение можно найти в терминах ПФ. Примеры расчетов, проведенных по формулам разд. 4, представлены в разд. 5. В заключении сформулированы основные результаты работы.

*Работа выполнена при частичной поддержке РФФИ (проект 13-07-00223).

[1]Институт проблем информатики Российской академии наук

[2]Институт проблем информатики Российской академии наук; Российский университет дружбы народов, rrazumchik@ieee.org

2 Описание системы

Рассмотрим функционирующую в непрерывном времени N-линейную ($N \geq 2$) СМО с накопителем неограниченной емкости, входящим пуассоновским потоком заявок интенсивности λ и экспоненциальным распределением времени обслуживания заявки каждым прибором с параметром μ.

При поступлении в систему всем заявкам присваивается порядковый номер. На выходе из СМО сохраняется порядок между заявками, установленный при входе в нее. Заявки, завершившие обслуживание и нарушившие установленный порядок, накапливаются на выходе системы в БП и покидают СМО только после того, как закончится обслуживание всех заявок с меньшими номерами. Такая СМО носит название системы с переупорядочением заявок.

Предполагается также выполнение необходимого и достаточного условия существования стационарного режима функционирования СМО

$$\tilde{\rho} = \frac{\rho}{N} < 1\,,$$

где $\rho = \lambda/\mu$.

3 Алгоритм нахождения совместного стационарного распределения

Предположим, что на приборах находится n, $n = \overline{1, N}$, заявок. Тогда заявкой первого уровня будем называть ту из них, которая в систему поступила последней, второго уровня — предпоследней, ..., n-го уровня — первой. При этом если $n = N$ (все приборы заняты), то находящиеся в БП заявки, поступившие между заявками второго и первого уровней, будем называть заявками первой очереди, заявки, поступившие между заявками третьего и второго уровней, — заявками второй очереди, ..., заявки, поступившие между заявками N-го и $(N-1)$-го уровней, — заявками $(N-1)$-й очереди. Если же $n < N$, то заявками первой очереди будем называть заявки из БП, поступившие после заявки первого уровня, заявками второй очереди — заявки, поступившие между заявками второго и первого уровней, и т. д.

При $n \geq N$ обозначим через $p_{n;i}^{(m)}$, $m = \overline{1, N-1}$, $i \geq 0$, стационарную вероятность того, что в системе на приборах и в накопителе находится n заявок, а в БП имеется в сумме i заявок первой, второй, ..., m-й очереди. Через $p_{n;i}^{(m)}$, $m = \overline{1, n}$, $i \geq 0$, обозначим аналогичную стационарную вероятность при $n = \overline{1, N-1}$. Через p_n, $n \geq 0$, обозначим стационарную вероятность того, что в системе на приборах и в накопителе (без учета числа заявок в БП) находится n заявок. Очевидно, что стационарные вероятности p_n определяются теми же самыми формулами, что и в обычной марковской СМО $M/M/N/\infty$ (см., например, [10]):

$$p_0 = \left(\sum_{i=0}^{N-1} \frac{\rho^i}{i!} + \frac{\rho^N}{(N-1)!(N-\rho)} \right)^{-1}; \tag{1}$$

$$p_i = \begin{cases} \dfrac{\rho^i}{i!} p_0\,, & i = \overline{1, N}\,, \\ \dfrac{\rho^i}{N!\, N^{i-N}} p_0 = \tilde{\rho}^{\,i-N} p_N\,, & i \geq N+1\,. \end{cases} \tag{2}$$

Наконец, через $p_{n;i}$, $n \geq 1$, $i \geq 0$, обозначим стационарную вероятность того, что в системе на приборах и в накопителе находится n заявок, а в БП — i заявок.

Используя принцип глобального баланса, можно выписать систему уравнений для вероятностей $p_{n;i}^{(m)}$. Для вероятностей $p_{n;i}^{(1)}$, $n \geq N$, $i \geq 0$, справедливы уравнения:

$$p_{n;0}^{(1)}(\lambda + N\mu) = p_{n-1;0}^{(1)}\lambda + p_{n+1}(N-1)\mu\,,\ n \geq N; \tag{3}$$

$$p_{n;i}^{(1)}(\lambda + N\mu) = p_{n-1;i}^{(1)}\lambda + p_{n+1;i-1}^{(1)}\mu\,, \quad n \geq N\,,\ i \geq 1\,. \tag{4}$$

Для вероятностей $p_{N-1;i}^{(1)}$, $i \geq 0$, справедливы уравнения:

$$p_{N-1;0}^{(1)}[\lambda + (N-1)\mu] = p_{N-2}\lambda + p_N(N-1)\mu\,; \tag{5}$$

$$p_{N-1;i}^{(1)}[\lambda + (N-1)\mu] = p_{N;i-1}^{(1)}\mu\,,\ i \geq 1\,. \tag{6}$$

Для вероятностей $p_{n;i}^{(1)}$, $n = \overline{1, N-2}$, $i \geq 0$, справедливы уравнения

$$p_{n;0}^{(1)}(\lambda + n\mu) = p_{n-1}\lambda + p_{n+1;0}^{(1)} n\mu,\ n = \overline{1, N-2}; \tag{7}$$

$$p_{n;i}^{(1)}(\lambda + n\mu) = p_{n+1;i}^{(1)} n\mu + p_{n+1;i-1}^{(2)}\mu\,, \quad n = \overline{1, N-2},\ i \geq 1. \tag{8}$$

Для остальных вероятностей $p_{n;i}^{(m)}$, $m = \overline{2, N-1}$, справедливы формулы:

$$p_{n;0}^{(m)}(\lambda + N\mu) = p_{n-1;0}^{(m)}\lambda + p_{n+1;0}^{(m-1)}(N-m)\mu\,, \quad n \geq N\,; \tag{9}$$

$$p_{n;i}^{(m)}(\lambda + N\mu) = p_{n-1;i}^{(m)}\lambda + p_{n+1;i}^{(m-1)}(N-m)\mu + p_{n+1;i-1}^{(m)} m\mu\,,\ n \geq N\,,\ i \geq 1\,; \tag{10}$$

$$p_{N-1;0}^{(m)}[\lambda + (N-1)\mu] = p_{N-2;0}^{(m-1)}\lambda = + p_{N;0}^{(m-1)}(N-m)\mu\,; \tag{11}$$

$$p_{N-1;i}^{(m)}[\lambda+(N-1)\mu] = p_{N-2;i}^{(m-1)}\lambda + \\ + p_{N;i}^{(m-1)}(N-m)\mu + p_{N;i-1}^{(m)}m\mu\,,\ \ i\geq 1\,; \quad (12)$$

$$p_{n;0}^{(m)}(\lambda+n\mu) = p_{n-1;0}^{(m-1)}\lambda + p_{n+1;0}^{(m)}(n-m+1)\mu\,, \\ n=\overline{m,N-2}\,; \quad (13)$$

$$p_{n;i}^{(m)}(\lambda+n\mu) = p_{n-1;i}^{(m-1)}\lambda + p_{n+1;i}^{(m)}(n-m+1)\mu + \\ + p_{n+1;i-1}^{(m+1)}m\mu\,,\ \ n=\overline{m,N-2}\,,\ \ i\geq 1\,. \quad (14)$$

Решение данной системы уравнений позволяет найти совместное стационарное распределение $p_{n;i}$ числа заявок на приборах и в накопителе и суммарного числа заявок в БП в виде следующих равенств:

$$p_{n;i} = p_{n;i}^{(N-1)}\,,\quad n\geq N\,,\ \ i\geq 0\,,$$
$$p_{n;i} = p_{n;i}^{(n)}\,,\qquad n=\overline{1,N-1}\,,\ \ i\geq 0\,.$$

Анализ системы (3)–(14) показал, что вычисление стационарных вероятностей $p_{n;i}^{(m)}$ можно проводить рекуррентным образом по следующему алгоритму.

Алгоритм 1 (**Алгоритм решения системы уравнений равновесия**).
Задать λ, μ и n.
Для $n\geq 0$ *рассчитать* p_n *по формулам* (1) и (2).
Рассчитать $p_{N-1;0}^{(1)}$ *по формуле* (5).
Для $n\geq N$ *рассчитать* $p_{n;0}^{(1)}$ *по формуле* (3).
Для $i\geq 1$
  рассчитать $p_{N-1;i}^{(1)}$ *по формуле* (6).
  для $n\geq N$ *рассчитать* $p_{n;i}^{(1)}$ *по формуле* (4).
Для $n=\overline{N-2,1}$ *рассчитать* $p_{n;0}^{(1)}$ *по формуле* (7).
Для $m=\overline{2,N-1}$
  рассчитать $p_{N-1;0}^{(m)}$ *по формуле* (11).
  для $n\geq N$ *рассчитать* $p_{n;0}^{(m)}$ *по формуле* (9);
  для $i\geq 1$
    рассчитать $p_{N-m;i}^{(1)}$ *по формуле* (8);
    если $m\neq 2$, *для* $j=\overline{2,m-1}$ *рассчитать* $p_{N-m+j-1;i}^{(j)}$ *по формуле* (14);
    рассчитать $p_{N-1;i}^{(m)}$ *по формуле* (12);
    для $n\geq N$ *рассчитать* $p_{n;i}^{(m)}$ *по формуле* (10);
  если $m\neq N-1$, *для* $m=\overline{N-2,m}$ *рассчитать* $p_{n;0}^{(m)}$ *по формуле* (13).

В связи с тем, что вычисление моментов после расчета вероятностей по представленному алгоритму может давать погрешности, в следующем разделе находятся формулы для совместного стационарного распределения в терминах ПФ.

4 Использование производящих функций

Система уравнений (3)–(14) допускает также решение с помощью ПФ. Для нахождения этого решения положим

$$f_m(u,z) = \lambda u^2 - (\lambda+N\mu)u + m\mu z\,,\ m=\overline{1,N-1}\,.$$

Обозначим через $u_m = u_m(z)$, $m=\overline{1,N-1}$, минимальное решение уравнения

$$f_m(u,z)=0\,,$$

т. е.

$$u_m = \frac{\lambda+N\mu-\sqrt{(\lambda+N\mu)^2-4m\lambda\mu z}}{2\lambda}\,.$$

Введем ПФ

$$P_n^{(m)}(z) = \sum_{i=0}^{\infty} z^i p_{n;i}^{(m)}\,,$$
$$0<z<1\,,\ \ n\geq 1\,,\ \ m=\overline{1,\min(n,N-1)}\,;$$

$$P^{(m)}(u,z) = \sum_{n=N}^{\infty} u^{n-N}P_n^{(m)}(z)\,,$$
$$0<u,z<1\,,\ \ m=\overline{1,N-1}\,,$$

и, кроме того, положим

$$P(u) = \sum_{n=N}^{\infty} u^{n-N}p_n = \frac{1}{1-\tilde{\rho}u}\,p_N\,.$$

Тогда, умножая (3) и (4) на z^i и суммируя по всем i от нуля до бесконечности, получаем:

$$(\lambda+N\mu)P_n^{(1)}(z) = \lambda P_{n-1}^{(1)}(z) + (N-1)\mu p_{n+1} + \\ + \mu z P_{n+1}^{(1)}(z)\,,\ \ n\geq N\,.$$

Умножая последнее выражение на u^{n-N} и суммируя по всем значениям $n\geq N$, после приведения подобных слагаемых имеем:

$$f_1(u,z)P^{(1)}(u,z) = \mu z P_N^{(1)}(z) - \\ - \lambda u P_{N-1}^{(1)}(z) - (N-1)\mu[P(u)-p_N]\,. \quad (15)$$

Теперь умножим (9) и (10) на z^i и просуммируем по всем значениям $i\geq 0$. В результате приходим к выражению:

$$(\lambda+N\mu)P_n^{(m)}(z) = \lambda P_{n-1}^{(m)}(z) + \\ + (N-m)\mu P_{n+1}^{(m-1)}(z) + m\mu z P_{n+1}^{(m)}(z)\,,\ \ n\geq N\,.$$

Умножая последнее выражение на u^{n-N}, после суммирования по всем $n\geq N$ получаем:

$$f_m(u,z)P^{(m)}(u,z) = m\mu z P_N^{(m)}(z) - \lambda u P_{N-1}^{(m)}(z) - \\ - (N-m)\mu[P^{(m-1)}(u,z) - P_N^{(m-1)}(z)]\,, \\ m = \overline{2, N-1}\,. \quad (16)$$

Из уравнений (5) и (6) после умножения на z^i и суммирования по всем значениям $i \geq 0$ находим:

$$P_{N-1}^{(1)}(z) = \frac{\lambda p_{N-2} + (N-1)\mu p_N}{\lambda + (N-1)\mu} + \\ + \frac{\mu z}{\lambda + (N-1)\mu} P_N^{(1)}(z)\,. \quad (17)$$

Действуя аналогичным образом с уравнениями (11) и (12), как и с уравнениями (5) и (6), приходим к выражению:

$$P_{N-1}^{(m)}(z) = \frac{\lambda P_{N-2}^{(m-1)}(z) + (N-m)\mu P_N^{(m-1)}(z)}{\lambda + (N-1)\mu} + \\ + \frac{m\mu z}{\lambda + (N-1)\mu} P_N^{(m)}(z)\,, \ \ m = \overline{2, N-1}\,. \quad (18)$$

Домножая уравнения (7) и (8) на z^i, после суммирования по всем значениям $i \geq 0$ имеем:

$$P_n^{(1)}(z) = \frac{\lambda p_{n-1} + n\mu P_{n+1}^{(1)}(z)}{\lambda + n\mu} + \frac{\mu z}{\lambda + n\mu} P_{n+1}^{(2)}(z)\,, \\ n = \overline{1, N-2}\,. \quad (19)$$

Наконец, производя аналогичные преобразования с уравнениями (13) и (14), получаем:

$$P_n^{(m)}(z) = \frac{\lambda P_{n-1}^{(m-1)}(z) + (n-m+1)\mu P_{n+1}^{(m)}(z)}{\lambda + n\mu} + \\ + \frac{m\mu z}{\lambda + n\mu} P_{n+1}^{(m+1)}(z)\,, \ \ m = \overline{2, N-2}\,, \\ n = \overline{m, N-2}\,. \quad (20)$$

Уравнения (15)–(20) позволяют находить выражения для всех ПФ $P_n^{(m)}(z)$, $m = \overline{1, N-1}$, $n = \overline{1, N-1}$, а также совместное стационарное распределение рекуррентным образом. Подставляя выражение для $P_{N-1}^{(1)}(z)$ из формулы (17) в формулу (15), получаем:

$$P^{(1)}(u,z) = \left(\left[\mu z - \frac{\lambda \mu z u}{\lambda + (N-1)\mu} \right] P_N^{(1)}(z) - \right. \\ - \left[\lambda u \frac{\lambda p_{N-2} + (N-1)\mu p_N}{\lambda + (N-1)\mu} + \right. \\ \left. \left. + (N-1)\mu[P(u) - p_N] \right] \right) \Big/ f_1(u,z)\,, \quad (21)$$

откуда из равенства нулю в точке $u_1(z)$ числителя и знаменателя правой части формулы (21) следует:

$$P_N^{(1)}(z) = (\lambda u_1(z)[\lambda p_{N-2} + (N-1)\mu p_N] + \\ + (\lambda + (N-1)\mu)(N-1)\mu\,[P(u_1(z)) - p_N]) \Big/ (\mu z\,[\lambda + \\ + (N-1)\mu - \lambda u_1(z)])\,.$$

Теперь, возвращаясь к формуле (17), получаем выражение для $P_{N-1}^{(1)}(z)$:

$$P_{N-1}^{(1)}(z) = ([\lambda p_{N-2} + (N-1)\mu p_N] + \\ + (N-1)\mu[P(u_1(z)) - p_N]) \Big/ (\lambda + (N-1)\mu - \\ - \lambda u_1(z))\,.$$

Далее из равенства (19) выражаем $P_{N-2}^{(1)}(z)$ через $P_{N-1}^{(2)}(z)$. Из равенства (18) выражаем $P_{N-1}^{(2)}(z)$ через $P_N^{(2)}(z)$. Подставляя полученное выражение для $P_{N-1}^{(2)}(z)$ в формулу (16), из равенства нулю в точке u_2 левой и правой части получившегося равенства находим $P^{(2)}(u,z)$. Затем из равенства (19) выражаем $P_{N-3}^{(1)}(z)$ через $P_{N-2}^{(2)}(z)$ и т. д.

Продолжая эту процедуру, можно найти соотношения для вычисления всех ПФ $P_n^{(m)}(z)$, $m = \overline{1, N-1}$, $n = \overline{1, N-1}$.

С каждым шагом выражение для очередной ПФ становится все сложнее, и в итоге при большом числе приборов выписать явный вид всех ПФ не удается. Тем не менее нахождение значений ПФ в каждой точке $z \neq 0$ можно свести к последовательному решению систем линейных уравнений. Для этого обозначим через $A_n(z)$, $n = \overline{2, N-1}$, матрицы размера $(n+1)\times(n+1)$, имеющие следующую структуру:

$$A_2(z) = \begin{pmatrix} 2\mu z & 0 & -2\mu z \\ -\lambda u_2(z) & -\mu z & \lambda + (N-1)\mu \\ 0 & \lambda + (N-2)\mu & -\lambda \end{pmatrix};$$

$$A_n(z) = \begin{pmatrix} n\mu z & 0 & -n\mu z & \cdots \\ -\lambda u_n(z) & 0 & \lambda + (N-1)\mu & \cdots \\ \vdots & \vdots & \vdots & \cdots \\ 0 & 0 & 0 & \cdots \\ 0 & 0 & 0 & \cdots \\ 0 & -\mu z & 0 & \cdots \\ 0 & \lambda + (N-n)\mu & 0 & \cdots \end{pmatrix} \\ \begin{pmatrix} \cdots & 0 & 0 \\ \cdots & 0 & 0 \\ \cdots & \vdots & \vdots \\ \cdots & -3\mu z & 0 \\ \cdots & \lambda + (N-n+2) & -2\mu z \\ \cdots & -\lambda & \lambda + (N-n+1)\mu \\ \cdots & 0 & -\lambda \end{pmatrix}, \\ n = \overline{3, N-1}\,.$$

Определим вектор-стро́ки $\vec{a}_n(z)$ и $\vec{b}_n(z)$ длины $(n+1)$ следующим образом:

$$\vec{a}_n(z) = \Big(P_N^{(n)}(z), P_{N-1}^{(n)}(z), \dots \\ \dots, P_{N-n+1}^{(2)}(z), P_{N-n}^{(1)}(z)\Big)\ ,\ \ n = \overline{2, N-1}\,;$$

$$\vec{b}_2(z) = \Big((N-2)\mu[P^{(1)}(u_2, z) - P_N^{(1)}(z)], \lambda p_{N-3} + \\ + (N-2)\mu P_{N-1}^{(1)}(z), (N-2)\mu P_N^{(1)}(z)\Big)\ ;$$

$$\vec{b}_n(z) = \Big((N-n)\mu[P^{(n-1)}(u_n, z) - P_N^{(n-1)}(z)], \\ \lambda p_{N-1-n} + (N-n)\mu P_{N-1-(n-2)}^{(1)}(z), \\ (N-n)\mu P_N^{(n-1)}(z), (N-n)\mu P_{N-1}^{(n-1)}(z), \dots, \\ (N-n)\mu P_{N-n+3}^{(3)}(z), (N-n)\mu P_{N-n+2}^{(2)}(z)\Big)\ , \\ n = \overline{3, N-1}\,.$$

Тогда алгоритм нахождения ПФ состоит в последовательном начиная с $n = 2$ решении системы линейных уравнений

$$\vec{a}_n(z) A_n(z) = \vec{b}_n(z)\,.$$

Из структуры матрицы $A_n(z)$, $n = \overline{3, N-1}$, видно, что она неприводима и обладает свойством диагонального преобладания т. е. перестановкой строк и столбцов можно добиться того, что в каждой строке модуль диагонального элемента будет либо строго больше, либо не меньше суммы модулей всех остальных элементов в строке. Покажем это. Если определить матрицы перестановки P_n^L и P_n^R размера $(n+1)\times(n+1)$ при $n = \overline{3, N-1}$ следующим образом:

$$P_n^L = \begin{pmatrix} 0\,0\cdots 0\,1 \\ 1\,0\cdots 0\,0 \\ \vdots\,\vdots\,\dots\,\vdots\,\vdots \\ 0\,0\cdots 0\,0 \\ 0\,0\cdots 1\,0 \end{pmatrix}\ ;\ \ P_n^R = \begin{pmatrix} 0\,1\cdots 0\,0 \\ 1\,0\cdots 0\,0 \\ \vdots\,\vdots\,\dots\,\vdots\,\vdots \\ 0\,0\cdots 1\,0 \\ 0\,0\cdots 0\,1 \end{pmatrix}\ ,$$

то матрица $P_n^L A_n(z) P_n^R$, $n = \overline{3, N-1}$, примет вид:

$$P_n^L A_n(z) P_n^R = \\ = \begin{pmatrix} \lambda + (N-n)\mu & 0 & 0 & \cdots \\ 0 & n\mu z & -n\mu z & \cdots \\ 0 & -\lambda u_n(z) & \lambda + (N-1)\mu & \cdots \\ \vdots & \vdots & \vdots & \cdots \\ 0 & 0 & 0 & \cdots \\ 0 & 0 & 0 & \cdots \\ -\mu z & 0 & 0 & \cdots \end{pmatrix}$$

$$\begin{pmatrix} \cdots & 0 & -\lambda \\ \cdots & 0 & 0 \\ \cdots & 0 & 0 \\ \cdots & \vdots & \vdots \\ \cdots & -3\mu z & 0 \\ \cdots & \lambda + (N-n+2)\mu & -2\mu z \\ \cdots & -\lambda & \lambda + (N-n+1)\mu \end{pmatrix}.$$

Легко видеть, что в каждой строке модуль диагонального элемента либо строго больше, либо не меньше суммы модулей всех остальных элементов в строке. Тогда, как вытекает из следствия 6.2.27 в [11], у матрицы $A_n(z)$ существует обратная и, значит, система $\vec{a}_n(z)A_n(z) = \vec{b}_n(z)$ при $z \neq 0$ имеет единственное решение.

5 Примеры расчетов

На основе полученных в разд. 4 результатов были проведены расчеты среднего и дисперсии числа заявок в БП, а также коэффициента корреляции числа заявок в накопителе и числа заявок в БП для различного числа приборов N и значений загрузки системы ρ/N. Напомним, что аналогичные показатели были рассчитаны в [9] по определению, на основе стационарных вероятностей, рассчитанных по приведенному выше алгоритму. Далее можно видеть, что результаты, полученные с помощью ПФ, как и ожидалось, полностью совпадают с результатами, представленными в [9].

На рис. 1 отражено поведение значения среднего числа заявок в БП в зависимости от загрузки системы ρ/N. Отметим, что полученные в предыдущих разделах результаты позволяют рассчитывать такие характеристики, как среднее число заявок только в первой очереди в БП, в сумме в первой и во второй очередях в БП (когда обе очереди существуют), в сумме в первой, второй,. . . , $(N-1)$-й очере-

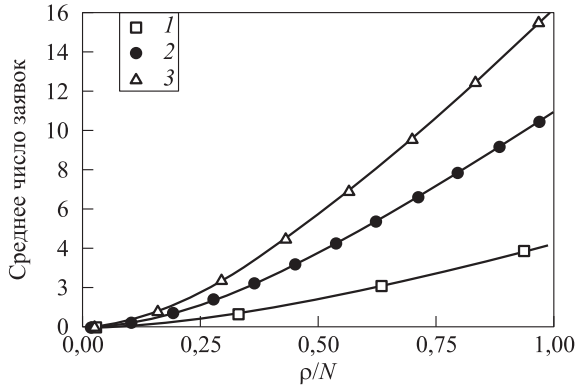


Рис. 1 Поведение среднего числа заявок в БП в зависимости от загрузки системы ρ/N: *1* — $N = 4$; *2* — 7; *3* — $N = 9$

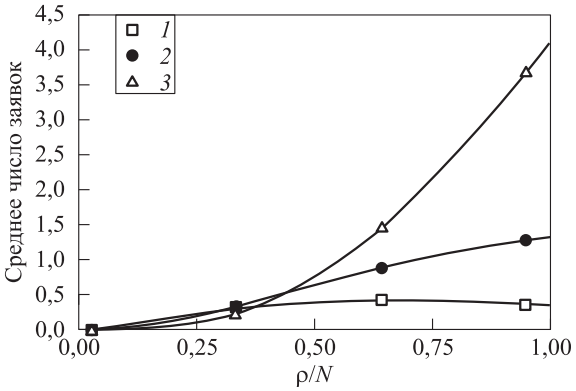


Рис. 2 Поведение среднего числа заявок в первой очереди в БП (*1*), в сумме в первой и во второй очередях в БП (*2*), в сумме в первой, второй и третьей очередях в БП (*3*) в зависимости от загрузки системы ρ/N. Число приборов $N = 4$

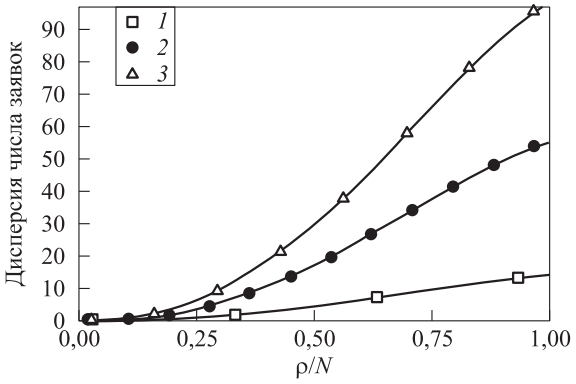


Рис. 3 Поведение дисперсии числа заявок в БП в зависимости от загрузки системы ρ/N: *1* — $N = 4$; *2* — 7; *3* — $N = 9$

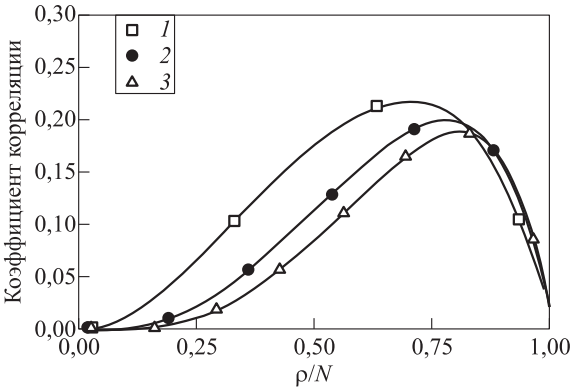


Рис. 4 Поведение коэффициента корреляции числа заявок в накопителе и числа заявок в БП в зависимости от загрузки системы ρ/N: *1* — $N = 4$; *2* — 7; *3* — $N = 9$

дях в БП (когда каждая из очередей существует). Поведение данных характеристик в зависимости от загрузки системы ρ/N для случая $N = 4$ представлено на рис. 2.

На рис. 3 и 4 изображено поведение дисперсии числа заявок в БП и поведение коэффициента корреляции числа заявок в накопителе и числа заявок в БП соответственно.

Во всех расчетах интенсивность обслуживания заявок μ принималась равной 1.

Анализируя графики на рис. 1–4, стоит отметить два момента. Среднее число заявок в БП не уходит в бесконечность с ростом загрузки (и даже при загрузке больше единицы), что следует из формулы Литтла. Число заявок в накопителе и число заявок в БП весьма слабо коррелированы, и с ростом числа приборов коэффициент корреляции уменьшается.

6 Заключение

В настоящей работе рассмотрена функционирующая в непрерывном времени многоканальная система обслуживания с накопителем бесконечной емкости и переупорядочением заявок. В систему поступает пуассоновский поток заявок, время обслуживания каждым прибором распределено по экспоненциальному закону с одним и тем же параметром. Для нахождения совместного стационарного распределения числа заявок в накопителе и суммарного числа заявок в БП получен рекуррентный алгоритм. Также показано, как можно находить совместное распределение в терминах ПФ, которые облегчают расчет его моментов.

Литература

1. *Boxma O., Koole G., Liu Z.* Queueing-theoretic solution methods for models of parallel and distributed systems // Performance Evaluation of Parallel and Distributed Systems Solution Methods, 1994. CWI Tract 105 and 106. P. 1–24.
2. *Dimitrov B.* Queues with resequencing. A survey and recent results // 2nd World Congress on Nonlinear Analysis, Theory, Methods, Applications Proceedings, 1997. Vol. 30. No. 8. P. 5447–5456.
3. *Huisman T., Boucherie R. J.* The sojourn time distribution in an infinite server resequencing queue with dependent interarrival and service times // J. Appl. Probab., 2002. Vol. 39. No. 3. P. 590–603.
4. *Xia Y., Tse D. N. C.* On the large deviations of resequencing queue size: 2-$M/M/1$ case // IEEE Trans. Inform. Theory, 2008. Vol. 54. No. 9. P. 4107–4118.
5. *Leung K., Li V. O. K.* A resequencing model for high-speed packet-switching networks // J. Comput. Commun., 2010. Vol. 33. No. 4. P. 443–453.

6. *Матюшенко С. И.* Стационарные характеристики двухканальной системы обслуживания с переупорядочением заявок и распределениями фазового типа // Информатика и её применения, 2010. Т. 4. Вып. 4. С. 67–71.
7. *De Nicola C., Pechinkin A. V., Razumchik R. V.* Stationary characteristics of homogenous Geo/Geo/2 queue with resequencing in discrete time // 27th European Conference on Modelling and Simulation Proceedings. —- Aalesund, 2013. P. 594–600.
8. *Pechinkin A. V., Caraccio I., Razumchik R. V.* Joint stationary distribution of queues in homogenous $M|M|3$ queue with resequencing // 28th European Conference on Modelling and Simulation Proceedings. — Brescia, 2014. P. 558–564.
9. *Pechinkin A. V., Caraccio I., Razumchik R. V.* On joint stationary distribution in exponential multiserver reordering queue // 12th Conference (International) on Numerical Analysis and Applied Mathematics Proceedings, 2014 (in press).
10. *Bocharov P. P., D'Apice C., Pechinkin A. V., Salerno S.* Queueing theory. — Urecht, Boston: VSP, 2004. 446 p.
11. *Horn R. A., Johnson C. R.* Matrix analysis. — 2nd ed. — Cambridge: Cambridge University Press, 2013. 662 p.

Поступила в редакцию 28.10.14

JOINT STATIONARY DISTRIBUTION OF THE NUMBER OF CUSTOMERS IN THE SYSTEM AND REORDERING BUFFER IN THE MULTISERVER REORDERING QUEUE

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

[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: The paper considers a continuous-time multiserver queueing system with buffer on infinite capacity and reordering. The Poisson flow of customers arrives at the system. Service times of customers at each server are exponentially distributed with the same parameter. Each customer obtains a sequential number upon arrival. The order of customers upon arrival should be preserved upon departure from the system. Customers whose service finished but which violated the order are kept in the reordering buffer of infinite capacity. A joint stationary distribution of the number of customers in the buffer, servers, and reordering buffer is obtained in terms of a computational algorithm and a generating function. A numerical example is provided.

Keywords: queueing system; reordering; infinite capacity; joint distribution

DOI: 10.14357/19922264140401

Acknowledgments

The research was partially financially supported by the Russian Foundation for Basic Research (project 13-07-00223).

References

1. Boxma O., G. Koole, and Z. Liu. 1994. Queueing-theoretic solution methods for models of parallel and distributed systems. *Performance Evaluation of Parallel and Distributed Systems Solution Methods*. CWI Tract 105 and 106:1–24.
2. Dimitrov, B. 1997. Queues with resequencing. A survey and recent results. *2nd World Congress on Nonlinear Analysis, Theory, Methods, Applications Proceedings*. 30(8):5447–5456.
3. Huisman, T., and R. J. Boucherie. 2002. The sojourn time distribution in an infinite server resequencing queue with dependent interarrival and service times. *J. Appl. Probab*. 39(3):590–603.
4. Xia, Y., and D. N. C. Tse. 2008. On the large deviations of resequencing queue size: 2-$M/M/1$ case. *IEEE Trans. Inform. Theory* 54(9):4107–4118.
5. Leung, K., and V. O. K. Li. 2010. A resequencing model for high-speed packet-switching networks. *J. Comput. Commun*. 33(4):443–453.
6. Matyushenko, S. I. 2010. Statsionarnye kharakteristiki dvukhkanal'noy sistemy obsluzhivaniya s pereuporyadochivaniem zayavok i raspredeleniyami fazovogo tipa [Stationary characteristics of the two-channel queueing system

with reordering customers and distributions of phase type]. *Informatika i ee Primemeniya — Inform. Appl.* 4(4):67–71.

7. De Nicola, C., A. V. Pechinkin, and R. V. Razumchik. 2013. Stationary characteristics of homogenous Geo/Geo/2 queue with resequencing in discrete time. *27th European Conference on Modelling and Simulation Proceedings*. Aalesund. 594–600.

8. Pechinkin, A. V., I. Caraccio, and R. V. Razumchik. 2014. joint stationary distribution of queues in homogenous $M|M|3$ queue with resequencing. *28th European Conference on Modelling and Simulation Proceedings*. Brescia. 558–564.

9. Pechinkin, A. V., I. Caraccio, and R. V. Razumchik. 2014 (in press). On joint stationary distribution in exponential multiserver reordering queue. *12th Conference (International) on Numerical Analysis and Applied Mathematics Proceedings*.

10. Bocharov, P. P., C. D'Apice, A. V. Pechinkin, and S. Salerno. 2004. *Queueing theory*. Urecht, Boston: VSP. 446 p.

11. Horn, R. A., and C. R. Johnson. 2013. *Matrix analysis*. Cambridge: Cambridge University Press. 662 p.

Received October 28, 2014

Contributors

Pechinkin Alexander V. (1946–2014) — Doctor of Science in physics and mathematics; principal scientist, Institute of Informatics Problems of the Russian Academy of Sciences, 44-2 Vavilov Str., Moscow 119333, Russian Federation

Razumchik Rostislav V. (b. 1984) — Candidate of Science (PhD) in physics and mathematics, senior scientist, Institute of Informatics Problems of the Russian Academy of Sciences, 44-2 Vavilov Str., Moscow 119333, Russian Federation; associate professor, Peoples' Friendship University of Russia, 6 Miklukho-Maklaya Str., Moscow 117198, Russian Federation; rrazumchik@ieee.org

ИНФОРМАТИКА И ЕЁ ПРИМЕНЕНИЯ, 2014. Т. 8. Вып. 4. С. 11–19

МОДИФИЦИРОВАННЫЙ СЕТОЧНЫЙ МЕТОД РАЗДЕЛЕНИЯ ДИСПЕРСИОННО-СДВИГОВЫХ СМЕСЕЙ НОРМАЛЬНЫХ ЗАКОНОВ*

В. Ю. Королев[1], А. Ю. Корчагин[2]

Аннотация: Описывается модифицированный двухэтапный сеточный метод разделения дисперсионно-сдвиговых смесей нормальных законов, представляющий собой альтернативу чистому EM (expectation-maximization) алгоритму. На первом этапе этого алгоритма строится дискретная аппроксимация для смешивающего распределения, на втором этапе подбирается абсолютно непрерывное распределение из заранее заданного семейства, например, обобщенных обратных гауссовских законов, ближайшее к дискретному распределению, полученному на первом этапе. Обсуждаются вопросы сходимости этого двухэтапного алгоритма. Доказана монотонность сеточного итерационного метода, используемого на первом этапе. Подробно обсуждается вопрос оптимального выбора параметров метода, прежде всего сетки, накидываемой на носитель смешивающего распределения. С этой целью предложены статистические оценки квантилей смешивающего распределения. Эффективность метода иллюстрируется примерами конкретных вычислений оценок параметров обобщенных гиперболических распределений.

Ключевые слова: смесь распределений вероятностей; дисперсионно-сдвиговая смесь нормальных законов; обобщенное гиперболическое распределение; EM-алгоритм; сеточный метод разделения смесей

DOI: 10.14357/19922264140402

1 Введение

При *практическом* решении задачи моделирования и исследования волатильности (изменчивости) хаотических стохастических процессов ключевым этапом является статистическое разделение смесей вероятностных распределений. Задача разделения смесей — статистического оценивания параметров смесей вероятностных распределений — в деталях разобрана, например, в книге [1].

Для решения задачи разделения смесей вероятностных распределений традиционно используются итерационные процедуры типа EM-алгоритма. К сожалению, классический EM-алгоритм обладает рядом серьезных недостатков при его применении к смесям нормальных законов, а именно: он демонстрирует крайнюю неустойчивость по отношению к исходным данным и начальным приближениям.

Для преодоления этих недостатков предложено много модификаций EM-алгоритма (см., например, [1]). Вместе с тем в указанной книге предложен и исследован принципиально новый — сеточный — метод приближенного решения задачи разделения смесей. В работе [2] подробно исследованы вопросы сходимости сеточных методов разделения смесей.

В соответствии с подходом к статистическому анализу хаотических стохастических процессов, в частности к решению задачи декомпозиции волатильности таких процессов, развитом в книге [1], в общем случае на практике приходится решать задачу разделения конечных смесей нормальных законов с произвольно большим числом неизвестных параметров (параметров компонент и их весов). И хотя в большинстве приложений возникают смеси не более чем с пятью–семью компонентами, даже при использовании таких смесей, скажем, в задачах анализа и прогнозирования финансовых рисков приходится моделировать траекторию движения точки в пространствах, размерность которых соответственно лежит в пределах от 14 (для пятикомпонентных смесей) до 20 (для семикомпонентных смесей), что существенно увеличивает вычислительные и временны́е ресурсы, необходимые для практического решения указанных задач.

Поскольку во многих ситуациях (например, при прогнозировании на основе высокочастотных дан-

*Работа поддержана Российским научным фондом (проект 14-11-00364).

[1]Факультет вычислительной математики и кибернетики Московского государственного университета им. М. В. Ломоносова; Институт проблем информатики Российской академии наук; victoryukorolev@yandex.ru

[2]Факультет вычислительной математики и кибернетики Московского государственного университета им. М. В. Ломоносова; sasha.korchagin@gmail.com

ных) эти задачи необходимо решать в режиме, близком к реальному времени, для создания эффективных методов статистического анализа на основе смешанных моделей на первый план выходит проблема снижения размерности решаемой задачи, т. е. параметрического пространства.

Одним из возможных подходов к снижению размерности является априорное сужение классов допустимых смесей. К примеру, при решении многих задач, связанных с анализом процессов атмосферной или плазменной турбулентности, а также процессов, описывающих эволюцию различных финансовых индексов, высочайшую адекватность продемонстрировали модели, основанные на дисперсионно-сдвиговых смесях нормальных законов. Класс таких смесей очень обширен и, в частности, включает в себя обобщенные гиперболические распределения, которые были введены О.-Е. Барндорфф-Нильсеном в 1977–1978 гг. как класс специальных сдвиг-масштабных смесей нормальных законов [3, 4]. Пусть $\alpha \in \mathbb{R}$, $\beta \in \mathbb{R}$. Если функцию распределения обобщенного гиперболического закона с параметрами α, β, ν, μ, λ обозначить $P_{GH}(x;\alpha,\beta,\nu,\mu,\lambda)$, то по определению

$$P_{GH}(x;\alpha,\beta,\nu,\mu,\lambda) = \int\limits_0^\infty \Phi\left(\frac{x-\beta-\alpha z}{\sqrt{z}}\right) p_{GIG}(z;\nu,\mu,\lambda)\,dz\,, \quad x \in \mathbb{R}\,, \quad (1)$$

где $\Phi(x)$ — стандартная нормальная функция распределения:

$$\Phi(x) = \int\limits_{-\infty}^{x} \varphi(z)\,dz\,, \quad \varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\,, \quad x \in \mathbb{R}\,;$$

$p_{GIG}(x;\nu,\mu,\lambda)$ — плотность обобщенного обратного гауссовского распределения:

$$p_{GIG}(x;\nu,\mu,\lambda) = \frac{\lambda^{\nu/2}}{2\mu^{\nu/2}K_\nu\left(\sqrt{\mu\lambda}\right)} x^{\nu-1} \exp\left\{-\frac{1}{2}\left(\frac{\mu}{x}+\lambda x\right)\right\}, \quad x > 0\,.$$

Здесь $\nu \in \mathbb{R}$;

$$\mu > 0\,,\ \lambda \geqslant 0\,, \text{ если } \nu < 0\,;$$
$$\mu > 0\,,\ \lambda > 0\,, \text{ если } \nu = 0\,;$$
$$\mu \geqslant 0\,,\ \lambda > 0\,, \text{ если } \nu > 0\,;$$

$K_\nu(z)$ — модифицированная бесселева функция третьего рода порядка ν:

$$K_\nu(z) = \frac{1}{2}\int\limits_0^\infty y^{\nu-1} \exp\left\{-\frac{z}{2}\left(y+\frac{1}{y}\right)\right\} dy\,, \quad z \in \mathbb{C}\,,\ \operatorname{Re} z > 0\,.$$

Обратим внимание, что в (1) смешивание происходит одновременно и по параметру сдвига, и по параметру масштаба, но так как эти параметры в (1) связаны жесткой зависимостью, так что параметр сдвига смешиваемого распределения пропорционален его дисперсии, то фактически смесь (1) является *однопараметрической* и поэтому называется *дисперсионно-сдвиговой* (см., например, [5]).

Другим примером дисперсионно-сдвиговых смесей нормальных законов являются обобщенные дисперсионные гамма-распределения, в которых смешивающими являются обобщенные гамма-распределения [6, 7].

В указанных семействах смесей число неизвестных параметров равно пяти или шести (если учитывать неслучайный сдвиг). Вместе с тем у подобных моделей имеются довольно серьезные теоретические обоснования: в работах [7, 8] показано, что указанные модели являются асимптотическими аппроксимациями в простой предельной схеме случайного суммирования и потому могут успешно применяться для анализа процессов типа остановленных случайных блужданий. Эти выводы подтверждены статистическим анализом высокочастотных финансовых данных, в результате которого выявлен синхронизированный характер изменения интенсивностей потоков заявок в системах электронных торгов, что естественно приводит к синхронизированному поведению параметров сдвига и диффузии в соответствующих моделях вида смесей нормальных законов [9].

2 Описание модифицированного сеточного метода разделения дисперсионно-сдвиговых смесей нормальных законов и его свойства

Оказывается, что сеточные методы разделения смесей довольно эффективны не только при разделении конечных смесей нормальных законов, но и при разделении произвольных дисперсионно-сдвиговых смесей нормальных законов. Поясним сказанное на примере задачи оценивания параметров обобщенных гиперболических распределений.

Для решения задачи оценивания параметров обобщенных гиперболических распределений тра-

диционно используется метод, предложенный в статье [10] и по сути являющийся классическим EM-алгоритмом, приспособленным к конкретной задаче, и, соответственно, наследующий присущие EM-алгоритмам недостатки.

Рассмотрим следующий альтернативный двухэтапный метод. На первом этапе на положительной полуприямой выделим основную часть носителя смешивающего распределения, т. е. ограниченный интервал, вероятность которого, вычисленная в соответствии со смешивающим распределением, практически равна единице. На этот интервал накинем конечную сетку, содержащую, возможно, очень много *известных* узлов $u_1, \ldots, u_K$. Считая параметр сдвига β равным нулю, приблизим искомое обобщенное гиперболическое распределение конечной смесью нормальных законов:

$$P_{GH}(x;\, \alpha, 0, \nu, \mu, \lambda) \approx \approx \sum_{i=1}^{K} p_i \Phi\left(\frac{x-\alpha u_i}{\sqrt{u_i}}\right), \quad x \in \mathbb{R}. \quad (2)$$

В смеси, стоящей в правой части соотношения (2), неизвестными являются только параметры $p_1, \ldots, p_{K-1}$ и α. Пусть $x_1, \ldots, x_n$ — анализируемая выборка значений случайной величины с оцениваемым обобщенным гиперболическим распределением. Итерационный процесс, определяющий сеточный EM-алгоритм для данной задачи, задается следующим образом. Пусть $p_1^{(m)}, \ldots, p_{K-1}^{(m)}$ и $\alpha^{(m)}$ — оценки параметров $p_1, \ldots, p_{K-1}$ и α на m-й итерации, $p_K^{(m)} = 1 - p_1^{(m)} - \cdots - p_{K-1}^{(m)}$. Обозначим

$$\varphi_{ij}^{(m)} = \frac{1}{\sqrt{u_i}} \varphi\left(\frac{x_j - \alpha^{(m)} u_i}{\sqrt{u_i}}\right);$$

$$g_{ij}^{(m)} = \frac{p_i^{(m)} \varphi_{ij}^{(m)}}{\sum\limits_{r=1}^{K} p_r^{(m)} \varphi_{rj}^{(m)}},$$

$$i = 1, \ldots, K; \;\; j = 1, \ldots, n.$$

Тогда, используя стандартные рассуждения, определяющие вычислительные формулы EM-алгоритма для параметров конечной смеси нормальных законов (см, например, [1, разд. 5.3.7–5.3.8]), следует положить

$$p_i^{(m+1)} = \frac{1}{n} \sum_{j=1}^{n} g_{ij}^{(m)}, \quad i = 1, \ldots, K. \quad (3)$$

Обозначим $\overline{x} = (1/n) \sum\limits_{j=1}^{n} x_j$. Используя соотношение (5.3.24) в [1], с учетом очевидного равенства $\sum\limits_{i=1}^{K} g_{ij}^{(m)} = 1$ можно заметить, что уточненная оценка параметра α имеет вид:

$$\alpha^{(m+1)} = \frac{\overline{x}}{\sum\limits_{i=1}^{K} u_i p_i^{(m+1)}}, \quad (4)$$

т. е. равна отношению генерального выборочного среднего и текущего эмпирического среднего смешивающего распределения, что вполне согласуется с тем, что в соответствии с приводимым ниже соотношением (5) в данном случае $\mathsf{E}X = \alpha \mathsf{E}U$.

В силу монотонности классического EM-алгоритма справедливо следующее утверждение.

Теорема 1. *Пусть узлы $u_1, \ldots, u_K$ сетки различны, неотрицательны и известны. Тогда итерационный процесс (3)–(4) является монотонным, т. е. каждая его итерация не уменьшает целевую сеточную функцию правдоподобия*

$$L(p_1, \ldots, p_K, \alpha; x_1, \ldots, x_n) = \prod_{j=1}^{n} \left[\sum\nolimits_{i=1}^{K} \frac{p_i}{\sqrt{u_i}} \varphi\left(\frac{x_j - \alpha^{(m)} u_i}{\sqrt{u_i}}\right)\right].$$

Замечание 1. В разд. 5.7.4 книги [1] показано, что при каждом фиксированном значении параметра α сеточная функция правдоподобия $L(p_1, \ldots, p_{K-1}, \alpha;\, x_1, \ldots, x_n)$ вогнута по аргументам $p_1, \ldots, p_{K-1}$. Поэтому на каждом шаге итерационного процесса вместо соотношения (3) можно использовать любой более быстрый алгоритм максимизации функции $L(p_1, \ldots, p_{K-1}, \alpha^{(m)};\, x_1, \ldots \ldots, x_n)$ по переменным $p_1, \ldots, p_{K-1}$. Например, оценки весов $p_1, \ldots, p_K$ можно искать методом условного градиента [1, 11].

Таким образом, на первом этапе получаются оценки параметра α и весов всех узлов u_i конечной сетки, накинутой на носитель смешивающего обобщенного обратного гауссовского распределения $P_{\mathrm{GIG}}(z;\, \nu, \mu, \lambda)$.

На втором этапе остается применить какой-либо стандартный метод подгонки обобщенного обратного гауссовского распределения $P_{\mathrm{GIG}}(z;\, \nu, \mu, \lambda)$ к эмпирическим данным типа гистограммы $(u_1, p_1), \ldots, (u_K, p_K)$. Например, параметры ν, μ и λ можно оценить, минимизируя соответствующую статистику хи-квадрат. Или же, например, можно решить задачу наименьших квадратов:

$$(\nu^*, \mu^*, \lambda^*) = \arg\min_{\nu,\mu,\lambda} \sum_{i=1}^{K} \left[p_i - \int_{(1/2)(u_{i-1}+u_i)}^{(1/2)(u_i+u_{i+1})} p_{GIG}(u;\, \nu, \mu, \lambda)\, du\right]^2,$$

где $u_0 = 0$; $u_{K+1} = \infty$.

На практике хорошие результаты показал подход с решением задачи наименьших квадратов. Для поиска параметров использовался алгоритм ns2sol, описанный в книге [12]. Указанный алгоритм доступен во многих статистических пакетах, отличается высоким быстродействием и возможностью при желании задавать разумные интервалы для поиска параметров.

3 О практическом выборе сетки на первом этапе модифицированного сеточного метода разделения дисперсионно-сдвиговых смесей нормальных законов

Естественно, что при использовании указанного двухэтапного метода в динамическом режиме крайне важным становится вопрос о выборе наиболее эффективных и быстродействующих численных процедур и их параметров. В частности, исключительную важность приобретает правильный выбор сетки на первом этапе. Рассмотрим этот вопрос подробнее.

Формально рассматриваемая задача выглядит так: по наблюдаемым значениям $x_1, \ldots, x_n$ требуется построить статистическую оценку верхней границы квантилей заданного порядка смешивающего закона так, чтобы как можно точнее оценить носитель смешивающего распределения.

В дальнейшем будем считать, что $x_1, \ldots, x_n$ — независимые реализации случайной величины $X = Y\sqrt{U} + \alpha U$, где Y — случайная величина со стандартным нормальным распределением, а U — независимая от нее случайная величина с обобщенным обратным гауссовским распределением. Тогда, очевидно, распределение случайной величины X имеет вид (1). Предположим, что у случайной величины U существуют моменты первых двух порядков. Тогда, как несложно видеть,

$$\mathsf{E}X = \mathsf{E}Y \cdot \mathsf{E}\sqrt{U} + \alpha \mathsf{E}U = \alpha \mathsf{E}U\,. \tag{5}$$

При этом по усиленному закону больших чисел с вероятностью единица $\overline{x} \longrightarrow \mathsf{E}X$ $(n \to \infty)$, так что при больших n справедливо приближенное равенство $\mathsf{E}X \approx \overline{x}$ и с учетом (5)

$$\mathsf{E}U \approx \frac{\overline{x}}{\alpha}\,. \tag{6}$$

Далее, очевидно,

$$\mathsf{E}X^2 = \mathsf{E}Y^2 \cdot \mathsf{E}U + 2\alpha \mathsf{E}X \cdot \mathsf{E}U^{3/2} + \\ + \alpha^2 \mathsf{E}U^2 = \mathsf{E}U + \alpha^2 \mathsf{E}U^2\,. \tag{7}$$

Поэтому, обозначив

$$m^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2\,,$$

получаем приближенное равенство $\mathsf{E}X^2 \approx m^2$, так что с учетом (6) и (7) имеем:

$$\mathsf{E}U^2 \approx \frac{1}{\alpha^2}\left(m^2 - \frac{\overline{x}}{\alpha}\right). \tag{8}$$

Если параметр α известен, то для определения верхней границы u^* сетки, накидываемой на носитель распределения случайной величины U, можно задать малое положительное число ε и воспользоваться требованием

$$\mathsf{P}(U \geqslant u^*) \leqslant \varepsilon\,. \tag{9}$$

А для гарантированного выполнения требования (9) можно использовать неравенство Маркова:

$$\mathsf{P}(U \geqslant u^*) \leqslant \frac{\mathsf{E}U^2}{(u^*)^2} \leqslant \varepsilon\,,$$

откуда с учетом (8)

$$(u^*)^2 \geqslant \frac{\mathsf{E}U^2}{\varepsilon} \approx \frac{1}{\alpha^2\varepsilon}\left(m^2 - \frac{\overline{x}}{\alpha}\right)$$

или

$$u^* \approx \frac{1}{\alpha\sqrt{\varepsilon}}\sqrt{m^2 - \frac{\overline{x}}{\alpha}}\,. \tag{10}$$

Если же параметр α, определяющий асимметрию распределения случайной величины X, неизвестен, то можно воспользоваться следующими рассуждениями. Обозначим

$$q_n = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}(x_i < 0)\,,$$

где $\mathbf{1}(A)$ — индикаторная функция множества (события) A. При этом по усиленному закону больших чисел с вероятностью единица $q_n \longrightarrow \mathsf{P}(X < 0)$ $(n \to \infty)$, так что при больших n справедливо приближенное равенство

$$q_n \approx \mathsf{P}(X < 0)\,. \tag{11}$$

Но

$$\mathsf{P}(X < 0) = \int_0^\infty \Phi\left(-\alpha\sqrt{u}\right) p_{\mathrm{GIG}}(u; \nu, \mu, \lambda)\, du = \\ = \mathsf{E}\Phi\left(-\alpha\sqrt{U}\right). \tag{12}$$

Предположим сначала, что $q_n < 1/2$. Если n достаточно велико, то можно с большой степенью уверенности утверждать, что тогда $\overline{x} > 0$ и $-\alpha < 0$, т. е. $\alpha > 0$ и, стало быть, на положительной полуоси значений аргумента u функция $\Phi(\alpha u)$ вогнута, т. е. выпукла вверх. Тогда из (11) и (12), дважды применяя неравенство Иенсена, в силу монотонности функции Φ получаем:

$$1 - q_n \approx 1 - \mathsf{E}\Phi\left(-\alpha\sqrt{\overline{U}}\right) = \mathsf{E}\Phi\left(\alpha\sqrt{\overline{U}}\right) \leqslant \leqslant \Phi\left(\alpha\mathsf{E}\sqrt{\overline{U}}\right) \leqslant \Phi\left(\alpha\sqrt{\mathsf{E}\overline{U}}\right). \quad (13)$$

Если теперь для $t \in (0,1)$ символом v_t обозначить t-квантиль стандартного нормального закона, то из (13) и (6) вытекает «приближенное неравенство»

$$v_{1-q_n} \leqslant \alpha\sqrt{\mathsf{E}U},$$

т. е.

$$\alpha \geqslant \frac{v_{1-q_n}}{\sqrt{\mathsf{E}U}} \approx \frac{v_{1-q_n}\sqrt{\alpha}}{\sqrt{\overline{x}}},$$

откуда получаем, что при достаточно больших n

$$\alpha \geqslant \frac{v_{1-q_n}^2}{\overline{x}}. \quad (14)$$

Если теперь задать малое положительное число ε, то для определения верхней границы u^* сетки, накидываемой на носитель распределения случайной величины U, можно воспользоваться требованием (9), для гарантированного выполнения которого с учетом (6) и (14) можно использовать неравенство Маркова:

$$\mathsf{P}(U \geqslant u^*) \leqslant \frac{\mathsf{E}U}{u^*} \approx \frac{\overline{x}}{\alpha u^*} \leqslant \frac{(\overline{x})^2}{v_{1-q_n}^2 u^*} \leqslant \varepsilon,$$

откуда окончательно вытекает оценка

$$u^* \approx \frac{(\overline{x})^2}{v_{1-q_n}^2 \varepsilon}. \quad (15)$$

В случае $q_n \geqslant 1/2$, если n достаточно велико, то можно с большой степенью уверенности утверждать, что $\overline{x} \leqslant 0$ и $-\alpha \geqslant 0$, т. е. на положительной

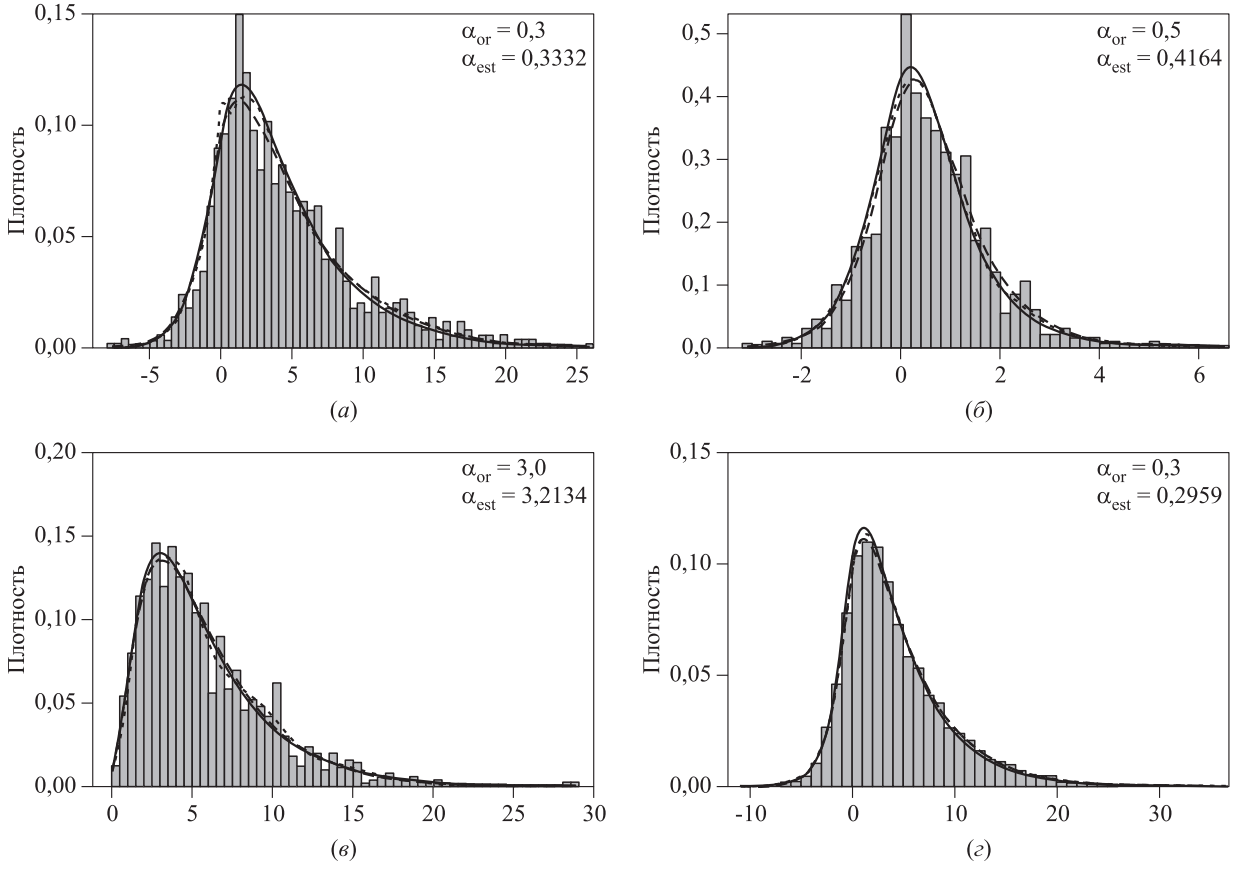


Рис. 1 Примеры применения модифицированного двухэтапного сеточного EM-алгоритма для подгонки обобщенного гиперболического распределения к искусственным данным, $\beta = 0$: (*а*) $n = 1000$, $\alpha = 0{,}3$, $\nu = 1{,}3$, $\mu = 1{,}6$, $\lambda = 0{,}2$; (*б*) $n = 1000$, $\alpha = 0{,}5$, $\nu = 1$, $\mu = 1$, $\lambda = 3$; (*в*) $n = 1000$, $\alpha = 3$, $\nu = 1{,}3$, $\mu = 1{,}6$, $\lambda = 2$; (*г*) $n = 10\,000$, $\alpha = 0{,}3$, $\nu = 1{,}3$, $\mu = 1{,}6$, $\lambda = 0{,}2$

полуоси значений аргумента u функция $\Phi(-\alpha u)$ вогнута, т. е. выпукла вверх. Тогда из (11) и (12), дважды применяя неравенство Иенсена, в силу монотонности функции Φ получаем

$$q_n \approx \mathsf{E}\Phi\left(-\alpha\sqrt{U}\right) \leqslant \Phi\left(-\alpha\sqrt{\mathsf{E}U}\right),$$

откуда вытекает «приближенное неравенство» $v_{q_n} \leqslant -\alpha\sqrt{\mathsf{E}U}$, т. е.

$$-\alpha \geqslant \frac{v_{q_n}}{\sqrt{\mathsf{E}U}} \approx \frac{v_{q_n}\sqrt{|\alpha|}}{\sqrt{|\overline{x}|}}$$

и при достаточно больших n

$$|\alpha| \geqslant \frac{v_{q_n}^2}{|\overline{x}|}. \qquad (16)$$

Для определения верхней границы u^* сетки, накидываемой на носитель распределения случайной величины U, снова зададим малое положительное число ε и потребуем, чтобы было справедливо условие (9), для гарантированного выполнения которого с учетом (6) и (16) используем неравенство Маркова и тот факт, что $\operatorname{sign}\overline{x} = \operatorname{sign}\alpha$ при достаточно больших n:

$$\mathsf{P}(U \geqslant u^*) \leqslant \frac{\mathsf{E}U}{u^*} \approx \frac{\overline{x}}{\alpha u^*} = \frac{|\overline{x}|}{|\alpha| u^*} \leqslant \leqslant \frac{(\overline{x})^2}{v_{q_n}^2\, u^*} \leqslant \varepsilon. \qquad (17)$$

В силу симметричности нормального распределения $v_t = -v_{1-t}$ для любого $t \in (0,1)$, поэтому $v_{q_n}^2 = v_{1-q_n}^2$ и в случае $q_n \geqslant 1/2$ соотношение (17) снова приводит к оценке (15).

Справедливости ради необходимо отметить, что оценки (10) и (15) являются завышенными, но они гарантируют, что $(1-\varepsilon)$-почти-весь носитель распределения случайной величины U будет лежать внутри интервала $[0, u^*]$.

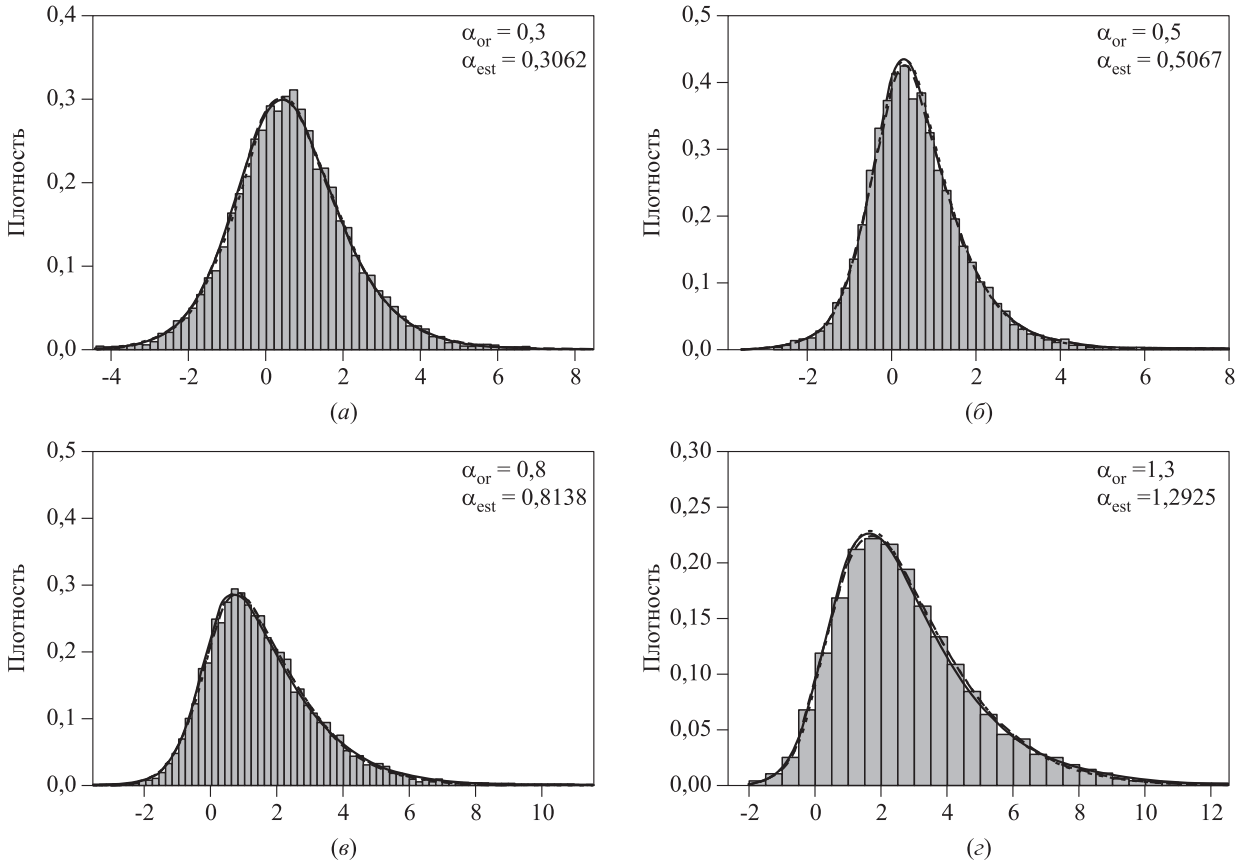


Рис. 2 Примеры применения модифицированного двухэтапного сеточного EM-алгоритма для подгонки обобщенного гиперболического распределения к искусственным данным, $n = 10\,000$, $\beta = 0$: (*а*) $\alpha = 0{,}3$, $\nu = 2$, $\mu = 2$, $\lambda = 2{,}5$; (*б*) $\alpha = 0{,}5$, $\nu = 1$, $\mu = 1$, $\lambda = 3$; (*в*) $\alpha = 0{,}8$, $\nu = 1{,}3$, $\mu = 1{,}6$, $\lambda = 2$; (*г*) $\alpha = 1{,}3$, $\nu = 2$, $\mu = 2$, $\lambda = 2{,}5$

4 Результаты численных экспериментов

Приводимые в данном разделе графики иллюстрируют качество работы модифицированного сеточного метода разделения дисперсионно-сдвиговых смесей нормальных законов на примере его применения к оцениванию параметров обобщенных гиперболических распределений с использованием указанного алгоритма выбора сетки с умеренным числом узлов $K = 40$. Для вычислений использовались искусственно сгенерированные выборки объемов $n = 1000$ и $n = 10\,000$ с разными наборами параметров, значения которых указаны на рисунках. На рис. 1 и 2 изображены гистограммы (серые столбики) и графики истинной плотности (штриховые линии), промежуточной оценки, полученной сеточным EM-алгоритмом (пунктирные линии) и итоговой оценки (непрерывные линии). На рис. 1 и 2 также указаны значения полученных оценок параметров. Как видно из приводимых рисунков, параметры α оцениваются очень точно. Точность оценок остальных параметров удовлетворительная и может быть повышена за счет использования более частых сеток и более чувствительных критериев остановки EM-алгоритма на первом этапе. Следует отметить, что даже в тех случаях, в которых наблюдаются заметные расхождения оценок параметров и их точных значений, оценки самих плотностей довольно точны.

Литература

1. *Королев В. Ю.* Вероятностно-статистические методы декомпозиции волатильности хаотических процессов. — М.: Изд-во Московского ун-та, 2011.
2. *Назаров А. Л.* Приближенные методы разделения смесей вероятностных распределений: Дисс. . . . канд. физ.-мат. наук. — М.: МГУ им. М. В. Ломоносова, 2013.
3. *Barndorff-Nielsen O.-E.* Exponentially decreasing distributions for the logarithm of particle size // Proc. Roy. Soc. Lond. A, 1977. Vol. 353. P. 401–419.
4. *Barndorff-Nielsen O.-E.* Hyperbolic distributions and distributions of hyperbolae // Scand. J. Statist., 1978. Vol. 5. P. 151–157.
5. *Barndorff-Nielsen O.-E., Kent J., Sørensen M.* Normal variance-mean mixtures and z-distributions // Int. Statist. Rev., 1982. Vol. 50. No. 2. P. 145–159.
6. Королев В. Ю., Соколов И. А. Скошенные распределения Стьюдента, дисперсионные гамма-распределения и их обобщения как асимптотические аппроксимации // Информатика и её применения, 2012. Т. 6. Вып. 1. С. 2–10.
7. *Закс Л. М., Королев В. Ю.* Обобщенные дисперсионные гамма-распределения как предельные для случайных сумм // Информатика и её применения, 2013. Т. 7. Вып. 1. С. 105–115.
8. *Королев В. Ю.* Обобщенные гиперболические распределения как предельные для случайных сумм // Теория вероятностей и ее применения, 2013. Т. 58. Вып. 1. С. 117–132.
9. *Королев В. Ю., Черток А. В., Корчагин А. Ю., Горшенин А. К.* Вероятностно-статистическое моделирование информационных потоков в сложных финансовых системах на основе высокочастотных данных // Информатика и её применения, 2013. Т. 7. Вып. 1. С. 12–21.
10. *Protassov R. S.* EM-based maximum likelihood parameter estimation for a multivariate generalized hyperbolic distribution with fixed λ // Statistics Computing, 2004. Vol. 14. P. 67–77.
11. *Королев В. Ю., Назаров А. Л.* Разделение смесей вероятностных распределений при помощи сеточных методов моментов и максимального правдоподобия // Автоматика и телемеханика, 2010. Вып. 3. С. 98–116.
12. *Dennis J. E., Schnabel R. B.* Numerical methods for unconstrained optimization and nonlinear equations. — Englewood Cliffs: Prentice-Hall, 1983. 378 p.

Поступила в редакцию 01.10.14

A MODIFIED GRID METHOD FOR STATISTICAL SEPARATION OF NORMAL VARIANCE-MEAN MIXTURES

V. Yu. Korolev[1,2] and A. Yu. Korchagin[1]

[1]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

Abstract: A modified two-stage grid method for statistical separation of normal variance-mean mixtures is described as an alternative to a pure EM (expectation-maximization) algorithm. At the first stage of this algorithm, a discrete approximation is constructed to the mixing distribution. At the second stage, the obtained discrete distribution is approximated by an absolutely continuous distribution from a predetermined family, say, by a generalized inverse Gaussian distribution. The convergence of this two-stage procedure is discussed. The monotonicity of the grid procedure used at the first stage is proved. The problem of the optimal choice of the parameters of the method is discussed in detail. First of all, the problem of the optimal choice of the grid thrown on the support of the mixing distribution is considered. Statistical estimators are proposed for the quantiles of the mixing law. The efficiency of the method is illustrated by examples of its application to the estimation of the parameters of generalized hyperbolic distributions.

Keywords: mixture of probability distributions; normal variance-mean mixture; generalized hyperbolic distribution; EM-algorithm; grid method of separation of mixtures

DOI: 10.14357/19922264140402

Acknowledgments

The research was supported by the Russian Science Foundation (project 14-11-00364).

References

1. Korolev, V. Yu. 2011. *Veroyatnostno-statisticheskie metody dekompozitsii volatil'nosti khaoticheskikh protsessov* [Probabilistic and statistical methods for the decomposition of volatility of chaotic processes]. Moscow: Moscow University Press. 510 p.
2. Nazarov, A. L. 2013. Priblizhennye metody razdeleniya smesey veroyatnostnykh raspredeleniy [Approximate methods for the decomposition of volatility of chaotic processes]. Ph.D. Thesis. Moscow: Moscow State University.
3. Barndorff-Nielsen, O. E. 1977. Exponentially decreasing distributions for the logarithm of particle size. *Proc. Roy. Soc. Lond. A* 353:401–419.
4. Barndorff-Nielsen, O. E. 1978. Hyperbolic distributions and distributions of hyperbolae. *Scand. J. Statist.* 5:151–157.
5. Barndorff-Nielsen, O. E., J. Kent, and M. Sørensen. 1982. Normal variance-mean mixtures and z-distributions. *Int. Statist. Rev.* 50(2):145–159.
6. Korolev, V. Yu., and I. A. Sokolov. 2012. Skoshennye raspredeleniya St'yudenta, dispersionnye gamma-raspredeleniya i ikh obobshcheniya kak asimptoticheskie approksimatsii [Skewed Student's distributions, variance gamma distributions, and their generalizations as asymptotic approximations]. *Informatika i ee Primeneniya — Inform. Appl.* 6(1):2–10.
7. Korolev, V. Yu., and L. M. Zaks. 2013. Obobshchennye dispersionnye gamma-raspredeleniya kak predel'nye dlya sluchaynykh summ [Generalized variance gamma distributions as limiting for random sums]. *Informatika i ee Primeneniya — Inform. Appl.* 7(1):105–115.
8. Korolev, V. Yu. 2013. Obobshchennye giperbolicheskie raspredeleniya kak predel'nye dlya sluchaynykh summ [Generalized hyperbolic distributions as limiting for random sums] *Theory Probab. Appl.* 58(1):117–132.
9. Korolev, V. Yu., A. V. Chertok, A. Yu. Korchagin, and A. K. Gorshenin. 2013. Veroyatnostno-statisticheskoe modelirovanie informatsionnykh potokov v slozhnykh finansovykh sistemakh na osnove vysokochastotnykh dannykh [Probability and statistical modeling of information flows in complex financial systems from high-frequency data]. *Informatika i ee Primeneniya — Inform. Appl.* 7(1):12–21.
10. Protassov, R. S. 2004. EM-based maximum likelihood parameter estimation for a multivariate generalized hyperbolic distribution with fixed λ. *Statistics Computing* 14:67–77.

11. Korolev, V. Yu., and A. L. Nazarov. 2010. Razdelenie smesey veroyatnostnykh raspredeleniy pri pomoshchi setochnykh metodov momentov i maksimal'nogo pravdopodobiya [Separation of mixtures using grid moment-based methods and maximum likelihood]. *Avtomatika i Telemekhanika* [Automatics and Telemechanics] 3:98–116.
12. Dennis, J. E., and R. B. Schnabel. 1983. *Numerical methods for unconstrained optimization and nonlinear equations*. Englewood Cliffs: Prentice-Hall. 378 p.

Received October 01, 2014

Contributors

Korolev Victor Yu. (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; leading scientist, Institute of Informatics Problems, Russian Academy of Sciences, 44-2 Vavilov Str., Moscow 119333, Russian Federation; victoryukorolev@yandex.ru

Korchagin Alexander Yu. (b. 1989) — PhD student, Faculty of Computational Mathematics and Cybernetics, M. V. Lomonosov Moscow State University, 1-52 Leninskiye Gory, GSP-1, Moscow 119991, Russian Federation; sasha.korchagin@gmail.com

ИНФОРМАТИКА И ЕЁ ПРИМЕНЕНИЯ, 2014. Т. 8. Вып. 4. С. 20–31

О ФОРМАЛИЗАЦИИ ПОНЯТИЯ ТОКСИЧНОСТИ ПОТОКА ЗАЯВОК НА ФИНАНСОВЫХ РЫНКАХ*

А. В. Черток[1]

Аннотация: Рассматривается микроструктурная модель потоков заявок на финансовых рынках. В качестве интегрального индикатора текущего состояния книги заявок используется дисбаланс потока заявок. Для анализа свойств текущего состояния книги заявок используется модель дисбаланса потока заявок, имеющая вид двустороннего процесса риска, известного в актуарной математике как процесс риска со случайными премиями. Исследуется понятие токсичности потока заявок на финансовых рынках. Понятие токсичности потока заявок на финансовых рынках формализуется с помощью вероятностей пересечения процессом дисбаланса потоков заявок фиксированных уровней. Вводятся понятия мгновенного профиля токсичности и байесовского и квантильного показателей токсичности. Эти показатели рассчитываются для двух модельных типов потоков заявок, в первом из которых заявки имеют единичный объем, во втором — объем заявок является случайным и имеющим показательное распределение.

Ключевые слова: финансовые рынки; книга заявок; поток заявок; дисбаланс потока заявок; неблагоприятный отбор; токсичность; пуассоновский процесс; обобщенный пуассоновский процесс; двусторонний процесс риска; процесс риска со случайными премиями; вероятность разорения

DOI: 10.14357/19922264140403

1 Введение

Активное развитие электронной торговли на финансовых рынках выявило необходимость анализа биржевых высокочастотных данных для более глубокого понимания рыночной микроструктуры, на которую оказали огромное влияние компании, занимающиеся автоматизированным высокочастотным трейдингом (они формируют до 70%–80% дневного оборота на ведущих мировых площадках). Эти высокочастотные системы, как правило, являются маркет-мейкерами — поставщиками ликвидности посредством размещения пассивных (лимитных) заявок на различных уровнях электронной книги заявок. Поставщик ликвидности, выставивший пассивную заявку, не имеет возможности влиять на время ее исполнения (разумеется, кроме как снять заявку). Маркет-мейкеры зачастую не прогнозируют в явном виде динамику рынка, а используют шумовую составляющую рыночных движений.

Степень эффективности деятельности маркет-мейкеров связана с контролем риска оказаться с большим количеством купленных или проданных контрактов, что напрямую зависит от их способности контролировать эффект неблагоприятного отбора (adverse selection) в отношении пассивных заявок.

Практики, как правило, описывают принцип неблагоприятного отбора как «естественную тенденцию слишком быстрого исполнения пассивных заявок в тех ситуациях, когда они должны исполняться медленно, и наоборот: исполняться слишком медленно в тех ситуациях, когда они должны исполниться быстро» [1]. Эта интуитивная формулировка согласуется с ранними микроструктурными моделями рынка [2–4], в которых информированные трейдеры получают преимущество над неинформированными участниками рынка. Поток заявок считается токсичным, когда происходит эффект неблагоприятного отбора маркет-мейкеров, поставляющих ликвидность.

В работе [5] предложена эмпирическая процедура оценки токсичности потока заявок на основе анализа информации о *сделках*. В предлагаемой статье рассматривается более точный подход к измерению токсичности рынка, использующий всю доступную информацию о *потоке заявок* (не только сами сделки, но также и постановки/снятия заявок) на основе аналитической модели процесса дисбаланса потока заявок, рассмотренной ранее в работах [6, 7].

*Работа выполнена при частичной поддержке РФФИ (проект 14-07-00041а).

[1]Факультет вычислительной математики и кибернетики Московского государственного университета им. М. В. Ломоносова; Euphoria Group LLC; a.v.chertok@gmail.com

2 Модель потока заявок

2.1 Терминология

На электронных рынках биржевая цена финансового инструмента в ее классическом понимании является результирующей, интегральной характеристикой системы торгов, которая описывается динамикой так называемой *книги заявок* (*limit order book*), представляющей собой информацию о всех актуальных на данный момент предложениях о покупке и продаже инструмента по различным ценам (рис. 1).

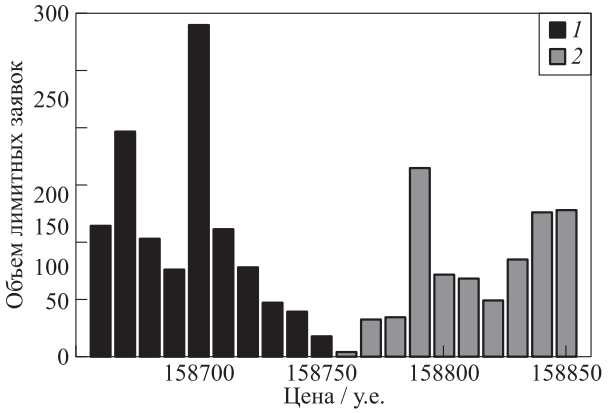


Рис. 1 Книга заявок в некоторый момент времени. Высота столбиков равна суммарному объему лимитных заявок на соответствующем ценовом уровне: *1* — покупки; *2* — продажи

Динамика книги заявок определяется тремя типами заявок, которые участники рынка могут отправить на рынок:

(1) *лимитная* заявка обозначает желание купить (продать) заданное количество акций по цене не выше (не ниже) заданной, при этом такая заявка немедленно добавляется в книгу заявок;

(2) *рыночная* заявка обозначает желание купить или продать заданное количество акций по лучшей цене, представленной в книге заявок, после чего немедленно происходит сделка;

(3) заявка *на отмену* обозначает намерение отменить существующую лимитную заявку, после чего она удаляется из книги заявок.

Более формально, в каждый момент времени информация о первых $d = 5$ уровнях книги заявок представляет собой массив

$$\text{book} = \left(b_1, a_1, v_1^b, v_2^b, \ldots, v_{10}^b, v_1^a, v_2^a, \ldots, v_{10}^a\right) ,$$

где b_1 — лучшая цена на покупку (бид) на текущий момент (кратная минимальному шагу цены δ); a_1 — лучшая цена на продажу (аск) на текущий момент (кратная минимальному шагу цены δ); $v_i^b \geqslant 0$ — суммарный объем заявок по цене b_i (при этом автоматически $b_i = b_1 - (i-1)\delta$); $v_i^a \geqslant 0$ — суммарный объем заявок по цене a_i (при этом автоматически $a_i = a_1 + (i-1)\delta$).

Всегда выполняется условие $b_1 < a_1$, так как иначе соответствующие заявки должны быть сведены в сделку, величина $p = (1/2)(b_1 + a_1)$ обычно называется *мидпрайсом*, а величина $s = a_1 - b_1$ называется *спредом*.

2.2 Динамика книги заявок

Потоки заявок моделируются с использованием независимых пуассоновских процессов — процессов восстановления с экспоненциальными распределениями интервалов между восстановлениями (как это сделано, например, в работах [8, 9]):

– лимитные заявки на покупку (продажу) приходят на ценовой уровень, расположенный на расстоянии i от лучшей котировки противоположного типа, в независимые моменты времени, имеющие экспоненциальное распределение с параметром $\lambda_i^+(\lambda_i^-)$ (эмпирические исследования [10, 11] показывают, что степенный закон $\lambda_i^{\pm} = k/i^{\alpha}$ является хорошей аппроксимацией);

– рыночные заявки на покупку (продажу) приходят в независимые моменты времени, имеющие экспоненциальное распределение с параметром $\mu^+(\mu^-)$;

– заявки на отмену лимитного ордера на покупку (продажу), находящегося на дистанции i от лучшей котировки того же типа, приходят с частотой $\theta_i^+(\theta_i^-)$.

Рассмотрим два пуассоновских процесса $N^+(t)$ и $N^-(t)$ с интенсивностями соответственно

$$\lambda^+ = \mu^+ + \sum_i \lambda_i^+ + \sum_i \theta_i^- ;$$
$$\lambda^- = \mu^- + \sum_i \lambda_i^- + \sum_i \theta_i^+ .$$

По своей сути процессы $N^+(t)$ и $N^-(t)$ соответствуют информации о числе заявок от покупателей и продавцов соответственно, пришедших к моменту времени t. Будем также считать, что объемы заявок от покупателей и продавцов — независимые одинаково распределенные величины X_i^+ и X_i^- с функциями распределения $G(t)$ и $F(t)$ соответственно и не зависят от процессов $N^+(t)$ и $N^-(t)$.

3 Процесс дисбаланса потока заявок и его связь с ценой

Понятие дисбаланса потока заявок введено в работе [12], окончательный вариант которой [13] опубликован в 2014 г. В работах [6, 7] независимо этот же процесс исследовался под названием *процесс обобщенной цены*.

В работах [6, 7] в качестве математической модели эволюции процесса дисбаланса потока заявок было предложено использовать двусторонний процесс риска — специальный обобщенный (compound) пуассоновский процесс. Следуя этому подходу, зафиксируем малый интервал времени $[0;T]$, в течение которого параметры распределений, описывающих объемы заявок, и интенсивности потоков заявок одного типа остаются постоянными и известными. Для $t \in [0,T]$ пусть $N^+(t)$ и $N^-(t)$ — количества заявок, пришедших от покупателей и продавцов соответственно в течение интервала времени $[0,t]$ — независимые пуассоновские процессы с интенсивностями $\lambda^+ > 0$ и $\lambda^- > 0$ ($\mathsf{E}N^+(t) = \lambda^+ t$, $N^+(0) = 0$, $\mathsf{E}N^-(t) = \lambda^- t$, $N^-(0) = 0$). Пусть X_i^+ и X_i^-, $i = 1,2,\ldots,$ — объемы заявок, поступающих от покупателей и продавцов соответственно – две независимые последовательности независимых и одинаково в каждой последовательности распределенных случайных величин с функциями распределения $G(x)$ и $F(x)$ соответственно, независимых от пуассоновских процессов $N^+(t)$ и $N^-(t)$. Положим

$$Q^+(t) = \sum_{i=1}^{N^+(t)} X_i^+ ; \;\; Q^-(t) = \sum_{j=1}^{N^-(t)} X_j^-$$

и определим процесс *дисбаланса потока заявок* $Q(t)$ как

$$Q(t) = Q^+(t) - Q^-(t)\,.$$

Этот процесс является намного более чувствительным индикатором (показателем) текущего состояния книги заявок, поскольку интервалы времени между последовательными изменениями состояний книги заявок обычно так малы, что изменения цены (мидпрайса) по сравнению с ними являются редкими событиями. Поэтому процесс цены является намного более грубым показателем, характеризующим книгу заявок, и дает грубое и весьма ограниченное описание динамики рынка. Вместе с тем процесс дисбаланса потоков заявок учитывает не только текущие значения наилучших цен покупки и продажи, но и влияние событий «в глубине» книги заявок и потому меняется существенно быстрее и позволяет интерполировать динамику рынка между изменениями цены, в частности отслеживать ситуации, связанные с токсичностью потоков заявок, т. е. чреватые необоснованными трендами в поведении цены (рис. 2).

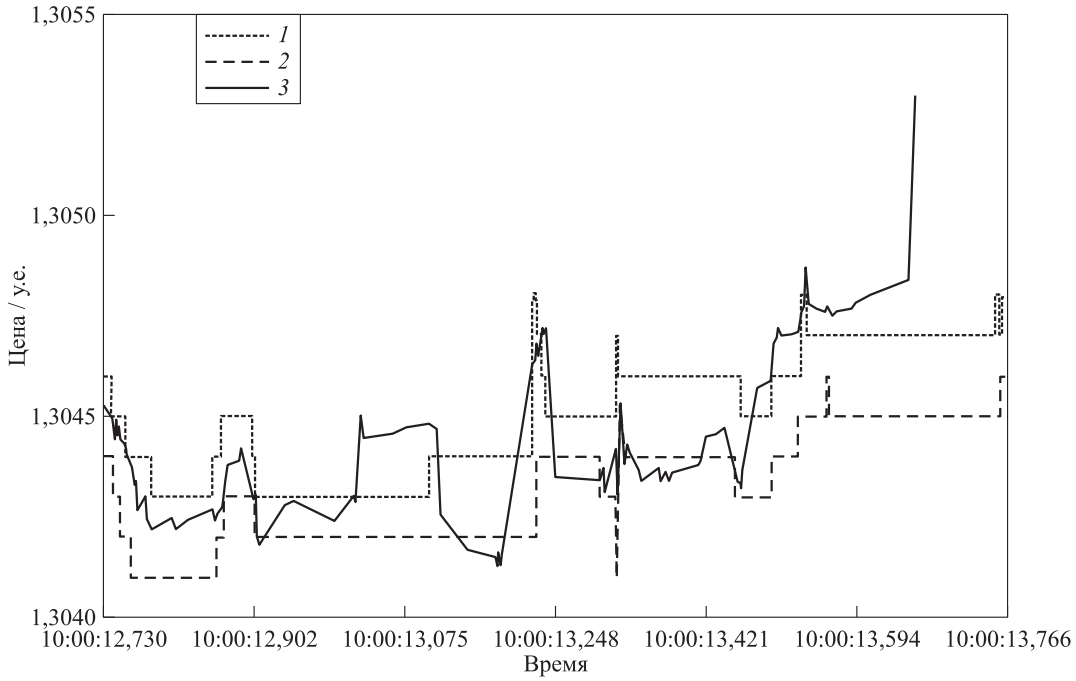


Рис. 2 Динамика лучшей цены покупки (*1*), лучшей цены продажи (*2*) и процесса дисбаланса потока заявок $Q(t)$ в течение 1 с (*3*) с момента 10:00:12,730 01.07.2014 (фьючерс на индекс РТС)

В работе [13] с помощью линейной модели

$$\frac{S(t+\Delta)-S(t)}{\delta} = c\frac{Q(t,t+\Delta)}{D(t)} + \epsilon(t)$$

было показано, что процесс дисбаланса потока заявок $Q(t)$ имеет сильную связь с высокочастотными изменениями цены финансового актива $S(t)$, построенной по ценам сделок, где δ — минимальный шаг цены (тик цены), $\epsilon(t)$ — белый шум и $D(t)$ — мера глубины книги заявок (количество заявок на лучшем биде/аске). Эмпирический анализ высокочастотных данных для американских акций подтверждает наличие линейной связи: коэффициент c оценивается между 0,1 и 1 и оказывается статистически значим в 98% случаев. Наличие такого рода связи позволяет напрямую исследовать свойства процесса дисбаланса потока заявок и соотносить их со свойствами процесса цены $S(t)$.

В работах [6, 7, 14] с помощью предельных теорем для двусторонних процессов риска были получены асимптотические аппроксимации для процесса дисбаланса потока заявок $Q(t)$ и его распределений.

4 Профиль мгновенной токсичности потока заявок

Как уже было сказано выше, поток заявок считается токсичным, когда он оказывается неблагоприятным для маркет-мейкеров, предоставляющих ликвидность в книге заявок. В работе [5] предложена процедура оценки токсичности потока заявок на основе анализа информации об интенсивности и направлении *сделок* (направление сделки определяется в зависимости от того, кто являлся ее инициатором — покупатель или продавец). В данной работе будет рассмотрен более точный подход к измерению токсичности потока заявок, использующий всю доступную информацию о заявках (не только сами сделки, но также и постановки/снятия заявок).

Чтобы формализовать понятие токсичности потока заявок, для начала рассмотрим процесс дисбаланса потока заявок $Q(t) = Q^+(t) - Q^-(t)$ в предположении, что $\mathsf{E}Q(t) > 0$, т. е.

$$\lambda^+ \mathsf{E}X_1^+ > \lambda^- \mathsf{E}X_1^-,$$

что означает преимущество покупателей над продавцами в рамках интервала $[0,T]$. Предположим, что $Q(0) = 0$.

Для $u > 0$ рассмотрим вероятность

$$\varphi_\pm(u,T) = \mathsf{P}\left(\inf_{0<t\leqslant T} Q(t) \geqslant -u\right),$$

т. е. вероятность того, что траектория процесса $Q(t)$ в течение интервала времени $[0,T]$ целиком будет находиться не ниже уровня $-u$, а также аналогичную предельную вероятность на бесконечном интервале времени:

$$\varphi_\pm(u) = \mathsf{P}\left(\inf_{t>0}\left(Q^+(t) - Q^-(t)\right) \geqslant -u\right) = \lim_{T\to\infty} \varphi_\pm(u,T).$$

Вероятность $\varphi_\pm(u)$ описывает вероятность того, что при *положительном* тренде процесс дисбаланса никогда не достигнет *отрицательного* уровня $-u$ при условии, что параметры потока заявок (λ^+, λ^-, $G(x)$ и $F(x)$) остаются неизменными.

Определение 1. Функцию $\varphi_\pm(u)$ будем называть *профилем мгновенной токсичности* потока заявок.

Введенная таким образом характеристика — профиль мгновенной токсичности потока заявок — формально совпадает с *вероятностью неразорения* в классической модели коллективного риска со случайными премиями, рассматривавшейся, например, в работах [15–17]. В некоторых источниках (см., в частности, [18]) справедливо отмечено, что интерпретация этого показателя именно как вероятности физического разорения страховой компании некорректна, поскольку изначальное предположение о неизменности основных параметров потоков страховых премий и страховых выплат в течение бесконечного интервала времени заведомо не выполняется. Тем не менее эта характеристика является удобным показателем текущей эффективности работы страховой компании и имеет смысл некоей оценки качества текущего состояния параметров страховой деятельности. Точно так же профиль мгновенной токсичности потока заявок является удобно интерпретируемым показателем неустойчивости текущего состояния потоков заявок.

Из работ [15, 16] следует

Лемма 1. *Функция профиля мгновенной токсичности потока заявок $\varphi_\pm(u)$ удовлетворяет интегральному уравнению*

$$\left(\lambda^+ + \lambda^-\right)\varphi_\pm(u) = \lambda^- \int_0^u \varphi_\pm(u-v)\,dF(v) + \lambda^+ \int_0^\infty \varphi_\pm(u+v)\,dG(v).$$

Если R — решение характеристического уравнения

$$\lambda^+\left(\mathsf{E}e^{-RX_1^+} - 1\right) + \lambda^-\left(\mathsf{E}e^{RX_1^-} - 1\right) = 0,$$

то

$$\varphi_{\pm}(u) = \frac{e^{-Ru}}{\mathsf{E}\{\, e^{-RQ(t)} \Big| \, \tau < \infty\}}\,,$$

при этом $\varphi_{\pm}(u) \geqslant 1 - e^{-Ru}$.

5 Токсичность потока заявок

Профиль токсичности представляет собой *функцию*, аргументом которой является уровень u. Это затрудняет сравнение токсичности потоков заявок на разных участков рынка, поскольку, вообще говоря, в множестве функций нельзя ввести отношение порядка. Поэтому хотелось бы иметь некий интегральный показатель токсичности, выражаемый одним числом. Для построения такого показателя можно воспользоваться одним из двух подходов.

5.1 Байесовский подход

Выделим некий «характеристический» уровень u_0, пересечение которого может иметь серьезные последствия. Пусть $w(x)$ — некоторая плотность распределения вероятностей, обладающая свойствами

$$\int\limits_0^\infty w(x)\,dx = 1\,; \quad \int\limits_0^\infty xw(x)\,dx = u_0\,. \tag{1}$$

Определение 2. Байесовским показателем *мгновенной токсичности* потока заявок $\theta_{\pm}^{(w)}$ называется величина

$$\theta_{\pm}^{(w)} = \theta_{\pm}^{(w)}(u_0) = \int\limits_0^\infty \varphi_{\pm}(u)w(u)\,du\,.$$

По сути показатель мгновенной токсичности потока заявок $\theta_{\pm}$ есть математическое ожидание «случайного» профиля мгновенной токсичности $\varphi_{\pm}(U)$, где U — неотрицательная случайная величина с плотностью распределения $w(x)$ и имеющая математическое ожидание u_0.

В случае, когда $\mathsf{E}Q(t) < 0$, т. е. $\lambda^+\mathsf{E}X_1^+ < \lambda^-\mathsf{E}X_1^-$, что означает преимущество продавцов над покупателями на интервале $[0,T]$, вместо $\varphi_{\pm}(u)$ будем рассматривать вероятность

$$\varphi_{\mp}(u) = \mathsf{P}(\sup_{t>0}\big(Q^+(t) - Q^-(t)\big) \leqslant u) = \lim_{T\to\infty} \varphi_{\mp}(u,T)\,,$$

которая описывает вероятность того, что при *отрицательном* тренде траектория процесса $Q(t)$ не превысит *положительный* уровень u при условии, что параметры потока заявок (λ^+, λ^-, $G(x)$ и $F(x)$) остаются неизменными.

В таком случае в качестве байесовского показателя **мгновенной токсичности** потока заявок θ возьмем величину

$$\theta_{\mp}^{(w)} = \theta_{\mp}^{(w)}(u_0) = \int\limits_0^\infty \varphi_{\mp}(u)w(u)\,du\,.$$

5.2 Квантильный подход

При условии $\mathsf{E}Q(t) > 0$ на промежутке $[0,T]$ зафиксируем некоторое $0 < \alpha < 1$.

Определение 3. Квантильным α-показателем мгновенной токсичности потока заявок называется такое минимальное значение $q_{\pm}$, при котором $\varphi_{\pm}(q_{\pm}) \geqslant \alpha$.

Таким образом, при наличии положительного тренда у процесса $Q(t)$ квантильный α-показатель мгновенной токсичности — это настолько минимальное значение $q_{\pm}$, что вероятность того, что траектория процесса $Q(t)$ на интервале $[0,T]$ целиком пройдет выше уровня $-q_{\pm}$, больше или равна α. Чем больше значение $q_{\pm}$, тем более токсичен поток заявок от покупателей.

По аналогии с предыдущим пунктом при наличии у процесса $Q(t)$ отрицательного тренда (т. е. при условии $\lambda^+\mathsf{E}X_1^+ < \lambda^-\mathsf{E}X_1^-$) α-квантильный показатель мгновенной токсичности $q_{\mp}$ определяется из уравнения $\varphi_{\mp}(q_{\mp}) \geqslant \alpha$. Чем больше значение $q_{\mp}$, тем более токсичен поток заявок от продавцов на интервале $[0,T]$.

6 Модели потоков заявок

В некоторых случаях удается напрямую вычислить профиль мгновенной токсичности потока заявок. Аналоги моделей, приведенных ниже, рассмотрены в работе [15] в рамках модели Крамера–Лундберга со стохастическими премиями.

6.1 Модель рынка с заявками единичного объема

Рассмотрим простейшую модель рынка, где потоки заявок имеют единичный объем, т. е.

$$\mathsf{P}(X_i^+ = 1) = \mathsf{P}(X_i^- = 1) = 1\,.$$

В этом случае

$$Q(t) = \sum_{i=1}^{N^+(t)} 1 - \sum_{i=1}^{N^-(t)} 1 = N^+(t) - N^-(t)\,.$$

Несмотря на очевидно идеальный характер такого примера, он имеет реальный практический смысл, поскольку при этом становится возможным учитывать чистые интенсивности потоков заявок и отслеживать влияние их соотношения (дисбаланса интенсивностей потоков заявок) на токсичность ситуации. Более того, в таком случае рассматриваемый процесс дисбаланса потоков заявок $Q(t)$ является простейшим процессом рождения и гибели, различные характеристики которого можно исследовать специально разработанными для этого методами.

Если $\lambda_+ > \lambda_-$, то покупатели преобладают над продавцами и характеристическое уравнение имеет вид:

$$\lambda^+ \left[e^{-R} - 1\right] + \lambda^- \left[e^R - 1\right] = 0\,,$$

откуда $e^R = \lambda^+/\lambda^-$ или $e^R = 1$. По лемме 1 для $u > 0$ имеем

$$\varphi_\pm(u) \geqslant 1 - \left(\frac{\lambda^-}{\lambda^+}\right)^u \;;\;\; \varphi_\pm(\infty) = 1\,.$$

Равенство $\varphi_\pm(u) = \varphi_\pm([u])$ очевидно. Для целых u интегральное уравнение переходит в разностное:

$$\lambda_1\varphi_\pm(u{+}1) - \left(\lambda^- + \lambda^+\right)\varphi_\pm(u) + \lambda\varphi_\pm(u{-}1) = 0\,, \quad (2)$$

откуда

$$\varphi_\pm(u) = C_1 + C_2\left(\frac{\lambda^-}{\lambda^+}\right)^u \;,\;\; C_1 = \varphi(\infty) = 1\,.$$

Константу C_2 найдем при подстановке в уравнение (2) $u = 0$:

$$\left(\lambda^- + \lambda^+\right)\varphi(0) = \lambda\varphi(1)\,,\;\; C_2 = -\frac{\lambda^-}{\lambda^+}\,,$$

откуда получаем, что для $u > 0$ профиль мгновенной токсичности имеет вид:

$$\varphi_\pm(u) = \varphi_\pm([u]) = 1 - \left(\frac{\lambda^-}{\lambda^+}\right)^{[u]+1} .$$

6.1.1 Байесовский показатель мгновенной токсичности

Коль скоро исследователь может сам назначать уровень u_0, относительно которого будут рассчитываться характеристики токсичности потока заявок, будем рассматривать u_0 на множестве натуральных чисел, а в качестве функции $w(u)$ можно выбрать функцию плотности вероятности распределения Пуассона (относительно считающей меры), также определенную на множестве натуральных чисел. Для удобства обозначим $r = \lambda^-/\lambda^+$, при этом $r < 1$ (рис. 3). В таком случае

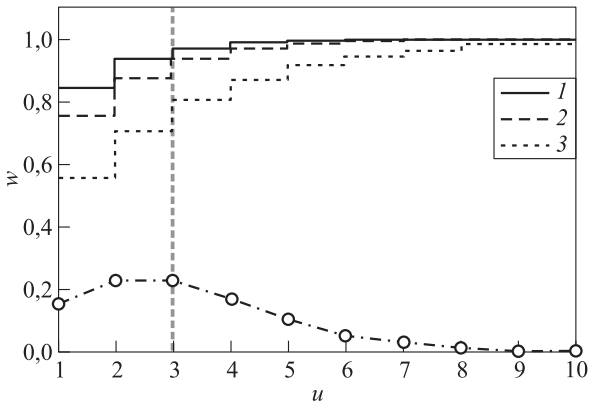


Рис. 3 Функция профиля токсичности потока заявок $\varphi_\pm(u)$ в модели рынка с единичными потоками заявок для разных значений $r = \lambda^-/\lambda^+$: *1* — $r = 2/5$; *2* — $1/2$; *3* — $r = 2/3$. Штрихпунктирная кривая: функция $w(u)$ — плотность (относительно считающей меры) пуассоновского распределения со средним $u_0 = 3$

$$\theta_\pm^{(w)}(u_0) = \int\limits_{\mathbb{N}} \varphi_\pm(u)w(u)\,du =$$

$$= \sum_{k=0}^{\infty}(1 - r^{k+1})\frac{u_0^k e^{-u_0}}{k!} = 1 - re^{-u_0}\sum_{k=0}^{\infty}\frac{(ru_0)^k}{k!} =$$

$$= 1 - re^{u_0(r-1)}\,.$$

На рис. 4, *а* изображен график токсичности в зависимости от значений $r = \lambda^-/\lambda^+$ для фиксированного значения $u_0 = 3$ в условиях положительного тренда ($\lambda^+ > \lambda^-$). Чем меньше значение r, тем токсичнее рынок. И напротив: рынок нетоксичен, когда $\lambda^+ = \lambda^-$, т. е. наблюдается баланс между покупателями и продавцами.

6.1.2 Квантильный показатель мгновенной токсичности

Для заданного $\alpha \in (0, 1)$ квантильный показатель мгновенной токсичности — это такое минимальное $q_\pm \in \mathbb{N}$, при котором

$$\varphi_\pm(q_\pm) = 1 - r^{q_\pm+1} \geqslant \alpha\,,$$

откуда

$$q_\pm(\alpha) = \left\lceil \frac{\ln(1-\alpha)}{\ln r} - 1 \right\rceil .$$

Заметим, что при $r = 1$ α-квантильный показатель мгновенной токсичности не определен и в таком случае полагается равным нулю.

На рис. 4, *б* на график нанесены различные значения квантильного показателя мгновенной токсичности потока заявок в зависимости от значения $r = \lambda^-/\lambda^+$ на промежутке $[0, T]$. Токсичность

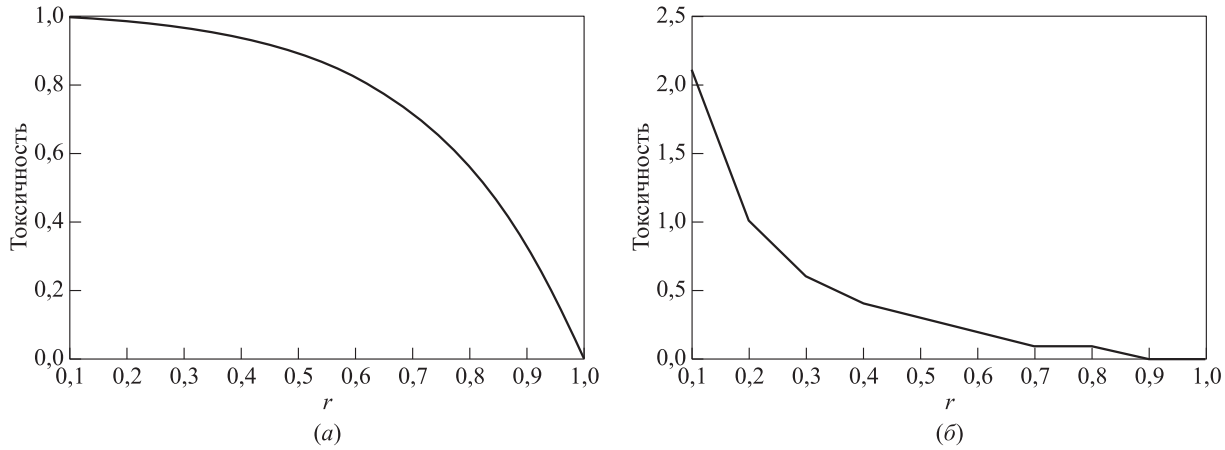


Рис. 4 Графики показателей токсичности в зависимости от значения $r = \lambda^-/\lambda^+$: (*а*) байесовский подход; (*б*) квантильный подход

покупателей максимальна при малых значениях r и близка к нулю при наличии баланса между покупателями и продавцами. Монотонность обоих графиков по r подтверждает обоснованность использования $\theta_\pm$ и $q_\pm$ в качестве показателей токсичности потока заявок в случае модели рынка с заявками единичного объема.

6.2 Модель рынка с экспоненциальными объемами заявок

Пусть объемы заявок покупателей и продавцов имеют экспоненциальное распределение (рис. 5), т. е.

$$G(t) = 1 - e^{-bt}; \quad F(t) = 1 - e^{-at}.$$

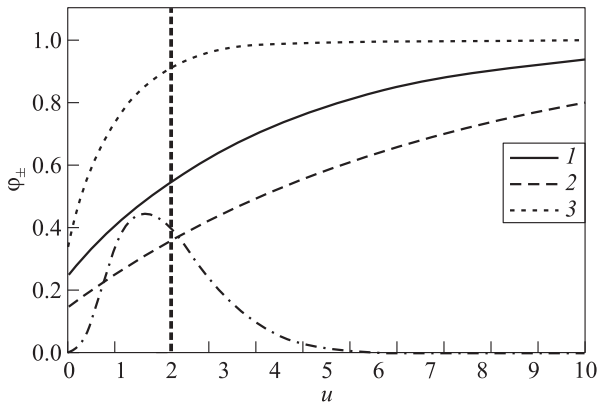


Рис. 5 Функция профиля токсичности $\varphi_\pm(u)$ в модели рынка с экспоненциальными объемами заявок для разных наборов $(\lambda^+, \lambda^-, b, a)$: *1* — (3, 1, 2, 1); *2* — (5, 2, 2, 1); *3* — (1, 1, 1, 3). Штрихпунктирная кривая: весовая функция $w(u)$ — плотность гамма-распределения $\Gamma(u_0^2, u_0^{-1})$ при $u_0 = 2$

В случае, когда покупатели преобладают над продавцами, т. е. $\lambda^+/b > \lambda^-/a$, характеристическое уравнение имеет вид:

$$\lambda^+ \left[\frac{b}{b+R} - 1\right] + \lambda^- \left[\frac{a}{a-R} - 1\right] = 0,$$

откуда $R = (\lambda^+ a - \lambda^- b)/(\lambda^+ + \lambda^-)$ или 0, а профиль мгновенной токсичности потока заявок [15]

$$\varphi_\pm(u) = \frac{(a+b)\lambda^-}{(\lambda^+ + \lambda^-)a} \exp\left(-\frac{\lambda^+ a - \lambda^- b}{\lambda^+ + \lambda^-}\, u\right).$$

В случае, когда продавцы преобладают над покупателями,

$$\varphi_\mp(u) = \frac{(a+b)\lambda^+}{(\lambda^+ + \lambda^-)b} \exp\left(-\frac{\lambda^- b - \lambda^+ a}{\lambda^+ + \lambda^-}\, u\right).$$

6.2.1 Байесовский показатель мгновенной токсичности

На множестве функций $w(u)$, удовлетворяющих условиям (1), рассмотрим функции, удовлетворяющие также условию

$$\int_0^\infty x^2 w(x)\, dx - \left(\int_0^\infty x w(x)\, dx\right)^2 = 1, \qquad (3)$$

т. е. обеспечивающие единичную дисперсию соответствующей случайной величины, имеющей функцию $w(u)$ в качестве плотности своего распределения.

Для вычисления байесовского показателя мгновенной токсичности возьмем в качестве $w(u)$ плотность гамма-распределения

$$w(u) = u^{k-1}\frac{e^{-u/\theta}}{\theta^k\,\Gamma(k)}\,,$$

где $\Gamma(k)$ — гамма-функция Эйлера:

$$\Gamma(k) = \int\limits_0^{+\infty} t^{k-1}e^{-t}\,dt\,.$$

Поскольку математическое ожидание и дисперсия случайной величины U, имеющей гамма-распределение, равны $k\theta$ и $k\theta^2$ соответственно, то с учетом условий (1) и (3) значения k и θ определяются из уравнений $k\theta = u_0$ и $k\theta^2 = 1$, откуда $k = u_0^2$ и $\theta = u_0^{-1}$.

Для удобства обозначим

$$\beta = \frac{(a+b)\lambda^-}{(\lambda^+ + \lambda^-)a} \text{ и } \gamma = \frac{\lambda^+ a - \lambda^- b}{\lambda^+ + \lambda^-} > 0\,. \qquad (4)$$

Байесовский показатель мгновенной токсичности равен

$$\theta_\pm^{(w)}(u_0) = \int\limits_0^\infty \varphi_\pm(u)w(u)\,du =$$

$$= \int\limits_0^\infty \left(1-\beta e^{-\gamma u}\right)u^{k-1}\frac{e^{-u/\theta}}{\theta^k\,\Gamma(k)}\,du =$$

$$= 1 - \frac{\beta}{\theta^k\Gamma(k)}\int\limits_0^\infty e^{-(\gamma+\theta^{-1})u}u^{k-1}\,du =$$

$$= \left[t = \frac{\theta\gamma+1}{\theta}u\right] = 1 -$$

$$- \frac{\beta}{(\theta\gamma+1)^k\Gamma(k)}\int\limits_0^\infty e^{-t}\frac{\theta^{k-1}}{(\theta\gamma+1)^{k-1}}\,t^{k-1}\frac{\theta}{\theta\gamma+1}\,dt =$$

$$= 1 - \frac{\beta}{(\theta\gamma+1)^k}\,.$$

После подстановки k и θ получаем значение показателя

$$\theta_\pm(u_0) = 1 - \frac{\beta}{\left(\gamma u_0^{-1}+1\right)^{u_0^2}}\,.$$

6.2.2 Квантильный показатель мгновенной токсичности

Для заданного $\alpha \in (0,1)$ квантильный показатель мгновенной токсичности — это такое минимальное $q_\pm$, при котором

$$\varphi_\pm(q_\pm) = 1 - \beta e^{-\gamma q_\pm} \geqslant \alpha\,.$$

Так как функция $\varphi_\pm$ является непрерывной по $q_\pm$, то данное неравенство может быть обращено в равенство, откуда получаем

$$q_\pm(\alpha) = \frac{\ln\beta - \ln(1-\alpha)}{\gamma}\,.$$

Заметим, что байесовский и квантильный показатели токсичности являются монотонными по каждой из величин β и γ.

7 Реальные данные

В данном разделе описывается структура данных о потоках заявок, на базе которых можно провести валидацию модели, предложенной в параграфе 6.2.

Далее оценим параметры потока заявок λ^+, λ^-, b и a в режиме скользящего окна и рассчитаем функции профиля мгновенной токсичности, а также показатели мгновенной токсичности потока заявок $\theta(t)$ и $q(t)$ в режиме реального времени и проанализируем адекватность полученных характеристик.

7.1 Описание данных

Рассматриваются данные о потоках всех заявок (лимитных, рыночных и заявок на отмену) на первые $d = 5$ уровней книги заявок фьючерса на индекс РТС (Российской торговой системы) за период с 1 по 30 июля 2014 г. Эти данные дают доступ к самой детальной информации о рыночных торгах в отличие от данных о сделках и котировках (TAQ, Trades and Quotes), которые часто используются для анализа высокочастотных данных и состоящих из цен и объемов сделок (что соответствует только рыночным заявкам в потоке всех заявок), а также информации о цене и объеме лучших котировок на покупку и продажу (т. е. только первый уровень книги заявок) с проставленными моментами времени.

В таблице приведен пример данных о потоке заявок для фьючерса на индекс РТС и о том, как выглядел срез книги заявок после прихода соответствующей заявки. Заметим, что на рынке FORTS присутствует всего два типа заявок: лимитные (L) и заявки на отмену (C), а механизм рыночных заявок участники рынка реализуют самостоятельно (отправляя лимитные заявки с ценами, гарантирующими их моментальное исполнение). Тем не менее имеется возможность оценить параметры потоков рыночных заявок в рамках предлагаемой модели, рассматривая для этого потоки лимитных заявок, которые приводили к сделкам.

Пример данных о потоке заявок для фьючерса на индекс РТС

Время	Тип	Опрерация	Цена, у.е.	Объем	b_1, у.е.	a_1, у.е.	v_1^b	v_1^a	v_2^b	v_2^a	v_3^b	v_3^a	v_4^b	v_4^a	v_5^b	v_5^a
10:02:36,444	L	Покупка	130 020	2	130040	130050	2	4	22	23	**54**	22	81	31	759	20
10:02:36,445	L	Продажа	130 070	1	130040	130050	2	4	22	23	**55**	22	81	31	759	20
10:02:36,465	C	Покупка	130 040	1	130040	130050	**1**	4	22	23	55	22	81	31	759	20
10:02:36,473	L	Покупка	130 050	3	130040	130050	1	**1**	22	23	55	22	81	31	759	20

7.2 Процедура оценки параметров

Разобьем один из рассматриваемых торговых дней (1 июля 2014 г.) на временны́е интервалы с шагом $\tau = 15$ с. При этом исключим интервалы времени в 5 мин торгов (с 10:00 до 10:05), а также в последние 5 мин торгов (с 18:40 до 18:45), поскольку они характеризуются аномальными всплесками волатильности, слабо поддающейся анализу в рамках представленной модели. Внутри каждого τ-интервала проведем оценку параметров λ^+, λ^-, b и a согласно модели рынка с экспоненциальными объемами заявок, предложенной в параграфе 6.2. Результат оценки параметров в режиме реального времени изображен на рис. 6.

7.3 Показатели токсичности

На основе оценок для λ^+, λ^-, b и a можно вычислить β и γ, а затем построить графики показателей мгновенной токсичности потока заявок $\theta(u_0)$ и $q(\alpha)$ для фиксированных u_0 и α в режиме реального времени (рис. 7) и идентифицировать участки, на которых деятельность покупателей или продавцов была токсичной. Прикладные исследования демонстрируют достаточную значимость данного показателя для своевременной идентификации участков неблагоприятного отбора маркет-мейкеров.

8 Заключение

В данной работе рассмотрена микроструктурная модель рынка, в которой потоки заявок моделируются пуассоновскими процессами с постоянными интенсивностями (такая аппроксимация возможна на небольших временны́х интервалах). В качестве интегрального индикатора текущего состояния книги заявок применялся дисбаланс потока заявок (order flow imbalance), который использует не только текущие значения наилучших цен покупки и продажи, но и влияние событий «в глубине» книги заявок и потому меняется существенно быстрее и позволяет интерполировать динамику рынка между изменениями цены, в частности отслеживать ситуации, связанные с токсичностью потока заявок. В рамках рассмотренной модели были введены такие понятия, как мгновенный профиль токсичности, а также байесовский и квантильний показатели токсичности, рассчитываемые на основе параметров, описывающих потоки всех заявок. Эти показатели рассчитываются для двух модельных типов потоков заявок, в первом из которых заявки имеют единичный объем, во втором — объем заявок является случайным и имеющим показательное распределение. Для последней из двух моделей была проведена валидация на реальных данных (фьючерс на индекс РТС) и были построены показатели токсичности в режиме реального

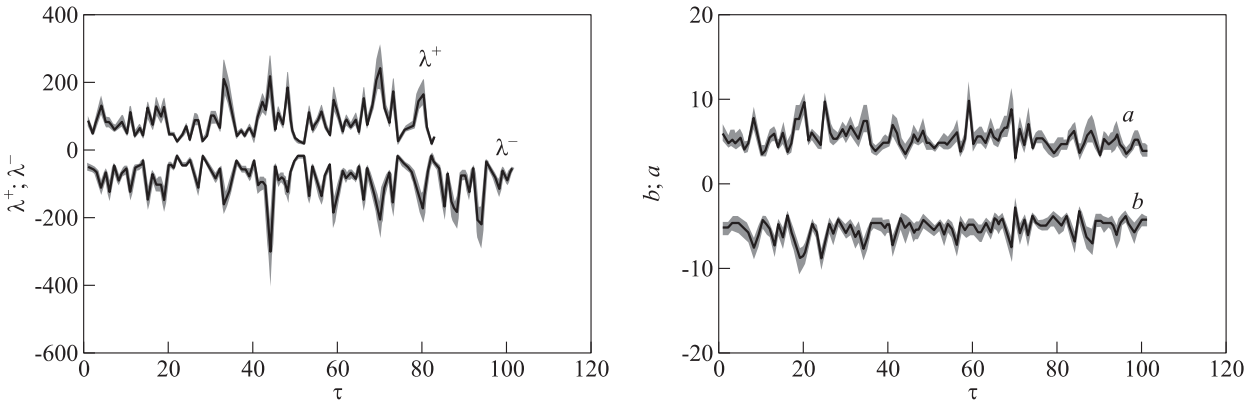


Рис. 6 Оценка параметров λ^+, λ^-, b и a в режиме реального времени (серый цвет — интервалы доверия), ось x — номер соответствующего τ-интервала (фьючерс на индекс РТС, дневная сессия 01.07.2014)

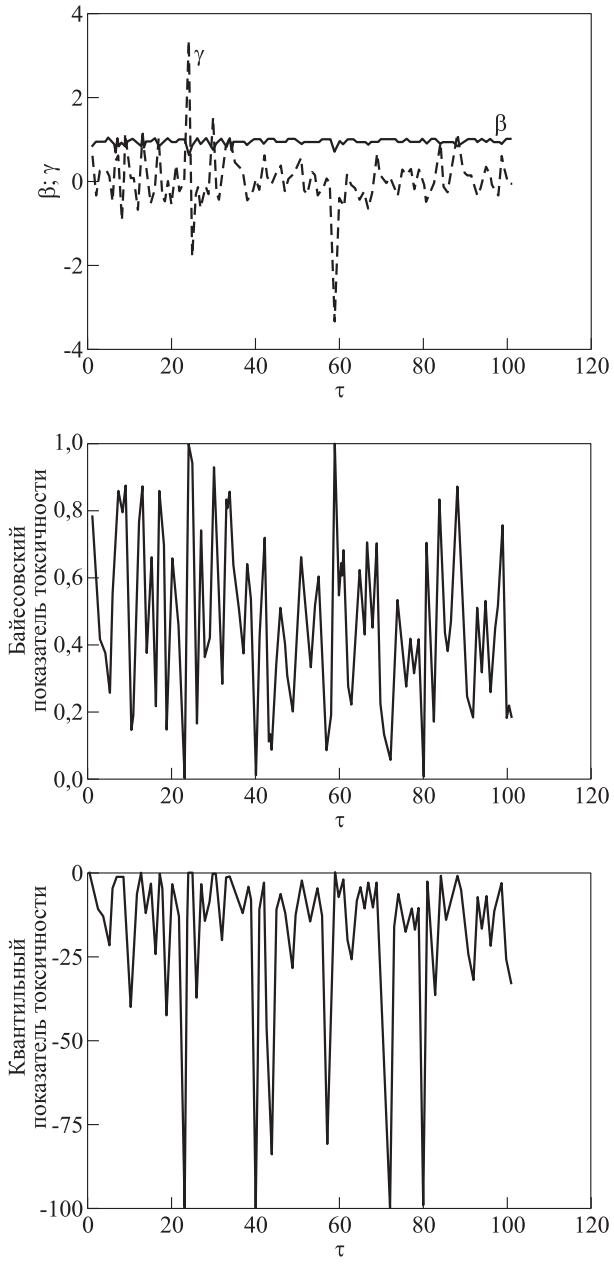


Рис. 7 Графики β и γ, рассчитанных по формулам (4), байесовского и квантильного показателей токсичности в режиме реального времени, ось x — номер сооветствующего τ-интервала (фьючерс на индекс РТС, дневная сессия 01.07.2014)

времени. Предложенная методика расчета показателей токсичности на основе информации о потоках всех заявок является перспективной и может быть распространена на модели рынка с неоднородными интенсивностями потоков заявок.

Автор статьи выражает огромную благодарность своему научному руководителю профессору Виктору Юрьевичу Королеву за ценные идеи и замечания, а также студентам факультета вычислительной математики и кибернетики МГУ им. М. В. Ломоносова Дарье Николайчук и Гелане Хазеевой за помощь в подготовке материала статьи.

Литература

1. *Jeria, D., Sofianos G.* Passive orders and natural adverse selection // Street Smart, September 4, 2008. No. 33.
2. *Glosten L. R., Milgrom P.* Bid, ask and transaction prices in a specialist market with heterogeneously informed traders // J. Financ. Econ., 1985. Vol. 14. P. 71–100.
3. *Kyle A. S.* Continuous auctions and insider trading // Econometrica, 1985. Vol. 53. P. 1315–1335.
4. *Easley D., O'Hara M.* Time and the process of security price adjustment // J. Financ., 1992. Vol. 47. P. 576–605.
5. *Easley D., Lopez de Prado M., O'Hara M.* Flow toxicity and liquidity in a high frequency world. *Rev. Financ. Stud.*, 2012. Vol. 25. No. 5. P. 1457–1493.
6. *Королев В. Ю., Черток А. В., Корчагин А. Ю., Горшенин А. К.* Вероятностно-статистическое моделирование информационных потоков в сложных финансовых системах на основе высокочастотных данных // Информатика и её применения, 2013. Т. 7. Вып. 1. С. 12–21.
7. *Chertok A., Korolev V., Korchagin A., Shorgin S.* Modeling high-frequency non-homogeneous order flows by compound Cox processes. January 14, 2014. http://ssrn.com/abstract=2378975.
8. *Cont R., Stoikov S., Talreja R.* A stochastic model for order book dynamics // Oper. Res., 2010. Vol. 58. No. 3. P. 549–563.
9. *Cont R., de Larrard A.* Price dynamics in a Markovian limit order market. Working paper. http://ssrn.com/abstract=1735338.
10. *Bouchaud J.-P., Mezard M., Potters M.* Statistical properties of stock order books: Empirical results and models // Quant. Financ., 2002. Vol. 2. P. 251–256.
11. *Zovko I., Farmer J. D.* The power of patience; A behavioral regularity in limit order placement // Quant. Financ., 2002. Vol. 2. P. 387–392.
12. *Cont R., Kukanov A., Stoikov S.* The price impact of order book events. March 01, 2011. http://ssrn.com/abstract=1712822.
13. *Cont R., Kukanov A., Stoikov S.* The price impact of order book events // J. Financ. Econom., 2014. Vol. 12. No. 1. P. 47–88.
14. *Korolev V., Chertok A., Zeifman A.* Functional limit theorems for order flow imbalance process. http://ssrn.com/abstract=1735338.
15. *Boykov A.* Cramer–Lundberg model with stochastic premiums // Theor. Probab. Appl., 2002. Vol. 47. No. 3. P. 549–553.
16. *Бойков А. В.* Стохастические модели капитала страховой компании и оценивание вероятности неразорения. Дисс. . . . канд. физ.-мат. наук. — М.: Мате-

матический институт им. В. А. Стеклова РАН, 2003. 83 с.

17. *Темнов Г. О.* Математические модели риска и случайного притока взносов в страховании. Дисс. . . . канд. физ.-мат. наук. — С.-Петербург: Санкт-Петербургский государственный архитектурно-строительный университет, 2004. 102 с.
18. *Королев В. Ю., Бенинг В. Е., Шоргин С. Я.* Математические основы теории риска. — 2-е изд., перераб. и доп. — М.: Физматлит, 2011. 620 с.

Поступила в редакцию 08.10.14

ON THE FORMALIZATION OF ORDER FLOW TOXICITY ON FINANCIAL MARKETS

A. V. Chertok[1,2]

[1]Faculty of Computational Mathematics and Cybernetics, M. V. Lomonosov Moscow State University; 1-52 Leninskiye Gory, GSP-1, Moscow 119991, Russian Federation
[2]Euphoria Group LLC, 9, bld. 1, of. 6 Arkhangelsky Lane, Moscow 101000, Russian Federation

Abstract: The paper considers the microstructural order flow model for financial markets. The order flow imbalance process is used as an integral indicator of the current state of the limit-order book. The model of order flow imbalance is used to analyze the properties of the current limit-order book state, which is considered as two-sided risk process with stochastic premiums. The concept of order flow toxicity on financial markets is studied. This concept is formalized with probabilities of crossing fixed levels by the order flow imbalance process. The paper introduces the concepts of the instantaneous toxicity profile and Bayesian and quantile indicators of toxicity. These indicators are calculated for two model types of order flows: the first one has unit volume orders and the second one consists of orders with random volume which has exponential distribution.

Keywords: financial markets; limit-order book; order flow; order flow imbalance; adverse selection; order flow toxicity; Poisson process; compound Poisson process; two-side risk process; risk process with stochastic premiums; ruin probability

DOI: 10.14357/19922264140403

Acknowledgments

The research was partly financially supported by the Russian Foundation for Basic Research (project 14-07-00041a).

References

1. Jeria, D., and G. Sofianos. September 4, 2008. Passive orders and natural adverse selection. *Street Smart* 33.
2. Glosten L. R., and P. Milgrom. 1985. Bid, ask and transaction prices in a specialist market with heterogeneously informed traders. *J. Financ. Econ.* 14:71–100.
3. Kyle, A. S. 1985. Continuous auctions and insider trading. *Econometrica* 53:1315–1335.
4. Easley, D., and M. O'Hara. 1992. Time and the process of security price adjustment. *J. Financ.* 47:576–605.
5. Easley, D., M. Lopez de Prado, and M. O'Hara. 2012. Flow toxicity and liquidity in a high frequency world. *Rev. Financ. Stud.* 25(5):1457–1493.
6. Korolev, V., A. Chertok, A. Korchagin, and A. Gorshenin. 2013. Veroyatnostno-statisticheskoe modelirovanie informatsionnykh potokov v slozhnykh finansovykh sistemakh na osnove vysokochastotnykh dannykh [Probability and statistical modeling of information flows in complex financial systems from high-frequency data]. *Informatika i ee Primeneniya — Inform. Appl.* 7(1):12–21.
7. Chertok, A., V. Korolev , A. Korchagin, and S. Shorgin. 2014. Modeling high-frequency non-homogeneous order flows by compound Cox processes. Available at: http://ssrn.com/abstract=2378975 (accessed January 14, 2014).
8. Cont, R., S. Stoikov, and R. Talreja. 2010. A stochastic model for order book dynamics. *Oper. Res.* 58(3):549–563.
9. Cont, R., and A. de Larrard. Price dynamics in a Markovian limit order market. Working paper. Available at: http://ssrn.com/abstract=1735338 (accessed February 2012).
10. Bouchaud, J.-P., M. Mezard, and M. Potters. 2002. Statistical properties of stock order books: Empirical results and models. *Quant. Financ.* 2:251–256.
11. Zovko, I., and J. D. Farmer. 2002. The power of patience; a behavioral regularity in limit order placement. *Quant. Financ.* 2:387–392.

12. Cont, R., A. Kukanov, and S. Stoikov. The price impact of order book events. Available at: http:// ssrn.com/abstract=1735338 (accessed March 01, 2011).
13. Cont, R., A. Kukanov, and S. Stoikov. 2014. The price impact of order book events. *J. Financ. Economet.* 12(1):47–88.
14. Korolev, V., A. Chertok, and A. Zeifman. Functional limit theorems for order flow imbalance process. Available at: http://ssrn.com/abstract=1735338 (accessed October 6, 2014).
15. Boykov, A. 2002. Cramer–Lundberg model with stochastic premiums. *Theor. Probab. Appl.* 47(3):549–553.
16. Boykov, A. 2003. Stokhasticheskie modeli kapitala strakhovoy kompanii i otsenivanie veroyatnosti nerazoreniya [Stochastic models of the capital of the insurance company and the evaluation of the probability of non-bankruptcy]. Ph.D. Diss. Moscow. 83 p.
17. Temnov, G. O. 2004. Matematicheskie modeli riska i sluchaynogo pritoka vznosov v strakhovanii [Mathematical models of risk and random inflow of contributions to insurance]. Ph.D. Diss. St. Petersburg. 102 p.
18. Korolev, V. Yu., V. E. Bening, and S. Ya. Shorgin. 2011. *Matematicheskie osnovy teorii riska* [Mathematical foundations of the risk theory]. Moscow: Fizmatlit. 620 p.

Received October 8, 2014

Contributor

Chertok Andrey V. (b. 1987) — junior scientist, Faculty of Computational Mathematics and Cybernetics, M. V. Lomonosov Moscow State University; 1-52 Leninskiye Gory, GSP-1, Moscow 119991, Russian Federation; Director General, Euphoria Group LLC, 9, bld. 1, of. 6 Arkhangelsky Lane, Moscow 101000, Russian Federation; a.v.chertok@gmail.com

ИНФОРМАТИКА И ЕЁ ПРИМЕНЕНИЯ, 2014. Т. 8. Вып. 4. С. 32–40

АСИМПТОТИЧЕСКИЕ СВОЙСТВА ОЦЕНКИ РИСКА В ЗАДАЧЕ ВОССТАНОВЛЕНИЯ ИЗОБРАЖЕНИЯ С КОРРЕЛИРОВАННЫМ ШУМОМ ПРИ ОБРАЩЕНИИ ПРЕОБРАЗОВАНИЯ РАДОНА*

А. А. Ерошенко[1], О. В. Шестаков[2]

Abstract: В последние годы вейвлет-методы, основанные на разложении проекций по специальному базису и последующей процедуре пороговой обработки, широко используются при решении задач реконструкции томографических изображений. Их привлекательность заключается, во-первых, в быстроте алгоритмов, а во-вторых, в возможности реконструкции локальных участков изображения по неполным проекционным данным, что имеет ключевое значение, например, для медицинских приложений, где пациента нежелательно подвергать лишней дозе облучения. Анализ погрешностей этих методов представляет собой важную практическую задачу, поскольку позволяет оценить качество как самих методов, так и используемого оборудования. В работе рассматривается задача оценки функции при обращении оператора Радона в модели с коррелированным шумом. Исследуются асимптотические свойства оценки риска при пороговой обработке коэффициентов вейвлет-вейглет-разложения функции изображения. Приводятся условия, при которых имеет место асимптотическая нормальность несмещенной оценки риска.

Ключевые слова: вейвлеты; линейный однородный оператор; преобразование Радона; пороговая обработка; несмещенная оценка риска; коррелированный шум; асимптотическая нормальность

DOI: 10.14357/19922264140404

1 Введение

Методы статистического анализа часто применяются для решения задач, в которых данные наблюдаются не напрямую, в частности для задач компьютерной томографии, связанных с обращением преобразования Радона. Томографические методы революционизировали медицинскую диагностику, поскольку позволили «увидеть» внутренние органы человека, не подвергая пациента опасности. Эти методы применяются также в геологии, астрономии, сейсмологии, электронной микроскопии, диагностике плазмы, химии и во многих других областях. Рассматривается следующая модель:

$$X = Rf + z\,,$$

где R — оператор Радона; f — искомая функция изображения; z — коррелированный шум с нулевым математическим ожиданием.

Для того чтобы «очистить» целевую функцию (изображение) от шума, используется пороговая обработка коэффициентов вейвлет-вейглет-разложения наблюдаемых данных. Наличие шума приводит к погрешностям. Оценки этих погрешностей (риска) и их свойства в моделях компьютерной томографии с независимым шумом исследовались в работе [1]. Показано, что при определенных условиях оценка риска обладает свойствами состоятельности и асимптотической нормальности. В данной работе исследуется асимптотическое поведение оценки риска в модели со стационарным коррелированным шумом.

2 Преобразование Радона: вейвлет-вейглет-разложение функции

Определим оператор Радона R как набор интегралов от функции f по всевозможным прямым плоскости:

$$(Rf)(s,\theta) = \int\limits_{L_{s,\theta}} f(x,y)\,dl\,,$$

где

$$L_{s,\theta} = \{(x,y)\colon x\cos\theta + y\sin\theta - s = 0\}\ .$$

*Работа выполнена при финансовой поддержке РНФ (проект 14-11-00364).

[1]Московский государственный университет им. М. В. Ломоносова, факультет вычислительной математики и кибернетики, кафедра математической статистики, aeroshik@gmail.com

[2]Московский государственный университет им. М. В. Ломоносова, факультет вычислительной математики и кибернетики, кафедра математической статистики; Институт проблем информатики Российской академии наук, oshestakov@cs.msu.su

Задача томографии — восстановить функцию по наборам ее линейных интегралов, т. е. восстановить f по Rf. Для решения этой задачи воспользуемся методом вейвлет-вейглет-разложения [2].

Пусть заданы $\phi(x)$ и $\psi(x)$ — отцовский и материнский вейвлеты. Тогда можно определить:

$$\left.\begin{aligned}\psi^{[1]}_{j,k_1,k_2}(x,y) &= 2^j\phi(2^jx-k_1)\psi(2^jy-k_2)\,;\\ \psi^{[2]}_{j,k_1,k_2}(x,y) &= 2^j\psi(2^jx-k_1)\phi(2^jy-k_2)\,;\\ \psi^{[3]}_{j,k_1,k_2}(x,y) &= 2^j\psi(2^jx-k_1)\psi(2^jy-k_2),\end{aligned}\right\} \quad (1)$$

семейство $\left\{\psi^{[\lambda]}_{j,k_1,k_2}\right\}_{\lambda,j,k_1,k_2}$ образует ортонормированный базис в $L^2(\mathbb{R}^2)$. Индекс j в (1) называется масштабом, а индексы k_1 и k_2 — сдвигами. Функция ψ должна удовлетворять определенным требованиям, однако ее можно выбрать таким образом, чтобы она обладала некоторыми полезными свойствами, например была дифференцируемой нужное число раз и имела заданное число M_0 нулевых моментов [3], т. е.

$$\int\limits_{-\infty}^{\infty} t^k\psi(t)\,dt = 0\,, \quad k=0,\ldots,M_0-1\,.$$

В данной работе предполагается, что используются вейвлеты Мейера [4] с достаточным количеством непрерывных производных.

Вейвлет-разложение функции f имеет вид:

$$f = \sum_{\lambda,j,k_1,k_2} \left\langle f\,,\ \psi^{[\lambda]}_{j,k_1,k_2}\right\rangle \psi^{[\lambda]}_{j,k_1,k_2}\,. \quad (2)$$

На практике в задаче томографии функция f обычно задана в дискретных отсчетах на круге. Без ограничения общности будем считать, что это круг единичного радиуса с центром в начале координат. Будем рассматривать «растянутую» версию функции $\overline{f}(Nx,Ny) = f(x,y)$. Для дискретных вейвлет-коэффициентов справедливо $f^{W^{[\lambda]}}_{j,k_1,k_2} \approx 2^J\left\langle f,\psi^{[\lambda]}_{j,k_1,k_2}\right\rangle$. Также потребуем, чтобы функция f была равномерно регулярной по Липшицу с некоторым параметром $0<\gamma<1$:

$$|f(x_1,y_1)-f(x_2,y_2)| \le C\Big(|x_1-x_2|^2+|y_1-y_2|^2\Big)^{\gamma/2},$$

где C — некоторая константа. Тогда существует константа A такая, что [4]:

$$\left|f^{W^{[\lambda]}}_{j,k_1,k_2}\right| \le \frac{A\cdot 2^J}{2^{j(\gamma+1)}}\,. \quad (3)$$

Для обращения оператора Радона определим вейглеты [2]:

$$\xi^{[\lambda]}_{j,k_1,k_2} = I^{-1}R\psi^{[\lambda]}_{j,k_1,k_2}\,, \quad (4)$$

где I — потенциал Рисса: $\hat{I^p}g(w) = |w|^{-p}\hat{g}(w)$. Для вейглетов справедливо соотношение:

$$\left[Rf,\xi^{[\lambda]}_{j,k_1,k_2}\right] = \left\langle f,\psi^{[\lambda]}_{j,k_1,k_2}\right\rangle\,,$$

которое позволяет использовать в разложении (2) только проекционные данные и тем самым предлагает метод обращения преобразования Радона.

3 Модель данных

Предположим, что проекционные данные измеряются при $(s,\theta)\in[-1,1]\times[0,\pi]$. Пусть $\{e_{i,j},\ i,j\in\mathbb{Z}\}$ — стационарный гауссовский процесс с ковариационной последовательностью $r_{k,p} = \mathrm{cov}(e_{i,j},e_{i+k,j+p})$. Для выборки с размерами $n=m$ модель проекционных данных с шумом выглядит следующим образом:

$$Y_{i,j} = Rf\left(-1+\frac{2i}{n},\frac{j\pi}{n}\right)+e_{i,j}\,,$$
$$i=1,\ldots,n\,,\ \ j=1,\ldots,n\,,\ \ n=2^J\,.$$

Структура ковариации шума для преобразования Радона должна отражать типичную ситуацию: на практике проекции для разных углов регистрируются независимо друг от друга. В рассматриваемой модели ошибок получаются независимые наблюдения в случае разных углов и стационарный гауссовский шум с нулевым математическим ожиданием, конечной дисперсией и ковариационной последовательностью $r_\delta \sim A\delta^{-\alpha}$ $(0<\alpha<1)$ для одинаковых углов.

Как и в одномерном случае [5], создаем для равномерных отсчетов наблюдаемое поле при $(s,\theta)\in$ $\in[-1,1]\times[0,\pi]$:

$$Y_n(s,\theta) = \frac{1}{n^2}\sum_{i=1}^{[n(s+1)/2]}\ \sum_{j=1}^{[n\theta/\pi]} Y_{i,j} = RF_n(s,\theta)+$$
$$+\frac{1}{n^2}\sum_{i=1}^{[n(s+1)/2]}\ \sum_{j=1}^{[n\theta/\pi]} e_{i,j}\,,$$

где

$$RF_n(s,\theta) = \frac{1}{n^2}\sum_{i=1}^{[n(s+1)/2]}\ \sum_{j=1}^{[n\theta/\pi]} Rf\left(-1+\frac{2i}{n},\frac{j\pi}{n}\right)$$

представляет собой суммарный «сигнал» поля.

Положим $H=1-\alpha/2$, $H\in(1/2,1)$ и определим дробное броуновское движение $\mathbf{B}_H(s)$ — гауссовский процесс на $\mathbb{R}$ с нулевым средним и ковариационной функцией

$$r(s,t) = \frac{V_H}{2}(|s|^{2H}+|s|^{2H}-|s-t|^{2H})\,,\ \ s,t\in\mathbb{R}\,,$$

где

$$V_H = \mathsf{D}(\mathbf{B}_H(1)) = \frac{-\Gamma(2-2H)\cos(\pi H)}{\pi H(2H-1)}\,.$$

Далее по аналогии с моделью, описанной в [5], рассматриваем непрерывную аппроксимацию:

$$Y(s,\theta) = RF(s,\theta) + n^{-(1+\alpha)/2}\tau \mathbf{B'}_H(s,\theta)\,, \qquad (5)$$

где $\tau = \sqrt{2A/((1-\alpha)(2-\alpha))}$ — нормировочный множитель, который без ограничения общности далее будем считать равным единице, $RF(s,\theta) = \int\limits_{-1}^{s}\int\limits_{0}^{\theta} Rf(t,q)\,dtdq$, а $\mathbf{B'}_H(s,\theta)$ — случайная функция, которая для каждого фиксированного угла θ представляет собой дробное броуновское движение $\mathbf{B}_H(s)$ и имеет некоррелированные приращения по θ.

Применяя к (5) вейглет-разложение, получаем:

$$\left[Y, \xi^{[\lambda]}_{j,k_1,k_2}\right] = \left[Rf, \xi^{[\lambda]}_{j,k_1,k_2}\right] + n^{-(1+\alpha)/2}\left[\mathbf{B'}_H, \xi^{[\lambda]}_{j,k_1,k_2}\right];$$
$$\left[Y, \xi^{[\lambda]}_{j,k_1,k_2}\right] = \left\langle f, \psi^{[\lambda]}_{j,k_1,k_2}\right\rangle + n^{-(1+\alpha)/2}\left[\mathbf{B'}_H, \xi^{[\lambda]}_{j,k_1,k_2}\right].$$

Переходя к дискретному вейглет-преобразованию и вспоминая, что $n = 2^J$, по аналогии с дискретным вейвлет-преобразованием получаем модель дискретных вейглет-коэффициентов:

$$X^{[\lambda]}_{j,k_1,k_2} = \mu^{[\lambda]}_{j,k_1,k_2} + 2^{(1-\alpha)J/2} e^{[\lambda]}_{j,k_1,k_2}\,,$$

где

$$\mu^{[\lambda]}_{j,k_1,k_2} = 2^J\left[Rf, \xi^{[\lambda]}_{j,k_1,k_2}\right];$$
$$e^{[\lambda]}_{j,k_1,k_2} = \left[\mathbf{B'}_H, \xi^{[\lambda]}_{j,k_1,k_2}\right] = \int \xi^{[\lambda]}_{j,k_1,k_2}\, d\mathbf{B'}_H\,.$$

Из (1) и (4) получаем выражения для преобразований Фурье вейглет-функций по первому аргументу:

$$\widehat{\xi}^{[1]}_{j,k_1,k_2}(w,\theta) = |w|\cdot 2^{-j} e^{i(k_1\cos\theta + k_2\sin\theta)2^{-j}w} \times$$
$$\times\, \widehat{\phi}(2^{-j}w\cos\theta)\widehat{\psi}(2^{-j}w\sin\theta)\,;$$
$$\widehat{\xi}^{[2]}_{j,k_1,k_2}(w,\theta) = |w|\cdot 2^{-j} e^{i(k_1\cos\theta + k_2\sin\theta)2^{-j}w} \times$$
$$\times\, \widehat{\psi}(2^{-j}w\cos\theta)\widehat{\phi}(2^{-j}w\sin\theta)\,;$$
$$\widehat{\xi}^{[3]}_{j,k_1,k_2}(w,\theta) = |w|\cdot 2^{-j} e^{i(k_1\cos\theta + k_2\sin\theta)2^{-j}w} \times$$
$$\times\, \widehat{\psi}(2^{-j}w\cos\theta)\widehat{\psi}(2^{-j}w\sin\theta).$$

Рассмотрим ковариацию коэффициентов модели, например для $(\lambda_1,\lambda_2) = (3,3)$. Проведем интегрирование по углу и воспользуемся свойствами $\mathbf{B}_H(s)$ [5]. Без ограничения общности считаем, что $j \geq i$. Обозначим $\Delta = j - i$. Имеем

$$\left|\operatorname{cov}\left(X^{[3]}_{j,k_1,k_2}, X^{[3]}_{i,l_1,l_2}\right)\right| =$$
$$= \left|\frac{1}{2\pi}\,2^{(1-\alpha)J}\iint \widehat{\xi}^{[3]}_{j,k_1,k_2}(w,\theta)\overline{\widehat{\xi}^{[3]}_{i,l_1,l_2}(w,\theta)} \times\right.$$
$$\left.\times\,|w|^{-(1-\alpha)}\,dwd\theta\right| = \left|\frac{1}{2\pi}\,2^{(1-\alpha)J}\iint 2^{-j-i}w^2 \times\right.$$
$$\times\, e^{i((k_1\cos\theta + k_2\sin\theta)2^{-j} - (l_1\cos\theta + l_2\sin\theta)2^{-i})w} \times$$
$$\times\, \widehat{\psi}(2^{-j}w\cos\theta)\widehat{\psi}(2^{-j}w\sin\theta) \times$$
$$\left.\times\,\overline{\widehat{\psi}(2^{-i}w\cos\theta)\widehat{\psi}(2^{-i}w\sin\theta)}|w|^{-(1-\alpha)}\,dwd\theta\right| =$$

(сделаем замену $w' = 2^{-i}w$ и перейдем от полярных координат к декартовым: $w_1 = w\cos\theta$, $w_2 = w\sin\theta$)

$$= \left|\frac{1}{2\pi}\,2^{(1-\alpha)J}\cdot 2^{i\alpha-\Delta} \times\right.$$
$$\times \iint e^{i((k_1 2^{-\Delta} - l_1)w_1 + (k_2 2^{-\Delta} - l_2)w_2)} \times$$
$$\times\,|w_1^2 + w_2^2|^{(1+\alpha)/2}\,\widehat{\psi}(2^{-\Delta}w_1)\widehat{\psi}(2^{-\Delta}w_2) \times$$
$$\left.\times\,\overline{\widehat{\psi}(w_1)\widehat{\psi}(w_2)}\,dw_1dw_2\right| \leq$$

(воспользуемся следующим свойством вейвлетов Мейера [5]: при любом натуральном M_0 существует константа $C_{M_0} > 0$ такая, что $|\widehat{\psi}(w)| \leq C_{M_0}|w|^{M_0}\mathbf{1}_{w\in\operatorname{supp}(\widehat{\psi})}$)

$$\leq \left|\frac{C_{M_0}^2}{2\pi}\,2^{(1-\alpha)J}\cdot 2^{i\alpha-\Delta} \times\right.$$
$$\times \int\limits_{\operatorname{supp}(\widehat{\psi})}\int\limits_{\operatorname{supp}(\widehat{\psi})} e^{i((k_1 2^{-\Delta} - l_1)w_1 + (k_2 2^{-\Delta} - l_2)w_2)} \times$$
$$\times\,|w_1^2 + w_2^2|^{(1+\alpha)/2}\cdot 2^{-2\Delta M_0}(w_1w_2)^{M_0} \times$$
$$\left.\times\,\overline{\widehat{\psi}(w_1)\widehat{\psi}(w_2)}\,dw_1dw_2\right| \leq$$

(предположим, что выбранный вейвлет Мейера имеет достаточное число непрерывных производных, чтобы функция $g(w_1,w_2) = |w_1^2 + w_2^2|^{(1+\alpha)/2} \times 2^{-2\Delta M_0}(w_1w_2)^{M_0}\overline{\widehat{\psi}(w_1)\widehat{\psi}(w_2)}$ имела M_1 непрерывных производных по w_1 и w_2, и воспользуемся свойствами обратного преобразования Фурье)

$$\leq 2^{(1-\alpha)J}\cdot 2^{i\alpha}\cdot 2^{-(1+2M_0)\Delta} \times$$
$$\times\,\frac{C''}{\left|k_1\cdot 2^{-\Delta} - l_1\right|^{M_1}\left|k_2\cdot 2^{-\Delta} - l_2\right|^{M_1}}$$

с некоторой константой $C'' > 0$.

Отдельно выделим случай $k_1\cdot 2^{-\Delta} = l_1$ и $k_2 \times 2^{-\Delta} = l_2$. Можно показать, что в этом случае

$$\left|\operatorname{cov}\left(X^{[\lambda]}_{j,k_1,k_2}, X^{[\lambda]}_{i,l_1,l_2}\right)\right| \leq C_e\cdot 2^{J(1-\alpha)+i\alpha-(2M_0+1)\Delta}$$

с некоторой константой $C_e > 0$. Аналогично рассматриваются случаи, когда $k_1 \cdot 2^{-\Delta} = l_1$ и $k_2 \cdot 2^{-\Delta} = l_2$ выполнены не одновременно. Варианты других комбинаций (λ_1, λ_2) рассматриваются аналогично.

Обозначим $M' = -(2M_1 - 2M_0 - 1)$ и выберем M_0 так, что $M' > 0$.

Тогда

$$\left|\operatorname{cov}\left(X^{[\lambda]}_{j,k_1,k_2}, X^{[\lambda]}_{i,l_1,l_2}\right)\right| \le$$
$$\le \begin{cases} C' \cdot 2^{J(1-\alpha)+i\alpha-(2M_0+1)\Delta} & \text{при } k_1 \cdot 2^{-\Delta} = l_1, \\ & \quad k_2 \cdot 2^{-\Delta} = l_2\,; \\ C' \dfrac{2^{(1-\alpha)J+i\alpha-M'\Delta}}{\left|k_2 - 2^{\Delta} l_2\right|^{M_1}} & \text{при } k_1 \cdot 2^{-\Delta} = l_1\,; \\ C' \dfrac{2^{(1-\alpha)J+i\alpha-M'\Delta}}{\left|k_1 - 2^{\Delta} l_1\right|^{M_1}} & \text{при } k_2 \cdot 2^{-\Delta} = l_2\,; \\ C' \dfrac{2^{(1-\alpha)J+i\alpha-M'\Delta}}{\left|k_1 - 2^{\Delta} l_1\right|^{M_1} \left|k_2 - 2^{\Delta} l_2\right|^{M_1}} & \text{иначе} \end{cases} \tag{6}$$

с некоторой константой $C' > 0$.

Дисперсия для коэффициентов модели имеет вид:

$$\sigma^2_{\lambda,j} = C_{\lambda,\alpha} \cdot 2^{(1-\alpha)J} \cdot 2^{j\alpha}\,, \tag{7}$$

где константа $C_{\lambda,\alpha}$ зависит от параметра α и выбранного вейвлет-базиса.

4 Пороговая обработка и оценка риска

Смысл пороговой обработки коэффициентов вейвлет-вейглет-разложения заключается в удалении достаточно маленьких коэффициентов, которые считаются шумом. Будем рассматривать так называемую мягкую пороговую обработку с порогом $T_{\lambda,j}$, зависящим от уровня j. К каждому коэффициенту применяется функция $\rho_{T_{\lambda,j}}(x) = \operatorname{sgn}(x)\left(|x| - T_{\lambda,j}\right)_+$, т. е. при такой пороговой обработке коэффициенты, которые по модулю меньше порога $T_{\lambda,j}$, обнуляются, а абсолютные величины остальных коэффициентов уменьшаются на величину порога. Погрешность (или риск) мягкой пороговой обработки определяется следующим образом:

$$R_J(f) = \sum_{j=0}^{J-1} \sum_{k_1=0}^{2^j-1} \sum_{k_2=0}^{2^j-1} \sum_{\lambda=1}^{3} \mathbb{E}\left(\mu^{[\lambda]}_{j,k_1,k_2} - \right.$$
$$\left. - \rho_{T_{\lambda,j}}\left(X^{[\lambda]}_{j,k_1,k_2}\right)\right)^2. \tag{8}$$

В [6] предложено использовать порог $T_{\lambda,j} = \sqrt{2\ln 2^{2j}}\sigma_{\lambda,j}$, названный универсальным. В дальнейшем будет использоваться именно такой вид порога. В выражении (8) присутствуют неизвестные величины $\mu^{[\lambda]}_{j,k_1,k_2}$, поэтому вычислить значение $R_J(f)$ нельзя. Однако его можно оценить. В качестве оценки риска используется следующая величина [4]:

$$\widehat{R}_J(f) =$$
$$= \sum_{j=0}^{J-1} \sum_{k_1=0}^{2^j-1} \sum_{k_2=0}^{2^j-1} \sum_{\lambda=1}^{3} F\left[\left(X^{[\lambda]}_{j,k_1,k_2}\right)^2, T_{\lambda,j}, \sigma_{\lambda,j}\right], \tag{9}$$

где $F(x,T,\sigma) = (x-\sigma^2)\mathbf{1}_{|x|\le T} + \left(\sigma^2 + T^2\right)\mathbf{1}_{|x|>T}$. Величина $\widehat{R}_J(f)$ является несмещенной оценкой для $R_J(f)$ [4]. В работе [1] исследовались асимптотические свойства оценки (9) в модели с независимым шумом. Было показано, что при определенных условиях гладкости эта оценка является состоятельной и асимптотически нормальной. Далее будет исследовано асимптотическое поведение оценки (9) в модели данных с долгосрочной зависимостью, рассматриваемой в данной работе.

5 Вспомогательные результаты

Введем следующее обозначение для последовательностей: $a_J \simeq b_J$, если $\lim\limits_{J\to\infty}(a_J/b_J) = 1$.

Лемма 1. *Для любых $0 < \alpha < 1$ и $\gamma > (1+\alpha)^{-1}$ выполняется $D_J^2 = \mathsf{D}\widehat{R}_J(f) \simeq \tilde{C} \cdot 2^{4J}$, где константа $\tilde{C}$ зависит от α и выбранного вейвлет-базиса, но не зависит от функции сигнала f.*

Доказательство. Поскольку $\gamma > 1/(1+\alpha)$, то $1/(1+\gamma) < (1+\alpha)/(2+\alpha)$. Выберем p'' так, что $1/(1+\gamma) < p'' < (1+\alpha)/(2+\alpha)$ и $p''J$ — целое число. Тогда в силу (3) $\mu^{[\lambda]}_{j,k_1,k_2} \to 0$ для всех j: $p''J \le j < J$ при $J \to \infty$. Разобьем выражение (9) на две суммы:

$$\widehat{R}_J(f) =$$
$$= \sum_{j=0}^{p''J} \sum_{k_1=0}^{2^j-1} \sum_{k_2=0}^{2^j-1} \sum_{\lambda=1}^{3} F\left[\left(X^{[\lambda]}_{j,k_1,k_2}\right)^2, T_{\lambda,j}, \sigma_{\lambda,j}\right] +$$
$$+ \sum_{j=p''J+1}^{J-1} \sum_{k_1=0}^{2^j-1} \sum_{k_2=0}^{2^j-1} \sum_{\lambda=1}^{3} F\left[\left(X^{[\lambda]}_{j,k_1,k_2}\right)^2, T_{\lambda,j}, \sigma_{\lambda,j}\right].$$

Поскольку существует константа $C_F > 0$ такая, что $F\left[\left(X^{[\lambda]}_{j,k_1,k_2}\right)^2, T_{\lambda,j}, \sigma_{\lambda,j}\right] \le C_F j \cdot 2^{(1-\alpha)J} \cdot 2^{j\alpha}$, то для первой суммы имеем:

$$\left|\sum_{j=0}^{p''J} \sum_{k_1=0}^{2^j-1} \sum_{k_2=0}^{2^j-1} \sum_{\lambda=1}^{3} F\left[\left(X^{[\lambda]}_{j,k_1,k_2}\right)^2, T_{\lambda,j}, \sigma_{\lambda,j}\right]\right| \le$$
$$\le C_F \sum_{j=0}^{p''J} \sum_{k_1=0}^{2^j-1} \sum_{k_2=0}^{2^j-1} \sum_{\lambda=1}^{3} j \cdot 2^{(1-\alpha)J} \cdot 2^{j\alpha} \le$$

$$\leq C_F \cdot 2^{(1-\alpha)J} \sum_{j=0}^{p''J} \sum_{\lambda=1}^{3} j \cdot 2^{j(2+\alpha)} \leq C'_F J \cdot 2^{J(1-\alpha+(\alpha+2)p'')} \tag{10}$$

с некоторой константой $C'_F > 0$. Далее,

$$\mathsf{D}\widehat{R}_J(f) = \sum_{j=0}^{J-1} \sum_{k_1=0}^{2^j-1} \sum_{k_2=0}^{2^j-1} \sum_{\lambda=1}^{3} \mathsf{D}F\left[\left(X^{[\lambda]}_{j,k_1,k_2}\right)^2, T_{\lambda,j}, \sigma_{\lambda,j}\right] +$$

$$+ \sum_{i=0}^{J-1} \sum_{l_1=0}^{2^i-1} \sum_{l_2=0}^{2^i-1} \sum_{\lambda_1=1}^{3} \sum_{j=0}^{J-1} \sum_{k_1=0}^{2^j-1} \sum_{k_2=0}^{2^j-1} \sum_{\lambda_2=1}^{3} \operatorname{cov}\left(F\left[\left(X^{[\lambda_1]}_{i,l_1,l_2}\right)^2, T_{\lambda_1,i}, \sigma_{\lambda_1,i}\right], F\left[\left(X^{[\lambda_2]}_{j,k_1,k_2}\right)^2, T_{\lambda_2,j}, \sigma_{\lambda_2,j}\right]\right), \tag{11}$$

где во второй сумме предполагается, что $(\lambda_1, i, l_1, l_2) \neq (\lambda_2, j, k_1, k_2)$.

Сумму дисперсий разобьем на две суммы:

$$\sum_{j=0}^{p''J} \sum_{k_1=0}^{2^j-1} \sum_{k_2=0}^{2^j-1} \sum_{\lambda=1}^{3} \mathsf{D}F\left[\left(X^{[\lambda]}_{j,k_1,k_2}\right)^2, T_{\lambda,j}, \sigma_{\lambda,j}\right] + \sum_{j=p''J+1}^{J-1} \sum_{k_1=0}^{2^j-1} \sum_{k_2=0}^{2^j-1} \sum_{\lambda=1}^{3} \mathsf{D}F\left[\left(X^{[\lambda]}_{j,k_1,k_2}\right)^2, T_{\lambda,j}, \sigma_{\lambda,j}\right].$$

Поскольку $p'' < (1+\alpha)/(2+\alpha)$, из (10) следует, что первая сумма имеет меньший порядок, чем вторая. Для второй в силу выбора порога и (7) справедливо

$$\sum_{j=p''J+1}^{J-1} \sum_{k_1=0}^{2^j-1} \sum_{k_2=0}^{2^j-1} \sum_{\lambda=1}^{3} \mathsf{D}F\left[\left(X^{[\lambda]}_{j,k_1,k_2}\right)^2, T_j, \sigma_j\right] \simeq \sum_{j=p''J+1}^{J-1} \sum_{k_1=0}^{2^j-1} \sum_{k_2=0}^{2^j-1} \sum_{\lambda=1}^{3} \mathsf{D}\left(X^{[\lambda]}_{j,k_1,k_2}\right)^2 =$$

$$= \sum_{j=p''J+1}^{J-1} \sum_{k_1=0}^{2^j-1} \sum_{k_2=0}^{2^j-1} C_1 \cdot 2\sigma_j^2\left(\sigma_j^2 + \mu^2_{j,k_1,k_2}\right) \simeq \sum_{j=p''J+1}^{J-1} \sum_{k_1=0}^{2^j-1} \sum_{k_2=0}^{2^j-1} 2C_1\sigma_j^4 \simeq C'_\alpha \cdot 2^{4J},$$

где C_1 и C'_α — некоторые положительные константы. Далее перейдем к сумме ковариаций в (11). Аналогично сумме дисперсий имеем:

$$\sum_{i=0}^{J-1} \sum_{l_1=0}^{2^i-1} \sum_{l_2=0}^{2^i-1} \sum_{\lambda_1=1}^{3} \sum_{j=0}^{J-1} \sum_{k_1=0}^{2^j-1} \sum_{k_2=0}^{2^j-1} \sum_{\lambda_2=1}^{3} \operatorname{cov}\left(F\left[\left(X^{[\lambda_1]}_{i,l_1,l_2}\right)^2, T_{\lambda_1,i}, \sigma_{\lambda_1,i}\right], F\left[\left(X^{[\lambda_2]}_{j,k_1,k_2}\right)^2, T_{\lambda_2,j}, \sigma_{\lambda_2,j}\right]\right) \simeq$$

$$\simeq \sum_{i=p''J+1}^{J-1} \sum_{l_1=0}^{2^i-1} \sum_{l_2=0}^{2^i-1} \sum_{\lambda_1=1}^{3} \sum_{j=p''J+1}^{J-1} \sum_{k_1=0}^{2^j-1} \sum_{k_2=0}^{2^j-1} \sum_{\lambda_2=1}^{3} \operatorname{cov}\left(\left(X^{[\lambda_1]}_{i,l_1,l_2}\right)^2, \left(X^{[\lambda_2]}_{j,k_1,k_2}\right)^2\right).$$

Известно, что $\operatorname{cov}\left(X^2, Y^2\right) = 4\mathbb{E}X\mathbb{E}Y\operatorname{cov}(X,Y) + 2\operatorname{cov}^2(X,Y)$, если вектор (X,Y) имеет двумерное нормальное распределение. Так же, как в работе [7], можно показать, что асимптотика первого слагаемого меньше второго, а для второго в силу (6) справедливо

$$\frac{1}{2} \sum_{i=p''J+1}^{J-1} \sum_{l_1=0}^{2^i-1} \sum_{l_2=0}^{2^i-1} \sum_{\lambda_1=1}^{3} \sum_{j=p''J+1}^{J-1} \sum_{k_1=0}^{2^j-1} \sum_{k_2=0}^{2^j-1} \sum_{\lambda_2=1}^{3} \operatorname{cov}^2\left(X^{[\lambda_1]}_{i,l_1,l_2}, X^{[\lambda_2]}_{j,k_1,k_2}\right) =$$

$$= \sum_{\lambda_1=1}^{3} \sum_{\lambda_2=1}^{3} \sum_{i=p''J+1}^{J-1} \sum_{l_1=0}^{2^i-1} \sum_{l_2=0}^{2^i-1} \sum_{\Delta=0}^{J-1-i} \sum_{k_1=\begin{cases} l_1+1, & \Delta=0; \\ 0, & \Delta>0. \end{cases}}^{2^{i+\Delta}-1} \sum_{k_2=\begin{cases} l_2+1, & \Delta=0; \\ 0, & \Delta>0. \end{cases}}^{2^{i+\Delta}-1} \operatorname{cov}^2\left(X^{[\lambda_1]}_{i,l_1,l_2}, X^{[\lambda_2]}_{j,k_1,k_2}\right) \leq$$

$$\leq \sum_{i=p''J+1}^{J-1} \sum_{l_1=0}^{2^i-1} \sum_{l_2=0}^{2^i-1} \left(\sum_{\delta_1=1}^{2^i-l_1} \sum_{\delta_2=1}^{2^i-l_2} \frac{C'^2 \cdot 2^{2J(1-\alpha)} 2^{2i\alpha}}{\delta_1^{2M_1}\delta_2^{2M_1}} + \right.$$

$$\left. + \sum_{\Delta=1}^{J-i-1} \sum_{k_1=0}^{2^{i+\Delta}-1} \sum_{k_2=0}^{2^{i+\Delta}-1} 2^{2J(1-\alpha)} \cdot 2^{2i\alpha} \cdot 2^{-2M'\Delta} \frac{C'^2 \mathbf{1}(k_1 \neq 2^\Delta l_1, k_2 \neq 2^\Delta l_2)}{\left|k_1 - 2^\Delta l_1\right|^{2M_1} \left|k_2 - 2^\Delta l_2\right|^{2M_1}} \right) =$$

$$= C'^2 \cdot 2^{2J(1-\alpha)} \sum_{i=p''J+1}^{J-1} 2^{2i\alpha} \sum_{l_1=0}^{2^i-1} \sum_{l_2=0}^{2^i-1} \left(\sum_{\delta_1=1}^{2^i-l_1} \frac{1}{\delta_1^{2M_1}} \sum_{\delta_2=1}^{2^i-l_2} \frac{1}{\delta_2^{2M_1}} + \right.$$

$$\left. + \sum_{\Delta=1}^{J-i-1} \sum_{k_1=0}^{2^{i+\Delta}-1} \sum_{k_2=0}^{2^{i+\Delta}-1} 2^{-2M'\Delta} \frac{\mathbf{1}(k_1 \neq 2^\Delta l_1, k_2 \neq 2^\Delta l_2)}{\left|k_1 - 2^\Delta l_1\right|^{2M_1} \left|k_2 - 2^\Delta l_2\right|^{2M_1}} \right) \simeq$$

$$\simeq C'^2 \cdot 2^{2J(1-\alpha)} \sum_{i=p''J+1}^{J-1} 2^{2i\alpha} \sum_{l_1=0}^{2^i-1} \sum_{l_2=0}^{2^i-1} \left(H_0^2 + \sum_{\Delta=1}^{J-i-1} 2^{-2M'\Delta} \sum_{k_1=0}^{2^{i+\Delta}-1} \frac{\mathbf{1}\left(k_1 \neq 2^\Delta l_1\right)}{\left|k_1 - 2^\Delta l_1\right|^{2M_1}} \sum_{k_2=0}^{2^{i+\Delta}-1} \frac{\mathbf{1}\left(k_2 \neq 2^\Delta l_2\right)}{\left|k_2 - 2^\Delta l_2\right|^{2M_1}} \right) \simeq$$

$$\simeq C'^2 \cdot 2^{2J(1-\alpha)} \sum_{i=p''J+1}^{J-1} 2^{2i(\alpha+1)} \left(H_0 + \sum_{\Delta=1}^{J-i-1} 2^{-2M'\Delta} H_1 \right) \simeq C'^2 \cdot 2^{2J(1-\alpha)} \sum_{i=p''J+1}^{J-1} 2^{2i(\alpha+1)} H_2 \simeq C''_\alpha \cdot 2^{4J},$$

где H_0, H_1, H_2 и C''_α — положительные константы (в случае $k_1 2^{-\Delta} = l_1$ или $k_2 2^{-\Delta} = l_2$ в приведенных выкладках вместо соответствующих слагаемых используется первая, вторая или третья оценки из (6)).

Объединяя результаты, получаем, что $\mathsf{D}\widehat{R}_J(f) \simeq \simeq \tilde{C} 2^{4J}$. Лемма доказана.

Докажем еще одно свойство эмпирических вейглет-коэффициентов. Говорят, что последовательность случайных величин $\{Y_i\}_{i=1}^\infty$ обладает свойством ρ-перемешивания, если для функции

$$\rho(m) = \sup_{i,j:|i-j|>m} |\mathrm{corr}\,(Y_i, Y_j)|$$

справедливо $\rho(m) \to 0$ при $m \to \infty$.

Лемма 2. *Последовательность* $\left\{ F\left[\left(X^{[\lambda]}_{j,k_1,k_2} \right)^2, T_{\lambda,j}, \sigma_{\lambda,j} \right] \right\}$, $\lambda = 1,2,3, j = 0,\ldots,J-1, k_1 = 1,\ldots,2^j, k_2 = 1,\ldots,2^j$, *обладает свойством ρ-перемешивания, причем для некоторой положительной константы C_ρ*

$$\rho(m) \leq \begin{cases} \dfrac{C_\rho}{(m+1)^{2M_1}} & \text{для элементов} \\ & \text{на одном уровне } (i=j); \\ \dfrac{C_\rho}{2^{(m+1)(2M'+\alpha)}} & \text{для элементов} \\ & \text{на разных уровнях}. \end{cases}$$

Доказательство. Сначала рассмотрим функцию $\rho(m)$ для разных коэффициентов $X^{[\lambda]}_{i,l_1,l_2}$ на одном уровне i.

Для некоторой константы $C_\rho > 0$ при $k_1 \neq l_1$ и $k_2 \neq l_2$ имеем:

$$\rho(m) = \sup_{\substack{0\leq i\leq J-1, \\ \lambda_1,\lambda_2=1,2,3, \\ k_1,k_2,l_1,l_2:\left[\begin{smallmatrix}|k_1-l_1|>m,\\ |k_2-l_2|>m.\end{smallmatrix}\right.}} \frac{\left|\mathrm{cov}\left(F\left[\left(X^{[\lambda_1]}_{i,l_1,l_2} \right)^2, T_{\lambda_1,i}, \sigma_{\lambda_1,i} \right], F\left[\left(X^{[\lambda_2]}_{i,k_1,k_2} \right)^2, T_{\lambda_2,i}, \sigma_{\lambda_2,i} \right] \right)\right|}{\sqrt{\mathsf{D}F\left[\left(X^{[\lambda_1]}_{i,l_1,l_2} \right)^2, T_{\lambda_1,i}, \sigma_{\lambda_1,i} \right] \mathsf{D}F\left[\left(X^{[\lambda_2]}_{i,k_1,k_2} \right)^2, T_{\lambda_2,i}, \sigma_{\lambda_2,i} \right]}} \leq$$

$$\leq \sup_{\substack{0\leq i\leq J-1, \\ k_1,k_2,l_1,l_2:\left[\begin{smallmatrix}|k_1-l_1|>m,\\ |k_2-l_2|>m.\end{smallmatrix}\right.}} C_\rho \frac{2^{2J(1-\alpha)} \cdot 2^{2i\alpha} \left|k_1 - l_1\right|^{-2M_1} \left|k_2 - l_2\right|^{-2M_1}}{\sqrt{\sigma_i^4 \sigma_i^4}} \leq$$

$$\leq \sup_{\substack{0\leq i\leq J-1, \\ k_1,k_2,l_1,l_2:\left[\begin{smallmatrix}|k_1-l_1|>m,\\ |k_2-l_2|>m.\end{smallmatrix}\right.}} C_\rho \frac{2^{2J(1-\alpha)} \cdot 2^{2i\alpha} \left|k_1 - l_1\right|^{-2M_1} \left|k_2 - l_2\right|^{-2M_1}}{2^{2J(1-\alpha)} \cdot 2^{2i\alpha}} \leq$$

$$\leq \sup_{\substack{0\leq i\leq J-1, \\ k_1,k_2,l_1,l_2:\left[\begin{smallmatrix}|k_1-l_1|>m,\\ |k_2-l_2|>m.\end{smallmatrix}\right.}} C_\rho \frac{1}{\left|k_1 - l_1\right|^{2M_1} \left|k_2 - l_2\right|^{2M_1}} \leq \frac{C_\rho}{(m+1)^{2M_1}}. \qquad (12)$$

Случаи $k_1 = l_1$ или $k_2 = l_2$ рассматриваются аналогично.

Далее рассмотрим функцию перемешивания для элементов

$$F\left[\left(X^{[\lambda_1]}_{i,l_1,l_2}\right)^2, T_{\lambda_1,i}, \sigma_{\lambda_1,i}\right], \quad F\left[\left(X^{[\lambda_2]}_{j,k_1,k_2}\right)^2, T_{\lambda_2,j}, \sigma_{\lambda_2,j}\right]$$

на разных уровнях $i, j : j > i, j - i = \Delta > 0$. Рассмотрим случай $\left|k_1 2^{-\Delta} \neq l_1\right|$, $\left|k_2 2^{-\Delta} \neq l_2\right|$:

$$\rho(m) = \sup_{\substack{j-i=\Delta>m,\\ \lambda_1,\lambda_2=1,2,3,\\ 0\le l_1,l_2\le 2^i-1, 0\le k_1,k_2\le 2^j-1,\\ |k_1\cdot 2^{-\Delta}\neq l_1|, |k_2\cdot 2^{-\Delta}\neq l_2|.}} \frac{\left|\operatorname{cov}\left(F\left[\left(X^{[\lambda_1]}_{i,l_1,l_2}\right)^2, T_{\lambda_1,i}, \sigma_{\lambda_1,i}\right], F\left[\left(X^{[\lambda_2]}_{j,k_1,k_2}\right)^2, T_{\lambda_2,j}, \sigma_{\lambda_2,j}\right]\right)\right|}{\sqrt{\mathsf{D}F\left[\left(X^{[\lambda_1]}_{i,l_1,l_2}\right)^2, T_{\lambda_1,i}, \sigma_{\lambda_1,i}\right]\mathsf{D}F\left[\left(X^{[\lambda_2]}_{j,k_1,k_2}\right)^2, T_{\lambda_2,j}, \sigma_{\lambda_2,j}\right]}} \le$$

$$\le \sup_{\substack{j-i=\Delta>m,\\ 0\le l_1,l_2\le 2^i-1, 0\le k_1,k_2\le 2^j-1,\\ |k_1\cdot 2^{-\Delta}\neq l_1|, |k_2\cdot 2^{-\Delta}\neq l_2|.}} C_\rho \frac{2^{2J(1-\alpha)}\cdot 2^{2i\alpha}\cdot 2^{-2M'\Delta}\left|\left(k_1 - 2^\Delta l_1\right)^{-2M_1}\left(k_2 - 2^\Delta l_2\right)^{-2M_2}\right|}{2^{J(1-\alpha)}\cdot 2^{i\alpha}\cdot 2^{J(1-\alpha)}\cdot 2^{(i+\Delta)\alpha}} \le$$

$$\le \sup_{\substack{j-i=\Delta>m,\\ 0\le l_1,l_2\le 2^i-1, 0\le k_1,k_2\le 2^j-1,\\ |k_1\cdot 2^{-\Delta}\neq l_1|, |k_2\cdot 2^{-\Delta}\neq l_2|.}} \frac{C_\rho\cdot 2^{-2M'\Delta}}{2^{\Delta\alpha}} \le \sup_{\substack{j-i=\Delta>m,\\ 0\le l_1,l_2\le 2^i-1, 0\le k_1,k_2\le 2^j-1,\\ |k_1\cdot 2^{-\Delta}\neq l_1|, |k_2\cdot 2^{-\Delta}\neq l_2|.}} \frac{C_\rho}{2^{\Delta(2M'+\alpha)}} = \frac{C_\rho}{2^{(m+1)(2M'+\alpha)}}. \qquad (13)$$

Аналогично рассматриваются случаи, когда выполнено хотя бы одно из равенств $\left|k_1\cdot 2^{-\Delta} = l_1\right|$, $\left|k_2\cdot 2^{-\Delta} = l_2\right|$.

Из (12) и (13) следует утверждение леммы.

6 Основная теорема

Докажем асимптотическую нормальность оценки риска.

Теорема. *Пусть $0 < \alpha < 1$ и функция f равномерно регулярна по Липшицу с параметром $\gamma > (1+\alpha)^{-1}$. Тогда при пороговой обработке с универсальным порогом $T_{\lambda,j}$ имеет место сходимость по распределению*:

$$\frac{\widehat{R}_J(f) - R_J(f)}{D_J} \Rightarrow \mathbf{N}(0,1), \quad J\to\infty, \qquad (14)$$

где $D_J^2 = \tilde{C}2^{4J}$, а константа $\tilde{C}$ зависит от α и выбранного вейвлет-базиса.

Доказательство. Из леммы 1 следует, что $\mathsf{D}\widehat{R}_J(f) \simeq D_J^2 = \tilde{C}\cdot 2^{4J}$. Разобьем выражение в (14) на две суммы, как и в лемме 1:

$$\frac{\widehat{R}_J(f) - R_J(f)}{D_J} = \frac{1}{D_j}\left(\sum_{j=0}^{p''J}\sum_{k_1=0}^{2^j-1}\sum_{k_2=0}^{2^j-1}\sum_{\lambda=1}^{3}\left(F\left[\left(X^{[\lambda]}_{j,k_1,k_2}\right)^2, T_{\lambda,j}, \sigma_{\lambda,j}\right] - \mathbb{E}F\left[\left(X^{[\lambda]}_{j,k_1,k_2}\right)^2, T_{\lambda,j}, \sigma_{\lambda,j}\right]\right)\right) +$$

$$+ \frac{1}{D_j}\left(\sum_{j=p''J+1}^{J-1}\sum_{k_1=0}^{2^j-1}\sum_{k_2=0}^{2^j-1}\sum_{\lambda=1}^{3}\left(F\left[\left(X^{[\lambda]}_{j,k_1,k_2}\right)^2, T_{\lambda,j}, \sigma_{\lambda,j}\right] - \mathbb{E}F\left[\left(X^{[\lambda]}_{j,k_1,k_2}\right)^2, T_{\lambda,j}, \sigma_{\lambda,j}\right]\right)\right),$$

где $1/(1+\gamma) < p'' < (1+\alpha)/(2+\alpha)$. Тогда вследствие (10) первая сумма стремится к нулю п. в.

Из леммы 2 следует, что последовательность $\left\{F\left[\left(X^{[\lambda]}_{j,k_1,k_2}\right)^2, T_{\lambda,j}, \sigma_{\lambda,j}\right]\right\}$, $\lambda = 1,2,3$, $j = 0,\dots,J-1$, $k_1 = 1,\dots,2^j$, $k_1 = 1,\dots,2^j$, обладает свойством ρ-перемешивания и, следовательно, обладает свойством α-перемешивания [8].

Далее, действуя, как в лемме 1, можно показать, что

$$\sup_{J>0}\frac{1}{D_J^2}\sum_{j=p''J+1}^{J-1}\sum_{k_1=0}^{2^j-1}\sum_{k_2=0}^{2^j-1}\sum_{\lambda=1}^{3}\mathbb{E}\left(F\left[\left(X^{[\lambda]}_{j,k_1,k_2}\right)^2, T_{\lambda,j}, \sigma_{\lambda,j}\right] - \mathbb{E}F\left[\left(X^{[\lambda]}_{j,k_1,k_2}\right)^2, T_{\lambda,j}, \sigma_{\lambda,j}\right]\right)^2 =$$

$$= \sup_{J>0}\frac{1}{D_J^2}\sum_{j=p''J+1}^{J-1}\sum_{k_1=0}^{2^j-1}\sum_{k_2=0}^{2^j-1}\sum_{\lambda=1}^{3}\mathsf{D}F\left[\left(X^{[\lambda]}_{j,k_1,k_2}\right)^2, T_{\lambda,j}, \sigma_{\lambda,j}\left(X^{[\lambda]}_{j,k_1,k_2}\right)^2\right] \le \sup_{J>0}\frac{\tilde{C}'_\alpha\cdot 2^{4J}}{\hat{C}\cdot 2^{4J}} \le \infty.$$

Также можно показать, что выполнено условие Линдеберга: для любого $\epsilon > 0$

$$\frac{1}{D_J^2}\sum_{j=p''J+1}^{J-1}\sum_{k_1=0}^{2^j-1}\sum_{k_2=0}^{2^j-1}\sum_{\lambda=1}^{3}\mathbb{E}\left(F\left[\left(X^{[\lambda]}_{j,k_1,k_2}\right)^2,\right.\right.$$
$$\left.T_{\lambda,j},\sigma_{\lambda,j}\right]-\mathbb{E}F\left[\left(X^{[\lambda]}_{j,k_1,k_2}\right)^2,T_{\lambda,j},\sigma_{\lambda,j}\right]\Bigg)^2\times$$
$$\times\,\mathbf{1}\left(\left|F\left[\left(X^{[\lambda]}_{j,k_1,k_2}\right)^2,T_{\lambda,j},\sigma_{\lambda,j}\right]-\right.\right.$$
$$\left.\left.-\,\mathbb{E}F\left[\left(X^{[\lambda]}_{j,k_1,k_2}\right)^2,T_{\lambda,j},\sigma_{\lambda,j}\right]\right|>\epsilon D_J\right)\to 0\,,$$
$$J\to\infty. \quad (15)$$

Действительно, поскольку

$$\left|F\left[\left(X^{[\lambda]}_{j,k_1,k_2}\right)^2,T_{\lambda,j},\sigma_{\lambda,j}\right]\right|\le T^2_{\lambda,j}=2\ln 2^{2j}\sigma^2_{\lambda,j}=$$
$$=\tilde{C}_{\lambda,\alpha}j\cdot 2^{J(1-\alpha)}\cdot 2^{j\alpha}$$

с некоторой константой $\tilde{C}_{\lambda,\alpha}>0$ и $D_J^2\simeq\hat{C}\cdot 2^{4J}$, то начиная с некоторого J все индикаторы в (15) обращаются в ноль.

Таким образом, выполнены все условия теоремы 2.1 из работы [9] и справедлива сходимость по распределению (14). Теорема доказана.

Литература

1. *Маркин А. В., Шестаков О. В.* Асимптотики оценки риска при пороговой обработке вейвлет-вейглет коэффициентов в задаче томографии // Информатика и её применения, 2010. Т. 4. Вып. 2. С. 36–45.
2. *Donoho D.* Nonlinear solution of linear inverse problems by wavelet-vaguelette decomposition // Appl. Comput. Harmon. Anal., 1995. Vol. 2. P. 101–126.
3. *Добеши И.* Десять лекций по вейвлетам / Пер. с англ. — Ижевск: НИЦ Регулярная и хаотическая динамика, 2001. 357 с. (*Daubechies I.* Ten lectures on wavelets. CBMF-NSF regional conference ser. in applied mathematics. SIAM, 1992. 369 p.)
4. *Mallat S.* A wavelet tour of signal processing. — Academic Press, 1999. 662 p.
5. *Johnstone I. M., Silverman B. W.* Wavelet threshold estimates for data with correlated noise // J. Roy. Stat. Soc. B, 1997. Vol. 59. P. 319–351.
6. *Kolaczyk E. D.* Wavelet methods for the inversion of certain homogeneous linear operators in the presence of noisy data. — Stanford: Stanford University, 1994. Ph.D. Dissertation. 163 p.
7. *Ерошенко А. А., Шестаков О. В.* Асимптотические свойства оценки риска при пороговой обработке вейвлет-коэффициентов в модели с коррелированным шумом // Информатика и её применения, 2014. Т. 8. Вып. 1. С. 36–44.
8. *Bradley R. C.* Basic properties of strong mixing conditions. A survey and some open questions // Probab. Surveys, 2005. Vol. 2. P. 107–144.
9. *Peligrad M.* On the asymptotic normality of sequences of weak dependent random variables // J. Theor. Probab., 1996. Vol. 9. No. 3. P. 703–715.

Поступила в редакцию 29.09.14

ASYMPTOTIC PROPERTIES OF RISK ESTIMATE IN THE PROBLEM OF RECONSTRUCTING IMAGES WITH CORRELATED NOISE BY INVERTING THE RADON TRANSFORM

A. A. Eroshenko[1] and O. V. Shestakov[1,2]

[1]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

Abstract: In recent years, wavelet methods based on the decomposition of projections in a special basis and the following thresholding procedure became widely used for solving the problems of tomographic image reconstruction. These methods are easily implemented through fast algorithms; so, they are very appealing in practical situations. Besides, they allow the reconstruction of local parts of the images using incomplete projection data, which is essential, for example, for medical applications, where it is not desirable to expose the patient to the redundant radiation dose. Wavelet thresholding risk analysis is an important practical task, because it allows determining the quality of techniques themselves and the equipment which is used. The present paper considers the problem of estimating the function by inverting the Radon transform in the model of data with correlated noise. The asymptotic properties of mean-square risk estimate of wavelet-vaguelette thresholding technique are studied. The conditions under which the unbiased risk estimate is asymptotically normal are given.

Keywords: wavelets; linear homogeneous operator; Radon transform; thresholding; unbiased risk estimate; correlated noise; asymptotic normality

DOI: 10.14357/19922264140404

Acknowledgments

The research was financially supported by the Russian Science Foundation (project 14-11-00364).

References

1. Markin, A. V., and O. V. Shestakov. 2010. Asimptotiki otsenki riska pri porogovoy obrabotke veyvlet-veyglet koeffitsientov v zadache tomografii [Asymptotic properties of risk estimate for wavelet-vaguelette coefficients thresholding in tomographic reconstruction problem]. *Informatika i ee Primeneniya — Inform. Appl.* 4(2):36–45.
2. Donoho, D. 1995. Nonlinear solution of linear inverse problems by wavelet-vaguelette decomposition. *Appl. Comput. Harmon. Anal.* 2:101–126.
3. Daubechies, I. 1992. *Ten lectures on wavelets*. CBMF-NSF regional conference ser. in applied mathematics. SIAM. 369 p.
4. Mallat, S. 1999. *A wavelet tour of signal processing*. Academic Press. 662 p.
5. Johnstone, I. M., and B. W. Silverman. 1997. Wavelet threshold estimates for data with correlated noise. *J. Roy. Stat. Soc. B* 59:319–351.
6. Kolaczyk, E. D. 1994. Wavelet methods for the inversion of certain homogeneous linear operators in the presence of noisy data. Stanford: Stanford University. Ph.D. Thesis. 163 p.
7. Eroshenko, A. A., and O. V. Shestakov. 2014. Asimptoticheskie svoystva otsenki riska pri porogovoy obrabotke veyvlet-koeffitsientov v modeli s korrelirovannym shumom [Asymptotic properties of wavelet thresholding risk estimate in the model of data with correlated noise]. *Informatika i ee Primeneniya — Inform. Appl.* 8(1): 36–44.
8. Bradley, R. C. 2005. Basic properties of strong mixing conditions. A survey and some open questions. *Probab. Surveys* 2:107–144.
9. Peligrad, M. 1996. On the asymptotic normality of sequences of weak dependent random variables. *J. Theor. Probab.* 9(3):703–715.

Received September 29, 2014

Contributors

Eroshenko Alexander A. (b. 1989) — 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; aeroshik@gmail.com

Shestakov Oleg V. (b. 1976) — Doctor of Science in physics and mathematics, associate professor, 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; oshestakov@cs.msu.su

ИНФОРМАТИКА И ЕЁ ПРИМЕНЕНИЯ, 2014. Т. 8. Вып. 4. С. 41–45

АНАЛИЗ МЕТОК В СКРЫТЫХ КАНАЛАХ*

А. А. Грушо[1], Н. А. Грушо[2], Е. Е. Тимонина[3]

Аннотация: Рассматривается класс скрытых каналов, построенных на основе меток. Предполагается, что выявление скрытого канала ведется контролером исключительно статистическими методами. Это значит, что редко встречающиеся и часто встречающиеся лингвистические конструкции для контролера неразличимы. Для него важно, чтобы передаваемая по каналу последовательность символов не содержала запретов, не соответствующих вероятностной модели легальных сообщений. Основная проблема при обеспечении невидимости таких каналов состоит в том, что при встраивании меток могут возникать запреты вероятностной меры, описывающей легальную передачу. В работе предложен метод построения меток, которые не могут выявляться контролером. Благодаря такому построению меток скрытый канал невидим.

Ключевые слова: скрытые каналы; информационная безопасность; метки, порождающие скрытый канал; невидимость меток; математические модели скрытых каналов

DOI: 10.14357/19922264140405

1 Введение

В работе рассматривается класс скрытых каналов, порожденных метками [1, 2]. В дальнейшем этот класс скрытых каналов будем называть классом A.

Скрытые каналы класса A устроены следующим образом. На приемном и передающем концах скрытого канала существуют синхронно работающие счетчики, которые фиксируют условное время. Чаще всего это время вычисляется в длинах сообщений между метками. Каждое сообщение определяется двумя метками — началом и концом сообщения, а длина сообщения между метками, или условное время, является кодом передаваемого сообщения. Метки должны однозначно идентифицироваться на приемном конце скрытого канала, в противном случае может возникнуть шум, т. е. сбой при приеме сообщения. С другой стороны, хорошо заметные метки являются признаком работы скрытого канала. Поэтому метки должны быть сделаны максимально невидимыми для контролирующего передачу субъекта.

Таким образом, каналы класса A предъявляют противоречивые требования к своему формированию: метки должны быть однозначно определяемыми и невидимыми. В работах [1, 2] рассмотрены примеры разрешения указанного противоречия. Так, в работе [1] предлагается метод передачи меток по ненаблюдаемому, синхронно работающему каналу. Если само сообщение представляет собой последовательность битов, то метки являются изменением формы передаваемого сигнала, которое в реально работающих технических устройствах не наблюдается. Однако при обработке сигналов в процессоре, где находится агент (программно-аппаратная сущность, присутствующая в микропроцессоре и являющаяся отправителем или получателем в работе скрытого канала) получателя сообщения, изменение формы сигнала легко обнаруживается на физическом уровне.

В работе [2] приведены примеры других способов построения скрытых каналов класса A в глобальной сети, основанных на скрытой синхронизации легальных каналов передачи сообщений. Кроме того, для сообщений, которые хорошо моделируются случайной равновероятной последовательностью, в этой работе описан метод построения невидимых скрытых меток.

В данной работе проводится исследование возможности построения меток для скрытых каналов, использующих модели сообщений, которые описываются случайными процессами, не являющимися случайными равновероятными последовательностями.

Поясним постановку задачи. В работе [3] показано, что даже при моделировании метки случайным процессом, полностью соответствующим модели передаваемого легального сообщения, возникают запреты [4, 5], однозначно выявляющие

*Работа частично поддержана РФФИ (проект 13-01-00215).

[1]Институт проблем информатики Российской академии наук; факультет вычислительной математики и кибернетики Московского государственного университета им. М. В. Ломоносова, grusho@yandex.ru

[2]Институт проблем информатики Российской академии наук, info@itake.ru

[3]Институт проблем информатики Российской академии наук, eltimon@yandex.ru

скрытый канал. При этом метки, определяющие сообщение из скрытого канала класса A, являются вставками в легальное сообщение. Запреты возникают в силу наличия зависимостей между элементами случайной последовательности, моделирующей легальное сообщение. Метка в форме вставки в случайную последовательность может нарушить эти зависимости, порождая запрет в случайной последовательности, который не может присутствовать в реальной реализации легального сообщения.

Основной результат работы заключается в демонстрации того, как согласовать статистические характеристики случайного процесса, моделирующего легальное сообщение, и свойства метки, которая является ключом, известным на приемном и передающем концах канала. Показано, что контролер, который реализует поиск запретов в случайной последовательности, находится в худшем положении, чем организатор скрытого канала, по двум причинам. Первая причина состоит в том, что организатор скрытого канала традиционно предполагается знающим алгоритмы контролера. Вторая причина кроется в том, что контролер для выявления скрытого канала вынужден использовать более глубокие зависимости легальных сообщений, чем использует организатор скрытого канала.

2 Математическая модель скрытого канала

Пусть $X = \{x_1, \ldots, x_r\}$ — конечное множество, X^n — декартово произведение X, X^∞ — множество всех последовательностей с элементами из X. Пусть $\mathcal{A}$ — это σ-алгебра на X^∞, порожденная цилиндрическими множествами; $\mathcal{A}$ также является борелевской σ-алгеброй в тихоновском произведении X^∞, где X имеет дискретную топологию [6, 7].

На $(X^\infty, \mathcal{A})$ определена вероятностная мера P. Предположим, что P_n является проекцией меры P на первые n координат последовательностей из X^∞.

Запретом [4, 5] в мере P_n называется вектор $\overline{x}_l \in X^l$, $l \le n$, такой что

$$P_n\left(\overline{x}_l \times X^{n-l}\right) = 0\,.$$

Пусть $\overline{x}_l \in X^l$ — запрет, а $\tilde{x}_{l-1}$ получена из $\overline{x}_l$ отбрасыванием последней координаты. Если $P_{l-1}(\tilde{x}_{l-1}) > 0$, то $\overline{x}_l$ называется *наименьшим запретом* [4, 5].

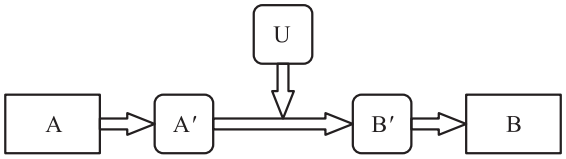
Если $\overline{x}_l$ является запретом в мере P_n, то для любых $l \le s \le n$ и для любых векторов $\overline{x}_s$, начинающихся с $\overline{x}_l$, имеем

$$P_s\left(\overline{x}_s\right) = 0\,.$$

Пусть $\overline{x}_s \in X^s$ и для любого $n \ge s$ и любого $\overline{x}_n \in X^n$, если $\overline{x}_s$ является частью $\overline{x}_n$, $P_n(\overline{x}_n) = 0$.

Вектор $\overline{x}_s$ называется *минимальным запретом* [8], если любой вектор $\overline{x}_n$, $P_n(\overline{x}_n) > 0$, после приписывания в конце недостающих знаков из вектора $\overline{x}_s$ становится наименьшим запретом.

Меру P будем называть моделью легального сообщения. Далее будем предполагать, что все запреты меры P определяются минимальными запретами и количество минимальных запретов конечно.



Модель скрытого канала

Рассмотрим следующую модель скрытого канала (см. рисунок). Компьютеры A и B связаны однонаправленным каналом от A к B. Выходная последовательность компьютера A проходит через агента A′ (программно-аппаратную сущность, функционирующую в компьютерной среде незаметно для пользователей и средств защиты), который строит скрытый канал и передает по нему вместе с легальной информацией имеющуюся у него скрываемую информацию. На приемном конце информация из канала прежде всего попадает агенту B′, который распознает скрытый канал, считывает скрытую информацию, уничтожает следы скрытого канала и передает компьютеру B легальную информацию. U — контролер передачи в канале.

Будем считать, что как противник, так и контролер, ищущий скрытые каналы, могут изучать свойства P статистическими методами. Противник – отправитель сообщения и противник – получатель сообщения имеют общие ключи в виде множества меток и некоторых ограничений на места их появления. Метки представляют собой набор векторов в алфавите X, имеющих для простоты одну длину k. Сообщения передаются кодом, который представляется длинами легального сообщения между двумя метками. Метка начала скрытого сообщения вставляется в легальное сообщение на одно из допустимых мест, определяемых ключом. Метка окончания скрытого сообщения ставится в фиксированном месте легального сообщения на расстоянии от последнего символа входной метки так, чтобы расстояние между метками соответствовало коду.

Допускается, что векторы, соответствующие меткам, могут появляться в легальном сообщении случайно. Это может привести к сбою скрытой пе-

редачи. В связи с этим на метки накладываются следующие ограничения:

– вероятность случайного появления метки должна быть как можно меньше;

– метки должны выбираться таким образом, чтобы не допустить появления запретов вероятностной меры P.

3 Построение меток для невидимого скрытого канала

Предположим, что все скрываемые сообщения имеют конечную длину не более N. Это значит, что начиная с некоторой длины $N_0 > N$ (с учетом ограничений на начало скрываемого сообщения) случайная последовательность, соответствующая легальному сообщению, будет строиться в соответствии с вероятностным распределением, аналогичным исходному процессу P.

Однако это не означает, что скрытый канал невозможно выявить. Если вставки порождают запреты вероятностной меры P [4, 5], то скрытый канал выявляется однозначно. Значит, вероятность выявления скрытого канала равна вероятности появления запрета меры P_{N_0}. Вероятность случайного появления метки длины k не меньше вероятности появления данного вектора в мере P_N. За исключением вырожденных случаев, можно считать, что с помощью выбора k эта вероятность может быть сделана как угодно малой и отвечает требованиям помехозащищенности скрытого канала.

Для простоты рассмотрим возможность скрытия одной метки. Обозначим ее через вектор $\overline{b} = (b_1, \ldots, b_k)$, а место ее вставки — t. Пусть минимальные запреты меры P представимы следующими векторами:

$$\left.\begin{array}{l} \overline{a}^{(1)} = \left(a_{i_1}^{(1)}, \ldots a_{i_{m_1}}^{(1)}\right); \\ \ldots\ldots\ldots\ldots\ldots\ldots \\ \overline{a}^{(s)} = \left(a_{i_1}^{(s)}, \ldots a_{i_{m_s}}^{(s)}\right), \end{array}\right\} \qquad (1)$$

где все $a_{i_j} \in X$. Для простоты далее положим $m_i = m$, $i = 1, \ldots, s$.

Определим возможности контролера U. Контролер U знает распределение P, знает все запреты меры P и может их отслеживать в реальном времени. Контролер не знает вставку $\overline{b}$ и место ее расположения t. Контролер U не анализирует семантику, поэтому для него маловероятные и высоковероятные последовательности одинаково являются допустимыми. Исходя из этих предположений, контролер U выявляет скрытый канал с помощью нахождения минимальных запретов в передаваемой последовательности.

Пусть $\overline{x} = x(1), x(2), \ldots, x(t-1), x(t), \ldots)$ — легальная последовательность, а $\overline{x}' = (x(1), x(2), \ldots$ $\ldots, x(t-1), b_1, \ldots, b_k, x(t), \ldots)$ — последовательность со вставкой. Возможность появления запрета в последовательности $\overline{x}'$ определяется тем, что хотя бы один из векторов

$$\left.\begin{array}{l} x(t-k+1), x(t-k+2), \ldots, x(t-1), b_1; \\ x(t-k+2), x(t-k+3), \ldots, x(t-1), b_1, b_2; \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ b_1, b_2, \ldots, b_k; \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ b_k, x(t), x(t+1), \ldots, x(t+k-1) \end{array}\right\} \qquad (2)$$

принадлежит (1). В других местах последовательности $\overline{x}'$ появление запретов невозможно по определению меры P. Если в последовательности $\overline{x}'$ не появилось запретов, то любой начальный участок этой последовательности является допустимым вектором и его вероятность больше 0 в мере P_N. Поэтому контролер U не может выявить скрытый канал. Отсюда следует следующее

Утверждение 1. Контролер U не выявляет скрытый канал тогда и только тогда, когда ни один из векторов (2) не содержится в множестве векторов (1).

Таким образом, для организации невидимого скрытого канала метка $\overline{b}$ должна быть выбрана так, чтобы выполнялось достаточное условие утверждения 1 при любом допустимом $\overline{x}$. При этом организатор скрытого канала может определить минимальные запреты только статистически, наблюдая за передаваемыми последовательностями. Для этого мера P должна удовлетворять дополнительным ограничениям. Например, случайная последовательность $\overline{x}$ должна быть эргодической.

Но даже если организатор скрытого канала знает минимальные запреты, построение метки, удовлетворяющей утверждению 1 для всех $\overline{x}$, является сложной задачей.

Однако можно предложить упрощенный метод решения поставленной задачи построения метки для невидимого скрытого канала. Пусть организатор скрытого канала нашел минимальные запреты. Рассмотрим множество C биграмм вида

$$C = \left\{\left(a_{i_{m-1}}^{(i)}, a_{i_m}^{(j)}\right),\ j = \overline{1,s}\right\} \bigcup \bigcup \left\{\left(a_{i_1}^{(j)}, a_{i_2}^{(j)}\right),\ j = \overline{1,s}\right\}.$$

Тогда справедливо следующее

Утверждение 2. Если для любой последовательности $\overline{x}$ в векторе $(x(t-1), b_1, b_2, \ldots, b_k, x(t))$ нет

биграмм из множества C, то скрытый канал не выявляем.

Доказательство. Если в векторе $(x(t-1), b_1, b_2, \ldots, b_k, x(t))$ не встречаются биграммы из множества C, то в последовательности $\overline{x}'$ не встречаются запреты из множества (1). В самом деле, появление запрета влечет появление биграмм из множества C. Если вставка не порождает запретов, то согласно утверждению 1 контролер не выявляет скрытого канала. Утверждение 2 доказано.

Однако создание метки, исключающей появление биграмм из множества C для всех последовательностей $\overline{x}$, может оказаться невозможным. В этом случае метку можно сделать сложной, а именно: метка $\overline{b}$ не должна содержать запрещенных биграмм из множества C. Для недопущения появления запрещенной биграммы на концах вставки можно ввести дополнительные символы до и после метки, которые не несут содержательной информации, но исключают возможность появления запрещенных биграмм на концах вставки. Эти незначащие символы выбираются для каждой конкретной последовательности $\overline{x}$.

Предположим, что организатор скрытого канала заранее скрытно выбрал момент t. Затем он статистически нашел все наименьшие запреты в окрестности этого момента времени, используя различные реализации передаваемых сообщений. Тогда, используя описанный метод перехода к биграммам, можно построить метку, которая не порождает наименьших запретов, связанных со вставкой метки в момент t. Такой канал будет невидим для U.

4 Заключение

Интенсивное использование многоагентных систем для скрытого контроля компьютерной среды требует разработки невидимых скрытых каналов для организации коллективной деятельности агентов. Уязвимость простых протоколов общения агентов может привести к полной нейтрализации систем контроля. Поэтому необходимо строить простые, доказано невидимые скрытые каналы. Рассмотренные в статье скрытые каналы просты, но могут легко выявляться при неправильной организации меток. Предложенный в работе метод позволяет делать метки невидимыми, когда их поиск ведется статистическими методами.

Литература

1. *Грушо Н.* Скрытые каналы, основанные на метках // Системы и средства информатики, 2013. Т. 23. № 1. С. 7–13.
2. *Грушо А., Грушо Н., Тимонина Е.* Скрытые каналы, порожденные метками, в дейтаграммах // Системы и средства информатики, 2013. Т. 23. № 2. С. 3–18.
3. *Grusho A., Grusho N., Timonina E.* Problems of modeling in the analysis of covert channels // Computer network security / Eds. I. Kotenko, V. Skormin. Lecture notes in computer science ser. — Berlin–Heidelberg: Springer-Verlag, 2010. Vol. 6258. P. 118–124. doi: 10.1007/978-3-642-14706-7_9.
4. *Грушо А., Тимонина Е.* Запреты в дискретных вероятностно-статистических задачах // Дискретная математика, 2011. Т. 23. Вып. 2. С. 53–58.
5. *Grusho A., Grusho N., Timonina E.* Consistent sequences of tests defined by bans // Springer proceedings in mathematics & statistics, optimization theory, decision making, and operation research applications. — New York – Heidelberg – Dordrecht – London: Springer, 2013. P. 281–291.
6. *Бурбаки Н.* Общая топология. Основные структуры / Пер. с франц. — М.: Наука, 1968. 272 с. (*Bourbaki N.* Topologie Générale. Chapitre 1: Structures topologiques. Chapitre 2: Structures uniformes. — Paris: Hermann, 1940. 129 p.)
7. *Прохоров Ю. В., Розанов Ю. А.* Теория вероятностей. — 2-е изд. — М.: Наука, 1973. 494 с.
8. *Грушо А., Грушо Н., Тимонина Е.* Статистические методы определения запретов вероятностных мер на дискретных пространствах // Информатика и её применения, 2013. Т. 7. Вып. 1. С. 54–57.

Поступила в редакцию 28.09.14

THE ANALYSIS OF TAGS IN COVERT CHANNELS

A. A. Grusho[1,2], N. A. Grusho[1], and E. E. Timonina[1]

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

[2]Faculty of Computational Mathematics and Cybernetics, M. V. Lomonosov Moscow State University, 1-52 Leninskiye Gory, GSP-1, Moscow 119991, Russian Federation

Abstract: The class of covert channels constructed on the basis of tags is considered. It is supposed that a covert channel is detected by the control subject using exclusively statistical techniques. It means that linguistic constructions

seldom or often met are indiscernible for the control subject. For control subject, it is important that the sequence transmitted through the channel does not contain the bans which do not correspond to the probability model of legal messages. The main problem for supporting invisibility of such channels consists in the fact that there can be bans of the probability measure describing a legal transmission when embedding tags. In the paper, the method for creating tags which cannot be detected by the control subject becomes suggested. Thanks to such creation of tags, the covert channel becomes invisible.

Keywords: covert channels; information security; covert channel generated by tags; "invisibility" of tags; mathematical models of covert channels

DOI: 10.14357/19922264140405

Acknowledgments

The paper was partly supported by the Russian Foundation for Basic Research (project 13-01-00215).

References

1. Grusho, N. 2013. Skrytye kanaly, osnovannye na metkakh [Covert channels generated by tags]. *Sistemy i Sredstva Informatiki — Systems and Means of Informatics* 23(1):7–13.
2. Grusho, A., N. Grusho, and E. Timonina. 2013. Skrytye kanaly, porozhdennye metkami, v deytagrammakh [Covert channels generated by tags in datagrams]. *Sistemy i Sredstva Informatiki — Systems and Means of Informatics* 23(2):3–18.
3. Grusho, A., N. Grusho, and E. Timonina. 2010. Problems of modeling in the analysis of covert channels. *Computer network security*. Eds. I. Kotenko and V. Skormin. Lecture notes in computer science ser. Berlin–Heidelberg: Springer-Verlag. 6258:118–124. doi: 10.1007/978-3-642-14706-7_9.
4. Grusho, A., and E. Timonina. 2011. Prohibitions in discrete probabilistic statistical problems. *Discrete Mathematics Application* 21(3):275–281.
5. Grusho, A., N. Grusho, and E. Timonina. 2013. Consistent sequences of tests defined by bans. *Springer proceedings in mathematics & statistics, optimization theory, decision making, and operation research applications*. New York – Heidelberg – Dordrecht – London: Springer. 281–291.
6. Bourbaki, N. 1940. *Topologie Générale*. Chapitre 1: Structures topologiques. Chapitre 2: Structures uniformes. Paris: Hermann. 129 p.
7. Prokhorov, U. V., and U. A. Rozanov. 1973. Teoriya veroyatnostey [Theory of probabilities]. 2nd ed. Moscow: Nauka. 494 p.
8. Grusho, A., N. Grusho, and E. Timonina. 2013. Statisticheskie metody opredeleniya zapretov veroyatnostnykh mer na diskretnykh prostranstvakh [Statistical techniques of bans determination of probability measures in discrete spaces]. *Informatika i ee Primeneniya — Inform. Appl.* 7(1): 54–57.

Received September 28, 2014

Contributors

Grusho Alexander A. (b. 1946) — Doctor of Science in physics and mathematics, Corresponding member of the Russian Academy of Cryptography; leading scientist, Institute of Informatics Problems, Russian Academy of Sciences, 44-2 Vavilov Str., Moscow 119333, Russian Federation; professor, Faculty of Computational Mathematics and Cybernetics, M. V. Lomonosov Moscow State University, 1-52 Leninskiye Gory, GSP-1, Moscow 119991, Russian Federation; grusho@yandex.ru

Grusho Nikolai A. (b. 1982) — Candidate of Science (PhD) in physics and mathematics, senior scientist, Institute of Informatics Problems, Russian Academy of Sciences, 44-2 Vavilov Str., Moscow 119333, Russian Federation; info@itake.ru

Timonina Elena E. (b. 1952) — Doctor of Science in technology, professor, leading scientist, Institute of Informatics Problems, Russian Academy of Sciences, 44-2 Vavilov Str., Moscow 119333, Russian Federation; eltimon@yandex.ru

ИНФОРМАТИКА И ЕЁ ПРИМЕНЕНИЯ, 2014. Т. 8. Вып. 4. С. 46–52

ВКЛЮЧЕНИЕ НОВЫХ ЗАПРЕТОВ В СЛУЧАЙНЫЕ ПОСЛЕДОВАТЕЛЬНОСТИ*

А. А. Грушо[1], Н. А. Грушо[2], Е. Е. Тимонина[3]

Аннотация: Рассматривается задача порождения одних вероятностных мер на пространстве бесконечных последовательностей над конечными алфавитами с σ-алгеброй, порожденной цилиндрическими множествами, из других вероятностных мер на этом пространстве. При этом новая вероятностная мера устроена так, чтобы определенным образом сокращать множество допустимых траекторий случайных последовательностей. Недопустимость траекторий определяется в терминах спецификаций наименьших запретов.

Ключевые слова: случайные последовательности; запреты вероятностных мер; порождение вероятностных мер; статистические задачи на случайных последовательностях

DOI: 10.14357/19922264140406

1 Введение

В системах контроля информационной безопасности и при поиске скрытых каналов часто возникает задача поиска аномалий в наблюдаемой случайной последовательности. Одним из основных инструментов в решении таких задач является математическая статистика. Однако при ненулевой ошибке принятия решения об аномалии наблюдение процесса порождает большое число ложных тревог [1], которые затрудняют или делают невозможным анализ причин аномалий. Для случайных процессов с дискретным временем и конечным множеством состояний найден подход для проверки последовательности гипотез о распределении проекций вероятностных мер на пространстве бесконечных последовательностей, при котором вероятность ложной тревоги всегда равна нулю. При этом с ростом размерностей вероятность правильного решения о наличии аномалий стремится к единице. Этот подход основан на запретах вероятностных мер [2, 3].

В предыдущих исследованиях [2, 3] было введено определение запрета для вероятностной меры на конечном пространстве. Запрет означает последовательность, имеющую нулевую вероятность на конечном пространстве. Было показано, что введение понятия запрета полезно для решения указанных выше задач. Запреты позволяют определять критические множества статистических критериев простейшим для вычисления способом [2]. Были доказаны необходимые и достаточные условия существования состоятельной последовательности критериев, в которых все критические множества статистических критериев определяются с помощью запретов [3].

Статистические критерии, основанные на запретах, обладают важными особенностями. Вероятности ложных решений равны нулю. Так как критические множества статистических критериев определяются только запретами, то эти критерии одинаковы для вероятностных мер, имеющих одинаковые множества запретов (т. е. это робастные критерии). В случае проверки гипотез состоятельность определяется тем условием, что вероятность появления хотя бы одного запрета при альтернативе стремится к единице.

Рассмотрим задачу внесения запретов в случайную последовательность.

Приведем следующий пример [4]. Для поиска скрытых каналов в случайную последовательность вносится запрет. При этом организаторы скрытого канала не знают о внесенном запрете, поэтому, когда в наблюдаемой последовательности наблюдается запрет, контролер определяет, что функционирует скрытый канал.

В работе [4] было показано, что другие статистические методы, отличные от методов, основанных на запретах, очень чувствительны к изменениям вероятностных моделей, поэтому статистические процедуры поиска скрытых каналов необходимо

*Работа частично поддержана РФФИ (проект 13-01-00215).

[1]Институт проблем информатики Российской академии наук; факультет вычислительной математики и кибернетики Московского государственного университета им. М. В. Ломоносова, grusho@yandex.ru

[2]Институт проблем информатики Российской академии наук, info@itake.ru

[3]Институт проблем информатики Российской академии наук, eltimon@yandex.ru

строить на робастных процедурах, учитывающих гетерогенный характер последовательностей передаваемых сообщений. Условиям робастности удовлетворяют статистические методы, основанные на запретах.

Пусть исходная случайная последовательность определена с помощью вероятностной меры P на пространстве бесконечных последовательностей с σ-алгеброй, порожденной цилиндрическими множествами. Проекции этой меры удовлетворяют условию согласованности, т. е. распределение очередного $(n+1)$-го знака в последовательности выражается через условное распределение появления этого знака при условии появления предыдущих n знаков последовательности и безусловного распределения этих n знаков. Внесение запрета не должно нарушать условие согласованности, иначе мера на пространстве бесконечных последовательностей может определяться некорректно. В рассматриваемом методе условных распределений для внесения запрета необходимо оперировать вероятностями всех начальных участков случайной последовательности.

В работе рассматривается другой способ внесения запрета в случайную последовательность. Этот метод основан на корректном определении некоторых множеств функций и учитывает спецификации наименьших запретов, которые должны присутствовать в новой построенной мере. В предлагаемом методе сохраняются условия согласованности, что позволяет использовать теорему Каратеодори [5] об однозначном продолжении меры.

Статья имеет следующую структуру. В разд. 2 приводятся определения и предыдущие результаты. Раздел 3 определяет условия для случая, когда вероятностная мера генерируется согласно заданной спецификации наименьших запретов. В разд. 4 приводится пример использования доказанных условий. В разд. 5 кратко анализируются условия существования состоятельной последовательности критериев для мер, построенных в примере.

2 Основные определения и предыдущие результаты

Пусть X_i, $i = 1, 2, \ldots$, — конечные множества, $\prod\limits_{i=1}^{n} X_i$ — декартово произведение X_i, $i = 1, 2, \ldots, n$, X^∞ — множество всех последовательностей, где i-й элемент принадлежит X_i. Пусть $\mathcal{A}$ — это σ-алгебра на X^∞, порожденная цилиндрическими множествами; $\mathcal{A}$ также является борелевской σ-алгеброй в тихоновском произведении X^∞, где X_i имеют дискретную топологию [6, 7].

На $(X^\infty, \mathcal{A})$ определена вероятностная мера P. Предположим, что P_n является проекцией меры P на пространство конечномерных векторов, порожденных первыми n координатами последовательностей из X^∞. Обозначим $X_n^\infty = \prod\limits_{i=n+1}^{\infty} X_i$. Ясно, что для каждого $B_n \subseteq \prod\limits_{i=1}^{n} X_i$

$$P_n(B_n) = P\left(B_n \times X_n^\infty\right).$$

Пусть D_n — носитель меры P_n:

$$D_n = \left\{ \vec{x}_n \in \prod_{i=1}^{n} X_i \;\; P_n\left(\vec{x}_n\right) > 0 \right\}.$$

Обозначим $\Delta_n = D_n \times X^\infty$. Последовательность Δ_n, $n = 1, 2, \ldots$, невозрастающая и

$$\Delta_P = \lim_{n\to\infty} \Delta_n = \bigcap_{n=1}^{\infty} \Delta_n .$$

Множество Δ_P замкнуто в топологии тихоновского произведения и является носителем меры P. Если $\overline{\omega}^{(k)} \in \prod\limits_{i=1}^{k} X_i$, то $\tilde{\omega}^{(k-1)}$ получается из $\overline{\omega}^{(k)}$ отбрасыванием последней координаты.

Определение 1. Запретом в мере P_n называется вектор $\overline{\omega}^{(k)} \in \prod\limits_{i=1}^{k} X_i$, $k \le n$, такой что

$$P_n\left(\overline{\omega}^{(k)} \times \prod_{i=k+1}^{n} X_i\right) = 0 .$$

Если $P_{k-1}\left(\tilde{\omega}^{(k-1)}\right) > 0$, то $\overline{\omega}^{(k)}$ называется наименьшим запретом.

Если $\overline{\omega}^{(k)}$ является запретом в мере P_n, тогда для любых $k \le s \le n$ и любых последовательностей $\overline{\omega}^{(s)}$, начинающихся с последовательности $\overline{\omega}^{(k)}$, имеем:

$$P_s\left(\overline{\omega}^{(s)}\right) = 0 .$$

Действительно, если $P_k\left(\overline{\omega}^{(k)}\right) = 0$, то

$$P\left(\overline{\omega}^{(k)} \times X_k^\infty\right) = 0 ;$$

$$P\left(\overline{\omega}^{(k)} \times \prod_{i=k+1}^{s} X_i \times X_s^\infty\right) = 0 .$$

Из этого следует, что

$$P_s\left(\overline{\omega}^{(s)}\right) = P\left(\overline{\omega}^{(s)} \times X_s^\infty\right) \le$$

$$\le P\left(\overline{\omega}^{(k)} \times \prod_{i=k+1}^{s} X_i \times X_s^\infty\right) = 0 .$$

Если существует $\overline{\omega}^{(n)} \in \prod\limits_{i=1}^{n} X_i$ такое, что $P_n\left(\overline{\omega}^{(n)}\right) = 0$, то существует наименьший запрет.

Пусть для всех n носители мер P_n совпадают с $\prod\limits_{i=1}^{n} X_i$. Тогда носитель меры P совпадает с X^∞. Предположим, что задана спецификация исходных наименьших запретов $\nu' = \{\nu'_n, n = 1, 2, \ldots\}$, где ν'_n — число наименьших запретов длины n в мере P. Пусть задана новая спецификация, определяемая дополнительными ограничениями $\nu = \{\nu_n, \nu_n \geq \nu'_n, n = 1, 2, \ldots\}$. Задача состоит в том, чтобы, используя меру P и спецификации ν и ν', построить вероятностную меру Q на пространстве $(X^\infty, \mathcal{A})$, у которой множество наименьших запретов обладает спецификацией ν. Для построения Q сначала построим согласованную систему вероятностных мер Q_n, $n = 1, 2, \ldots$, на пространствах, определяемых проекциями X^∞ на первые n координат. Эти меры определяют аддитивную меру на алгебре цилиндрических множеств, которая по теореме Каратеодори будет однозначно определять меру Q на $(X^\infty, \mathcal{A})$. Далее будем обозначать через D'_n, $n = 1, 2, \ldots$, $D'_n \subseteq \prod\limits_{i=1}^{n} X_i$, носители мер P_n, через d'_n — мощности этих носителей, а через D_n, $n = 1, 2, \ldots$, $D_n \subseteq \prod\limits_{i=1}^{n} X_i$, носители мер Q_n и через d_n — мощности этих носителей.

В работе [3] доказано, что числа d_n, $n = 1, 2, \ldots$, однозначно связаны со спецификацией ν следующими соотношениями:

$$\nu_1 \prod_{i=2}^{n} m_i + \cdots + \nu_{n-1} m_n + \nu_n + d_n = \prod_{i=1}^{n} m_i \,. \quad (1)$$

для всех $n = 1, 2, \ldots$

Таким образом, необходимо построить согласованное семейство вероятностных мер $\{Q_n\}$, мощности носителей которых однозначно определены соотношениями (1).

3 Порождение вероятностных мер с заданной спецификацией наименьших запретов

Пусть $\{D_n,\ D_n \subseteq D'_n,\ D_n \subseteq \prod\limits_{i=1}^{n} X_i,\ n = 1, 2, \ldots\}$ — некоторое семейство множеств, удовлетворяющих (1), $\overline{x}_n$ — произвольный элемент $\prod\limits_{i=1}^{n} X_i$. Для любого n, $n = 1, 2, \ldots$, определим функцию

$$g_{n+1} : D_{n+1} \to \prod_{i=1}^{n} X_i$$

следующим образом. Для любых $\overline{x}_{n+1} \in D_{n+1}$, $\overline{x}_{n+1} = \overline{x}_n x$, где $\overline{x}_n \in \prod\limits_{i=1}^{n} X_i$, $x \in X_{n+1}$, определяем

$$g_{n+1}\left(\overline{x}_{n+1}\right) = \overline{x}_n \,.$$

Кроме (1) на множества $\{D_n\}$ наложим следующие два ограничения, связанных с функциями g_n. Для любого n, $n = 1, 2, \ldots$, и любых $\overline{x}_{n+1} \in D_{n+1}$

$$g_{n+1}\left(\overline{x}_{n+1}\right) \in D_n \,; \quad (2)$$

$$g_{n+1} : D_{n+1} \xrightarrow{\text{на}} D_n \,. \quad (3)$$

По аналогии с функциями g_n определяются функции h_n для последовательности множеств D'_n так, что для любого n, $n = 1, 2, \ldots$, и любых $\overline{x}_n x \in D'_{n+1}$

$$h_{n+1}\left(\overline{x}_n x\right) = \overline{x}_n \,.$$

Лемма. *Пусть P — вероятностная мера на $(X^\infty, \mathcal{A})$, ν' — спецификация наименьших запретов, $\{D'_n\}$ — носители мер P_n, $d'_n = |D'_n|$, $n = 1, 2, \ldots$ Тогда для ν', $\{D'_n\}$, $\{d'_n\}$, $\{h_n\}$ выполняются соотношения* (1)–(3).

Доказательство. Выполнение (1) доказано в [3]. По определению $\forall\, \overline{x}_{n+1} = \overline{x}_n x \in D'_{n+1}$

$$h_{n+1}\left(\overline{x}_{n+1}\right) = \overline{x}_n \,.$$

По определению D'_{n+1} имеем:

$$P_n\left(\overline{x}_{n+1}\right) > 0 \,;$$

$$0 < P_{n+1}\left(\overline{x}_{n+1}\right) = P\left(\overline{x}_{n+1} \times X^\infty_{n+1}\right) =$$

$$= P\left(\overline{x}_n \times \{x\} \times X^\infty_{n+1}\right) \leq$$

$$\leq P\left(\overline{x}_n \times X_n \times X^\infty_{n+1}\right) = P\left(\overline{x}_n \times X^\infty_n\right) = P_n\left(\overline{x}_n\right) .$$

Отсюда следует, что $\overline{x}_n \in D'_n$. Это доказывает соотношение (2).

Если $\overline{x}_n \in D'_n$ и $\forall x \in X_{n+1}$ выполняется

$$P_{n+1}\left(\overline{x}_n x\right) = 0 \,,$$

то $P_{n+1}\left(\overline{x}_n, X_{n+1}\right) = 0$, что противоречит согласованности мер P_n и P_{n+1}:

$$P_{n+1}\left(\overline{x}_n, X_{n+1}\right) = P_n\left(\overline{x}_n\right)$$

и предположению, что $P_n\left(\overline{x}_n\right) > 0$. Значит, существует такое x, что $P_{n+1}\left(\overline{x}_n x\right) > 0$, т. е. $\overline{x}_n x \in D'_{n+1}$. Это доказывает соотношение (3).

Возьмем произвольную последовательность сюръективных функций $f_n : D'_n \to D_n$, $n = 1, 2, \ldots$

Каждая такая функция порождает на D_n и, следовательно, на $\prod\limits_{i=1}^{n} X_i$ вероятностную меру Q_n с носителем D_n.

Тогда для всех n функции g_{n+1} и меры Q_{n+1} порождают на $\prod_{i=1}^{n} X_i$ вероятностные меры Q'_n с носителями D_n (это следует из сюръективности f_{n+1} и g_{n+1}).

Теорема 1. *Пусть задано произвольное семейство вероятных мер $\{Q_n\}$ с носителями $\{D_n\}$ и семейство функций $\{g_n\}$, удовлетворяющих условиям* (2) *и* (3). *Семейство вероятностных мер $\{Q_n\}$ является согласованным семейством тогда и только тогда, когда для всех n выполняются равенства $Q_n = Q'_n$.*

Доказательство. Докажем достаточность. Для согласованности мер достаточно, чтобы для любых $\overline{x} \in \prod_{i=1}^{n} X_i$

$$Q_n(\overline{x}_n) = Q_{n+1}(\overline{x}_n, X_{n+1}) .$$

Из конечности вероятностных схем

$$Q_{n+1}(\overline{x}_n, X_{n+1}) = \sum_{x \in X_{n+1}} Q_{n+1}(\overline{x}_n x) .$$

По определению

$$Q'_n(\overline{x}_n) = Q_{n+1}\left(g_{n+1}^{-1}(\overline{x}_n)\right) = \sum_{(\overline{x}_n x) \in D_{n+1}} Q_{n+1}(\overline{x}_n x) =$$

$$= \sum_{x \in X_{n+1}} Q_{n+1}(\overline{x}_n x) = Q_{n+1}(\overline{x}_n, X_{n+1}) .$$

По условию теоремы

$$Q'_n(\overline{x}_n) = Q_n(\overline{x}_n)$$

для любых $\overline{x}_n \in \prod_{i=1}^{n} X_i$. Отсюда следует, что для любых $\overline{x}_n \in \prod_{i=1}^{n} X_i$

$$Q_n(\overline{x}_n) = Q_{n+1}(\overline{x}_n, X_{n+1}) .$$

Достаточность доказана. Докажем необходимость. Если $\{Q_n\}$ — согласованное семейство вероятностных мер, то для любых $\overline{x}_n \in \prod_{i=1}^{n} X_i$

$$Q_{n+1}(\overline{x}_n, X_{n+1}) = Q_n(\overline{x}_n) .$$

Кроме того, для любых $\overline{x}_n \in \prod_{i=1}^{n} X_i$

$$Q'_n(\overline{x}_n) = Q_{n+1}\left(g_{n+1}^{-1}(\overline{x}_n)\right) =$$

$$= Q_{n+1}(\overline{x}_n, X_{n+1}) = Q_n(\overline{x}_n) .$$

Теорема доказана.

Семейство функций $\{f_n\}$ и вероятностная мера P порождают семейство вероятностных мер $\{Q_n\}$ с носителями $\{D_n\}$. Пусть функции $\{g_n\}$ удовлетворяют условия (2) и (3). Тогда справедливо следующее утверждение.

Следствие 1. Семейство функций $\{f_n\}$ и вероятностная мера P порождают единственную вероятностную меру Q тогда и только тогда, когда для всех $n = 1, 2, \ldots$ выполняется равенство $Q_n = Q'_n$.

Теорема 2. *Для согласованности множества вероятностных мер $\{Q_n\}$, порожденных функциями $\{f_n\}$ и проекциями меры P, достаточно, чтобы функции $\{g_n\}$ удовлетворяли соотношениям* (2) *и* (3) *и для всех n были коммутативны следующие диаграммы*:

$$\begin{array}{ccc} D'_n & \xleftarrow{h_{n+1}} & D'_{n+1} \\ f_n \downarrow & & \downarrow f_{n+1} \\ D_n & \xleftarrow{g_{n+1}} & D_{n+1} \end{array} \qquad (4)$$

Доказательство. Каждая функция f_n и мера P_n на D'_n порождают на D_n вероятностное распределение Q_n. В силу согласованности проекций меры P каждая функция h_{n+1} порождает меру P_n из меры P_{n+1}. Поэтому можно считать, что мера Q_n порождена из меры P_{n+1} с помощью композиции отображений $(f_n * h_{n+1})$.

В свою очередь, функция f_{n+1} и мера P_{n+1} порождают распределение вероятностей Q_{n+1} на D_{n+1}. Эта мера и функция g_{n+1} порождают меру Q'_n на D_n, т. е. мера Q'_n на D_n порождена из меры P_{n+1} с помощью композиции функций $(g_{n+1} * f_{n+1})$. По условию (4) функции $(f_n * h_{n+1}))$ и $(g_{n+1} * f_{n+1})$ совпадают. Следовательно, эти функции и мера P_{n+1} порождают на D_n одну и ту же меру, т. е. $Q_n = Q'_n$. Отсюда и из теоремы 1 следует согласованность семейства вероятностных мер $\{Q_n\}$. Теорема доказана.

4 Пример порождения вероятностных мер с заданной спецификацией наименьших запретов

Пусть $\nu = \{\nu_i = 1, i = 1, 2, \ldots\}$, $X_m = \{0, 1, \ldots, m-1\}$, $m > 2$, и P — равномерная мера на X^∞. Пусть элементы $\prod_{i=1}^{n} X_i$ лексикографически упорядочены.

Приведем пример построения меры Q со спецификацией наименьших запретов ν с помощью подхода, описанного в разд. 3.

В каждом множестве $B_n \subseteq \prod_{i=1}^{n} X_i$ есть наименьший вектор $\overline{x}_n$ с точки зрения лексикографического порядка. Требуемую меру будем строить индуктивно. В D_1 наименьшим запретом будем считать 0.

Функция f_1 отображает X_1 в $X_1\backslash\{0\}$. Предположим, что определены D_n и f_n. Определим D_{n+1}. Пусть $\overline{x}_n^0$ — наименьший элемент в D_n. В множестве $D_n \times X_{n+1}$ определим наименьший запрет — $\left(\overline{x}_n^0 0\right)$. Положим

$$D_{n+1} = (D_n \times X_{n+1}) \setminus \{(\overline{x}_n^0 0)\} .$$

Построим сюръективную функцию

$$f_{n+1} : \prod_{i=1}^{n+1} X_i \overset{\text{на}}{\to} D_{n+1} .$$

Для любых $(\overline{x}_n x) \in \prod\limits_{i=1}^{n+1} X_i$, кроме тех, у которых $f_n(\overline{x}_n) = \overline{x}_n^0$ и $x = 0$, положим

$$f_{n+1}(\overline{x}_n x) = (f_n(\overline{x}_n), x) \in D_n \times X_{n+1} .$$

Обозначим $\overline{y}_n^{(i)}$, $i = 1, \ldots, k$, все элементы множества $f_n^{-1}\left(\overline{x}_n^0\right)$. Определим

$$f_{n+1}\left(\overline{y}_n^{(i)}, 0\right) = \left(\overline{x}_n^0, 1\right) \in D_n \times X_{n+1} .$$

Отметим, что $\left(\overline{x}_n^0, 1\right)$ — наименьший элемент в D_{n+1}. По определению f_n — это отображение $\prod\limits_{i=1}^{n} X_i$ на D_n. Поэтому f_{n+1} отображает X^{n+1} на $D_{n+1} = (D_n \times X_{n+1})\backslash\{\overline{x}_n^0, 0\}$. По построению D_{n+1} из D_n и из (1) следует, что $\nu_{n+1} = 1$ и этот элемент равен $\left(\overline{x}_n^0, 0\right)$.

Докажем коммутативность диаграмм (4). По построению f_n и h_{n+1} для любых $x \in X_{n+1}$

$$h_{n+1}\left(\overline{y}_n^{(i)}, x\right) = \overline{y}_n^{(i)} ,$$

поэтому

$$(f_n * h_{n+1})\left(\overline{y}_n^{(i)}, x\right) = \overline{x}_n^0 .$$

При $\overline{x}_n \neq \overline{y}_n^{(i)}$, $i = 1, \ldots, k$,

$$(f_n * h_{n+1})(\overline{x}_n, x) = f_n(\overline{x}_n) .$$

Далее

$$f_{n+1}\left(\overline{y}_n^{(i)}, 0\right) = \left(\overline{x}_n^0, 1\right) \in D_{n+1}, \quad i = 1, \ldots, k;$$

$$(f_{n+1} * g_{n+1})\left(\overline{y}_n^{(i)}, 0\right) = f_n\left(\overline{y}_n^{(i)}\right) = \overline{x}_n^0 .$$

Для элементов $\left(\overline{y}_n^{(i)}, x\right)$, $x \neq 0$, $i = 1, \ldots, k$,

$$f_{n+1}\left(\overline{y}_n^{(i)}, x\right) = \left(f_n\left(\overline{y}_n^{(i)}\right), x\right) = \left(\overline{x}_n^0, x\right) \in D_{n+1} .$$

Поэтому при $x \neq 0$

$$(f_{n+1} * g_{n+1})\left(\overline{y}_n^{(i)}, x\right) = \overline{x}_n^0 .$$

При $\overline{x}_n \neq \overline{y}_n^{(i)}$, $i = 1, \ldots, k$, по определению f_{n+1} имеем, что для любых $x \in X_{n+1}$

$$(f_{n+1})(\overline{x}_n, x) = (f_n(\overline{x}_n) x) \in D_{n+1} .$$

Тогда по построению

$$(g_{n+1} * f_{n+1}) = (f_n * h_{n+1}) .$$

Коммутативность диаграмм (4) доказана.

Отсюда следует существование меры Q на $(X^\infty, \mathcal{A})$ со спецификацией наименьших запретов $\nu = \{\nu_i = 1,\ i = 1, 2, \ldots\}$.

5 Применение к анализу состоятельности

Из соотношений (1) получаем соотношения:

$$d_{n+1} - m_{n+1} d_n + \nu_{n+1} = 0 ,\ n = 1, 2, \ldots \qquad (5)$$

Пусть P — равномерная мера на $(X^\infty, \mathcal{A})$. Тогда отношение $d_n \Big/ \prod\limits_{i=1}^{n} m_i$ есть вероятность множества D_n в мере P_n.

Для спецификации $\nu = \{\nu_i = 1,\ i = 1, 2, \ldots\}$ из (5) для некоторого положительного ε получаем следующие соотношения при $m_n \geq 3$, $n = 1, 2, \ldots$:

$$\frac{d_n}{\prod\limits_{i=1}^{n} m_i} > \varepsilon > 0 .$$

При $n \to \infty$ предел этой вероятности равен вероятности P носителя Δ_Q меры Q

$$P(\Delta_Q) \geq \varepsilon > 0 .$$

Из необходимых и достаточных условий [2] существования состоятельных последовательностей критериев, определяемых запретами, для проверки гипотез $H_{0,n} : Q_n$ против $H_{1,n} : P_n$ следует, что таких последовательностей критериев нет.

6 Заключение

Получены условия корректного построения дополнительных ограничений на случайную последовательность с помощью спецификации наименьших запретов. Корректное построение стохастических моделей позволяет использовать в анализе статистических данных хорошо разработанный аппарат теории случайных последовательностей и процессов.

Литература

1. *Axelson S.* The base-rate fallacy and its implications for the difficulty of intrusion detection // 6th ACM Conference on Computer and Communications Security Proceedings. — New York: ASM, 1999. P. 1–7.
2. *Грушо А., Тимонина Е.* Запреты в дискретных вероятностно-статистических задачах // Дискретная математика, 2011. Т. 23. Вып. 2. С. 53–58.
3. *Grusho A., Grusho N., Timonina E.* Consistent sequences of tests defined by bans // Springer proceedings in mathematics & statistics, optimization theory, decision making, and operation research applications. — New York – Heidelberg – Dordrecht – London: Springer, 2013. P. 281–291.
4. *Grusho A., Grusho N., Timonina E.* Problems of modeling in the analysis of covert channels // Computer network security, 2010. Lecture notes in computer science ser. Vol. 6258. P. 118–124. doi: 10.1007/978-3-642-14706-7_9.
5. *Неве Ж.* Математические основы теории вероятностей / Пер. с англ. — М.: Мир, 1969. 309 с. (*Neveu J.* Bases mathematiques du calcul des probabilites. Paris: Masson, 1964. 203 p.)
6. *Бурбаки Н.* Общая топология. Основные структуры / Пер. с франц. — М.: Наука, 1968. 272 с. (*Bourbaki N.* Topologie Générale. Chapitre 1: Structures topologiques. Chapitre 2: Structures uniformes. — Paris: Hermann, 1940. 129 p.)
7. *Прохоров Ю. В., Розанов Ю. А.* Теория вероятностей. — М.: Наука, 1993. 496 с.

Поступила в редакцию 31.10.14

SWITCHING ON OF NEW BANS IN RANDOM SEQUENCES

A. A. Grusho[1,2], N. A. Grusho[1], and E. E. Timonina[1]

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

[2]Faculty of Computational Mathematics and Cybernetics, M. V. Lomonosov Moscow State University, 1-52 Leninskiye Gory, GSP-1, Moscow 119991, Russian Federation

Abstract: The problem of generating one probability measure on space of the infinite sequences on finite alphabets with σ-algebra generated by cylindrical sets out of another probability measure on this space is considered. A new probability measure is arranged to reduce the set of admissible trajectories of random sequences definitely. Inadmissibility of trajectories is defined in terms of specifications of the smallest bans. If a specification of the smallest bans is given, then the powers of support of projections of the new measure can be determined. It gives conditions to construct several sets of functions. These functions and projections of the initial measure define a set of measures on finite spaces which define the only probability measure on the space of infinite sequences.

Keywords: random sequences; bans of probability measures; generation of probability measures; statistical problems on random sequences

DOI: 10.14357/19922264140406

Acknowledgments

The paper was partially supported by the Russian Foundation for Basic Research (project 13-01-00215).

References

1. Axelsson, S. 1999. The base-rate fallacy and its implications for the difficulty of intrusion detection. *6th Conference on Computer and Communications Security Proceedings*. New York: ASM. 1–7.
2. Grusho, A., and E. Timonina. 2011. Prohibitions in discrete probabilistic statistical problems. *Discrete Mathematics Applications* 21(3):275–281.
3. Grusho, A., N. Grusho, and E. Timonina. 2013. Consistent sequences of tests defined by bans. *Springer proceedings in mathematics & statistics, optimization theory, decision making, and operation research applications*. New York – Heidelberg – Dordrecht – London: Springer. 281–291.
4. Grusho, A., N. Grusho, and E. Timonina. 2010. Problems of modeling in the analysis of covert channels. *Computer network security*. Lecture notes in computer science ser. 6258:118–124. doi: 10.1007/978-3-642-14706-7_9.
5. Neveu, J. 1964. *Bases mathematiques du calcul des probabilites*. Paris: Masson. 203 p.
6. Bourbaki, N. 1940. *Topologie Générale*. Chapitre 1: Structures topologiques. Chapitre 2: Structures uniformes. Paris: Hermann, 1940. 129 p.
7. Prokhorov, Yu. V., and Yu. A. Rozanov. 1993. *Teoriya veroyatnostey* [Theory of probabilities]. Moscow: Nauka. 496 p.

Received October 31, 2014

Contributors

Grusho Alexander A. (b. 1946) — Doctor of Science in physics and mathematics, Corresponding member of the Russian Academy of Cryptography; leading scientist, Institute of Informatics Problems, Russian Academy of Sciences, 44-2 Vavilov Str., Moscow 119333, Russian Federation; professor, Faculty of Computational Mathematics and Cybernetics, M. V. Lomonosov Moscow State University, 1-52 Leninskiye Gory, GSP-1, Moscow 119991, Russian Federation; grusho@yandex.ru

Grusho Nikolai A. (b. 1982) — Candidate of Science (PhD) in physics and mathematics, senior scientist, Institute of Informatics Problems, Russian Academy of Sciences, 44-2 Vavilov Str., Moscow 119333, Russian Federation; info@itake.ru

Timonina Elena E. (b. 1952) — Doctor of Science in technology, professor, leading scientist, Institute of Informatics Problems, Russian Academy of Sciences, 44-2 Vavilov Str., Moscow 119333, Russian Federation; eltimon@yandex.ru

ИНФОРМАТИКА И ЕЁ ПРИМЕНЕНИЯ, 2014. Т. 8. Вып. 4. С. 53–57

ОБ ОПТИМАЛЬНОЙ ДОСТАВКЕ ГРУЗОВ ТРАНСПОРТНЫМ СРЕДСТВОМ С УЧЕТОМ ЗАВИСИМОСТИ СТОИМОСТИ ПЕРЕВОЗОК ОТ ЗАГРУЗКИ ТРАНСПОРТНЫХ СРЕДСТВ ПО НЕСКОЛЬКИМ ЦИКЛИЧЕСКИМ МАРШРУТАМ*

Е. М. Бронштейн[1], П. А. Зелёв[2]

Аннотация: Рассматривается задача построения маршрута доставки грузов потребителям от одного производителя (базы, склада) транспортным средством (ТС) с минимальными затратами на перевозки. При этом учитывается зависимость стоимости транспортировки от загрузки ТС и качества дороги. Предполагается, что ТС может возвращаться на базу для дозагрузки. Построена соответствующая математическая модель. Для случая линейной зависимости стоимости проезда от загрузки получена линейная целочисленная модель. Для решения поставленной задачи наряду с точным алгоритмом предложена модификация известного эвристического алгоритма Кларка–Райта. Проведен вычислительный эксперимент.

Ключевые слова: эвристический алгоритм; построение маршрута; транспортировка; задача маршрутизации

DOI: 10.14357/19922264140407

1 Введение

Систематическое изучение оптимизационных задач транспортной логистики (Vehicle Routing Problem, VRP) началось с работы [1]. За последние полвека поставлено множество задач этого типа, развиты как точные, так и эвристические методы решения. Частным случаем подобных задач является известная задача коммивояжера. Классификация оптимизационных задач транспортной логистики приведена, например, в [2].

Задача, рассматриваемая в данной работе, перекликается с CVRP (Capacitated VRP), в этих задачах учитываются ограничения на вместимость ТС (см., например, [3]). В большинстве случаев при постановке задач маршрутизации стоимость транспортировки является функцией, зависящей только от расстояния между городами.

В работе [4] предлагается при постановке задач маршрутизации учитывать ряд дополнительных факторов, часть из которых рассматривается в настоящей работе. В частности, предполагается, что стоимость транспортировки зависит от загрузки ТС и состояния дороги.

Рассматриваемая задача является NP-трудной, как обобщение задачи коммивояжера, в связи с чем точные методы применимы лишь при малых размерностях задач (менее 10), а значит, актуальной задачей является разработка эвристических методов ее решения.

2 Постановка задачи

Груз следует доставить потребителям из пункта производства ТС грузоподъемностью Q. Известны значения q_i — потребности в грузе i-го пункта потребления ($i = 1, \ldots, n$), n — число пунктов потребления. Для единообразия будем считать пункт производства (базу, склад) нулевым пунктом. Предполагается, что ТС, доставив груз в некоторые пункты, возвращается на базу и загружается для доставки другим потребителям.

Дороги между пунктами характеризуются двумя показателями: расстоянием l_{ij} ($i, j = 0, \ldots, n$) и коэффициентом сложности дороги k_{ij} между пунктами i и j. Расстояние l_{ij} может зависеть от направления движения (например, есть дороги с односторонним движением), коэффициент сложности дороги влияет на расход топлива, его величина также может зависеть от направления маршрута (пример: подъем в гору или спуск с горы) [5].

Каждый пункт посещается в точности один раз (не допускается split delivery [6]). Очевидно, что задача имеет решение тогда и только тогда, когда вместимость ТС не меньше потребности в грузе

*Работа выполнена при поддержке РФФИ (проект 13-01-00005).

[1]Уфимский государственный авиационный технический университет, bro-efim@yandex.ru

[2]Уфимский государственный авиационный технический университет, pz1988@yandex.ru

в каждом из пунктов потребления. Число циклов, за которые происходит доставка, не превосходит n. Будем считать, что оно равно n, но среди них могут быть пустые.

Предполагается известной зависимость $f(q)$ расхода топлива на единицу пути дороги стандартного качества.

Функция расхода топлива в зависимости от веса перевозимого груза является возрастающей и (как правило) вогнутой на всем интервале области определения функции $[0;\, Q]$.

Введем булевы переменные X_{ij}^t $(i, j, t = 0, \ldots, n)$, равные 1 тогда и только тогда, когда следующим после i-го пункта на пути следования ТС в t-м цикле является j-й пункт. Должны выполняться следующие условия:

$$\sum_{i=0}^{n} X_{i0}^t = \sum_{j=0}^{n} X_{0j}^t = 1\,,\ \ t = 1, \ldots, n\,; \tag{1}$$

$$\sum_{t=1}^{n}\sum_{i=0}^{n} X_{ij}^t = 1\,,\ \ j = 1, \ldots, n\,; \tag{2}$$

$$\sum_{t=1}^{n}\sum_{i=0}^{n} X_{ij}^t = 1\,,\ \ i = 1, \ldots, n\,. \tag{3}$$

Ограничение (1) означает наличие в каждом цикле одного пункта потребления (или базы для нулевого маршрута), из которого ТС попадает непосредственно на базу, и ровно одного пункта потребления (или базы для нулевого маршрута), в который ТС выезжает с базы; ограничения (2) и (3) описывают условие однократного посещения каждого пункта потребления.

Введем целочисленные переменные v_i^t $(i, t = 1, \ldots, n))$, имеющие смысл номеров потребителей в порядке прохождения в t-м цикле, содержащем все пункты, за исключением начального:

$$1 \le v_i^t \le \sum_{i=0}^{n}\sum_{j=0}^{n} X_{ij}^t\,,\ \ i, t = 1, \ldots, n\,; \tag{4}$$

$$\left(v_i^t - v_j^t\right) + nX_{ij}^t \le n - 1\,,\ \ i, j, t = 1, \ldots, n\,. \tag{5}$$

Из условий (4) и (5) следует отсутствие подциклов.

Введем, наконец, булевы переменные Z_{is}^t, равные 1 тогда и только тогда, когда $v_s^t > v_i^t$ $(i, s, t = 1, \ldots, n)$. Ограничения на введенные переменные имеют следующий вид:

$$nZ_{is}^t \ge v_s^t - v_i^t\,,\ \ i, s, t = 1, \ldots, n\,. \tag{6}$$

Это условие обеспечивает выполнение равенства $Z_{is}^t = 1$ при $v_s^t > v_i^t$.

Для обеспечения равенства $Z_{is}^t = 0$ при $v_s^t \le v_i^t$ заметим, что число величин v_i^t, меньших v_s^t, равно $v_s^t - 1$. Таким образом, выполнение нужного свойства обеспечивается равенствами:

$$\sum_{i=1}^{n} Z_{is}^t = v_s^t - 1\,,\, s\,, t = 1, \ldots, n\,. \tag{7}$$

Ограничение на вместимость:

$$\sum_{i=0}^{n}\sum_{j=0}^{n} q_i X_{ij}^t \le Q\,,\ \ t = 1, \ldots, n\,. \tag{8}$$

Целью является минимизация расходов на транспортировку:

$$R(X) = \\ = \sum_{t=1}^{n}\sum_{j=0}^{n}\sum_{i=0}^{n} X_{ij}^t f\left(\sum_{s=0}^{n} Z_{is}^t q_i\right) l_{ij} k_{ij} \to \min\,. \tag{9}$$

3 Линейная целочисленная модель

Предположим, что стоимость транспортировки является линейной функцией от массы перевозимого груза:

$$f(q) = vq + w\,.$$

Тогда при сохранении условий (1)–(8) целевая функция (9) примет вид:

$$R(X) = v\sum_{t=1}^{n}\sum_{s=0}^{n}\sum_{j=0}^{n}\sum_{i=0}^{n} X_{ij}^t Z_{is}^t q_s l_{ij} k_{ij} + \\ + w\sum_{t=1}^{n}\sum_{j=0}^{n}\sum_{i=0}^{n} X_{ij}^t l_{ij} k_{ij} \to \min\,. \tag{10}$$

Для приведения целевой функции к линейной форме введем булевы переменные $P_{ijs}^t = X_{ij}^t Z_{is}^t$ $(i, j, s, t = 0, \ldots, n)$. Это условие равносильно выполнению следующей системы неравенств:

$$P_{ijs}^t \ge X_{ij}^t + Z_{is}^t - 1\,,\ \ i, j, s, t = 0, \ldots, n\,; \tag{11}$$

$$P_{ijs}^t \le X_{ij}^t\,,\qquad\qquad i, j, s, t = 0, \ldots, n\,; \tag{12}$$

$$P_{ijs}^t \le Z_{is}^t\,,\qquad\qquad i, j, s, t = 0, \ldots, n\,. \tag{13}$$

Тогда целевая функция (10) примет вид:

$$R(X) = v\sum_{t=1}^{n}\sum_{s=0}^{n}\sum_{j=0}^{n}\sum_{i=0}^{n} P_{ijs}^t q_s l_{ij} k_{ij} + \\ + w\sum_{t=1}^{n}\sum_{j=0}^{n}\sum_{i=0}^{n} X_{ij}^t l_{ij} k_{ij} \to \min\,. \tag{14}$$

Как число переменных задачи (1)–(8), (11)–(14), так и число ограничений имеют порядок $O(n^4)$.

4 Модифицированный алгоритм Кларка–Райта

Для решения задачи был модифицирован эвристический алгоритм Кларка–Райта [7, 8]. Достоинствами метода являются его простота, надежность и гибкость.

Опишем основные идеи алгоритма.

Метод Кларка–Райта является итерационным, причем на каждой итерации осуществляется попытка сращивания двух циклических маршрутов по определенным правилам. Модификация метода Кларка–Райта заключается в возможности варьирования глубины сращивания маршрутов p, а также в учете зависимости стоимости перевозки от направления прохождения цикла. В классическом алгоритме Кларка–Райта глубина $p = 0$.

Изначально генерируются n циклов вида $0–i–0$ ($i = \overline{1,n}$). На каждом из последующих шагов рассматриваются всевозможные допустимые пары циклов (т. е. такие, для которых суммарная загрузка не превосходит Q), для каждой из которых строится не более 8 циклов при $p = 0$ и не более $16p$ циклов при $p \geq 1$ следующим образом: каждый из новых циклов состоит из начального или конечного отрезка (длиной не более p) одного из циклов, затем проходится второй цикл (без нулевой вершины), затем продолжается движение по первому циклу. При этом учитывается возможность изменения направления движения по каждому из циклов на противоположное, поскольку задача не предполагается симметричной и загрузка ТС зависит от порядка прохождения пунктов.

Например, сращивание циклов $a^1_1, a^1_2, \ldots, a^1_{k_1}$ и $a^2_1, a^2_2, \ldots, a^2_{k_2}$ (с исключенной нулевой вершиной) при единичных начальном и конечном отрезках дает следующие циклы:

$$0, a^1_1, a^2_1, a^2_2, \ldots, a^2_{k_2}, a^1_2, \ldots, a^1_{k_1}, 0;$$
$$0, a^1_1, a^2_{k_2}, a^2_{k_2-1}, \ldots, a^2_1, a^1_2, \ldots, a^1_{k_1}, 0;$$
$$0, a^1_{k_1}, a^1_{k_1-1}, \ldots, a^1_2, a^2_1, a^2_2, \ldots, a^2_{k_2}, a^1_1, 0;$$
$$0, a^1_{k_1}, a^1_{k_1-1}, \ldots, a^1_2, a^2_{k_2}, a^2_{k_2-1}, \ldots, a^2_1, a^1_1, 0\,.$$

Еще четыре цикла возникнут при добавлении части второго цикла между пунктами $a^1_{k_1}$ и $a^1_{k_1-1}$. Затем первый и второй циклы можно поменять местами, тогда получим еще 8 циклов.

В каждом случае вычисляется разность между суммой расходов на доставку грузов по отдельным циклам и расходов на доставку грузов по их объединению и выбирается объединенный цикл, для которого эта разность максимальная. При этом в модификации алгоритма Кларка–Райта для каждой пары циклов перебираются все возможные варианты их сращивания при всех значениях глубины сращивания от 0 до заданного пользователем значения p. Процесс прекращается при невозможности дальнейшего сращивания циклов или отсутствии пары циклов, сращивание которых выгодно [7, 8].

5 Вычислительный эксперимент

Задача (1)–(8), (11)–(14) решалась двумя способами: точным методом и модифицированным алгоритмом Кларка–Райта, которые были реализованы в среде Scilab-4.1.2. Поскольку на задачах с размерностями $n \geq 8$ время решения линейной целочисленной задачи оказалось весьма продолжительным (в некоторых случаях более 45 ч), сравнение алгоритмов производилось для случайно сгенерированных задач при размерностях $n = \overline{4,8}$. Для каждой размерности было решено по 20 задач каждым из алгоритмов.

Генерация тестовых примеров осуществлялась для значений грузоподъемности ТС $Q \in [1{,}5,\ 5]$ (т); элементы матрицы L протяженности дорог между пунктами генерировались из диапазона $L_{ij} \in \in [1,\ 100]$, $i, j = \overline{1,n}$ (км); элементы матрицы коэффициентов сложности дорог между пунктами генерировались целыми из диапазона $[1,\ 5]$; вектор потребностей пунктов в грузе генерировался из диапазона $q_i \in [300,\ 1500]$ (кг). Функция расхода топлива для используемого типа ТС задана полиномом первой степени от величины загрузки ТС в виде $f(q) = 15 + 3q$.

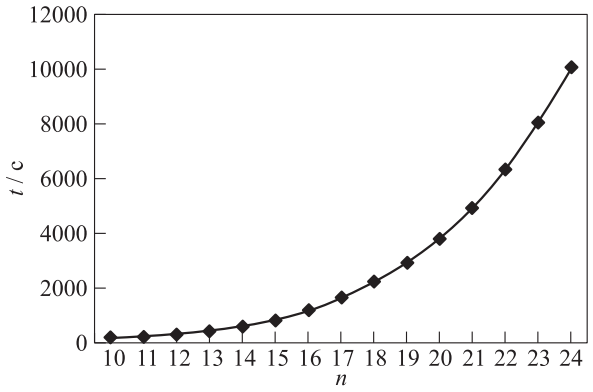
Все сгенерированные задачи были решены при помощи точного алгоритма, а также с помощью модифицированного алгоритма Кларка–Райта при значении параметра $p = 3$. Для каждого значения $n = \overline{4,8}$ в таблице приведены результаты численного эксперимента — средние значения отклонения от оптимального значения расхода топлива, среднее время работы точного алгоритма и модифицированного алгоритма Кларка–Райта. Также приведены сведения о доле задач, решение которых алгоритмом Кларка–Райта совпало с оптимальным.

Как видно из таблицы, модифицированный алгоритм Кларка–Райта на небольших размерностях показывает хорошие результаты. Вычислительное время, затрачиваемое точным алгоритмом, оказывается неприемлемо велико (более 27 ч уже при $n = = 8$), в то время как вычислительное время, затрачиваемое модифицированным алгоритмом Кларка–Райта, остается приемлемым при существенном увеличении размерности

Рисунок иллюстрирует зависимость времени работы эвристического алгоритма от размерности за-

Результаты численного эксперимента

n	Точный алгоритм	Модифицированный алгоритм Кларка–Райта		
	Среднее время работы алгоритма, с	Среднее время работы алгоритма, с	Среднее отклонение расхода топлива, %	Совпадения результатов с оптимумом, %
4	2,70	17,57	3,08	85
5	13,74	20,82	4,47	75
6	213,85	24,36	3,04	65
7	6050,91	46,40	9,15	50
8	100 084	54,60	14,26	25



Зависимость вычислительного времени работы модификации алгоритма Кларка–Райта от размерности задач

дачи. По горизонтали указаны значения n — размерности задачи, по вертикали — среднее время вычисления (в секундах). Как видно, при увеличении размерности время работы модифицированного алгоритма Кларка–Райта возрастает, однако точный алгоритм на идентичных размерностях работает несравнимо дольше.

6 Заключение

Рассматриваемая проблема является актуальной прикладной задачей. Полученные результаты позволяют сделать вывод о целесообразности использования модификации алгоритма Кларка–Райта для решения задачи транспортировки грузов с учетом загрузки ТС для снижения отклонения получаемых результатов от точно найденного оптимума.

Литература

1. *Dantzig G. B., Ramser J. H.* The truck dispatching problem // Management Sci., 1959. No. 1. P. 80–91.
2. *Бронштейн Е. М., Заико Т. А.* Детерминированные оптимизационные задачи транспортной логистики // Автоматика и телемеханика, 2010. № 10. С. 133–147.
3. *Ralphs T. K., Kopman L., Pulleyblank W. R., Trotter L. E., Jr.* On the capacitated vehicle routing problem // Math. Program. Ser. B, 2003. Vol. 94. P. 343–359.
4. *Kara I., Kara B. Y., Kadri Yetis M.* Energy minimizing vehicle routing problem // Combinatorial optimization and applications / Eds. Dress A. W. M., Xu Y., Zhu B. Lecture notes in computer science ser. — Springer, 2007. Vol. 4616. P. 62–71.
5. *Зелёв П. А., Бронштейн Е. М.* Задача транспортной логистики с учетом зависимости расходов на транспортировку от загрузки транспортного средства // Логистика и управление цепями поставок, 2010. № 4. С. 39–45.
6. *Dror M., Laporte G., Trudeau P.* Vehicle routing with split deliveries // Discrete Appl. Math., 1994. Vol. 50. P. 239–254.
7. *Clarke G., Right J. W.* Scheduling of vehicles from a central depot to a number of delivery points // Oper. Res., 1963. No. 11. P. 568–581.
8. *Бронштейн Е. М., Зелёв П. А.* Задача маршрутизации транспортного средства с учетом зависимости стоимости перевозок от загрузки // Информационные технологии, 2014. № 4. С. 33–37.

Поступила в редакцию 06.02.14

ABOUT OPTIMUM DELIVERY OF FREIGHTS BY THE VEHICLE TAKING INTO ACCOUNT DEPENDENCE OF COST OF TRANSPORTATIONS ON LOADING OF VEHICLES ON SEVERAL CYCLIC ROUTES

E. M. Bronshtein and P. A. Zelyov

Ufa State Aviation Technical University, 12 K. Marx Str., Ufa 450000, Russian Federation

Abstract: The problem of creation of a route of freights delivery from one producer (base, a warehouse) to consumers by the vehicle with the minimum costs of transportations is considered. Dependence of cost of transportation on loading of the vehicle and quality of the road is thus considered. It is supposed that the vehicle can come back to the base for additional charge. The corresponding mathematical model is constructed; for a case of linear dependence of fare from loading, the linear integer model is received. For the solution of an objective along with the exact algorithm, modification of the known heuristic algorithm of Clark and Right is suggested. Computing experiment has been made.

Keywords: heuristic algorithm; creation of a route; transportation; problem of routing

DOI: 10.14357/19922264140407

Acknowledgments

The work was financially supported by the Russian Foundation for Basic Research (project 13-01-00005).

References

1. Dantzig, G. B., and J. H. Ramser. 1959. The truck dispatching problem. *Management Sci.* 1:80–91.
2. Bronshtein, E. M., and T. A. Zaiko. 2010. Deterministic optimizational problems of transportation logistics. *Automation Remote Control* 10:2132–2144.
3. Ralphs, T. K., L. Kopman, W. R. Pulleyblank, and L. E. Trotter, Jr. 2003. On the capacitated vehicle routing problem. *Math. Program. Ser. B.* 94:343–359.
4. *Kara, I., B. Y. Kara, and M. Kadri Yetis*. 2007. Energy minimizing vehicle routing problem. *Combinatorial optimization and applications*. Eds. A. W. M. Dress, Y. Xu, and B. Zhu. Lecture notes in computer science ser. 4616:62–71.
5. Zelev, P. A., and E. M. Bronshteyn. 2010. Zadacha transportnoy logistiki s uchetom zavisimosti raskhodov na transportirovku ot zagruzki transportnogo sredstva [The problem of transportation logistics, taking into account transportation costs, depending on the load of the vehicle]. *Logistika i Upravlenie Tsepyami Postavok* [Logistics and Supply Chain Management] 4:39–45.
6. Dror, M., G. Laporte, and P. Trudeau. 1994. Vehicle routing with split deliveries. *Discrete Appl. Math.* 50:239–254.
7. Bronshteyn, E. M., and P. A. Zelev. 2014. Zadacha marshrutizatsii transportnogo sredstva s uchetom zavisimosti stoimosti perevozok ot zagruzki [Vehicle routing problem, taking into account the cost of transportation, depending on the load]. *Informatsionnye Tekhnologii* [Information Technologies] 4:33–37.
8. Clarke, G., and J. W. Right. 1963. Scheduling of vehicles from a central depot to a number of delivery points. *Oper. Res.* 11:568–581.

Received February 6, 2014

Contributors

Bronshtein Efim M. (b. 1946) — Doctor of Science in physics and mathematics; professor, Ufa State Aviation Technical University, 12 K. Marx Street, Ufa 450000, Russian Federation; bro-efim@yandex.ru

Zelyov Pavel A. (b. 1988) — PhD student, Ufa State Aviation Technical University, 12 K. Marx Street, Ufa 450000, Russian Federation; pz1988@yandex.ru

МЕТОД ПОВЫШЕНИЯ ЭФФЕКТИВНОСТИ РЕШЕНИЯ ЗАДАЧ ВЕРОЯТНОСТНОЙ ВЕРИФИКАЦИИ ВЫЧИСЛИТЕЛЬНЫХ И ТЕЛЕКОММУНИКАЦИОННЫХ СИСТЕМ*

А. М. Миронов[1], С. Л. Френкель[2]

Аннотация: Рассматривается проблема снижения трудоемкости вероятностной верификации при проектировании вычислительных систем. Поставленная цель достигается редукцией вероятностных систем переходов (ВСП), моделирующих проектируемые системы. Верификация ВСП заключается в вычислении истинностных значений формул вероятностной темпоральной логики (PCTL, Probabilistic Computational Tree Logic) в начальных состояниях ВСП. Редукция ВСП выполняется по алгоритму удаления эквивалентных состояний, в результате работы которого получается такая ВСП, у которой все свойства, выражаемые формулами логики PCTL, совпадают со свойствами исходной ВСП.

Ключевые слова: верификация; вероятностные системы переходов; вероятностная темпоральная логика; редукция вероятностных моделей

DOI: 10.14357/19922264140408

1 Введение

1.1 Постановка задачи

Необходимость в вероятностной верификации проектов цифровых систем возникает либо при проектировании систем со стохастическим поведением, например многоканальных телекоммуникационных систем, либо в случаях, когда у разработчика есть основание полагать, что проектируемая система в рабочем режиме может быть подвержена различным не специфицированным при проектировании ошибкам и случайным сбоям как внутренней природы, так и инициированных внешними воздействиями. Поскольку точно локализация и функциональные последствия наличия таких ошибок априори не известны, их можно попытаться характеризовать вероятностью проявления в результатах работы и, соответственно, говорить о вероятностной верификации. Наиболее распространенным подходом к формальной вероятностной верификации является Probabilistic Model Checking [1, гл. 11], дополняющий проверку соответствия формальной спецификации проектируемой системы ее свойствам (properties) вычислением вероятностей выполнения этих свойств. В данном случае проектируемые системы описываются моделью ВСП, которая используется в алгоритмах формальной верификации, основанных на проверке моделей (Model Checking). Одна из главных проблем использования Probabilistic Model Checking, как и прочих формальных методов спецификации, состоит в их вычислительной сложности, и поэтому снижение размера соответствующих моделей, в частности матриц тех или иных переходов, является важнейшим фактором реализуемости соответствующих методов.

В настоящей работе рассматривается задача редукции ВСП, целью которой является понижение сложности верификации свойств ВСП, выражаемых формулами вероятностной темпоральной логики PCTL. Вероятностные системы переходов представляют собой один из наиболее широко используемых классов моделей дискретных динамических систем. Понятие ВСП является обобщением понятия цепи Маркова [2], которое нашло широкое применение в естественных и гуманитарных науках. Понятие ВСП можно рассматривать также как частный случай понятия вероятностного автомата [3]. Главной отличительной особенностью понятия ВСП от понятий цепи Маркова и вероятностного автомата является наличие выразительного логического формализма, позволяющего эффективно описывать различные свойства поведения ВСП. В качестве такого формализма выступает вероятностная темпоральная логика PCTL [4,

*Работа выполнена при частичной поддержке РФФИ (проект 12-07-00109).

[1]Институт проблем информатики Российской академии наук, amironov66@gmail.com

[2]Институт проблем информатики Российской академии наук; Московский государственный технический университет радиотехники, электроники и автоматики (МГТУ МИРЭА), fsergei@mail.ru

5], которая представляет собой вероятностный аналог темпоральной логики ветвящегося времени CTL [6], использующейся для спецификации свойств параллельных и распределенных программ, и является эффективным инструментом для описания различных свойств дискретных вероятностных динамических систем.

Формулы логики PCTL могут отражать различные вероятностные аспекты поведения анализируемых систем, к числу которых относятся, например, частота выполнения тех или иных действий или переходов в анализируемых системах, вероятность отказа компонентов анализируемых систем, вероятностный характер взаимодействия анализируемой системы с ее окружением, например: частота поступления входных запросов или сообщений, частота получения искаженных сообщений (для протоколов передачи сообщений в компьютерных сетях) и т. п.

В данной работе уточняются основные формулировки и демонстрируется на новом примере решение задачи, сформулированной в [7], а именно: преобразование ВСП проектируемой системы в эквивалентную ВСП с меньшим числом состояний. Под эквивалентностью понимается, что результаты верификации исходной и редуцированной модели будут одинаковы.

Некоторые подходы к редукции ВСП изучались в различных работах по вероятностной верификации, однако в этих исследованиях были рассмотрены лишь частные методы редукции ВСП, такие как редукция частичных порядков [8, 9] и редукция, основанная на понятии симметрии множества состояний ВСП [10, 11]. Данные методы можно эффективно использовать лишь для ВСП достаточно специального вида: как правило, это вероятностные модели параллельных и распределенных программ.

1.2 Современное состояние проблемы вероятностной верификации

Первые алгоритмы вероятностной верификации были предложены в 1980-е гг. в работах [12–14]. Данные алгоритмы были предназначены для верификации качественных вероятностных свойств (т. е. таких, которые выполняются с вероятностью 1 или 0). Затем эти алгоритмы были обобщены на случай верификации количественных вероятностных свойств (в спецификации таких свойств могло присутствовать любое значение вероятности). Эти алгоритмы были изложены в работах [4, 15, 16]. Программные реализации этих алгоритмов были представлены в работах [17, 18].

Первые промышленные системы вероятностной верификации были разработаны в 2000-х гг. [19, 20]. Эти системы успешно применяются во многих областях, таких как анализ распределенных алгоритмов, телекоммуникационные протоколы, компьютерная безопасность, криптографические протоколы, моделирование биологических процессов. С использованием этих систем верификации были обнаружены уязвимости и аномальные поведения анализируемых систем (подробнее см. в [21]). При помощи систем вероятностной верификации могут быть вычислены такие характеристики программных систем, как, например, вероятность вторжения злоумышленника в компьютерную сеть, математическое ожидание времени отклика веб-сервиса и другие количественные и качественные характеристики.

Наиболее популярной практической системой вероятностной верификации в настоящее время является система PRISM [22], разработанная на факультете компьютерных наук Оксфордского университета (Великобритания) в группе Quantitative Analysis and Verification под руководством Марты Квятковской. Информация о деятельности этой группы представлена на веб-сайте http://qav.comlab.ox.ac.uk/.

2 Вероятностные системы переходов

2.1 Понятие вероятностной системы переходов

Предположим, что задано конечное множество AP, элементы которого называются **атомарными утверждениями**. Ниже запись 2^{AP} обозначает множество всех подмножеств AP.

Вероятностная система переходов (называемая также в англоязычной литературе **Discrete Time Markov Chain**) — это четверка D вида

$$D = (S, s^0, P, L)\,,$$

компоненты которой имеют следующий смысл:

(1) S — множество, элементы которого называются **состояниями** ВСП D;

(2) $s^0 \in S$ — выделенное состояние, называемое **начальным состоянием** ВСП D;

(3) P — функция вида $P : S \times S \to [0, 1]$, называемая **функцией перехода** ВСП D и удовлетворяющая условию: $\forall\, s \in S \quad \sum\limits_{s' \in S} P(s, s') = 1$. Для каждой пары $(s_1, s_2) \in S \times S$ число $P(s_1, s_2)$

понимается как вероятность того, что если в текущий момент времени D находится в состоянии s_1, то через один такт времени D будет находиться в состоянии s_2. Если $P(s_1, s_2) > 0$, то будем называть тройку $(s_1, s_2, P(s_1, s_2))$ **переходом** из s_1 в s_2 с вероятностью $P(s_1, s_2)$. Ниже запись $s_1 \xrightarrow{a} s_2$ является другим обозначением перехода (s_1, s_2, a);

(4) L — функция вида $L : S \to 2^{\mathrm{AP}}$, называемая **оценкой**, которая имеет следующий смысл: для каждого состояния $s \in S$ и каждого атомарного утверждения $p \in \mathrm{AP}$ утверждение p считается **истинным** в s, если $p \in L(s)$, и **ложным** в s, если $p \notin L(s)$.

Вероятностную систему переходов удобно рассматривать как помеченный граф, вершинами которого являются состояния, помеченные элементами множества 2^{AP}: каждая вершина $s \in S$ имеет метку $L(s)$ и для каждой пары $(s_1, s_2) \in S \times S$ такой, что $P(s_1, s_2) > 0$, граф содержит ребро из s_1 в s_2 с меткой $P(s_1, s_2)$.

2.2 Случайные функции

Пусть X и Y — два конечных множества.

Случайной функцией (СФ) из X в Y называется произвольная функция f вида

$$f : X \times Y \to [0, 1] \tag{1}$$

такая, что $\forall\, x \in X \quad \sum\limits_{y \in Y} f(x, y) = 1$.

Для любых $x \in X$ и $y \in Y$ значение $f(x, y)$ можно интерпретировать как вероятность того, что СФ f отображает x в y.

Случайная функция (1) называется **детерминированной**, если для каждого $x \in X$ существует единственный $y \in Y$, такой что $f(x, y) = 1$. Если f — детерминированная СФ вида (1) и x, y — такие элементы X и Y соответственно, что $f(x, y) = 1$, то будем говорить, что f **отображает** x **в** y.

Если f — СФ из X в Y, то будем обозначать этот факт записью $f : X \underset{r}{\to} Y$. Будем называть X **областью определения** СФ f, а Y — **областью значений** СФ f.

Для каждого конечного множества X запись id_X обозначает детерминированную СФ $X \to X$, которая отображает каждый $x \in X$ в x.

2.3 Матрицы, соответствующие случайным функциям

Если СФ f имеет вид $f : X \underset{r}{\to} Y$ и на множествах X и Y заданы упорядочения их элементов, которые имеют вид $(x_1, \ldots, x_m)$ и $(y_1, \ldots, y_n)$ соответственно, то СФ f можно представить в виде матрицы (обозначаемой тем же символом f)

$$f = \begin{pmatrix} f(x_1, y_1) & \cdots & f(x_1, y_n) \\ \vdots & \ldots & \vdots \\ f(x_m, y_1) & \cdots & f(x_m, y_n) \end{pmatrix}. \tag{2}$$

Ниже будем отождествлять СФ f с матрицей (2).

Будем предполагать, что для каждого конечного множества X, являющегося областью определения или областью значений какой-либо из рассматриваемых СФ, на X задано фиксированное упорядочение его элементов. Таким образом, для каждой рассматриваемой СФ соответствующая ей матрица определена однозначно.

Для каждой СФ $f : X \underset{r}{\to} Y$ и произвольных $x \in X$, $y \in Y$ будем называть

- строку $(f(x, y_1), \ldots, f(x, y_n))$ матрицы f — **строкой** x;
- столбец $\begin{pmatrix} f(x_1, y) \\ \vdots \\ f(x_m, y) \end{pmatrix}$ матрицы f — **столбцом** y.

Если f и g — СФ вида $f : X \underset{r}{\to} Y$, $g : Y \underset{r}{\to} Z$, то их **композицией** называется СФ $f \cdot g : X \underset{r}{\to} Z$, определяемая следующим образом:

$$\forall\, x \in X \quad (f \cdot g)(x) \stackrel{\mathrm{def}}{=} \sum_{y \in Y} f(x, y) \cdot g(y, z)\,. \tag{3}$$

По определению произведения матриц из (3) следует, что матрица $f \cdot g$ является произведением матриц f и g.

2.4 Случайные функции, соответствующие вероятностным системам переходов

Пусть задана ВСП $D = (S, s^0, P, L)$ и список элементов множества S имеет вид $(s_1, \ldots, s_n)$.

Будем использовать следующие обозначения.

1. Символ $\mathbf{1}$ означает множество, состоящее из одного элемента, который будем обозначать символом e.
2. Для каждого состояния $s \in S$ запись I_s обозначает детерминированную СФ вида $I_s : \mathbf{1} \underset{r}{\to} S$, отображающую элемент $e \in \mathbf{1}$ в состояние s ВСП D.
3. Для каждого $n \geq 0$ обозначим записью P^n СФ вида $P^n : S \underset{r}{\to} S$, определяемую индуктивно: $P^0 \stackrel{\mathrm{def}}{=} \mathrm{id}_S$ и $\forall\, n \geq 0 \;\; P^{n+1} \stackrel{\mathrm{def}}{=} P^n \cdot P$. Нетрудно видеть, что матрицы, соответствующие СФ P^i,

имеют следующий вид: P^0 — единичная матрица и $\forall\, n > 0$ матрица P^n является n-й степенью матрицы P.

Для любых $n \geq 0$, $s_1, s_2 \in S$ число $P^n(s_1, s_2)$ можно понимать как вероятность того, что если в текущий момент времени ВСП D находится в состоянии s_1, то через n тактов времени D будет находиться в состоянии s_2.

3 Логика PCTL

Логика PCTL — это темпоральная логика, предназначенная для формального описания свойств ВСП. Логика PCTL была введена Х. Ханссоном и Б. Джонссоном в работе [4].

3.1 Формулы логики PCTL

В определении понятия формулы логики PCTL будем использовать множество AP атомарных утверждений, введенное в разд. 2.

Формулы логики PCTL делятся на два класса: StateFm — **формулы состояний** — и PathFm — **формулы путей**. Формулы из классов StateFm и PathFm будем обозначать символами φ и α соответственно (возможно, с индексами), а формулу произвольного вида — символом f (возможно, с индексом).

Классы StateFm и PathFm определяются следующим образом.

StateFm:

1. Каждое атомарное утверждение p из AP является формулой из StateFm.
2. Символы $\top$ и $\bot$ является формулами из StateFm. Данные символы обозначают тождественно истинное и тождественно ложное утверждение соответственно.
3. Если φ_1 и φ_2 — формулы из StateFm, то следующие знакосочетания являются формулами из StateFm: $\neg\varphi_1$; $\varphi_1 \wedge \varphi_2$; $\varphi_1 \vee \varphi_2$; $\varphi_1 \rightarrow \varphi_2$; $\varphi_1 \leftrightarrow \varphi_2$.
4. Если
   - $\triangle$ — функциональный символ, которому соответствует функция (обозначаемая тем же символом) вида
   $$\triangle : [0,1] \times [0,1] \rightarrow \{0,1\};$$
   - a — число из $[0,1]$;
   - α — формула из PathFm,

   то знакосочетание $\mathcal{P}_{\triangle a}\alpha$ является формулой из StateFm.

PathFm:

1. Если f — формула логики PCTL, то знакосочетание $\mathbf{X}f$ является формулой из PathFm.
2. Если φ_1 и φ_2 — формулы из StateFm, то следующие знакосочетания являются формулами из PathFm:
   (а) $\varphi_1 \mathbf{U}^{\leq n} \varphi_2$, где n — натуральное число;
   (б) $\varphi_1 \mathbf{U} \varphi_2$.
3. Если α — формула из PathFm, то знакосочетание $\neg\alpha$ является формулой из PathFm.

В записи формул из PathFm могут использоваться символы $\mathbf{F}$ и $\mathbf{G}$, которые являются сокращением знакосочетаний $\top\mathbf{U}$ и $\neg\mathbf{F}\neg$ соответственно (т. е., например, знакосочетания $\mathbf{F}\alpha$ и $\mathbf{G}^{\leq n}\alpha$ обозначают формулы $\top\mathbf{U}\alpha$ и $\neg\mathbf{F}^{\leq n}\neg\alpha$ соответственно).

3.2 Значения формул логики PCTL в состояниях вероятностных систем переходов

Пусть $D = (S, s^0, P, L)$ — некоторая ВСП.

Для каждого состояния $s \in S$ и каждой формулы f логики PCTL определено **значение** формулы f в состоянии s, которое обозначается записью $s(f)$, и

(1) если $f \in \mathrm{StateFm}$, то $s(f) \in \{0,1\}$ и
   - в случае $s(f) = 1$ формула f считается истинной в s;
   - в случае $s(f) = 0$ формула f считается ложной в s;

(2) если $f \in \mathrm{PathFm}$, то значение $s(f)$ является числом из $[0,1]$ и интерпретируется как вероятность того, что формула f истинна в состоянии s.

Для каждой формулы f логики PCTL будем обозначать записью $S(f)$ вектор-столбец $\begin{pmatrix} s_1(f) \\ \vdots \\ s_n(f) \end{pmatrix}$.

Значения формул логики PCTL в состояниях ВСП определяются индукцией по структуре формул в соответствии с излагаемыми ниже правилами. В одних из этих правил определяется значение $s(f)$, в других — определяется вектор-столбец $S(f)$ целиком. В этих определениях будем использовать следующие обозначения:

- для любых векторов $U = \begin{pmatrix} u_1 \\ \vdots \\ u_n \end{pmatrix}$, $V = \begin{pmatrix} v_1 \\ \vdots \\ v_n \end{pmatrix}$ из $[0,1]^n$ записи $\max(U, V)$ и $U \circ V$ обозначают

векторы $\begin{pmatrix}\max(u_1,v_1)\\ \vdots \\ \max(u_n,v_n)\end{pmatrix}$ и $\begin{pmatrix}u_1\cdot v_1\\ \vdots \\ u_n\cdot v_n\end{pmatrix}$ соответственно.

– если A и B — матрицы порядков $n\times n$ и $n\times$ $\times 1$ соответственно с компонентами из $[0,1]$, то запись $[A^*\cdot B]$ обозначает матрицу, получаемую заменой всех ненулевых компонентов A и B на 1 и вычислением $(\sum\limits_{i\geq 0} A^i)\cdot B$, где сложение понимается как дизъюнкция (т. е. $\sum\limits_{i\geq 0} A^i$ является конечной).

Правила определения значений формул логики PCTL в состояниях ВСП имеют следующий вид:

– Для каждого $p\in$ AP
$$s(p)\stackrel{\text{def}}{=}\begin{cases}1, & \text{если } p\in L(s)\,;\\ 0 & \text{иначе}\,;\end{cases}$$
– $s(\top)\stackrel{\text{def}}{=}1,\ \ s(\bot)\stackrel{\text{def}}{=}0$;
– $s(\neg f)\stackrel{\text{def}}{=}1-s(f)$, $s(\varphi_1\wedge\varphi_2)\stackrel{\text{def}}{=}s(\varphi_1)\cdot s(\varphi_2)$ и т. д. (т. е. значения формул коммутируют с булевыми операциями);
– $s(\mathcal{P}_{\triangle a}\alpha)\stackrel{\text{def}}{=}\triangle(s(\alpha),a)$;
– $S(\mathbf{X}f)\stackrel{\text{def}}{=}P\cdot S(f)$;
– пусть $\alpha_n=\varphi_1\mathbf{U}^{\leq n}\varphi_2$ (где $n\geq 0$). Тогда
$$S(\alpha_0)\stackrel{\text{def}}{=}S(\varphi_2)\,;$$
$$\forall\, n>0\ S(\alpha_n)\stackrel{\text{def}}{=}\max\left(S(\varphi_2),\ S(\varphi_1)\circ S(\mathbf{X}\alpha_{n-1})\right)\,.$$
– пусть $\alpha=\varphi_1\mathbf{U}\varphi_2$. Тогда $S(\alpha)$ определяется системой линейных уравнений
$$S(\alpha)=$$
$$=\max\left(S(\varphi_2),[P^*\cdot S(\varphi_2)]\circ S(\varphi_1)\circ(P\cdot S(\alpha))\right).$$

4 Метод редукции вероятностных систем переходов

4.1 Задача редукции вероятностных систем переходов

Если анализируемая ВСП имеет большой размер, то анализ ее свойств, выражаемых формулами логики PCTL (т. е. вычисление значений формул логики PCTL в состояниях этой ВСП), может быть связан с трудновыполнимыми требованиями к вычислительным ресурсам, с использованием которых производится этот анализ. В связи с этим представляет большую актуальность проблема редукции ВСП, т. е. удаления части состояний и переходов анализируемой ВСП с таким расчетом, чтобы получившая ВСП была эквивалентна исходной в следующем смысле: для каждой формулы состояний f логики PCTL формула f истинна в начальном состоянии исходной ВСП тогда и только тогда, когда она истинная в начальном состоянии редуцированной ВСП.

Основная идея предлагаемого в настоящей работе метода редукции ВСП основана на понятии эквивалентности состояний ВСП: будем называть состояния эквивалентными, если значения всех формул логики PCTL в этих состояниях совпадают. Алгоритм редукции ВСП представляет собой вычисление классов эквивалентности состояний анализируемой ВСП и удаление эквивалентных состояний.

4.2 Эквивалентность вероятностных систем переходов

Пусть заданы две ВСП:
$$D_i=(S_i,s_i^0,P_i,L_i)\ \ (i=1,2)\,. \tag{4}$$

Будем называть состояния $s_1\in S_1$ и $s_2\in S_2$ **эквивалентными**, если для каждой формулы f логики PCTL верно равенство $s_1(f)=s_2(f)$.

Если состояния s_1 и s_2 эквивалентны, то будем обозначать это записью $s_1\sim s_2$.

Будем называть ВСП D_1 и D_2 вида (4) **эквивалентными**, если $s_1^0\sim s_2^0$. Если ВСП D_1 и D_2 эквивалентны, то будем обозначать этот факт записью $D_1\sim D_2$.

Если ВСП D_1 и D_2 совпадают и S — множество их состояний, то бинарное отношение на S, состоящее из всех пар (s_1,s_2) таких, что $s_1\sim s_2$, является отношением эквивалентности. Будем обозначать это отношение символом $\sim$.

Отношение $\sim$ может быть найдено при помощи алгоритма, излагаемого в параграфе 5.3.

4.3 Редукция вероятностных систем переходов

4.3.1 Задача редукции вероятностных систем переходов

Пусть задана ВСП $D=(S,s^0,P,L)$.

Задача редукции ВСП D заключается в построении ВСП D', которая эквивалентна D и число состояний которой меньше, чем число состояний ВСП D.

Излагаемый в настоящем пункте алгоритм редукции ВСП является вероятностным обобщением алгоритма редукции детерминированных автоматов. Идея данного алгоритма основана на отождествлении неразличимых состояний ВСП:

(1) алгоритм вычисляет классы $S_1,\ldots,S_k$ разбиения множества S, соответствующего эквивалентности $\sim$;

(2) ВСП D преобразуется путем удаления состояний в классах $S_1, \ldots, S_k$ (и соответствующего переопределения функции перехода) до тех пор, пока не останется по одному состоянию в каждом из этих классов.

В результате этих удалений получается искомая ВСП D'.

4.3.2 *Построение разбиения множества состояний редуцируемой вероятностной системы переходов*

Разбиение множества S состояний ВСП $D = (S, s^0, P, L)$, соответствующее отношению эквивалентности $\sim$, вычисляется следующим образом:

(1) вычисляется разбиение Σ^0, соответствующее отношению эквивалентности $\{(s_1, s_2) \in S \times S \mid L(s_1) = L(s_2)\}$;

(2) затем работает цикл, состоящий из следующих шагов.

Пусть для некоторого $i \geq 0$ определены

- отношение эквивалентности ρ^i;
- соответствующее ему разбиение Σ^i, которое состоит из классов $S_1^i, \ldots, S_k^i$.

Обозначим записями $\Sigma_1^i, \ldots, \Sigma_k^i$ строки матрицы π^i, соответствующей детерминированной СФ $\pi^i : S \to \Sigma^i$, и $\varphi_1^{\Sigma^i}, \ldots, \varphi_k^{\Sigma^i}$ — список формул таких, что $\forall\, j = 1, \ldots, k \quad S(\varphi_j^{\Sigma^i}) = \Sigma_j^i$.

Определим отношение эквивалентности ρ^{i+1} на S:

$$\rho^{i+1} \stackrel{\text{def}}{=} \rho^i \cap \Big\{(s_1, s_2) \in S^2 \mid \forall\, j = 1, \ldots, k \quad s_1(\mathbf{X}\varphi_j^{\Sigma^i}) = s_2(\mathbf{X}\varphi_j^{\Sigma^i})\Big\}\,.$$

Разбиение Σ^{i+1}, соответствующее отношению ρ^{i+1}, можно построить следующим образом:

- вычисляются вектор-столбцы

$$S(\mathbf{X}\varphi_j^{\Sigma^i}) = P \cdot \Sigma_j^i \qquad (5)$$

(каждый из которых, как нетрудно видеть, является суммой некоторых столбцов матрицы P: для каждого $j = 1, \ldots, k$ вектор-столбец (5) является суммой таких столбцов s матрицы P, для которых $s \in S_j$);

- классы разбиения Σ^{i+1} получаются путем измельчения классов разбиения Σ^i: в один и тот же класс разбиения Σ^{i+1} попадают такие состояния, для которых соответствующие им компоненты векторов (5) совпадают для каждого $j = 1, \ldots, k$.

Возможны два случая:

(а) $\Sigma^{i+1} = \Sigma^i$. В этом случае искомое разбиение $\sim$ найдено: оно совпадает с Σ^i;

(б) $\Sigma^i \neq \Sigma^{i+1}$. В этом случае увеличиваем i на 1 и возвращаемся в начало цикла (т. е. выполняем шаг 2 с увеличенным значением i).

Нетрудно видеть, что таких возвращений может быть не больше числа элементов множества S (так как разбиение Σ^{i+1} является измельчением разбиения Σ^i).

4.3.3 *Удаление эквивалентных состояний из вероятностных систем переходов*

Пусть ВСП $D = (S, s^0, P, L)$ содержит пару эквивалентных состояний s_1, s_2, где $s_1 \neq s^0$. Определим ВСП

$$D_1 \stackrel{\text{def}}{=} (S_1, s^0, P_1, L_1)\,, \qquad (6)$$

где $S_1 \stackrel{\text{def}}{=} S \setminus \{s_1\}$, $\forall\, s, s' \in S_1$;

$$P_1(s, s') \stackrel{\text{def}}{=} \begin{cases} P(s, s') + P(s, s_1), & \text{если } s' = s_2\,; \\ P(s, s'), & \text{если } s' \neq s_2, \end{cases}$$

$\forall\, s \in S_1$ $L_1(s) \stackrel{\text{def}}{=} L(s)$.

Таким образом, матрица P_1 получается из матрицы P прибавлением к столбцу s_2 столбца s_1 и удалением строки s_1 и столбца s_1, а матрица L_1 получается из матрицы L удалением строки s_1.

Будем говорить что ВСП (6) получается из ВСП D путем **удаления состояния s_1, эквивалентного состоянию s_2**. По определению ВСП (6) каждое ее состояние является также и состоянием ВСП D.

Для каждого $s \in S_1$ и каждой формулы f логики PCTL будем обозначать записями $s_D(f)$ и $s_{D_1}(f)$ значения формулы f в состоянии s в ВСП D и D_1 соответственно и записями $S_D(f)$ и $S_{D_1}(f)$ — вектор-столбцы значений формулы f в состояниях ВСП D и D_1 соответственно.

Теорема 1. *Пусть ВСП* (6) *получается из ВСП D путем удаления состояния s_1, эквивалентного состоянию s_2. Тогда* $\forall\, s \in S_1 \;\; s_{D_1}(f) = s_D(f)$.

4.3.4 *Описание алгоритма редукции вероятностной системы переходов*

Теорема 1 является обоснованием излагаемого ниже алгоритма редукции ВСП $D = (S, s^0, P, L)$. Этот алгоритм имеет следующий вид.

1. Вычисляется разбиение множества состояний ВСП D, соответствующее отношению эквивалентности $R \stackrel{\text{def}}{=} \sim$ (для этого выполняются действия, изложенные в п. 5.3.2).

2. Искомая ВСП D' строится путем удаления состояний из ВСП D и переопределения функции перехода и отношения R следующим образом:

   (а) если отношение R содержит пару (s_1, s_2), такую что $s_1 \neq s_2$ и $s_2 \neq s^0$, то выберем произвольную такую пару (s_1, s_2) и преобразуем компоненты ВСП D описываемым ниже образом. Будем излагать данное преобразование в терминах графа, соответствующего ВСП D (данный граф будем обозначать тем же символом D):

   (i) если граф D содержит ребро с началом в некоторой вершине s и с концом s_2, то данное ребро удаляется, а к метке ребра с началом в s и с концом в s_1 прибавляется число, равное метке удаленного ребра. Данная операция выполняется до тех пор, пока имеются ребра с концом в s_2;

   (ii) вершина s_2 удаляется и, кроме того, удаляются все ребра, выходящие из этой вершины;

   (iii) из R удаляются все пары, содержащие s_2, и осуществляется переход к шагу 2а;

   (б) если каждая пара, входящая в R, имеет вид (s, s), то работа завершается.

5 Пример редукции вероятностной системы переходов

В этом разделе рассматривается пример редукции вероятностной модели протокола передачи сообщений через ненадежный канал связи, в котором пересылаемые сообщения могут пропадать или искажаться. Протокол представляет собой систему, состоящую из двух агентов — отправителя и получателя, а также канала, в который помещаются сообщения, пересылаемые от одного агента другому. Предполагается, что факт искажения получаемых сообщений может быть установлен, и если исходное сообщение не может быть восстановлено из искаженного, то отправитель получает сигнал о необходимости повторной посылки этого сообщения. Как только сообщение успешно доходит до получателя, отправителю посылается сигнал подтверждения успешного получения и он переходит к отправке следующего сообщения. Предполагается, что сигналы и подтверждения отправителю не пропадают и не искажаются в канале.

Графовая модель этого протокола имеет вид, представленный на рис. 1.

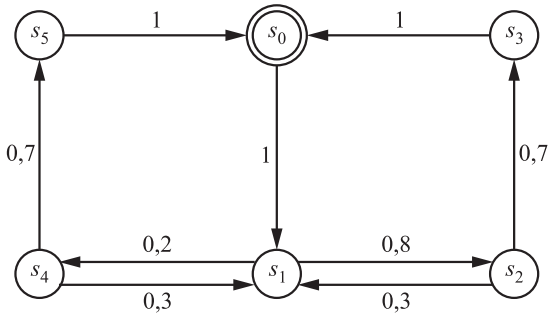


Рис. 1 Графовая модель протокола передачи сообщений через ненадежный канал связи

Переходы в этом графе имеют следующий смысл.

1. Переход $s_0 \xrightarrow{1} s_1$ заключается в получении отправителем от внешнего источника сообщения, которое должно быть передано через канал получателю.
2. Переход $s_1 \xrightarrow{0,8} s_2$ заключается в помещении сообщения в канал отправителем, причем сообщение в канале не искажается.
3. Переход $s_1 \xrightarrow{0,2} s_4$ заключается в помещении сообщения в канал отправителем, причем сообщение в канале искажается.
4. Переход $s_2 \xrightarrow{0,3} s_1$ заключается в потере неискаженного сообщения в канале и посылке отправителю сигнала о необходимости повторной передачи.
5. Переход $s_4 \xrightarrow{0,3} s_1$ заключается в посылке отправителю сигнала о том, что исходное сообщение не может быть восстановлено из искаженного сообщения и должно быть передано повторно.
6. Переход $s_2 \xrightarrow{0,7} s_3$ заключается в передаче неискаженного сообщения из канала получателю.
7. Переход $s_4 \xrightarrow{0,7} s_5$ заключается в восстановлении исходного сообщения из искаженного и передаче восстановленного сообщения получателю.
8. Переходы $s_3 \xrightarrow{1} s_0$ и $s_5 \xrightarrow{1} s_0$ заключаются в получении сообщения получателем и посылке им отправителю уведомления о том, что получение сообщения было выполнено успешно.

Одно из свойств протокола, представленного моделью рис. 1, заключается в том, что каждое сообщение, полученное отправителем от внешнего источника, будет с вероятностью $\geq 0{,}9$ доставлено получателю не более чем через 5 единиц времени. Для формального представления этого свойства будем полагать, что множество АР атомарных

утверждений состоит из одной переменной p и эта переменная принимает в состоянии s_0 (см. рис. 1) значение 1, а в остальных состояниях — значение 0. Таким образом, множество 2^{AP} состоит из двух элементов: $\emptyset$ и $\{p\}$. Будем обозначать эти элементы символами 0 и 1 соответственно.

Формула логики PCTL, соответствующая указанному выше свойству, имеет следующий вид:

$$\mathbf{G}((\neg p) \to \mathcal{P}_{\geq 0.9}(\mathbf{F}^{\leq 5} p))\,, \tag{7}$$

где символ $\geq$ обозначает функцию вида

$$\geq\colon [0,1] \times [0,1] \to \{0,1\}\,,$$

которая сопоставляет паре $(a,b) \in [0,1] \times [0,1]$ элемент 1, если $a \geq b$, и 0 иначе.

Анализируемая ВСП получается из графа, представленного на рис. 1 приписыванием к каждой его вершине s метки $L(s)$, которая равна 1, если $s = s_0$, и 0 иначе. Для вычисления значения формулы (7) в состояниях этой ВСП можно использовать описанный выше метод редукции.

Матрицу P, соответствующую данной ВСП, представим в виде следующей таблицы:

	s_0	s_1	s_2	s_3	s_4	s_5
s_0	0	1	0	0	0	0
s_1	0	0	0,8	0	0,2	0
s_2	0	0,3	0	0,7	0	0
s_3	1	0	0	0	0	0
s_4	0	0,3	0	0,7	0	0
s_5	1	0	0	0	0	0

Вычисление эквивалентности $\sim$ для анализируемой ВСП происходит следующим образом.

1. Вычисляется отношение эквивалентности ρ^0, которое состоит из всех пар $(s_1, s_2) \in S \times S$, удовлетворяющих равенству $L(s_1) = L(s_2)$.

   По предположению значение p в s_0 равно 1 и в каждом $s \in S$ (где S — множество состояний анализируемой ВСП), таком что $s \neq s_0$, значение p равно 0, т. е. $L(s_0) = 1$ и $\forall\, s \in S \setminus \{s_0\}$ $L(s) = 0$. Следовательно, Σ^0 состоит из двух классов:

   $$\{s_0\}, \quad \{s_1, s_2, s_3, s_4, s_5\}\,. \tag{8}$$

2. Матрица π^0, соответствующая детерминированной СФ $\pi^0 : S \to \Sigma^0$, имеет вид:

s_0	1	0
s_1	0	1
s_2	0	1
s_3	0	1
s_4	0	1
s_5	0	1

   Затем вычисляется матрица $P \cdot \pi^0$. Данная матрица будет иметь следующий вид:

s_0	0	1
s_1	0	1
s_2	0	1
s_3	1	0
s_4	0	1
s_5	1	0

(9)

   По матрице (9) нетрудно вычислить отношение ρ^1 и соответствующее ему разбиение Σ^1. Из определения отношения ρ^1 непосредственно следует, что состояния s и s' находятся в одном и том же классе разбиения Σ^1 тогда и только тогда, когда они оба находятся в одном и том же классе из списка (8) и, кроме того, строки матрицы (9), соответствующие состояниям s и s', совпадают.

   Разбиение Σ^1 будет состоять из трех классов (измельчится второй класс в (8), а первый класс останется тем же), эти классы имеют следующий вид:

   $$\{s_0\}, \quad \{s_1, s_2, s_4\}, \quad \{s_3, s_5\}. \tag{10}$$

3. Затем вычисляется матрица π^1, соответствующая детерминированной СФ $\pi^1 : S \underset{r}{\to} \Sigma^1$. Она выглядит следующим образом:

s_0	1	0	0
s_1	0	1	0
s_2	0	1	0
s_3	0	0	1
s_4	0	1	0
s_5	0	0	1

   Произведение $P \cdot \pi_1$ выглядит так:

s_0	0	1	0
s_1	0	1	0
s_2	0	0,3	0,7
s_3	1	0	0
s_4	0	0,3	0,7
s_5	1	0	0

   После этого, действуя так же, как и в предыдущем пункте, вычисляем классы разбиения Σ^2, соответствующего эквивалентности ρ^2. Таких классов будет четыре (измельчится второй класс в (10), а первый и третий классы останутся теми же), эти классы имеют следующий вид:

   $$\{s_0\}, \quad \{s_1\}, \quad \{s_2, s_4\}, \quad \{s_3, s_5\}. \tag{11}$$

4. Затем вычисляется матрица π^2, соответствующая детерминированной СФ $\pi^2 : S \underset{r}{\to} \Sigma^2$:

s_0	1	0	0	0
s_1	0	1	0	0
s_2	0	0	1	0
s_3	0	0	0	1
s_4	0	0	1	0
s_5	0	0	0	1

Произведение $P \cdot \pi_2$ имеет следующий вид:

s_0	0	1	0	0
s_1	0	0	1	0
s_2	0	0,3	0	0,7
s_3	1	0	0	0
s_4	0	0,3	0	0,7
s_5	1	0	0	0

Далее, действуя так же, как и в предыдущем пункте, вычисляем классы разбиения Σ^3, соответствующего эквивалентности ρ^3. Нетрудно проверить, что классы разбиения Σ^3 будут иметь точно такой же вид, что и классы эквивалентности разбиения Σ^2. Это означает, что искомое разбиение множества S на классы эквивалентных состояний построено, оно имеет вид (11).

Теперь можно приступить к удалению избыточных состояний (так, чтобы среди оставшихся состояний было ровно по одному состоянию из каждого класса эквивалентности (11)). Нетрудно видеть, что можно удалить состояния s_4 и s_5. После удаления данных состояний граф, представленный на рис. 1 примет вид, представленный на рис. 2.

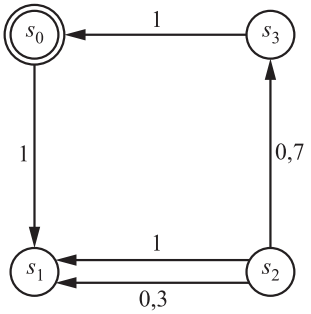


Рис. 2 Модифицированный граф рис. 1

Таким образом, задача вычисления значений формулы (7) в состояниях ВСП (см. рис. 1) сводится к задаче вычисления значений формулы (7) в состояниях ВСП (см. рис. 2), что требует выполнения меньшего числа операций, чем задача вычисления значений формулы (7) в состояниях исходной ВСП.

6 Заключение

В настоящей работе изложен алгоритм редукции ВСП, идея которого заключается в удалении избыточных состояний. Результатом применения этого алгоритма является ВСП, число состояний которой не превосходит числа состояний исходной ВСП и все свойства которой, выражаемые формулами логики PCTL, совпадают со свойствами исходной ВСП. Идея алгоритма заключается в построении последовательности вложенных разбиений множества состояний исходной ВСП. Алгоритм построения последовательности разбиений множества состояний заканчивает свою работу, когда эта последовательность стабилизируется. Редукция ВСП выполняется методом удаления эквивалентных состояний и переопределения вероятностей перехода. На примере показана возможность применения предложенного алгоритма к задаче вероятностной верификации протокола передачи сообщений через ненадежный канал связи, в котором пересылаемые сообщения могут пропадать или искажаться, с возможной коррекцией искажения. Отметим, что в результате такой редукции может получиться ВСП, которая, хотя и не содержит различных эквивалентных состояний, тем не менее, может не являться минимальной по числу состояний среди всех ВСП, эквивалентных исходной ВСП. В связи с этим встает вопрос об алгоритме нахождения минимальной по числу состояний ВСП, эквивалентной заданной ВСП, и исследовании единственности такой минимальной ВСП (с точностью до подходящим образом сформулированного понятия изоморфизма). Также представляет интерес исследование проблем минимизации других классов моделей, связанных с вероятностной верификацией, в частности минимизации марковских решающих процессов.

Литература

1. *Карпов Ю. Г.* Model checking. Верификация параллельных и распределенных программных систем. — СПб.: БХВ-Петербург, 2010. 560 с.
2. *Кемени Дж., Снелл Дж.* Конечные цепи Маркова. — М.: Наука, 1970. 225 с.
3. *Бухараев Р. Г.* Основы теории вероятностных автоматов. — М.: Наука, 1985. 288 с.
4. *Hansson H., Jonsson B.* A logic for reasoning about time and reliability // Formal Aspects Computing, 1994. Vol. 6. No. 5. P. 512–535.
5. *Kwiatkowska M., Parker D.* Advances in probabilistic model checking // NATO Science for Peace and Security Series. Information and Communication Security, 2012. Vol. 33. P. 126–151.

6. *Кларк Э. М., Грамберг О., Пелед Д.* Верификация моделей программ. Model Checking. — М.: МЦНМО, 2002. 416 с.
7. *Миронов А. М., Френкель С. Л.* Минимизация вероятностных моделей программ // Фундаментальная и прикладная математика, 2014. Т. 19. Вып. 1. С. 121–163.
8. *Baier C., Groesser M., Ciesinski F.* Partial order reduction for probabilistic systems // 1st Conference (International) on Quantitative Evaluation of Systems (QEST'04) Proceedings. — IEEE Computer Society Press, 2004. P. 230–239.
9. *D'Argenio P., Niebert P.* Partial order reduction on concurrent probabilistic programs // 1st Conference (International) on Quantitative Evaluation of Systems (QEST'04) Proceedings. — IEEE Computer Society Press, 2004. P. 240–249.
10. *Kwiatkowska M., Norman G., Parker D.* Symmetry reduction for probabilistic model checking // Computer aided verification / Eds. T. Ball, R. B. Jones. Lecture notes in computer science ser. — Springer, 2006. Vol. 4144. P. 234–248.
11. *Donaldson A., Miller A.* Symmetry reduction for probabilistic model checking using generic representatives // Automated technology for verification and analysis / Eds. S. Graf, W. Zhang. Lecture notes in computer science ser. — Springer, 2006. Vol. 4218. P. 9–23.
12. *Hart S., Sharir M., Pnueli A.* Termination of probabilistic concurrent programs // ACM Trans. Programming Languages Syst., 1983. Vol. 5. No. 3. P. 356–380.
13. *Vardi M.* Automatic verification of probabilistic concurrent finite state programs // 26th Annual Symposium on Foundations of Computer Science (FOCS'85) Proceedings. — IEEE Computer Society Press, 1985. P. 327–338.
14. *Courcoubetis C., Yannakakis M.* Verifying temporal properties of finite state probabilistic programs // 29th Annual Symposium on Foundations of Computer Science (FOCS'88) Proceedings. — IEEE Computer Society Press, 1988. P. 338–345.
15. *Bianco A., de Alfaro L.* Model checking of probabilistic and nondeterministic systems // Foundations of software technology and theoretical computer science / Ed. P. S. Triagarejan. Lecture notes in computer science ser. — Springer, 1995. Vol. 1026. P. 499–513.
16. *Baier C., Haverkort B., Hermanns H., Katoen J.-P.* Model-checking algorithms for continuous-time Markov chains // IEEE Trans. Software Eng., 2003. Vol. 29. No. 6. P. 524–541.
17. *Hansson H.* Time and probability in formal design of distributed systems. — Elsevier, 1994. 304 p.
18. *Baier C., Clarke E., Hartonas-Garmhausen V., Kwiatkowska M., Ryan M.* Symbolic model checking for probabilistic processes // Automata, languages and programming / Eds. P. Degano, R. Gorrieri, A. Marinetti-Spaccamela. Lecture notes in computer science ser. — Springer, 1997. Vol. 1256. P. 430–440.
19. *Hermanns H., Katoen J.-P., Meyer-Kayser J., Siegle M.* A Markov chain model checker // Tools and algorithms for the construction and analysis of systems / Eds. S. Graf, M. I. Schwartzbach. Lecture notes in computer science ser. — Springer, 2000. Vol. 1785. P. 347–362.
20. *De Alfaro L., Kwiatkowska M., Norman G., Parker D., Segala R.* Symbolic model checking of probabilistic processes using MTBDDs and the Kronecker representation // Tools and algorithms for the construction and analysis of systems / Eds. S. Graf, M. I. Schwartzbach. Lecture notes in computer science ser. — Springer, 2000. Vol. 1785. P. 395–410.
21. *Kwiatkowska M., Norman G., Parker D.* Probabilistic model checking in practice: Case studies with PRISM // ACM SIGMETRICS Performance Evaluation Review, 2005. Vol. 32. No. 4. P. 16–21.
22. *Kwiatkowska M., Norman G., Parker D.* PRISM 4.0: Verification of probabilistic real-time systems // Computer aided verification / Eds. G. Gopalakrishnan, S. Qadeer. Lecture notes in computer science ser. — Springer, 2011. Vol. 6806. P. 585–591.

Поступила в редакцию 05.11.14

A METHOD OF ENHANCING PROBABILISTIC VERIFICATION EFFICIENCY FOR COMPUTER AND TELECOMMUNICATION SYSTEMS

A. M. Mironov[1] and S. L. Frenkel[1,2]

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

[2]Moscow Institute of Radio, Electronics, and Automation (MIREA), 78 Prosp. Vernadskogo, Moscow 119454, Russian Federation

Abstract: The paper considers the problem of reduction of probabilistic transition systems (PTS) in order to reduce the complexity of model checking of such systems. The problem of model checking of a PTS is to calculate truth values of formulas of temporal probabilistic computational tree logic (PCTL) in the initial state of the PTS. The

paper introduces the concept of equivalence of states of a PTS and represents an algorithm for removing equivalent states. The result of this algorithm is a PTS such that all its properties expressed by formulas of PCTL coincide with those of the original PTS.

Keywords: verification; model checking; probabilistic transition systems; probabilistic temporal logic; reduction of probabilistic models

DOI: 10.14357/19922264140408

Acknowledgments

The research was partially supported by the Russian Foundation for Basic Research (project 12-07-00109).

References

1. Karpov, Yu. G. 2010. *Model checking. Verification of parallel and distributed systems*. St. Petersburg.: BHV-Peterburg. 560 p.
2. Kemeny, J. G., and J. L. Snell. 1976. *Finite Markov chains*. New York – Berlin – Heidelberg – Tokyo: Springer-Verlag. 225 p.
3. Bukharaev, R. G. 1985. *Foundations of probabilistic automata theory*. Moscow: Nauka. 288 p.
4. Hansson, H., and B. Jonsson. 1994. A logic for reasoning about time and reliability. *Formal Aspects Computing* 6(5):512–535.
5. Kwiatkowska, M., and D. Parker. 2012. Advances in probabilistic model checking. *NATO Science for Peace and Security Series, Information and Communication Security* 33:126–151.
6. Clarke, E. M., O. Grumberg, and D. Peled. 1999. *Model checking*. MIT Press. 314 p.
7. Mironov, A. M., and S. L. Frenkel. 2014. Minimization of probabilistic models of programs. *Fundamental Applied Mathematics* 19(1):121–163.
8. Baier, C., M. Groesser, and F. Ciesinski. 2004. Partial order reduction for probabilistic systems. *1st Conference (International) on Quantitative Evaluation of Systems (QEST'04) Proceedings*. IEEE Computer Society Press. 230–239.
9. D'Argenio, P., and P. Niebert. 2004. Partial order reduction on concurrent probabilistic programs. *1st Conference (International) on Quantitative Evaluation of Systems (QEST'04) Proceedings*. IEEE Computer Society Press. 240–249.
10. Kwiatkowska, M., G. Norman, and D. Parker. 2006. Symmetry reduction for probabilistic model checking. *Computer aided verification*. Eds. T. Ball and R. B. Jones. Lecture notes in computer science ser. 4144:234–248.
11. Donaldson, A., and A. Miller. 2006. Symmetry reduction for probabilistic model checking using generic representatives. *Automated technology for verification and analysis*. Eds. S. Graf and W. Zhang. Lecture notes in computer science ser. 4218:9–23.
12. Hart, S., M. Sharir, and A. Pnueli. 1983. Termination of probabilistic concurrent programs. *ACM Trans. Programming Languages Syst.* 5(3):356–380.
13. Vardi, M. 1985. Automatic verification of probabilistic concurrent finite state programs. *26th Annual Symposium on Foundations of Computer Science (FOCS'85) Proceedings*. IEEE Computer Society Press. 327–338.
14. Courcoubetis, C., and M. Yannakakis. 1988. Verifying temporal properties of finite state probabilistic programs. *29th Annual Symposium on Foundations of Computer Science (FOCS'88) Proceedings*. IEEE Computer Society Press. 338–345.
15. Bianco, A., and L. de Alfaro. 1995. Model checking of probabilistic and nondeterministic systems. *Foundations of software technology and theoretical computer science*. Ed. P. S. Triagarejan. Lecture notes in computer science ser. 1026:499–513.
16. Baier, C., B. Haverkort, H. Hermanns, and J.-P. Katoen. 2003. Model-checking algorithms for continuous-time Markov chains. *IEEE Trans. Software Engineering* 29(6):524–541.
17. Hansson, H. 1994. *Time and probability in formal design of distributed systems*. Elsevier. 304 p.
18. Baier, C., E. Clarke, V. Hartonas-Garmhausen, M. Kwiatkowska, and M. Ryan. 1997. Symbolic model checking for probabilistic processes. *Automata, languages and programming*. Eds. P. Degano, R. Gorrieri, and A. Marinetti-Spaccamela. Lecture notes in computer science ser. 1256:430–440.
19. Hermanns, H., J.-P. Katoen, J. Meyer-Kayser, and M. Siegle. 2000. A Markov chain model checker. *Tools and algorithms for the construction and analysis of systems*. Eds. S. Graf and M. I. Schwartzbach. Lecture notes in computer science ser. 1785:347–362.
20. De Alfaro, L., M. Kwiatkowska, G. Norman, D. Parker, and R. Segala. 2000. Symbolic model checking of probabilistic processes using MTBDDs and the Kronecker representation. *Tools and algorithms for the construction and analysis of systems*. Eds. S. Graf and M. I. Schwartzbach. Lecture notes in computer science ser. 1785:395–410.
21. Kwiatkowska, M., G. Norman, and D. Parker. 2005. Probabilistic model checking in practice: Case studies with PRISM. *ACM SIGMETRICS Performance Evaluation Review*. 32(4):16–21.
22. Kwiatkowska, M., G. Norman, and D. Parker. 2011. PRISM 4.0: Verification of probabilistic real-time systems. *Computer aided verification*. Eds. G. Gopalakrishnan and S. Qadeer. Lecture notes in computer science ser. 6806:585–591.

Received November 5, 2014

Contributors

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 Vavilov Str., Moscow 119333, Russian Federation; amironov66@gmail.com

Frenkel Sergey L. (b. 1951) — Candidate of Science (PhD) in technology, senior scientist, Institute of Informatics Problems, Russian Academy of Sciences, 44-2 Vavilov Str., Moscow 119333, Russian Federation; associate professor, Moscow Institute of Radio, Electronics, and Automation (MIREA), 78 Prosp. Vernadskogo, Moscow 119454, Russian Federation; fsergei@mail.ru

ИНФОРМАТИКА И ЕЁ ПРИМЕНЕНИЯ, 2014. Т. 8. Вып. 4. С. 70–77

ФАЛЬШТЕКСТЫ: КЛАССИФИКАЦИЯ И МЕТОДЫ ОПОЗНАНИЯ ТЕКСТОВЫХ ИМИТАЦИЙ И ДОКУМЕНТОВ С ПОДМЕНОЙ АВТОРСТВА*

М. Ю. Михеев[1], Н. В. Сомин[2], И. В. Галина[3], О. В. Золотарев[4], Е. Б. Козеренко[5], Ю. И. Морозова[6], М. М. Шарнин[7]

Аннотация: Современное текстовое пространство, включая Интернет, огромно и постоянно пополняется новыми текстами. Все текстовые документы можно разбить на две большие группы: «добросовестные тексты» и то, что можно назвать «фальштекстами». К настоящему времени индустрия фальштекстов приобрела столь массовое распространение, что возникает настоятельная потребность изучения этого явления и разработки действенных методов обнаружения подобных текстовых документов. Цель настоящей статьи состоит в том, чтобы дать адекватное описание понятия фальштекста как информационного и лингвистического феномена и предложить некоторые подходы к опознанию таких текстов.

Ключевые слова: порождение текста; обработка естественного языка; статистический анализ языковых объектов; плагиат; типология фальштекстов

DOI: 10.14357/19922264140409

1 Что такое фальштекст?

Задача исследования, представленного в данной статье, состоит в том, чтобы найти поддающиеся проверке признаки фальштекстов и по ним сделать выводы о степени их легитимности. Для этого постараемся выяснить смысл понятий «добросовестный текст» и «фальштекст». Под добросовестным текстом (текстовым документом) будем понимать текст, несущий какую-либо содержательную информацию. К добросовестным текстовым документам относятся тексты различных жанров: художественная, научная, инженерная литература, материалы в средствах массовой информации, сообщения на форумах Интернета, в социальных сетях и прочие тексты, имеющие хождение в современном потоке документов и относимые к точно определенному автору или источнику.

Создание документа, имеющего достаточно высокую степень признания в обществе, требует значительных трудовых и финансовых затрат. Поэтому у недобросовестных авторов возникает соблазн добиться аналогичного социального эффекта с помощью текстов, порождение которых требует гораздо меньших усилий. Подобные тексты и будем называть фальштекстами.

Материалом исследований, проводимых авторами, являются только естественно-языковые тексты, которые рассчитаны на прочтение их людьми (а не программами). Тексты на искусственных языках (например, на языках программирования) не рассматриваются. Термин «добросовестный текст» относится к способу порождения текста, а не к информации, содержащейся в том или ином высказывании, и ее истинности.

Таким образом, постановка проблемы фальштекстов не выходит за рамки анализа естественно-языкового текстового пространства.

Для опознания фальштекстов применяются уже наработанные научные методы, в частности методы статистического и структурного анализа лингвистических объектов, теории информации.

Итак, фальштекст — это текстовый документ, который маскируется его создателем под добросовестный текстовый документ, при этом делаются попытки изменить текст для получения нужного составителю социального и экономического эффекта,

*Работа выполнена при финансовой поддержке РФФИ (проект 13-06-00402).

[1]Научно-исследовательский вычислительный центр Московского государственного университета им. М. В. Ломоносова; Институт проблем информатики Российской академии наук, m-miheev@rambler.ru

[2]Институт проблем информатики Российской академии наук, somin@post.ru

[3]Институт проблем информатики Российской академии наук, irn_gl@mail.ru

[4]Российский новый университет, ol-zolot@yandex.ru

[5]Институт проблем информатики Российской академии наук, kozerenko@mail.ru

[6]Институт проблем информатики Российской академии наук, miss-yulia-morozova@yandex.ru

[7]Институт проблем информатики Российской академии наук, mc@keywen.com

однако его «фальшивость» можно определить путем текстологического анализа.

Использование фальштекстов может быть обусловлено разными целями: это похищение чужого текста, или плагиат; борьба за высокие рейтинги интернет-сайтов (а значит, высокую посещаемость сайтов и, соответственно, лучшую рекламу представленных на них товаров). Могут быть и более сложные случаи, например фальсификация, т. е. выдача своего собственного текста за чужой, принадлежащий знаменитому автору.

Следует подчеркнуть, что какие бы цели ни ставились, фальштексты, безусловно, весьма нежелательны. В ряде случаев запрет на использование фальштекстов носит правовой характер, например в случае плагиата или контрафакта, которые нарушают или ущемляют авторские права. Но и в тех случаях, когда правовых запретов нет, фальштексты существенно ухудшают информационную эффективность, засоряя информационное пространство текстами низкого качества и дезориентируя поисковые системы.

К настоящему времени индустрия фальштекстов приобрела столь массовое распространение, что возникает настоятельная потребность изучения этого явления и разработки действенных методов обнаружения подобных текстовых документов. Фактически мы становимся свидетелями возникновения новой отрасли информатики (пока еще не имеющей общепринятого названия), предметом изучения которой являются фальштексты.

В настоящей статье рассматриваются некоторые первоначальные аспекты проблематики фальштекстов. Объектом рассмотрения являются любые естественно-языковые тексты, доступные в электронном виде. Предметом является исключительно информационный аспект проблемы — изучение фальштекстов как информационного и лингвистического феномена.

2 Типология фальштекстов

Типологию фальштекстов начнем с наиболее распространенного типа — плагиата, т. е. присваивания авторства на чужой текст. Обычная цель плагиатора — незаслуженно присвоить себе определенную квалификацию или популярность. Круг плагиаторов широк и разнообразен: тут и копирование (полное или частичное) чужих художественных или научных произведений, тут и (возникающее как следствие первого) незаслуженное присвоение ученых степеней и званий, тут и изначально не замечаемое некритичное «списывание» студенческих курсовых работ. Плагиат помимо того, что обесценивает квалификационные звания, резко ухудшает эффективность учебного процесса, просто является обманом, разновидностью кражи и потому недопустим.

Плагиат — достаточно сложное явление, и поэтому сейчас в обществе нет четкого понимания, что считать плагиатом. Жестко установить границу плагиата невозможно, и на этот счет существуют разные мнения. Это понятие обязательно будет «плавать» в зависимости от моральных установок общества и его технических возможностей. Кроме того, невооруженным глазом видно, что степени плагиата различны: от дословного копирования до заимствования только нюансов самой идеи. Виды и степени плагиата, конечно, требуют тщательного изучения и более подробной их кодификации. Но назрела необходимость введения и различения по крайней мере двух уровней плагиата: «жесткого» плагиата, под которым понимается грубое копирование достаточно больших фрагментов текста, и «мягкого» плагиата, когда ворується идея посредством тем или иным способом скорректированного (и отличного от оригинала) текста.

Плагиат следует отличать от контрафакта — несанкционированного копирования чужих текстов (без присвоения авторства). Текстовый контрафакт сейчас очень распространен, поскольку он может существенно повысить рейтинг интернет-ресурса. Поэтому его выявление и борьба с ним также являются весьма актуальной задачей.

Другой большой класс фальштекстов — «генеранты», т. е. тексты, практически полностью сгенерированные компьютером. Обычная цель введения генерантов — повышение рейтинга в поисковых системах. Генеранты иногда предназначают для прочтения людьми — в этом случае они должны представлять собой осмысленные тексты. Однако по большей части их назначение — ввести в заблуждение программную систему. Для этого в соответствии с алгоритмами поисковых машин применяется либо генерация многочисленных текстов с определенными ключевыми словами (это так называемые «дорвеи»), либо встраивание таких блоков ключевых слов в посещаемые страницы. Аналогичные манипуляции выполняются и со ссылками на сайты. Если первый тип генерантов (для людей-потребителей контента) можно назвать «осмысленными генерантами», то второй тип получил название «сгенерированный шум», поскольку подобные тексты сильно зашумляют информационное пространство. Отметим, что все приемы веб-спама или «сгенерированного шума»: «скрытый текст», «клоакинг», «ссылочный спам», «дорвей», «редирект», «свопинг» и др. [1–4], — хорошо известны программистам, занимающимся продви-

жением сайтов, как «черные» и «серые» методы их раскрутки. К сожалению, столь некорректные методы сейчас приобретают все большее и большее распространение.

Отметим, что методики машинной генерации фальштекстов часто сочетаются с плагиатом (контрафактом). С одной стороны, сам плагиат (а точнее, его сокрытие) часто реализуется путем автоматической замены некоторых слов на синонимы (с целью затруднить обнаружение плагиата). А с другой стороны, осмысленные генеранты часто используют фразы или целые предложения из имеющихся в Интернете источников. Поэтому имеет смысл рассматривать еще один тип текстов: «гибриды», сочетающие в себе характеристики плагиатов и генерантов. Отметим, что гибриды, если их качество достаточно высоко, могут приносить пользу и не являться фальштекстами. Например, к таким гибридам, генерирующим вполне осмысленные тексты, принадлежит разработанная одним из авторов настоящей статьи система автоматического построения энциклопедий Keywen [5].

3 Пространство оценки фальштекстов

Для опознавания фальштекстов и установления их типов необходима разработка критериев и характеристик, допускающих их объективное определение. Явление фальштекстов настолько нетривиально, что выбрать единственный универсальный критерий для оценки таких текстов невозможно. Поэтому предлагается оценивать фальштексты с помощью системы критериев, куда входят: оригинальность, натуральность, лингвистичность и информативность (сокращенно ОНЛИ). Рассмотрим координаты этой системы и взаимоотношение ее компонент.

Два первых критерия — оригинальность и натуральность — образуют двумерную систему координат.

Оригинальность есть свойство текстового документа текстуально или идейно не зависеть от других имеющихся в информационном пространстве текстовых документов. Наиболее важной тут является оригинальность информации, новизна содержания, которые и должны выражаться в текстуальной уникальности.

Оригинальность может быть представлена в виде непрерывной «шкалы», измеряемой в долях или процентах, поскольку подавляющее большинство текстов оригинально лишь в какой-то степени и неизбежно содержит определенную долю заимствованных фрагментов. Причем под фрагментами понимаются не отдельные слова, а достаточно длинные части текста.

Разумеется, далеко не каждое заимствование текста является плагиатом или контрафактом. В огромном числе случаев копирование чужого текста не влечет негативных последствий или даже желательно, так как подтверждает актуальность исходного текста. Поэтому в критерий оригинальности помимо процента заимствований входят признаки ситуации, сигнализирующие о наличии или отсутствии плагиата или контрафакта. Например, факт публикации чужого текста под именем и фамилией публикатора без ссылки на источник является наиболее очевидным признаком наличия плагиата. Поскольку эти признаки влекут правовые последствия, то их идентификация является одной из важных задач изучения фальштекстов. С другой стороны, как уже указывалось, у плагиата нет четкой границы, и потому сформулировать их раз и навсегда не представляется возможным.

Натуральность выражает качество текста в смысле его приближения к естественному, написанному человеком художественному, научному или техническому тексту. Натуральность противопоставляется некорректности, искусственности текста, что связано с недавно появившимися и ныне широко применяющимися методами автоматической компьютерной генерации текстов.

По мнению авторов, натуральность является важнейшей характеристикой именно в контексте рассмотрения фальштекстов. Она позволяет идентифицировать и отсеивать многие виды некорректных генерантов. Дело в том, что качественное автоматическое порождение текста является одной из труднейших научных проблем и требует для своей реализации больших ресурсов, в частности финансовых. Но, как утверждается в работе [1], порождение генерантов выгодно только в случае затраты на них незначительных средств. Поэтому натуральность генерантов неизбежно является невысокой (хотя, вероятно, постепенно будет увеличиваться), что и может служить их идентифицирующим признаком. Отметим, что натуральность, как и оригинальность, представляется непрерывной шкалой.

Вводя понятие натуральности, авторы преследовали цель дать интегральный критерий качества текста. Анализируя этот критерий, можно заметить, что натуральность — сложное понятие, имеющее по крайней мере две составляющие: лингвистическую и информативно-содержательную. Поэтому натуральность имеет смысл представлять как суперпозицию двух частных критериев — лингвистичности и информативности.

Лингвистичность — это языковая правильность текста, соответствие законам языка, включая лексику, морфологию, синтаксис, семантику, прагматику. Степень нарушения лингвистичности может быть разная: начиная с отсутствия общей темы и потери связности сообщения при сохранении смысла отдельных предложений и вплоть до текста, содержащего бессвязный набор лексем.

Информативность есть свойство текста представлять собой осмысленное для человека сообщение. Отметим, что в понятие информативности входят далеко не все содержательные параметры сообщения. В частности, сообщение может быть истинным или ложным, положительным или негативным, провокационным или нейтральным — все это в понятие информативности не входит, а входит лишь информационная насыщенность текста. Разбирая понятие информативности, многие авторы указывают на его субъективный характер — информативность текста зависит от уровня знаний воспринимающего [6]. Безусловно, это так. Но этот факт не умаляет интегральных оценок информационной насыщенности текста, таких как индекс цитируемости.

Таким образом, система критериев ОНЛИ, на основании которых предполагается изучение фальштекстов, выглядит так, как это показано на рис. 1.

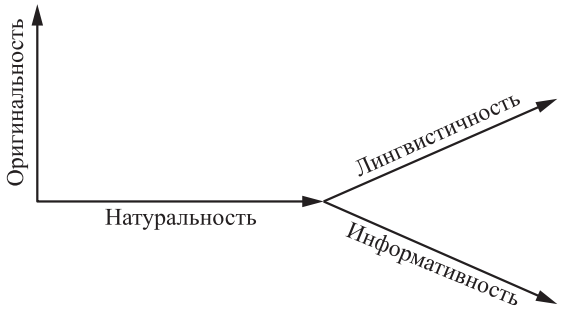


Рис. 1 Система критериев оценки фальштекстов ОНЛИ

Следует отметить, что система ОНЛИ обладает значительной гибкостью при ее приложении к различным частным задачам исследования фальштекстов. Так, применяя отдельные критерии, можно независимо исследовать: плагиат и контрафакт (оригинальность), общую приближенность текста к «нормальному, человеческому» (натуральность), языковую правильность текста (лингвистичность) и информационную ценность текста (информативность). Можно рассматривать тексты в двумерном пространстве оригинальность–натуральность, оригинальность–лингвистичность,

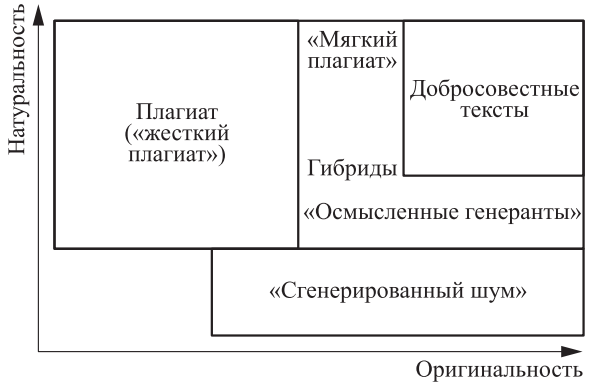


Рис. 2 Основные типы фальштекстов в пространстве оригинальность–натуральность

оригинальность–информативность. Наконец, можно абстрагироваться от проблем копирования и плагиата, исследовать текст в более привычных для лингвиста координатах лингвистичность–информативность. Например, рассмотрение проблемы в координатах оригинальность–натуральность позволяет наглядно представить основные типы фальштекстов (рис. 2.) При этом каждый тип текстов занимает свою нишу.

Добросовестные тексты характеризуются и высокой степенью оригинальности, и достаточно высокой натуральностью. Некорректные сгенерированные тексты («сгенерированный шум») отличаются от добросовестных текстов гораздо меньшей степенью натуральности. И наконец, плагиаты («жесткие») отличаются от добросовестных текстов гораздо меньшей степенью оригинальности.

Отметим, что границы между типами нельзя считать жесткими — типы отчасти перекрывают друг друга и плавно переходят один в другой. Так, между фальштекстами и добросовестными текстами лежит «нейтральная полоса» — тексты, занимающие промежуточное положение. Эту полосу, с одной стороны, занимают «мягкие плагиаты», а с другой — «осмысленные генеранты». Поскольку для создания плагиатов часто используется программная генерация текстов, то возникают всевозможные гибриды.

Необходимо указать, что все предложенные критерии, хотя и являются интуитивно ясными, но эксплицировать их и привязать к определенным легко вычисляемым параметрам документа оказывается непростой задачей. Поэтому необходима разработка конкретных методов и систем оценки этих критериев. В двух последующих разделах настоящей статьи предлагаются некоторые соображения на этот счет.

4 Методы оценки оригинальности текстов

Прежде всего, отметим, что оригинальность текста выявляет любая система обнаружения плагиата (СОП), так что в настоящий момент оценка оригинальности текстов неразрывно связана с созданием СОП. Рассмотрим несколько примеров.

Распространенная у нас в России система «eTXT Антиплагиат» [7] оценивает оригинальность в процентах. Принцип работы системы — сравнение текста с базой данных оригинальных текстов. Этот принцип далек от совершенства, поскольку поддерживать в актуальном состоянии такую базу текстов — задача трудновыполнимая. Хотя, как утверждается, системы, основанные на принципе сопоставления с базой данных, успешно работают. Примером может служить интернет-сервис Turnitin, разработанный в Университете Роберта Гордона из Абердина (Шотландия). Общий объем базы текстов — 25 ТБ, причем база пополняется со скоростью 40 млн новых страниц в день [8].

Имеются системы, которые осуществляют поиск текстов-аналогов прямо в Интернете, например: Advego Plagiatus (http://advego.ru/plagiatus), Praide Unique Content Analyser II (http://www.nado.su/downloads.html), онлайн-сервисы www.miratools.ru, www.istio.com.

Отметим, что параллельно с этим существует также и несколько «антисистем», помогающих недобросовестным авторам преодолеть СОП. К таким программам относятся, например, программы Smartrewriter pro (http://www.райхан.рф/raskrutka-bloga/predlagayu-vam-programmu-dlya-xoroshego-rerajta/) и Анти Плагиат Killer (http:// antiplagius.ru/), нацеленная на «обман» системы «Антиплагиат». Чтобы замаскировать плагиат под оригинальный текст, производятся следующие преобразования текстовых фрагментов:

- замена слова на синонимичное ему слово;
- добавление новых слов;
- удаление слов;
- разбиение одного предложения на два;
- объединение двух предложений в одно.

В целом можно сделать вывод, что сейчас ведется нешуточная война между разработчиками систем антиплагиата и плагиаторами, которые изобретают все более изощренные способы «уникализации» текста. Пока «плагиатчикам» удается работать с опережением.

Поскольку работа по выявлению плагиата связана с множеством тонких моментов, требующих вмешательства специалиста высокой квалификации, то имеет смысл говорить о рабочем месте «антиплагиатора», т. е. о комплексе программных средств, объединенных в единую систему, которая давала бы возможность специалисту воспользоваться теми или иными инструментами для наиболее точного определения степени оригинальности документа [9].

5 Методы определения лингвистичности и информативности текстов

Как уже указывалось выше, натуральность текста является обобщающим критерием, который разбивается на два более специальных: лингвистичность и информативность.

Насколько известно авторам, каких-то завершенных реально действующих систем, оценивающих лингвистичность текста, не существует. Однако методы лингвистики активно используются для выявления некорректных генерантов. Так, в работе [10] используется ряд лингвистических методов для обнаружения «дорвеев» — текстов, имитирующих натуральные тексты, но являющихся искусственными и содержащих большое количество ключевых слов для повышения поискового рейтинга определенной рекламной страницы.

Оценка информативности текста является весьма сложной задачей. Для ее решения может быть привлечена следующая гипотеза: документ имеет тем более высокую информативность, чем больше имеется ссылок на него или чем чаще он подвергается корректному цитированию. Конечно, эта гипотеза не гарантирует выявления всех информативных документов. Однако их основную массу можно определить с ее помощью и, более того, можно оценить степень информативности конкретного документа. На основе этой гипотезы одним из авторов настоящей работы был разработан метод оценки информативности, реализованный в автоматизированной системе построения текстовых энциклопедий Keywen [11]. Суть метода — в отыскании в среде Интернета множества текстовых документов, связанных с определенной предметной областью, и вычислении их рейтингов на основе корректного их цитирования. Этот рейтинг и принимается за оценку информативности текста. Такой метод сродни установлению рейтинга сайтов в поисковых системах с использованием индекса цитирования, но касается он не сайтов, а отдельных документов. Несомненно, такой подход может дать в большинстве случаев вполне приемлемый

результат, поскольку публикация научных статей в открытом доступе в Интернете приобретает все бóльшую популярность. Отчасти это связано с тем, что, как показал исследователь из Вильнюса Альгирдас Аушра [8], такие работы цитируются намного активнее, чем доступные по платной подписке или в библиотеке.

Следует, однако, отметить, что цитируемость документа далеко не всегда идентична его информационной ценности. Например, только что опубликованная статья, содержащая новые научные идеи, может по этой методике получить низкий рейтинг. Априори ценность этих новых идей неизвестна, и поэтому должно пройти определенное время, за которое научное сообщество оценит их и «проголосует» либо за них — увеличением индекса цитируемости, либо против них, оставив статью без ссылок. Отсюда видна ограниченность рейтингового подхода, который оценивает информативность в статике, не учитывая динамики информационных процессов. В связи с этим встает проблема определения времени «отстоя» документа, т. е. временно́го промежутка, за который научное сообщество «проголосует» за него своими ссылками.

6 Перспективы дальнейших исследований

Можно указать на несколько направлений дальнейших исследований фальштекстов.

1. Необходимо продолжить исследование их типологии. Сегодня информационное сообщество имеет дело только с описанными в статье разновидностями фальштекстов. Но не исключено, что в будущем появятся другие разновидности — ведь ввести в заблуждение человека (или систему) и получить от этого определенный выигрыш можно самыми разными способами.
2. Увеличение числа типов фальш-документов потребует введения новых критериев оценки в дополнение к уже предложенным.
3. Необходимо разработать эффективные методы оценки уже выявленных критериев ОНЛИ. В частности, тут могут быть использованы методы компьютерного семантического анализа [5].
4. Дальнейшей задачей могло бы стать лингвистическое и информационное описание методов, которыми пользуются «заказчики» и «исполнители» фальштекстов, и выработка рекомендаций для их эффективного выявления и устранения.
5. Необходима разработка практически удобных и эффективных программных средств обнаружения и обезвреживания фальштекстов. В будущем можно ждать появления как систем, работающих автоматически, так и «рабочих мест антиплагиатора», предоставляющих набор инструментов, способствующих выявлению фальштекстов.

Проблема фальшдокументов только ставится на повестку дня. Будем надеяться, что, как в свое время с компьютерными вирусами, эта проблема будет со временем удовлетворительно решена.

Литература

1. *Дёнди З., Гарсия-Молина Г.* Таксономия веб-спама. http://wseob.ru/seo/web-spam-taxonomy.
2. *Hall S.* On postmodernism and articulation: An interview with Stuart Hall // J. Communication Inquiry, 1986. No. 5. P. 35–60. doi: 10.1177/019685998601000204.
3. *Baggaley J., Spencer B.* The mind of a plagiarist // Learning Media Technol., 2005. Vol. 30. No. 1. P. 55–62.
4. *Selber S. A., Johnson-Eilola J.* Plagiarism, originality, assemblage // Comput. Composition, 2007. Vol. 24. No. 4. P. 375–403.
5. Keywen: Encyclopedia of keywords, key-phrases & key ideas. www.keywen.com.
6. *Шрейдер Ю. А.* Об одной модели семантической информации // Проблемы кибернетики. — М.: Наука, 1965. Вып. 13. С. 233–240.
7. eTXT Антиплагиат: Система проверки уникальности текста. http://www.etxt.ru/antiplagiat.
8. *Аушра А.* Научная электронная библиотека как средство борьбы с плагиатом // J. Educ. Technol. Soc., 2006. Vol. 9. Iss. 3. P. 270–276.
9. *Дягилев В. В., Цхай А. А., Бутаков С. В.* Архитектура сервиса определения плагиата, исключающая возможность нарушения авторских прав // Вестник НГУ. Сер. Информационные технологии, 2011. Т. 9. Вып. 3. С. 23–29.
10. *Павлов А. С., Добров Б. В.* Метод определения массово порождаемых неестественных текстов // Компьютерная лингвистика и интеллектуальные технологии: По мат-лам ежегодной Междунар. конф. «Диалог». — М.: РГГУ, 2010. Вып. 9(16). С. 368–374.
11. *Кузнецов И. П., Шарнин М. М., Мацкевич А. Г.* Интеллектуальные механизмы семантического поиска в сети Интернет // Системы и средства информатики, 2012. Т. 22. № 2. С. 129–145.

Поступила в редакцию 01.11.14

FALSE TEXTS: CLASSIFICATION AND METHODS OF IDENTIFICATION OF TEXT DOCUMENTS WITH IMITATIONS AND SUBSTITUTION OF AUTHORSHIP

M. Yu. Mikheev[1,2], N. V. Somin[2], I. V. Galina[1], O. V. Zolotaryev[3], E. B. Kozerenko[2], Yu. I. Morozova[2], and M. M. Charnine[2]

[1]Research Computer Center, M. V. Lomonosov Moscow State Uviversity (MGU NIVC), 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]Russian New University, 22 Radio Str., Moscow 105005, Russian Federation

Abstract: Modern textual space, including the Internet, is enormous and is constantly updated with new texts. All text documents can be divided into two large groups: "good texts" and that might be called "false texts." So far, the industry of false texts flow production has become so massive that there is an urgent need to study this phenomenon and to develop effective methods of detection of such text documents. The purpose of the paper is to give an adequate description of the concept of false text as information and linguistic phenomenon and suggest some approaches to the identification of such texts.

Keywords: text generation; natural language processing; statistical analysis of language objects; plagiarism; typology of false texts

DOI: 10.14357/19922264140409

Acknowledgments

The research was financially supported by the Russian Foundation for Basic Research (project 13-06-00402).

References

1. Gyongyi, Z., and H. Garcia-Molina. 2007. Taksonomiya web-spama [The taxonomy of web spam]. Available at: http://wseob.ru/seo/web-spam-taxonomy (accessed November 01, 2014).
2. Hall, S., and L. Grossberg. 1986. On postmodernism and articulation: An interview with Stuart Hall. *J. Communication Inquiry* 5:35–60. doi: 10.1177/019685998601000204.
3. Baggaley, J., and B. Spencer. 2005. The mind of a plagiarist. *Learning Media Technol.* 30(1):55–62.
4. Selber, S. A., and J. Johnson-Eilola. 2007. Plagiarism, originality, assemblage. *Comput. Composition* 24(4):375–403.
5. Keywen: Encyclopedia of keywords, key-phrases & key ideas. Available at: www.keywen.com (accessed November 01, 2014).
6. Shrader, J. A. 1965. Ob odnoy modeli semanticheskoy informatsii [On one model of semantic information]. *Problemy kibernetiki* [Problems of cybernetics]. Moscow: Nauka. 13:233–240.
7. eTXT Antiplagiat: Sistema proverki unikal'nosti teksta [eTXT Antiplagiarism: The system to verify the uniqueness of the text]. Available at: http://www.etxt.ru/antiplagiat (accessed November 01, 2014).
8. Aushra, A. 2006. Nauchnaya elektronnaya biblioteka kak sredstvo bor'by s plagiatom [Scientific electronic library as a means of combating plagiarism]. *J. Educ. Technol. Soc.* 9(3):270–276.
9. Diaghilev, V. V., A. A. Tskhai, and S. V. Butakov. 2011. Arkhitektura servisa opredeleniya plagiata isklyuchayuschaya vozmozhnost' narusheniya avtorskikh prav [Architecture of the service for detection of plagiarism which excludes the possibility of copyright infringement]. *Vestnik NGU. Ser. Informatsionnye Tekhnologii* [Herald of the NSU. Information technology ser.]. 9(3):23–29.
10. Pavlov, A. S., and B. V. Dobrov. 2010. Metod opredeleniya massovo porozhdaemykh neestestvennykh tekstov [Mass generation unnatural texts detection method]. *Komp'uternaya lingvistika i intellektual'nye tekhnologii: Po mat-lam Ezhegodnoi Mezhdunar. konf. "Dialog"* [Computational Linguistics and Intelligent Technologies: based on the Proceedings of the "Dialog" Annual Conference]. Moscow: RSUH. 9(16):368–374.
11. Kuznetsov, I. P., M. M.Sharnin, and A. G. Matskevich. 2012. Intellektual'nye mekhanizmy semanticheskogo poiska v seti Internet [Intellectual mechanisms for semantic searching in Internet]. *Sistemy i Sredstva Informatiki — Systems and Means of Informatics* 22(2):129–145.

Received November 01, 2014

Contributors

Mikheev Michael Yu. (b. 1957) — Doctor of Sciences in philology, leading scientist, Research Computer Center, M. V. Lomonosov Moscow State Uviversity (MGU NIVC), 1-52 Leninskiye Gory, GSP-1, Moscow 119991, Russian Federation; leading scientist, Institute of Informatics Problems, Russian Academy of Sciences, 44-2 Vavilov Str., Moscow 119333, Russian Federation; m-miheev@rambler.ru

Somin Nicolay V. (b. 1947) — Candidate of Science (PhD) in physics and mathematics, leading scientist, Institute of Informatics Problems, Russian Academy of Sciences, 44-2 Vavilov Str., Moscow 119333, Russian Federation; somin@post.ru

Galina Irina V. (b. 1967) — Senior scientist, Institute of Informatics Problems, Russian Academy of Sciences, 44-2 Vavilov Str., Moscow 119333, Russian Federation; irn_gl@mail.ru

Zolotaryev Oleg V. (b. 1959) — Candidate of Science (PhD) in technology, associate professor, Russian New University, 22 Radio Str., Moscow 105005, Russian Federation; ol-zolot@yandex.ru

Kozerenko Elena B. (b. 1959) — Candidate of Science (PhD) in linguistics, Head of Laboratory, Institute of Informatics Problems, Russian Academy of Sciences, 44-2 Vavilov Str., Moscow 119333, Russian Federation; kozerenko@mail.ru

Morozova Yulia I. (b. 1984) — scientist, Institute of Informatics Problems, Russian Academy of Sciences, 44-2 Vavilov Str., Moscow 119333, Russian Federation; miss-yulia-morozova@yandex.ru

Charnine Mikhail M. (b. 1959) — Candidate of Science (PhD) in technology, senior scientist, Institute of Informatics Problems, Russian Academy of Sciences, 44-2 Vavilov Str., Moscow 119333, Russian Federation; mc@keywen.com

ВИЗУАЛИЗАЦИЯ РЕЗУЛЬТАТОВ ДЛЯ МЕТОДА СКОЛЬЗЯЩЕГО РАЗДЕЛЕНИЯ СМЕСЕЙ*

А. К. Горшенин[1]

Аннотация: Метод скользящего разделения смесей (CPC-метод) представляет собой мощный инструмент анализа стохастических процессов различной природы. Именно на основании экспертной оценки результатов, полученных в ходе работы итерационных процедур CPC-метода, был получен ряд важных результатов в физике турбулентной плазмы, произведено уточнение математических моделей функционирования финансовых рынков. Зачастую каждая группа исследователей готовит результаты в удобном для себя формате, что затрудняет экспертам сравнение и интерпретацию результатов, особенно если речь идет о тестировании одной модели на принципиально разнородных выборках из отличных между собой предметных областей. В настоящей работе представлено удобное для исследователя-эксперта средство визуального отображения оценок параметров моделей, не зависящее от используемых для расчетов методов.

Ключевые слова: метод скользящего разделение смесей; пользовательский интерфейс; смеси нормальных распределений; вероятностные модели; интеллектуальный анализ данных

DOI: 10.14357/19922264140410

1 Введение

Для обнаружения и отслеживания изменений во времени в структуре формирующих стохастических процессов успешно используется так называемый CPC-метод [1]. Применение данного метода для проведения исследований в области физики турбулентной плазмы (см., например, работу [2]), данных с финансовых рынков (см., например, статью [3]) позволило получить ряд важных, принципиально новых результатов (в частности, впервые удалось определить число процессов, которые формируют ионно-звуковую турбулентность [4]). Получение подобных результатов на основании анализа портретов волатильности [1] было бы невозможным без экспертных оценок специалистов из областей — поставщиков данных. В то же время при реализации общих принципов обработки с помощью CPC-метода каждый исследователь придерживается своих представлений об удобстве отображения результатов, что, безусловно, не способствует универсальности графического вывода для различных предметных областей. В настоящей работе предлагается новое средство визуализации, учитывающее все особенности и тонкости CPC-метода, предлагающее максимальную наглядность и упрощающее интерпретацию результатов для широкого круга специалистов в области интеллектуального анализа данных.

2 Основы метода скользящего разделения смесей

В рамках CPC-метода традиционно используется многомерная интерпретация во времени такого важного параметра стохастических процессов, как волатильность. Предполагается, что волатильность может быть разложена на так называемые динамическую и диффузионную компоненты, которые позволяют отслеживать появление различных по своей природе эффектов в модели (например, трендовую составляющую и совокупное влияние значительного числа случайных факторов на основной процесс). С помощью CPC-метода каждая из указанных составляющих может быть представлена в виде совокупности различных компонент, обычно соответствующих определенным реальным процессам.

Предполагается, что моделирование стохастических процессов осуществляется с помощью обобщенных процессов Кокса со скачками, имеющими конечную дисперсию, так как данные модели являются в определенном смысле наилучшими для аппроксимации неоднородных хаотических потоков. Таким образом, задача статистической реконструкции распределений процессов сводится к оценке параметров неизвестного смешивающего распределения. Для корректности решения данной задачи,

*Работа выполнена при частичной финансовой поддержке гранта Президента Российской Федерации МК-4103.2014.9.

[1]Институт проблем информатики Российской академии наук; Московский государственный технический университет радиотехники, электроники и автоматики; agorshenin@ipiran.ru

а также с целью упрощения вида конечной модели, проводится аппроксимация конечными сдвиг-масштабными смесями нормальных законов. При этом такое приближение неизвестного распределения естественным образом приводит к многомерной интерпретации волатильности (подробнее об этом — в книге [1]).

Итак, в рамках CPC-метода предполагается, что неизвестное распределение некоторого процесса Z можно представить в виде:

$$F_Z(x) = \sum_{i=1}^{k} p_j F_i(x)\,, \tag{1}$$

где

$$F_i(x) = \frac{1}{\sigma_i\sqrt{2\pi}} \int\limits_{-\infty}^{x} \exp\left\{-\frac{(t-a_i)^2}{2\sigma_i^2}\right\} dt,$$

$$x \in \mathbb{R}\,,\ a_i \in \mathbb{R}\,, \sigma_i > 0\,; \tag{2}$$

$$\sum_{i=1}^{k} p_i = 1\,,\ \ p_i \geqslant 0\,. \tag{3}$$

Модель вида (1) называется конечной смесью распределений $F_i(x)$. В частности, если $F_i(x)$ определяются формулами вида (2), то говорят о конечных смесях нормальных законов. Параметры $p_1, \ldots, p_k$ называются весами компонент $F_1(x), \ldots$ $\ldots, F_k(x)$, при этом предполагается справедливость условия (3). Параметр k в приведенных выше формулах — количество компонент смеси.

Правую часть (1) с учетом (2) и обозначения $\Phi(x)$ для стандартной нормальной функция распределения можно переписать в виде:

$$F_Z(x) = \mathbb{P}(Z < x) = \sum_{i=1}^{k} p_i \Phi\left(\frac{x-a_i}{\sigma_j}\right) = \mathbb{E}\Phi\left(\frac{x-V}{U}\right),$$

где пара случайных величин U, V имеет дискретное распределение

$$\mathbb{P}((U,V) = (\sigma_i, a_i)) = p_i, \quad i = 1, \ldots, k\,.$$

Волатильность естественно отождествить с величиной

$$\mathbb{D}Z = \mathbb{D}V + \mathbb{E}U^2 \tag{4}$$

(или $\sqrt{\mathbb{D}Z}$; о справедливости самого представления (4) подробнее см. книгу [1]). При этом величина $\mathbb{D}V$ в формуле (4) зависит только от весов p_i и параметров сдвига a_i компонент, а потому характеризует ту часть волатильности, которая обусловлена наличием локальных трендов, т. е. динамическую компоненту волатильности. Величина $\mathbb{E}U^2$ в (4) зависит только от весов p_i и параметров масштаба («коэффициентов диффузии») σ_i компонент и потому характеризует диффузионную компоненту волатильности.

Статистические закономерности поведения стохастических процессов зачастую изменяются во времени нерегулярным образом, результатом чего является отсутствие универсального смешивающего закона. Таким образом, чтобы изучить динамику изменения статистических закономерностей в поведении исследуемого хаотического процесса, задача статистического разделения конечных смесей нормальных законов должна быть последовательно решена на интервалах времени, постоянно сдвигающихся в направлении «астрономического» времени. Такие интервалы в рамках CPC-метода принято называть окнами и проводить оценивание неизвестных параметров на каждом из положений окна. Обычно размер окна (т. е. количество составляющих элементов) выбирается заранее и не изменяется в процессе работы. Для получения наиболее точной картины окно на каждом шаге «сдвигается» только на один элемент выборки. Это позволяет отследить момент формирования или исчезновения той или иной компоненты.

Несомненным достоинством CPC-метода является выявление объективно существующих диффузионных и динамических компонент волатильности. Таким образом, сначала автоматически выделяются компоненты, формирующие стохастический процесс (со статистической оценкой их параметров), и лишь затем возникает задача поиска соответствия выделенных компонент предметной области. Это позволяет избегать предположения о малой значимости (или наоборот — значительном вкладе) того или иного явления на этапе априорного анализа.

3 Требования к графическому выводу метода скользящего разделения смесей

При практической реализации CPC-метода возникает задача отображения изменяющихся во времени динамической и диффузионной компонент. Помимо численного значения соответствующих параметров на каждом шаге каждая оценка имеет свой вес, который не связан с абсолютной величиной отображаемого параметра, но который также необходимо продемонстрировать на графике. При работе с CPC-методом на первый план зачастую выходит экспертная оценка результатов, полученных в автоматическом режиме, а потому качество

графического вывода приобретает первостепенное значение.

Для отображения результатов традиционно [1] используются двумерные графики, в которых каждой точке по оси абсцисс соответствует текущее положение окна, а по оси ординат откладываются значения оценок, полученных на данном шаге. Для изображения весов на графиках используется цветовая шкала с плавной градацией от темно-синего до темно-красного, при этом по специальному правилу каждому весу из сегмента $[0, 1]$ ставится в соответствие цвет по шкале RGB. Однако при работе с реальными данными часто возникает множество оценок с небольшими весами, которые скорее относятся к «шумам», т. е. погрешностям вычислений, и не несут значительной смысловой нагрузки, но затрудняют комфортное восприятие основных компонент. Для преодоления такой сложности можно использовать различные способы. Например, осуществлять отбрасывание компонент с некоторым весом при выводе графиков. Но такая «фильтрация» вывода, очевидно, может существенно изменить сами результаты, так как далеко не во всех ситуациях можно корректно установить порог отсечения для весов. Возможно при рисовании точек использовать такой параметр, как прозрачность, однако при типографской печати таких графиков могут возникать определенные сложности, также ухудшающие восприятие информации. В настоящей статье предложено решение проблемы, основанное на определении размера выводимой точки в зависимости от веса соответствующего параметра, а именно: размер каждой точки при рисовании задается формулой $\lceil p_i^{(m)} \cdot \mathrm{Size}_{\max} \rceil$, где $p_i^{(m)}$ обозначает вес компоненты с номером i на m-м итерационном шаге, а $\mathrm{Size}_{\max}$ — некоторое заранее заданное максимальное значение размера выводимой точки. Кроме того, максимальный размер $\mathrm{Size}_{\max}$ варьируется для разных размеров выборок: для более длинных рядов разумно использовать меньшее значение, чтобы отдельные точки не сливались друг с другом (особенно это важно при рисовании объектов с близкими значениями оценок параметров и весов).

4 Описание функциональных возможностей программного продукта

Перейдем к описанию возможностей средства визуализации результатов для CPC-метода. Начальный экран при запуске приложения представлен на рис. 1.

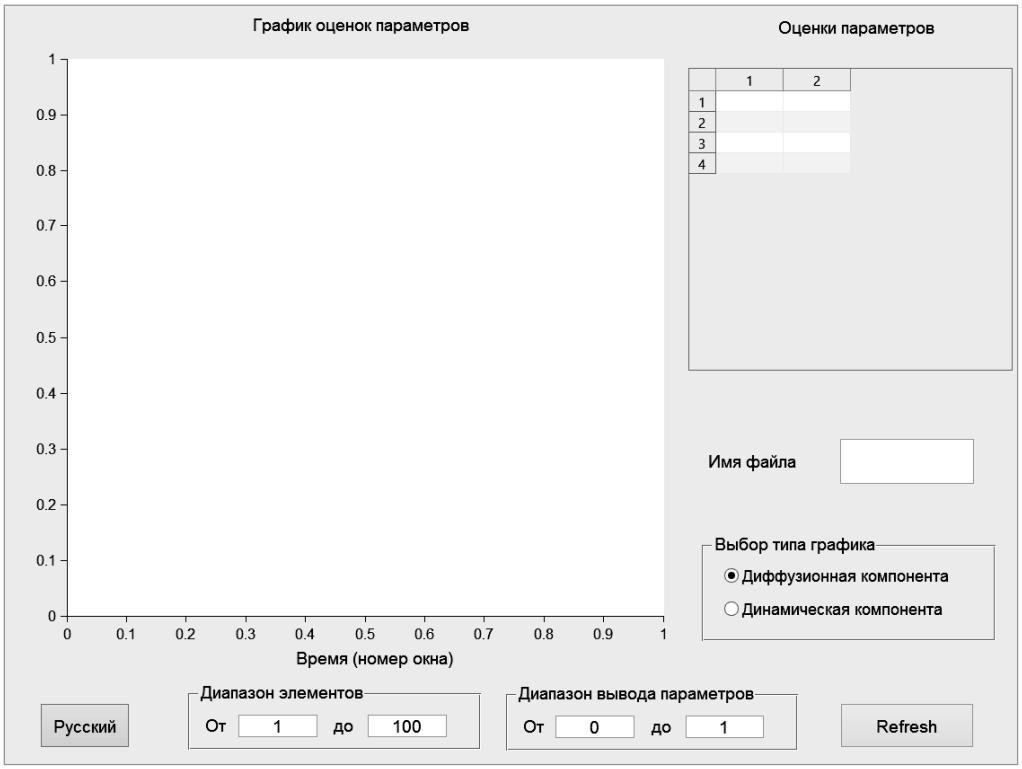


Рис. 1 Вид начального экрана приложения

Область «График оценок параметров» («Figure») предназначена для непосредственной визуализации оценок, полученных с помощью какого-либо метода. В начале работы обе оси промаркированы в диапазоне от 0 до 1 с шагом 0,1. Однако при отрисовке актуального графика обозначения будут автоматически выбраны в соответствии с данными.

Кнопка с надписью «Русский» («English») позволяет выбирать язык интерфейса (по умолчанию установлено отображение на русском языке, нажатие на кнопку изменяет все надписи на англоязычные варианты), содержимое остальных полей при переключении не изменяется, область вывода графика не перерисовывается, не происходит повторного вывода значений оценок.

Блок «Диапазон элементов» («Gap for windows») задает область вывода оценок по временно́й оси, соответствующей количеству сдвигов окна в СРС-методе. Например, если есть необходимость рассмотреть крупнее отдельную область или анализируемый ряд слишком большой (оценки сливаются), то можно отобразить только часть данных. В качестве значений по умолчанию используется диапазон от первого до сотого элемента.

Блок «Диапазон вывода параметров» («Parameter gap») задает область вывода значений оценок. Для разных данных получаются разные по порядку оценки, кроме того, динамическая компонента может принимать и отрицательные значения. Для удобства масштабирования и корректности вывода параметров и предназначен данный блок. В качестве значений по умолчанию используется диапазон от 0 до 1.

Таблица «Оценки параметров» («Estimations of parameters») отображает полный набор оцененных параметров из блока «Диапазон элементов», который хранится в файле, адрес (имя) которого задается в поле «Имя файла» («Filename»).

По умолчанию предполагается, что отображается диффузионная компонента («Diffusive component»), однако в блоке «Выбор типа графика» («Type of figure») это можно изменить, выбрав динамическую компоненту («Dynamic component»). Нажатие кнопки «Refresh» осуществит корректное обновление (либо первичное изображение) графика в соответствии с выбранными настройками с помощью специально разработанного для СРС-метода алгоритма рисования.

5 Примеры

В настоящем разделе рассмотрим примеры применения разработанного средства визуализации для реальных данных. В данном случае не станем уделять внимание предметным областям, которые выступали в роли поставщиков выборок, а сосредоточимся исключительно на демонстрации описанных выше возможностей программы.

На рис. 2 изображена диффузионная компонента для некоторого ряда с названием Params. Вывод осуществляется для положений окон от 4000 до 9000, при этом границы параметров установлены от 0 до $1{,}1 \cdot 10^{-4}$.

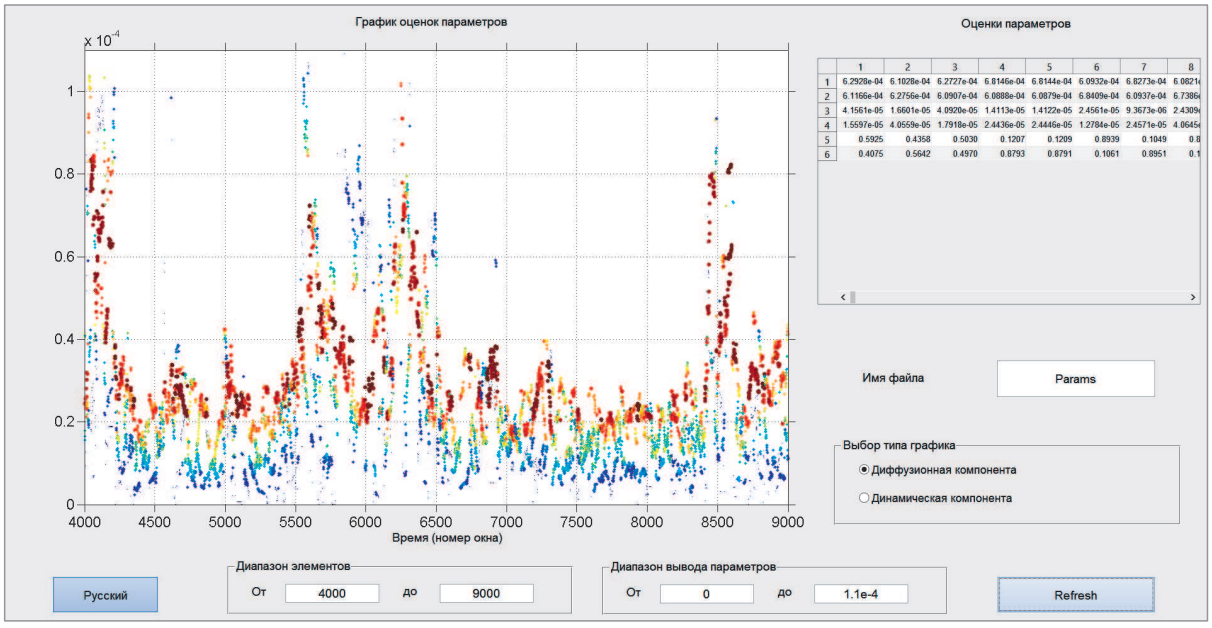


Рис. 2 Пример вывода диффузионной компоненты для некоторого ряда Params

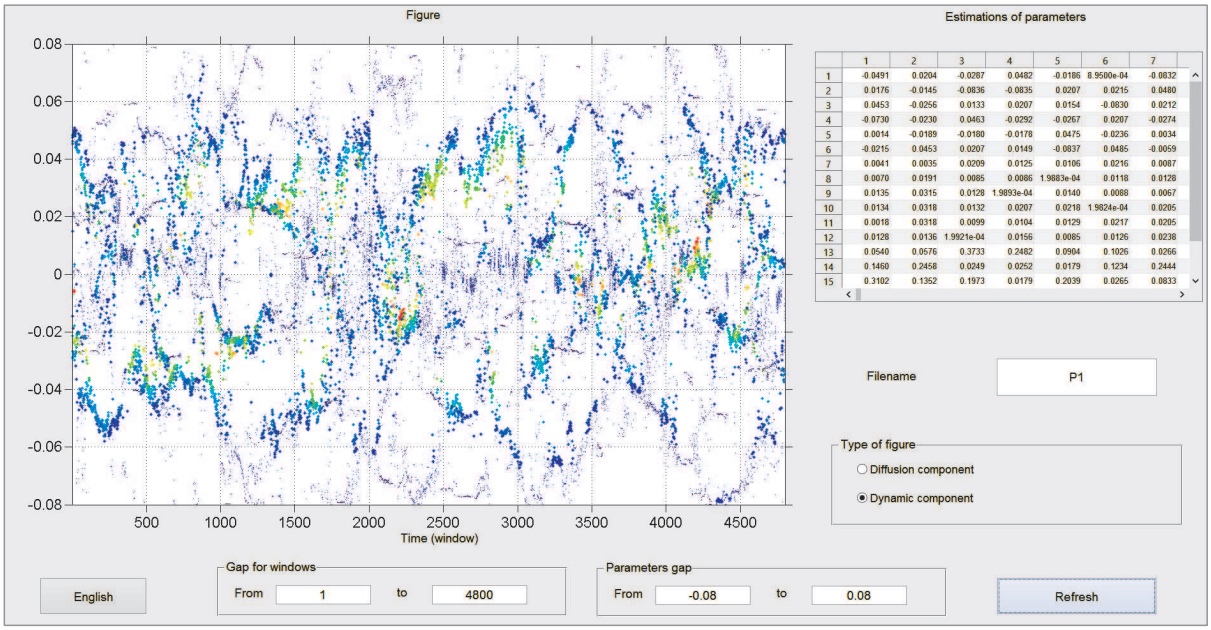


Рис. 3 Пример вывода динамической компоненты для некоторого ряда P1 (англоязычный вариант интерфейса)

В данной ситуации расчеты проводились с высокой точностью (с пороговым значением 10^{-12} для критерия останова), для максимального количества компонент в смеси использовалась величина $k = 2$. Визуально четко выделяются две формирующие компоненты — темно-красная и составленная из точек разных цветов (голубой, зеленый, желтый, оранжевый). Шумовые компоненты с очень маленькими весами практически отсутствуют, например можно выделить всего несколько небольших (напомним, что вес и размер точки для рисования в разработанном средстве визуализации тесно связаны) отметок между делениями 5500 и 6000 на оси абсцисс. В такой ситуации какая-либо фильтрация данных для улучшения визуального восприятия не требуется. В окне «Оценки параметров» отображается табличное представление параметров, где первые две строки соответствуют значениям для динамической компоненты, третья и четвертая — диффузионной, а пятая и шестая строки содержат изменяющиеся во времени значения весов.

На рис. 3 изображена динамическая компонента волатильности для другого ряда с именем `P1`. Вывод осуществляется от первого положения окна до значения 4800, при этом границы параметров определяются диапазоном от $-0{,}08$ до $0{,}08$. Также отметим, что рис. 3 демонстрирует вариант вывода с англоязычными надписями.

Для данного ряда в качестве максимального количества компонент использовалось значение $k = 6$, а вычисления проводились с точностью 10^{-5}. На графике присутствуют от двух до трех основных компонент, при этом общая картина существенно дополняется шумовыми составляющими. Однако использование точек небольшого размера для значений параметров с малыми весами позволяет анализировать приведенный график без проведения какой-либо фильтрации достаточно уверенно. Таблица в окне «Оценки параметров» построена по тому же принципу, как и для предшествующего графика, но теперь диапазон строк для динамической компоненты — 1–6, для диффузионной — 7–12, а строки 13–18 соответствуют весам.

6 Заключение

Средство визуализации создано с помощью встроенного языка программирования системы `MATLAB` и ориентировано на работу с файлами формата `MAT`. Однако не представляет каких-либо сложностей экспортировать данные и из текстовых файлов `TXT`, таблиц `CSV`. Это позволяет осуществлять расчеты с помощью методов, реализованных с помощью произвольных языков программирования (например, специализированных, достаточно низкоуровневых — для повышения скорости вычислений), а затем работать с удобным средством визуализации.

Это особенно важно для групп исследователей, осуществляющих обработку различными (причем и с программной, и с математической точки зрения) способами, так как общие итоговые графики будут унифицированы за счет единого интерфейса.

Предусмотрена возможность автоматического сохранения области графика в файл PNG, что удобно для представления результатов в рамках презентаций, журнальных публикаций. Таким образом, разработанное средство визуализации можно использовать не только в качестве решения для проведения интеллектуального анализа данных, но и вспомогательного инструмента подготовки отчетов по научным и практическим исследованиям. Кроме того, описанное в работе средство визуализации обладает интуитивно понятным визуальным интерфейсом, а потому подойдет как для продвинутых пользователей, так и для тех, кто только начинает изучение возможностей СРС-метода.

Для разработанного программного кода в Роспатенте получено свидетельство государственной регистрации программ для ЭВМ «Средство визуализации результатов для метода скользящего разделения смесей» (автор Горшенин А. К., № 2014661369 от 29.10.2014).

В качестве возможного пути дальнейшего развития созданного продукта можно отметить возможность использования интерфейса в качестве базового модуля для построения информационной технологии, интегрирующей методы визуализации с инструментарием оценивания неизвестных параметров модели. Разработка такого пакетного решения представляет собой совершенно самостоятельную и в достаточной степени трудоемкую задачу.

Литература

1. *Королев В. Ю.* Вероятностно-статистические методы декомпозиции волатильности хаотических процессов. — М.: Изд-во Моск. ун-та, 2011. 512 с.
2. *Skvortsova N. N., Batanov G. M., Malakhov D. V., Petrov A. E., Saenko V. V., Sarksyan K. A., Kharchev N. K., Kholnov Yu. V., Korolev V. Yu., Zhukov Yu. V., Rey M., Merkulov A. S., Shatalin S. V., Lashkul S. I., Vekshina E. O., Popov A. Yu.* Estimation of dynamic and diffusive components in edge turbulent particle fluxes in the L-2M stellarator and the FT-2 tokamak // 21st IAEA Fusion Energy Conference. Chengdu, 2006. IAEA-CN-149, PD/P6-3.
3. *Горшенин А. К., Королев В. Ю., Турсунбаев А. М.* Медианные модификации EM- и SEM-алгоритмов для разделения смесей вероятностных распределений и их применение к декомпозиции волатильности финансовых временных рядов // Информатика и её применения, 2008. Т. 2. Вып. 4. С. 12–47.
4. *Батанов Г. М., Горшенин А. К., Королев В. Ю., Малахов Д. В., Скворцова Н. Н.* Эволюция вероятностных характеристик низкочастотной турбулентности плазмы в микроволновом поле // Математическое моделирование, 2011. Т. 23. № 5. С. 35–55.

Поступила в редакцию 13.11.14

A VISUALIZATION OF ESTIMATORS IN THE METHOD OF MOVING SEPARATION OF MIXTURES

A. K. Gorshenin[1,2]

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

[2]Moscow Institute of Radio, Electronics, and Automation (MIREA), 78 Prosp. Vernadskogo, Moscow 119454, Russian Federation

Abstract: The method of moving separation of mixtures (MSM method) is a powerful tool for analyzing different stochastic processes. Using the MSM method within the experts' conclusions for the results obtained by iterative numerical procedures, a number of important results were achieved in the physics of turbulent plasma, a few mathematical models for the functioning of financial markets were refined. In most cases, research teams present results in a form which is convenient just for themselves, and it is difficult for experts to compare and interpret results, especially, in the case when the model is tested on fundamentally dissimilar samples from different subject areas. The paper presents a visualization tool for displaying parameter estimates independently of the used numerical methods. The tool is convenient for researchers and experts.

Keywords: method of moving separation of mixtures; user interface; normal mixtures; probabilistic models; data mining

DOI: 10.14357/19922264140410

Acknowledgments

The research was partially financially supported by the President Grant for Government Support of Young Russian Scientists MK-4103.2014.9.

References

1. Korolev, V. Yu. 2011. *Veroyatnostno-statisticheskie metody dekompozitsii volatil'nosti khaoticheskikh protsessov* [Probabilistic and statistical methods of decomposition of volatility of chaotic processes]. Moscow: Moscow University Publishing House. 512 p.
2. Skvortsova, N. N., G. M. Batanov, D. V. Malakhov, A. E. Petrov, V. V. Saenko, K. A. Sarksyan, N. K. Kharchev, Yu. V. Kholnov, V. Yu. Korolev, Yu. V. Zhukov, M. Rey, A. S. Merkulov, S. V. Shatalin, S. I. Lashkul, E. O. Vekshina, and A. Yu. Popov. 2006. Estimation of dynamic and diffusive components in edge turbulent particle fluxes in the L-2M stellarator and the FT-2 tokamak. *21st IAEA Fusion Energy Conference*. Chengdu. IAEA-CN-149, PD/P6-3.
3. Gorshenin, A. K., V. Yu. Korolev, D. V. Malakhov, and A. M. Tursunbaev. 2008. Mediannye modifikatsii EM- i SEM-algoritmov dlya razdeleniya smesey veroyatnostnykh raspredeleniy i ikh primenenie k dekompozitsii volatil'nosti finansovykh vremennykh ryadov [Median modification of EM- and SEM-algorithms for separation of mixtures of probability distributions and their application to the decomposition of volatility of financial time series]. *Informatika i ee Primeneniya — Inform. Appl.* 2(4):12–47.
4. Batanov, G. M., A. K. Gorshenin, V. Yu. Korolev, D. V. Malakhov, and N. N. Skvortsova. 2011. Evolyutsiya veroyatnostnykh kharakteristik nizkochastotnoy turbulentnosti plazmy v mikrovolnovom pole [The evolution of probability characteristics of low-frequency plasma turbulence]. *Matematicheskoe Modelirovanie* [Mathematical Modeling] 23(5):35–55.

Received November 13, 2014

Contributor

Gorshenin Andrey K. (b. 1986) — Candidate of Science (PhD) in physics and mathematics, senior scientist, Institute of Informatics Problems, Russian Academy of Sciences, 44-2 Vavilov Str., Moscow 119333, Russian Federation; associate professor, Moscow Institute of Radio, Electronics, and Automation (MIREA), 78 Prosp. Vernadskogo, Moscow 119454, Russian Federation; agorshenin@ipiran.ru

ИНФОРМАТИКА И ЕЁ ПРИМЕНЕНИЯ, 2014. Т. 8. Вып. 4. С. 85–93

ОБ ЭРГОНОМИЧЕСКИХ ЗАВИСИМОСТЯХ МЕЖДУ ПАРАМЕТРАМИ СИТУАЦИОННОГО ЗАЛА С ИСПОЛЬЗОВАНИЕМ ИЗОГНУТОГО КОЛЛЕКТИВНОГО ЭКРАНА

А. А. Зацаринный[1], К. Г. Чупраков[2]

Аннотация: Рассмотрен подход к определению зависимостей между параметрами ситуационного зала: размерами помещения, числом наблюдателей, информационной емкостью контента (количеством знаков) и шириной экрана. Эти зависимости позволяют рассчитать неизвестный параметр ситуационного зала при известных других с выполнением требований государственных и международных стандартов по эргономике рабочих мест. Предложенные формулы применимы и для изогнутых экранов, определяемых в рамках статьи углом кривизны β (для плоского экрана $\beta = 0$). Данный параметр может быть интерпретирован как угол наклона между дисплеями в полиэкране. Наличие этого параметра позволяет оценить эффективность использования изогнутых экранов в составе систем отображения информации коллективного использования. Предложен общий подход к определению количества рабочих мест для коллективного экрана, который может быть применен для их различных взаимных расположений.

Ключевые слова: изогнутый экран коллективного пользования; ситуационный зал; диспетчерский пункт; эргономические зависимости; область комфортного наблюдения; угол кривизны экрана; видеостена; полиэкран; эффективность; оправданность цены

DOI: 10.14357/19922264140411

1 Введение

Опыт создания ситуационных центров (СЦ) обозначил целый ряд проблем, которые препятствуют разработке и внедрению новых технологий в процесс управления [1–4]. Эти трудности могут возникать в результате недостаточного применения системного подхода при проектировании, когда имеющийся функционал прикладных средств не подкреплен достаточной технологической или технической базой и наоборот. Значительное число проблем создания и внедрения СЦ вызвано человеческим фактором: плохой мотивацией персонала на обучение новым процессам и слабой заинтересованностью первого лица [5, 6].

Важно понимать, что любая проблема, будь то техническая или организационная, может стать узким местом в обеспечении функционирования СЦ в требуемых режимах [3, 7]. Для выявления таких узких мест необходимы системные оценки эффективности СЦ, которые должны учитывать оценки всех его отдельных компонентов [8]. При этом в рамках технического и эксплуатационного компонентов важную роль играют эргономические показатели, так как они определяют эффективность пользовательского интерфейса в СЦ. Эргономические требования к системам отображения информации являются важными с точки зрения не только обеспечения комфортных условий работы, но и более эффективного использования пространственного и материального ресурса [9, 10] .

Одним из наиболее заметных направлений в развитии средств отображения информации является применение так называемых изогнутых экранов (curved screen), реализуемых на основе технологии OLED (organic light-emitted diode), широко применяемой для мобильных устройств. Отметим, что идея таких «изогнутых» или вообще неплоских экранов не нова: помимо кинотеатров неплоские экраны можно встретить и в диспетчерских пунктах. Производители профессиональных полиэкранов и видеостен предусматривают возможность создания изогнутых экранов за счет взаимного поворота между отдельными дисплеями. Вместе с тем ощутимых преимуществ, которые дают изогнутые экраны, производители и их маркетологи сформулировать не смогли [11].

В настоящей статье сделана попытка восполнить этот пробел. Рассмотрены методические подходы к определению зависимостей между размерами помещения, размерами экрана, числом наблюдателей и характеристиками отображаемого контента, сформулированные в работах [9, 10], применительно к изогнутому экрану. Полученные

[1] Институт проблем информатики Российской академии наук, azatsarinny@ipiran.ru
[2] Институт проблем информатики Российской академии наук, chkos@rambler.ru

зависимости позволяют получать оценки эффективности применения изогнутых экранов в системах коллективного отображения информации СЦ. Более того, в отличие от известного оценочного подхода по уже реализованным комплексам в [12] предлагаемый в статье подход предусматривает построение системы «помещение–экран–наблюдатели» в строгом соответствии с нормативно-технической базой, определяемой существующими государственными и международными стандартами в части эргономических норм.

2 Общий подход. Термины и определения

Создание и оборудование ситуационного зала должно среди прочих требований опираться на существующие стандарты по эргономике, действующие на территории РФ. Большинство указаний, содержащихся в стандартах, основано на особенностях человеческого восприятия и потому может служить практическим руководством при оснащении помещений. Это относится и к средствам отображения информации. Подход, сформулированный в [9, 10] и используемый в данной статье, опирается на стандарты [13–19].

О существовании связей между основными параметрами системы «помещение–дисплей–наблюдатели» можно судить на основании следующего примера. Увеличение информационной емкости контента за счет уменьшения символов незамедлительно приведет к уменьшению проектного расстояния наблюдения. Это, в свою очередь, уменьшит рабочую площадь наблюдения, а следовательно, и количество персонала, который может работать в нем одновременно.

На основании рекомендаций, сформулированных в [13–19] и выделенных в [9, 10], можно приступить к формированию взаимосвязей между основными параметрами системы «помещение–экран–наблюдатели». Далее в работе будут использованы следующие термины и обозначения:

D — проектное расстояние наблюдения, измеряется в метрах. Это расстояние или диапазон расстояний между экраном и глазами наблюдателей, при котором изображение соответствует требованиям разборчивости и удобочитаемости;

N — количество людей, которые должны одновременно работать с коллективным экраном, получая с него визуальную информацию в комфортных условиях;

Q — диаметр помещения, ограниченного стенами. В большинстве случаев помещение является прямоугольным, а в рамках статьи, обобщенно, — выпуклым. Параметр Q — максимальное из расстояний между двумя произвольными точками помещения, измеряется в метрах. В некоторых задачах Q может быть ограничено искусственным образом ввиду особенностей геометрии помещения, группировки наблюдателей. Отметим разницу между проектным расстоянием наблюдения D и диаметром помещения Q. Первый параметр характеризуется свойствами системы «экран–наблюдатели», а второй — исключительно свойствами помещения. Ясно, что D не может превосходить Q;

W — максимальное расстояние между двумя точками экрана в горизонтальной плоскости, «плоская» ширина экрана. Для плоского экрана W соответствует его ширине, а для кривого — расстоянию между его краями. Измеряется в метрах;

I — необходимая статичная информационная емкость отображаемого контента — максимальное количество знаков или символов, которые должен отобразить дисплей в одном кадре или неподвижном изображении. Определяется на основании задач ситуационного зала и иных приложений и объемов отображаемого контента. Измеряется в количестве отображаемых знаков;

α — максимальный стягиваемый угол (угловой размер экрана) по горизонтали, ограничиваемый уровнем концентрации наблюдения;

изогнутый экран — экран, сечение которого в горизонтальной плоскости является дугой некоторой окружности. Также под изогнутым экраном будем понимать полиэкран, составленный из плоских дисплеев одинакового размера и расположенных друг к другу под одним углом (дискретное приближение к окружности). В рамках данной работы в вертикальной плоскости экран считается плоским;

β — угол, определяющий кривизну (изогнутость) экрана. Определяется как половина дуги, которую стягивает экран на соответствующей ему окружности. Такое определение эквивалентно углу наклона между отдельными дисплеями в случае полиэкрана из плоских дисплеев. Положительные значения β соответствуют случаю, когда наблюдатели расположены в той же части пространства, что и центр окружности, на которую ложится горизонтальное сечение экрана, или экран, «выпуклый от наблюдате-

ля». Отрицательные значения β, соответственно, — когда экран «выпуклый к наблюдателю». Если $\beta = 0$, то экран плоский;

ОКН — область комфортного наблюдения — область пространства, где выполнены рекомендации международных и государственных стандартов по оборудованию рабочих мест наблюдения с коллективного экрана.

3 Построение области комфортного наблюдения и исследование ее свойств

По аналогии с результатами [9, 10] с учетом ограничений [13–19] ОКН в случае изогнутого экрана будет также являться фигурой пересечения нескольких областей (рис. 1).

Рассмотрим такую систему координат, ось абсцисс которой проходит через крайние точки экрана, а ось ординат является осью симметрии для экрана и направлена в сторону наблюдателей. Единица измерения по обеим осям равна 1 м. С помощью средств аналитической геометрии вычислим координаты основных точек, которые будут участвовать в дальнейших оценках:

O_1 — центр левого круга: $O_1\left((D/2)\sin\beta - W/2;\ (D/2)\cos\beta\right)$;

O_2 — центр правого круга: $O_2\left(W/2 - (D/2)\sin\beta;\ (D/2)\cos\beta\right)$;

l_1 — прямая, выходящая из левого края экрана под углом α к OY;

l_2 — прямая, выходящая из правого края экрана под углом α к OY;

K — левый край экрана: $K\left(-W/2;\ 0\right)$;

L — правый край экрана: $L\left(W/2;\ 0\right)$;

M — точка пересечения прямых l_1 и l_2:

$$M\left(0;\ (W/2)\operatorname{ctg}\left(\alpha+\beta\right)\right);$$

A — верхняя точка пересечения кругов:

$$A\left(0;\ (D/2)\left(\cos\beta + \sqrt{1-(C-\sin\beta)^2}\right)\right);$$

B — нижняя точка пересечения кругов:

$$B\left(0;\ (D/2)\left(\cos\beta - \sqrt{1-(C-\sin\beta)^2}\right)\right);$$

F — точка пересечения прямой l_2 и правого круга, отличная от L: $F\left(-(D/2)\times\right.$ $\left.\times\left(2\sin(\alpha+\beta)\cos\alpha - C\right);\ D\cos(\alpha+\beta)\cos\alpha\right)$;

E — точка пересечения прямой l_1 и левого круга, отличная от K: $E\left((D/2)\times\right.$ $\left.\times\left(2\sin(\alpha+\beta)\cos\alpha - C\right);\ D\cos(\alpha+\beta)\cos\alpha\right)$, где C — отношение «плоской» ширины экрана к проектному расстоянию наблюдения:

$$C = \frac{W}{D}. \qquad (1)$$

Как будет показано далее, оно зависит от информационной емкости контента I, отношения k высоты экрана к ширине, отношения p ширины знака к его высоте и угла ψ, стягиваемого одним символом.

Криволинейный четырехугольник $AFME$ и есть ОКН, параметры которой позволят оценить количество рабочих мест, которые можно расположить в ней, а также ширину экрана, которая при

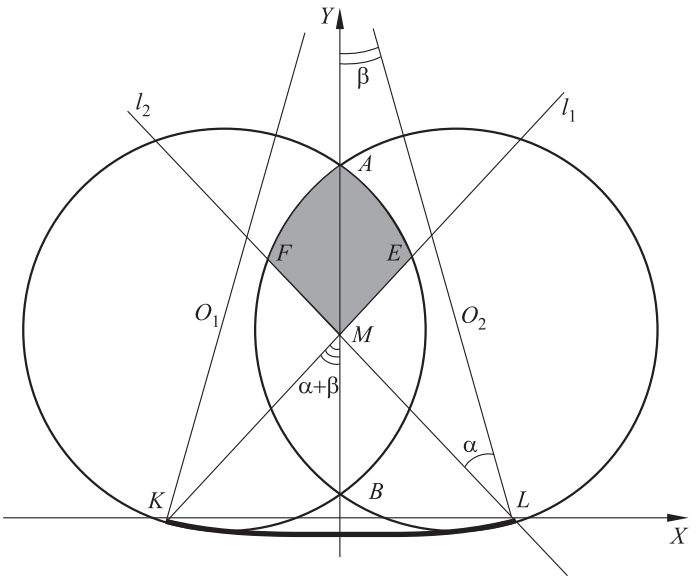


Рис. 1 Область комфортного наблюдения экрана

этом потребуется. Согласно методике, сформулированной в [9, 10], для оценки количества рабочих мест и ширины экрана необходимо рассчитать площадь ОКН и ее периметр, но сначала необходимо определить условия, при которых эта область будет непустой и будет приобретать различные формы. Четырехугольник будет вырожденным, если круги с центрами O_1 и O_2 не будут пересекаться, т. е. точки A и B не будут существовать. Это условие эквивалентно неравенству

$$1 - (C - \sin\beta)^2 < 0\,. \tag{2}$$

Далее полагаем, что это неравенство выполнено.

Рассмотрим взаимное расположение точек A, B и M. Возможны три принципиально разных случая:

1. ***Точка M находится выше точки A.*** Это условие эквивалентно системе неравенств:
$$\left.\begin{aligned} C &> \cos\beta\,\mathrm{tg}\,(\alpha+\beta) = R_1\,; \\ C &> 2\sin(\alpha+\beta)\cos\alpha = R_2\,. \end{aligned}\right\} \tag{3}$$
В этом случае решений нет — ОКН вырождена.
2. ***Точка M принадлежит отрезку*** $[\boldsymbol{A}, \boldsymbol{B}]$. Это условие эквивалентно неравенству:
$$C \le R_2\,. \tag{4}$$
В этом случае ОКН ограничена дугами $\overset{\frown}{AE}$, $\overset{\frown}{AF}$ и отрезками MF, ME (основной случай).
3. ***Точка M находится ниже точки B.*** Это условие эквивалентно системе:
$$\left.\begin{aligned} C &> R_2\,; \\ C &< R_1\,. \end{aligned}\right\} \tag{5}$$
В этом случае ОКН ограничена дугами $\overset{\frown}{AB}$, принадлежащими левому и правому кругу. Решение этой системы существует, когда $R_1 > R_2$ или $2\alpha + \beta > \pi/2$.

В случае 2 площадь и периметр криволинейного четырехугольника могут быть рассчитаны по формулам:

$$\begin{aligned} S_{\text{ОКН}} = \frac{D^2}{4}\Big[\Big(&\cos\beta - C\,\mathrm{ctg}\,(\alpha+\beta) + \\ &+\sqrt{1-(C-\sin\beta)^2}\Big)\,(2\sin(\alpha+\beta)\cos\alpha - C) + \\ &+\{2\alpha+\beta-\arcsin(C-\sin\beta) - \\ &-\sin(2\alpha+\beta-\arcsin(C-\sin\beta))\}\Big]\,; \end{aligned} \tag{6}$$

$$\begin{aligned} P_{\text{ОКН}} = D\Big[&\cos\alpha - \frac{C}{\sin(\alpha+\beta)} + 2\alpha + \beta - \\ &-\arcsin(C-\sin\beta)\Big]\,. \end{aligned} \tag{7}$$

В случае 3 площадь и периметр ОКН могут быть рассчитаны по формулам:

$$\begin{aligned} S_{\text{ОКН}} = \frac{D^2}{4}\Big[&2\arcsin\sqrt{1-(C-\sin\beta)^2} - \\ &-2(C-\sin\beta)\sqrt{1-(C-\sin\beta)^2}\Big]\,; \end{aligned} \tag{8}$$

$$P_{\text{ОКН}} = 2D\arcsin\sqrt{1-(C-\sin\beta)^2}\,. \tag{9}$$

Замечание 1. Данные формулы выполнимы при любом взаимном расположении окружностей, OY и прямых l_1 и l_2 в рамках ограничений, заданных случаями.

Замечание 2. Формулы (2)–(9) действительны и для отрицательных β, т. е. случаев, когда экран выпуклый к наблюдателям.

Замечание 3. В случае 2 при $\beta = 0$ (экран плоский) формулы (6) и (7) приобретают вид результатов, полученных в [9, 10], но в отличие от них являются точными, так как не пренебрегают малыми слагаемыми, которые для неплоского случая могут стать существенными.

3.1 Количество рабочих мест в области комфортного наблюдения

Расчет количества рабочих мест, которые могут быть размещены в области комфортного наблюдения, должен опираться на способ их размещения. Например, для построения ситуационного зала, где в центре будет находиться овальный стол, расчет будет заключаться в определении максимального размера такого стола при заданных ограничениях по характеристикам экрана и информационной емкости контента, отображаемого на нем. В [9, 10] рассмотрен способ равномерно-плотного распределения рабочих мест — «сеточный». Согласно методике, предложенной в [9, 10], количество рабочих мест в области комфортного наблюдения может быть оценено сверху следующей величиной:

$$N = B + \Gamma = \frac{\sqrt{2}}{1{,}8^2}\,S + \frac{P}{3{,}6} + 1\,. \tag{10}$$

Физический смысл единицы, входящей одним из слагаемых в эту формулу, в том, что даже если ОКН вырождена в точку и ее площадь и периметр равны нулю, то все равно в эту точку можно посадить хотя бы одного наблюдателя. Также обратим внимание на то, что используемая в получении этой оценки формула Пика по своей сути смешивает размерности — выводит безразмерную величину из м^2 и м.

Формула (10) дает оценку сверху для случая расположения рабочих мест равномерно плотно —

в вершинах треугольной сетки. Для других случаев расположения рабочих мест необходимо определить другую подходящую функцию

$$N = N(S_{\text{ОКН}}, P_{\text{ОКН}}).$$

Наличие такой функции для других случаев расстановки позволит использовать оценки (6)–(9), предложенные в данной статье.

3.2 Оценка максимальной площади области комфортного наблюдения и максимального количества рабочих мест в этой области

Вследствие ограниченности помещения проектное расстояние D не может превышать диаметра помещения Q, поэтому ввиду оценки

$$C \le \frac{\sqrt{I}}{193} \tag{11}$$

(см. формулу (21) в [9]) для случая 2 формулы (6) и (7) могут быть преобразованы следующим образом:

$$S_{\text{ОКН}} = \frac{Q^2}{4}\left[\left(\cos\beta - \frac{\sqrt{I}}{193}\operatorname{ctg}(\alpha+\beta) + \sqrt{1-\left(\frac{\sqrt{I}}{193}-\sin\beta\right)^2}\right)\times \right.$$
$$\left.\times\left(2\sin(\alpha+\beta)\cos\alpha - \frac{\sqrt{I}}{193}\right) + \{\theta - \sin(\theta)\}\right]; \tag{12}$$

$$P_{\text{ОКН}} = Q\left[\cos\alpha - \frac{\sqrt{I}}{193}\sin(\alpha+\beta) + \theta\right].$$

Здесь

$$\theta = 2\alpha + \beta - \arcsin\left(\frac{\sqrt{I}}{193} - \sin\beta\right).$$

Для случая 3 формулы для площади (8) и периметра (9) приобретут соответственно вид:

$$S_{\text{ОКН}} = \frac{Q^2}{4}\left[2\arcsin\sqrt{1-\left(\frac{\sqrt{I}}{193}-\sin\beta\right)^2} - \right.$$
$$\left. - 2\left(\frac{\sqrt{I}}{193}-\sin\beta\right)\sqrt{1-\left(\frac{\sqrt{I}}{193}-\sin\beta\right)^2}\right];$$

$$P_{\text{ОКН}} = 2Q\arcsin\sqrt{1-\left(\frac{\sqrt{I}}{193}-\sin\beta\right)^2}.$$

Далее количество рабочих мест в обоих случаях может быть посчитано по формуле (10).

Всплески, отображенные на рис. 2, происходят в области, где выполнены условия для случая 1, т. е. ОКН является вырожденной, поэтому эти всплески не представляют интереса для исследований. Сами невырожденные случаи 2 и 3 соответствуют гладким областям графика. Кроме того, график показывает, что применение кривых экранов действительно может быть эффективным для увеличения площади и периметра ОКН, а значит, и количества рабочих мест в ней. Например, для случая $\alpha = 45°$ прирост количества рабочих мест при увеличении угла β с 0° до 15° при информационной емкости контента 9000 знаков может достигать 40%.

3.3 Оценка минимальной ширины активной поверхности дисплея

Очевидно, что чем меньше размеры экрана, тем при прочих равных условиях меньше его стоимость,

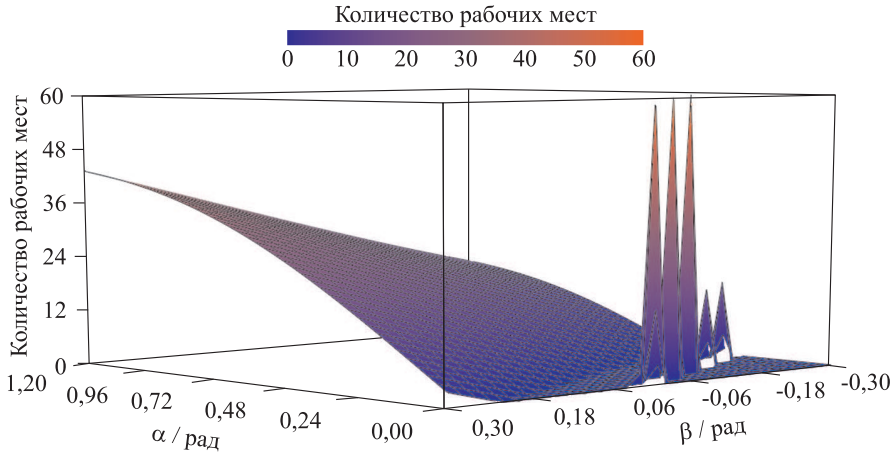


Рис. 2 Зависимость количества рабочих мест от углов α и β при $I = 4000$ и $Q = 10$

но при этом снижаются и функциональные возможности экрана. Поэтому необходимо найти такие минимальные размеры экрана, при которых обеспечивается достаточность решения требуемого перечня функциональных задач.

Пусть известна информационная емкость контента I. Рассмотрим два принципиально разных случая:

(1) число наблюдателей неизвестно, необходимо оценить размеры экрана (его ширину), позволяющие эффективно использовать пространство помещения;

(2) число наблюдателей известно, требуется оценить минимальные размеры коллективного экрана, достаточные для одновременной работы всех наблюдателей с выполнением эргономических требований.

Случай 1. Из соотношений (1) и (11) следует, что

$$W_{\min} = QC \geq QC_{\min} = \frac{Q\sqrt{I}}{193}\,.$$

Случай 2. Согласно методике, предложенной в [9, 10], и формулам (11) и (12) для системы неравенств (4) получаем:

$$W_{\min} = \; = 2C_{\min}\sqrt{S_{\min}}\left(\left(\cos\beta - \right.\right.$$

$$\left. - \left(\frac{\sqrt{I}}{193}\right)\,\mathrm{ctg}\,(\alpha+\beta) + \sqrt{1-\left(\frac{\sqrt{I}}{193} - \sin\beta\right)^2}\right) \times$$

$$\left. \times \left(2\sin(\alpha+\beta)\cos\alpha - \frac{\sqrt{I}}{193}\right) + \{\theta - \sin(\theta)\}\right)^{-1/2} =$$

$$= \frac{2}{193}\sqrt{1{,}94I(N-2)}\left(\left(\cos\beta - \left(\frac{\sqrt{I}}{193}\right) \times\right.\right.$$

$$\times\ \mathrm{ctg}\,(\alpha+\beta) + \sqrt{1-\left(\frac{\sqrt{I}}{193} - \sin\beta\right)^2}\Bigg) \times$$

$$\left. \times \left(2\sin(\alpha+\beta)\cos\alpha - \frac{\sqrt{I}}{193}\right) + \{\theta - \sin(\theta)\}\right)^{-1/2}.$$

Для системы неравенств (5) минимальная ширина экрана может быть посчитана по формуле:

$$W_{\min} =$$

$$= \frac{2}{193}\Big(1{,}94I(N-2)\Big/\Big(2\arcsin\sqrt{1-(C-\sin\beta)^2} -$$

$$- 2(C-\sin\beta)\sqrt{1-(C-\sin\beta)^2}\Big)\Big)^{1/2}.$$

Найденная по этим формулам ширина — это расстояние между крайними точками экрана. Таким образом, действительная ширина экрана с учетом его кривизны может быть рассчитана умножением на коэффициент $\beta/\sin\beta$:

$$W_{\text{экр}} = W_{\min}\frac{\beta}{\sin\beta}\,.$$

Здесь при $\beta = 0$ используем предел

$$\lim_{\beta\to 0}\left(\frac{\beta}{\sin\beta}\right) = 1\,,$$

т. е. функция ширины экрана непрерывна и в окрестности точки $\beta = 0$.

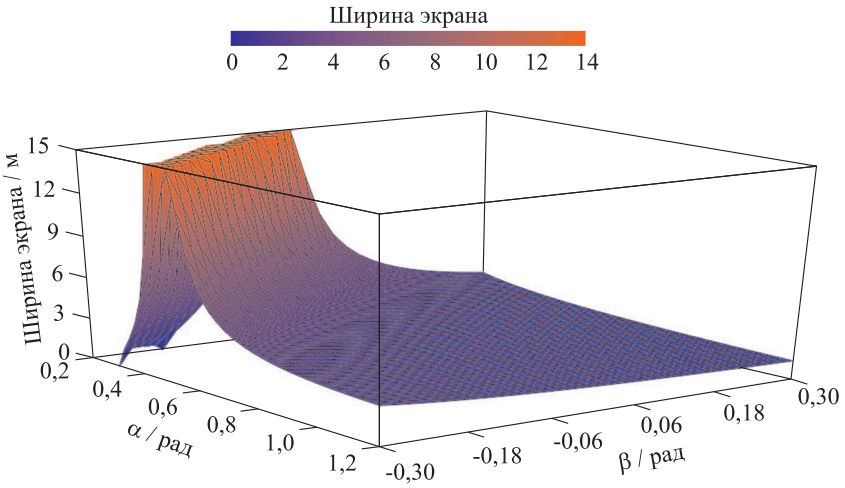


Рис. 3 График «плоской» ширины экрана при $I = 4000$, $N = 10$ от углов α и β

На рис. 3 показан график зависимости ширины экрана от углов α и β. Основные потребности в большой ширине экрана начинаются в тот момент, когда экран становится выпуклым в сторону наблюдателей, а в остальном для 10 наблюдателей понадобится экран шириной около 1,5–2 м в зависимости от угла наблюдения α.

4 Заключение

Полученные в статье результаты являются развитием методических подходов, предложенных в [9, 10], применительно к изогнутому экрану за счет введения нового параметра — угла кривизны β, который для случая полиэкрана представляет собой угол между двумя его дисплеями. Важно, что случай $\beta = 0$ полностью отвечает результатам, полученным в литературе для плоских экранов.

Предложен общий подход к расчету максимального количества рабочих мест на основании формулы , формализация которой для разных расположений может позволить использовать полученные в данной статье готовые формулы.

Показано, что использование изогнутых экранов в качестве коллективных может увеличить количество рабочих мест с соблюдением эргономических требований за счет увеличения площади и периметра ОКН.

Использование изогнутых экранов может быть особенно эффективным в условиях жестких требований к ситуационному залу (например, маленький допустимый угол наблюдения, большая информационная емкость контента или большое количество наблюдателей). Именно в таких случаях эффект от применения изогнутых экранов (увеличение числа рабочих мест) может оказаться наибольшим за счет расположения дисплеев полиэкрана под определенным углом друг к другу.

Полученные зависимости позволяют оценить максимальную дополнительную стоимость изогнутого экрана относительно плоского такой же ширины при реализации проектов с коллективным экраном.

Литература

1. *Зацаринный А. А., Сучков А. В., Босов А. В.* Ситуационные центры в современных информационно-телекоммуникационных системах специального назначения // Ведомственные корпоративные сети и системы (ВКСС Connect!), 2007. № 5(44). С. 64–76.
2. *Зацаринный А. А. Шабанов А. П.* Исследование и разработка методического обеспечения и технологических решений по управлению производительностью контрольно-технологических трактов // Информационные технологии в науке, социологии, экономике и бизнесе (IT + S&E'10): Мат-лы XXXVII Междунар. конф. // Открытое образование, 2010. № 6. Приложение. С. 44–45.
3. *Зацаринный А. А., Шабанов А. П.* Ситуационные центры: информация–процессы–организация // Электросвязь, 2011. № 6. С. 42–46.
4. *Зацаринный А. А., Козлов С. В., Сучков А. П.* Особенности проектирования и функционирования ситуационных центров // Системы высокой доступности, 2012. Т. 8. № 1. С. 12–21.
5. *Зацаринный А. А.* Организационные принципы системного подхода к разработке, проектированию и внедрению современных информационно-телекоммуникационных сетей // Ведомственные корпоративные сети и системы (ВКСС Connect!), 2007. № 1(40). С. 60–67.
6. *Зацаринный А. А., Шабанов А. П.* Эффективность ситуационных центров и человеческий фактор // Вестник Московского ун-та имени С. Ю. Витте. Сер. 1: Экономика и управление, 2013. № 3. С. 43–53.
7. *Зацаринный А. А., Ионенков Ю. С., Шабанов А. П.* Методический подход к оценке эффективности ситуационных центров // Фундаментальные и прикладные исследования, разработка и применение высоких технологий в промышленности и экономике: Сб. статей 15-й Междунар. науч.-практич. конф. — СПб.: СПбГТУ, 2013. Т. 2. С. 37–39.
8. *Зацаринный А. А., Шабанов А. П.* Системные аспекты эффективности ситуационных центров // Вестник Московского ун-та имени С. Ю. Витте. Сер. 1: Экономика и управление, 2013. № 2. С. 110–123.
9. *Чупраков К. Г.* Исследование и разработка методов построения систем отображения информации для ситуационного центра. Дисс. . . . канд. техн. наук. — М.: ИПИ РАН, 2010. 214 с.
10. *Чупраков К. Г.* К вопросу о размещении коллективных средств отображения в ситуационном зале с заданными параметрами // Информатика и её применения, 2010. Т. 4. Вып. 4. С. 89–96.
11. *Золотов Е.* Кривое ТВ: кому выгодно гнуть телевизор. http://www.computerra.ru/100617/curved-tv.
12. *Новикова Е. В., Переверзев Б. Л., Лавренюк С. Ю.* Метод расчета зоны оптимальной видимости при работе с экранами коллективного пользования // Информационные технологии в проектировании и производстве, 2011. № 3. С. 104–109.
13. ГОСТ 21958-76. Зал и кабины операторов. Взаимное расположение рабочих мест. — М.: Изд-во стандартов, 1976. 7 с.
14. ГОСТ 12.2.032-78. Система стандартов безопасности труда. Рабочее место при выполнении работ сидя. Общие эргономические требования. — М.: Изд-во стандартов, 2001. 9 с.
15. ГОСТ 12.2.033-78. Система стандартов безопасности труда. Рабочее место при выполнении работ стоя. Об-

щие эргономические требования. —М.: Изд-во стандартов, 2001. 9 с.
16. ГОСТ 26387-84. Система «Человек–машина». Термины и определения. — М.: Стандартинформ, 2006. 6 с.
17. ГОСТ 27833-88. Средства отображения информации. Термины и определения. — М.: Изд-во стандартов, 1988. 11 с.
18. ГОСТ Р ИСО 9241-3-2003. Эргономические требования при выполнении офисных работ с использованием видеодисплейных терминалов. — М.: Изд-во стандартов, 2003. 39 с.
19. ГОСТ Р 52324-2005. (ИСО 13406-2:2001). Эргономические требования к работе с визуальными дисплеями, основанными на плоских панелях. — М.: Стандартинформ, 2005. 13 с.

Поступила в редакцию 12.10.14

REGARDING ERGONOMIC DEPENDENCES BETWEEN SITUATIONAL HALL PARAMETERS USING COLLECTIVE CURVED SCREEN

A. A. Zatsarinnyy and K. G. Chuprakov

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

Abstract: The paper presents an approach to determining dependences between such parameters of a situational hall as measurements of the hall, quantity of people working with the screen, information capacity of the content (the quantity of symbols), and screen width. These dependences make it possible to calculate an unknown parameter of a situational hall using known parameters satisfying requirements of the Russian and International ergonomic standards. The presented formulas are applicable to the case of curved screens by using the angle of curvature β (for a flat screen, $\beta = 0$). This parameter may be interpreted as an angle between displays in a polyscreen or a videowall. This parameter makes it possible to evaluate the efficiency of curved screens as a collective screen compared to the flat screens. The paper also suggests an approach to estimating the quantity of workplaces that may be used for their different interpositions.

Keywords: collective curved screen; situational hall; dispatch room; ergonomic dependences; comfort observation area; curve angle; videowall, polyscreen; efficiency; price justification

DOI: 10.14357/19922264140411

References

1. Zatsarinnyj, A. A., A. V. Suchkov, and A. V. Bosov. 2007. Situatsionnye tsentry v sovremennykh informatsionno-telekommunikatsionnykh sistemakh spetsial'nogo naznacheniya [Situational centers in modern information-telecommunicational network of special purposes]. *VKSS Connect! (Vedomstvennye korporativnye seti i sistemy)* [VKSS Connect! (Departmental Corporate Networks and Systems] 5(44):64–76.
2. Zatsarinnyj, A. A., and A. P. Shabanov 2010. Issledovanie i razrabotka metodicheskogo obespecheniya i tekhnologicheskikh resheniy po upravleniyu proizvoditel'nost'yu kontrol'no-tekhnologicheskikh traktov [Investigation and development of methodical base and technologic solutions concerning the management of control and technologic tract performance]. *Prilozhenie k zhurnalu "Otkrytoe obrazovanie." Mat-ly XXXVII Mezhdunar. konf. i diskussionnogo nauchnogo kluba "Informatsionnye Tekhnologii v Nauke, Sotsiologii, Ekonomike i Biznese" IT + SE'10* [Appendix to magazine "Open Education." 37th Conference (International) and Discussion Club "Informational Technologies in Science, Education, Telecommunications, and Business" Proceedings]. Yalta. 44–45.
3. Zatsarinnyj, A. A., and A. P. Shabanov. 2011. Situatsionnye centry: Informatsiya–protsessy–organizatsiya. [Situational centers: Information–procecces–organization]. Electrosvyaz' [Telecommunications] 6:42–46.
4. Zatsarinnyj, A. A., S. V. Kozlov, and A. P. Suchkov. 2012. Osobennosti proektirovaniya i funktsionirovaniya situatsionnykh tsentrov [Peculiarity of situational centers design and operation]. *Sistemy Vysokoy Dostupnosti* [Systems of High Accessability]. Moscow: Radiotechnika Publ. 8(1):12–21.
5. Zatsarinnyj, A. A. 2007. Organizatsionnye printsipy sistemnogo podkhoda k razrabotke, proektirovaniyu i vnedreniyu sovremennykh informatsionno-telekommunikatsionnykh setey [Organizational principles of systemic approach to development, design, and implementation of modern information-telecommunicational networks]. *VKSS Connect! (Vedomstvennye korporativnye seti i sistemy)* [VKSS Connect! (Departmental Corporate Networks and Systems] 1(40):60–67.

6. Zatsarinnyj, A. A., and A. P. Shabanov 2013. Effektivnost' situatsionnykh tsentrov i chelovecheskiy faktor [Situational centers efficiency and the human factor]. *Vestnik Moskovskogo Un-ta imeni S. Yu. Vitte. Ser. 1: Ekonomika i Upravlenie* [Herald of S. Y. Vitte Moscow University. Ch. 1: Economics and Management] 3:43–53.
7. Zatsarinnyj, A. A., Yu. S. Ionenkov, and A. P. Shabanov. 2013. Metodicheskiy podkhod k otsenke effektivnosti situatsionnykh tsentrov [Methodical approach to situational centers efficiency evaluation]. *Sb. statey 15-y Mezhdunar. nauch.-praktich. konf. "Fundamental'nye i Prikladnye Issledovaniya, Razrabotka i Primenenie Vysokikh Tekhnologiy v Promyshlennosti i Ekonomike"* [15th Scientific and Practical Conference (International) "Fundamental and Applied Investigations, Development and Applications of Fine Technologies in Industry and Economics Proceedings]. Ed. A. P. Kudinova. St. Petersburg, Russia: Polytechnical University Publ. 2(1):37–39.
8. Zatsarinnyj, A. A., and A. P. Shabanov. 2013. Sistemnye aspekty effektivnosti situatsionnykh tsentrov [Systemic peculiarities of situational centers efficiency]. *Vestnik Moskovskogo Un-ta imeni S. Yu. Vitte* [Herald of S. Yu. Vitte Moscow University. Chapter 1: Economics and Management]. 2:110–123.
9. Chuprakov, K. G. 2010. Issledovanie i razrabotka metodov postroeniya sistem otobrazheniya informatsii dlya situatsionnogo tsentra [Investigations and development of methods of visualization systems creation for situational center]. PhD. Thesis. Moscow. 214 p.
10. Chuprakov, K. G. 2010. K voprosu o razmeshchenii kollektivnykh sredstv otobrazheniya v situatsionnom zale s zadannymi parametrami [On collective display facilities placed in a situational hall with prescribed parameters]. *Informatika i ee Primeneniya — Inform. Appl.* 4(4):89–96.
11. Zolotov, E. 2014. Krivoe TV: Komu vygodno gnut' televizor [Curved TV: Who benefit from bending TV]. Available at: http://www.computerra.ru/100617/curved-tv/ (accessed August 27, 2014).
12. Novikova, E. V., B. L. Pereverzev, and S. Yu. Lavrenyuk. 2011. Metod rascheta zony optimal'noy vidimosti pri rabote s ekranami kollektivnogo pol'zovaniya [A calculation method for area of optimal visibility during work with collective screen]. *Informatsionnye Tekhnologii v Proektirovanii i proizvodstve* [Informational Technologies in Design and Industry] 3:104–109.
13. GOST 21958-76. 1976. *Zal i kabiny operatorov. Vzaimnoe raspolozhenie rabochikh mest* [Hall and operator's cabins. Mutual disposition of workplaces]. Moscow: Standard Pubs. 7 p.
14. GOST 12.2.032-78. 2001. *Sistema standartov bezopasnosti truda. Rabochee mesto pri vypolnenii rabot sidya. Obshchie ergonomicheskie trebovaniya* [A system of work safety standards. Workplace for sitting operations. Main ergonomic requirements]. Moscow: Standard Publs. 9 p.
15. GOST 12.2.033-78. 2001. *Sistema standartov bezopasnosti truda. Rabochee mesto pri vypolnenii rabot stoya. Obshchie ergonomicheskie trebovaniya* [A system of work safety standards. Workplace for standing operations. Main ergonomic requirements]. Moscow: Standard Publs. 9 p.
16. GOST 26387-84. 2006. *Sistema "Chelovek–mashina." Terminy i opredeleniya* [A system "human-machine." Terms and definitions]. Moscow: StandardInform Publ. 6 p.
17. GOST 27833-88. 1990. *Sredstva otobrazheniya informatsii. Terminy i opredeleniya* [Information visualization means. Terms and definitions]. Moscow: Standards Publs. 11 p.
18. GOST R ISO 9241-3-2003. 2003. *Ergonomicheskie trebovaniya pri vypolnenii ofisnykh rabot s ispol'zovaniem videodispleynykh terminalov* [Ergonomic requirement during office works operations using videodisplay terminals]. Moscow: Standard Publs. 39 p.
19. GOST R 52324-2005 (ISO 13406-2:2001). 2005. *Ergonomicheskie trebovaniya k rabote s vizual'nymi displeyami, osnovannymi na ploskikh panelyakh* [Ergonomic requirements for work with visual displays based on flat panels]. Moscow: Standardinform Publ. 13 p.

Received October 12, 2014

Contributors

Zatsarinnyy Alexander A. (b. 1951) — Doctor of Science in technology, professor, Deputy Director, Institute of Informatics Problems, Russian Academy of Sciences, 44-2 Vavilov Str., Moscow 119333, Russian Federation; azatsarinny@ipiran.ru

Chuprakov Konstantin G. (b. 1985) — Candidate of Sciences (PhD) in technology, leading mathematician, Institute of Informatics Problems, Russian Academy of Sciences, 44-2 Vavilov Str., Moscow 119333, Russian Federation; chkos@rambler.ru

ИНФОРМАТИКА И ЕЁ ПРИМЕНЕНИЯ, 2014. Т. 8. Вып. 4. С. 94–109

МЕТОДЫ РАЗРЕШЕНИЯ СУЩНОСТЕЙ И СЛИЯНИЯ ДАННЫХ В ETL-ПРОЦЕССЕ И ИХ РЕАЛИЗАЦИЯ В СРЕДЕ HADOOP*

А. Е. Вовченко[1], Л. А. Калиниченко[2], Д. Ю. Ковалев[3]

Аннотация: При интеграции данных из совокупности исходных коллекций важной задачей является извлечение сущностей, их трансформация и загрузка в интегрированное хранилище. Такие действия являются частью ETL-процесса (extract–transform–loading). Под сущностью здесь понимается некоторое цифровое представление объекта реального мира (например, информация о персонах). При извлечении сущностей возникает проблема их разрешения: из различных ресурсов можно извлечь различную информацию об одном и том же объекте реального мира. Проблема разрешения сущностей ориентирована на решение таких задач, как идентификация сущностей, выявление дубликатов, удаление дубликатов, установление связей между сущностями, сопоставление сущностей с некоторым шаблонным образцом и др. После разрешения сущностей следует этап их слияния — формирование интегрированных сущностей (содержащих информацию из всех связанных сущностей). Слияние сущностей является заключительным этапом интеграции данных. В работе дан обзор методов разрешения и слияния сущностей. Рассматриваются вопросы адаптации таких методов для применения в ETL-процессе при интеграции больших данных в Hadoop. Также рассматриваются способы программирования методов разрешения и слияния сущностей как частей ETL-процесса. В качестве языка программирования используется HIL (High-Level Integration Language) — декларативный язык, ориентированный на разрешение и интеграцию сущностей в Hadoop-инфраструктуре.

Ключевые слова: интеграция данных; ETL; разрешение сущностей; слияние сущностей; большие данные; Hadoop; Jaql; HIL

DOI: 10.14357/19922264140412

1 Введение

В течение нескольких последних лет информатика стала играть все возрастающую роль в широком наборе научных дисциплин, особенно заметную из-за существенных проблем, вызванных взрывоподобным ростом данных в таких науках. X-информатика образовалась как совокупность академических дисциплин, направленных на применение средств информатики для федерализации, организации и анализа данных в конкретных областях науки с интенсивным использованием данных (НИИД): X = астро-, био-, гео-, нейро- и пр.

Сложность использования данных в НИИД усугубляется еще и вследствие естественной разнородности моделей обрабатываемых данных, в которых представляют тексты, графы, структурированную и слабоструктурированную информацию и пр. Разнообразие обрабатываемой информации вызывается, в частности, не только большим числом источников поступления обрабатываемой информации, но и разнообразием объектов исследования, непрерывным и быстрым совершенствованием инструментов, вызывающим адекватные изменения структуры и содержания накапливаемой информации. Это приводит к необходимости использования неоднородной, распределенной информации, накопленной в течение значительного периода наблюдений технологически различными инструментами.

Для анализа больших объемов накапливаемых данных используются современные распределенные инфраструктуры обработки массивных данных (например, Hadoop [1, 2]). Основной особенностью подобных инфраструктур является почти линейная горизонтальная масштабируемость (производительность системы растет линейно относительно числа узлов кластера), а также высокая отказоустой-

* Работа выполнена при поддержке РФФИ (проекты 13-07-00579, 14-07-00548), ИПИ РАН (Тема 38.25 «Спецификация и решение задач анализа данных в концептуальных терминах предметных областей с интенсивным использованием данных» государственного задания ФГБУН ИПИ РАН) и Президиума РАН (Программа фундаментальных исследований Президиума РАН № 16 «Фундаментальные проблемы системного программирования»).

[1] Институт проблем информатики Российской академии наук, alexey.vovchenko@gmail.com

[2] Институт проблем информатики Российской академии наук; Московский государственный университет им. М. В. Ломоносова, факультет вычислительной математики и кибернетики, leonidk@synth.ipi.ac.ru

[3] Институт проблем информатики Российской академии наук, dm.kovalev@gmail.com

чивость (отказ любого узла кластера не должен влиять на работоспособность системы в целом).

Главным достоинством подобных инфраструктур является возможность анализировать и обрабатывать разноструктурированные данные, например реляционные, XML, JSON, NoSQL, текстовые и др. При этом возникает проблема интеграции информации, извлекаемой из таких разноструктурированных данных.

Процесс интеграции данных (рассматриваемый здесь как ETL-процесс) можно представить состоящим из следующих этапов:

- сопоставление схем;
- интеграция схем;
- трансформация данных;
- разрешение сущностей (Entity Resolution [3–5]);
- слияние сущностей (Data Fusion [6]).

В данной работе рассматриваются последние два этапа. При интеграции сырых разноструктурированных данных задачей ETL-процесса является извлечение сущностей из исходных коллекций, их разрешение, трансформация и загрузка в интегрированное хранилище. Под сущностью здесь понимается некоторое цифровое представление объекта реального мира (например, информация о персонах). При извлечении сущностей возникает проблема их разрешения: из разных ресурсов можно извлечь разную информацию об одном и том же объекте реального мира. В общем случае под термином разрешения сущностей (entity resolution (ER) [3–5, 7–10]) понимают извлечение информации об одной и той же сущности реального мира из разнообразных структурированных, слабоструктурированных и неструктурированных коллекций данных и приведение извлеченных данных к унифицированному представлению. При этом применяются методы извлечения, сопоставления, группирования, связывания устранения дублирования различных представлений информации. Подходы к разрешению сущностей рассматриваются в разд. 2. Слияние сущностей является заключительным этапом интеграции данных. Под слиянием сущностей [6, 11–13] понимается образование интегрированного представления информации об одной и той же же сущности реального мира, полученной из разных источников. Операции и процедуры, используемые при слиянии сущностей, приведены в разд. 3.

В разд. 4 приводится описание средств программирования в среде Hadoop, а также обоснование выбора в работе средств (языки Jaql и HIL). Наконец, в разд. 5 показаны примеры программирования методов разрешения сущностей и слияния данных как части ETL-процесса в Hadoop.

2 Методы разрешения сущностей

Этап разрешения сущностей важен для сохранения исходной информации в интегрированной коллекции. Кроме того, важно обнаруживать дубликаты сущностей, поскольку, например, увеличение числа узлов и ребер в сетевых задачах может существенно удлинять время работы простейших алгоритмов (например, поиска кратчайшего пути). Алгоритмы разрешения сущностей (включая поиск дубликатов — duplicate detection) часто используются и во всевозможных поисковых системах, таких как Google, Amazon и др. Аналогичная проблема встает остро в различных агрегаторах информации (например, новостных агрегаторах).

В общем случае процесс разрешения сущностей включает следующие этапы [10]:

- подготовка данных;
- выбор методов сопоставления значений;
- определение методов разрешения пар сущностей;
- определение зависимостей (constraints).

2.1 Методы сопоставления значений

Важным действием для успешного разрешения сущностей является подготовка данных, которая включает нормализацию схем и нормализацию данных. Нормализация схем — непростая задача, подробное ее рассмотрение выходит за рамки данной работы и содержится в работах по интеграции данных в традиционных архитектурах или в инфраструктурах больших данных. Ниже представлен пример списка действий, которые могут быть отнесены к нормализации схем:

- сопоставление атрибутов схем (например, «контактный телефон» и «мобильный телефон»);
- слияние атрибутов (например, «полный адрес» образуется из атрибутов «город», «индекс», «улица», . . .);
- слияние множественных значений и списков (например, «контактные телефоны» и «основной номер телефона», «дополнительный номер телефона») и др.

Нормализация данных может включать приведение к строчному или заглавному регистру; удаление разделителей; поиск и исправление опечаток; поиск сокращений и аббревиатур и замена их на

полные стандартные формы; использование словарей для нормализации строк и многое другое. Нормализация данных, так же как и нормализация схем, не рассматривается в данной статье.

Для сопоставления сущностей важно определиться с выбором метода оценки сходства (similarity) значений. Рассматриваются как булевы, так и вещественные меры сходства. Следующий список включает примеры часто используемых методов оценки сходства значений:

- эквивалентность булевых предикатов;
- вычисление функции сходства простых значений (расстояние Левенштейна [14], алгоритм Смита–Ватермана [14]);
- вычисление функции сходства множеств (коэффициент Жаккара [14], коэффициент сходства Дайса [14], коэффициент Адара [15]);
- вычисление функции сходства векторов (коэффициент косинусов [14], статистическая мера TFIDF (term frequency – inverse document frequency) [14]);
- оценка сходства на основе выравнивания (сходство Джаро–Винклера [14], статистическая мера Soft-TFIDF [16], расстояние Монг–Элкана [17]);
- оценка сходства фонетических данных (алгоритм сравнения двух строк по их звучанию Soundex [14]);
- оценка сходства, основанная на переводе (может использоваться для нормализации аббревиатур);
- оценка сходства, основанная на знаниях о предметной области.

Также существуют специальные методы для определения сходства отношений. Меры, используемые для отношений, обычно основаны на сходстве множеств и предполагают использование функций вычисления сходства множеств.

2.2 Методы сопоставления пар сущностей

2.2.1 Традиционные методы сопоставления пар сущностей

Пусть даны две коллекции объектов с атрибутами: author, venue, paper. Значение некоторой меры сходства (одной из приведенных в подразд. 2.1) для конкретного атрибута будем обозначать

$$\text{XX-match-score}(\text{author-match-score}, \text{venue-match-score}, \text{paper-match-score}) .$$

Традиционным методом сравнения объектов является подсчет сходства некоторым алгоритмом (см. подразд. 2.1) для каждого из атрибутов независимо. Затем реализуется подсчет взвешенной суммы.

Например:

$$0{,}5\text{autor-match-score} + 0{,}2\text{venue-match-score} + 0{,}3\text{paper-match-score} .$$

Недостатком такого подхода является сложность выбора весов для каждого из атрибутов и сложность выбора порога сходства сущностей.

Другой метод предполагает задание булевого обобщенного правила, где условия накладываются на каждый атрибут независимо.

Например:

$$(\text{author-match-score} > 0{,}7 \text{ AND venue-match-score} > 0{,}8) \text{ OR } (\text{paper-match-score} > 0{,}9 \text{ AND venue-match-score} > 0{,}9) .$$

Недостатком этого подхода является сложность формулирования подобных правил вручную.

2.2.2 Методы машинного обучения для сопоставления пар сущностей

Для сопоставления пар сущностей применяют также специальные методы машинного обучения, которые позволяют автоматизировать процесс формулирования критериев для сопоставления сущностей. Использование таких методов основано на применении теории Феллеги и Сантера [14] для связывания сущностей. Рассмотрим этот подход подробнее.

Пусть даны коллекции A и B.

Пусть r — это пара $r(x, y)$, где $x \in A$, $y \in B$.

Пусть $\gamma = \gamma(r)$ — это вектор сравнения, например:

$$\gamma(r) = \{x.\text{author} = y.\text{author}, x.\text{venue} = y.\text{venue}, x.\text{paper} = y.\text{paper}\} ,$$

$\gamma(r) = \{\text{true}, \text{false}, \text{true}\}$ — пример, в случае если $x.\text{author} = y.\text{author}$, $x.\text{venue} \neq y.\text{venue}$, $x.\text{paper} = y.\text{paper}$.

Пусть M — множество всех пар, являющихся дубликатами.

Пусть U — множество всех пар, не являющихся дубликатами.

Тогда правило для определения сходства сущностей можно описать следующей формулой:

$$R(r) = \frac{m(\gamma)}{u(\gamma)} = \frac{P(\gamma | r \in M)}{P(\gamma | r \in U)} .$$

Определим два порога T_I и t_U, такие что

- $R(r) \leq t_I$ — объекты (пара) не являются дубликатами;

– $R(r) > t_I$ AND $R(r) < t_u$ — невозможно определить, являются ли объекты (пара) дубликатами или нет;
– $R(r) \geq t_u$ — объекты (пара) являются дубликатами.

Правилом связывания, обозначаемым $L(t_I, t_U)$, называется пара порогов t_I и t_U.

При подобном подходе учитываются стандартные для задачи проверки статистических гипотез ошибки первого и второго рода. Ошибки первого рода — два объекта не являются дубликатами ($r(x, y) \in U$), однако правило L относит их к дубликатам. Ошибки первого рода обозначаются буквой μ и их можно описать формулой:

$$\mu = P(L_{\text{match}}|U) = \sum_{\gamma} u(\gamma) P\left(L_{\text{match}}|\gamma\right) .$$

Ошибки второго рода — два объекта являются дубликатами ($r(x, y) \in M$), однако правило L определяет, что это не дубликаты. Ошибки второго рода означаются буквой λ и их можно описать формулой:

$$\lambda = P(L_{\text{nonmatch}}|M) = \sum_{\gamma} m(\gamma) P\left(L_{\text{nonmatch}}|\gamma\right).$$

Оптимальным правилом связывания $L^*(t_I^*, t_U^*)$ называется такое правило, которое соответствует ограничениям на ошибки первого и второго рода для правила, а также ограничения на неопределенности. Эти ограничения выражаются следующими формулами:

– ограничения на ошибки:

$$P\left(L^*_{\text{match}}|U\right) \leq \mu\,; \qquad P\left(L^*_{\text{nonmatch}}|M\right) \leq \lambda\,;$$

– ограничения на неопределенности:

$$P\left(L^*_{\text{uncertain}}|U\right) \leq P\left(L_{\text{uncertain}}|U\right)\,;$$
$$P\left(L^*_{\text{uncertain}}|M\right) \leq P\left(L_{\text{uncertain}}|M\right)\,.$$

Нахождение оптимального правила является основной задачей при использовании теории Феллеги и Сантера. Классические (переборные) методы при таком подходе работают неэффективно, поэтому нахождение оптимального правила достигается с помощью средств машинного обучения, например можно использовать наивный байесовский классификатор. Одна из основных проблем при этом заключается в том, что для вычисления $P(\gamma|r \in M)$ и $P(\gamma|r \in U)$ необходимы знания о том, какие объекты являются дубликатами, а какие нет (знания о множествах M и U).

Для разрешения сущностей применяются различные реализации подходов, основанных на алгоритмах машинного обучения и использовании теории Феллеги и Сантера, например использование:

– деревьев решений [18];
– метода опорных векторов [19, 20];
– ансамблей классификаторов [21];
– метода условных случайных полей [22].

К недостаткам этих подходов можно отнести несбалансированность результирующих классифицированных множеств (так, в результате образуется значительно больше непохожих объектов, чем похожих), а также высокую вероятность того, что объект не будет причислен ни к какому классу (из-за неопределенности). Но оба этих недостатка могут быть устранены путем тонкой настройки алгоритмов. Ключевой проблемой при использовании методов машинного обучения при сравнении пар сущностей является выбор обучающего множества.

Выделяют следующие методы классификации сущностей, не требующие построения обучающей выборки:

– обучение без учителя или с частичным привлечением учителя [14, 23];
– методы с активным обучением;
– ансамбли классификаторов [24, 25];
– доказуемая оптимизация точности/полноты [26, 27];
– краудсорсинг [28, 29].

Подводя итог методам разрешения пар сущностей, выделим методы, основанные на мерах сходства, и методы, основанные на использовании машинного обучения. Общим недостатком первой группы методов является сложность формулирования критериев сходства (подбор весов или явных формул). Методы машинного обучения лишены этого недостатка в силу своей структуры, но при этом ключевой проблемой является выбор обучающего множества, да и сами методы значительно сложнее. Перспективными (но все еще мало изученными) представляются методы машинного обучения, не требующие изначального определения обучающего множества, такие как методы, основанные на активном обучении и краудсорсинге.

2.3 Использование ограничений

После того как определен метод разрешения конкретных пар сущностей, можно определить зависимости. Далее представлены примеры зависимостей, используемых для установления сходства сущностей:

– транзитивность: если M1 и M2 похожи и M2 и M3 похожи, то и M1 и M3 похожи;

– эксклюзивность: если M1 и M2 похожи, то M3 не может быть похож на M2;
– функциональные зависимости: если M1 и M2 похожи, то M3 и M4 должны быть похожи.

Транзитивность часто используется в методах удаления дубликатов, а эксклюзивность используется в методах установления связей между сущностями.

В заключение можно отметить, что разрешение сущностей является быстро развиваемой областью. Исследуются новые меры сходства [14], ведутся работы по применению перспективных методов машинного обучения [24–29]. Развивается применение функциональных зависимостей при очистке данных (data cleaning) [30–32]. Ведутся работы по построению сущностей с наиболее представительными данными (включающими данные из разнообразных дубликатов — методы канонизации сущностей [33]). Также исследуются методы, в которых решения по сходству двух сущностей принимаются на основе анализа совокупности сущностей, применения вероятностных логик сходства, латентной модели Дирихле [34–36].

3 Методы слияния данных

Под слиянием данных [6, 12, 13] понимается образование интегрированного представления информации об одной и той же сущности реального мира, полученной из разных источников данных. Процесс слияния данных включает следующие задачи: слияние записей о сущностях, разрешение возможных конфликтов, обнаружение и удаление ошибочных данных. Методы слияния данных, кратко рассмотренные в данном разделе, исследованы в Потсдамском университете [13]. Различные аспекты проблемы слияния данных представлены на рис. 1.

3.1 Типы конфликтов при слиянии данных

Различают два типа конфликтов: конфликты, вызванные неопределенными значениями, и конфликты, вызванные противоречивыми значениями.

Неопределенность означает, что в одном источнике данных содержатся неизвестные значения (null), а в другом — известные. Проблема заключается в том, что семантика неопределенных значений (null) может сильно отличаться. Различают три варианта: неизвестные значения, несуществующие значения (например, атрибут «имя супруга» всегда будет null для неженатых), скрытые значения (такие данные, которые по каким-то причинам не позволено видеть).

Противоречивость значений означает появление двух различных ненулевых (not null) значений. Возможны различные стратегии обработки подобных конфликтов, о чем рассказывается в следующем подразделе.

3.2 Стратегии разрешения конфликтов

Различают следующие подходы к разрешению конфликтов:

– игнорирование конфликтов;
– избегание конфликтов;
– разрешение конфликтов.

Стратегия игнорирования конфликтов предполагает извлечение всей доступной информации.

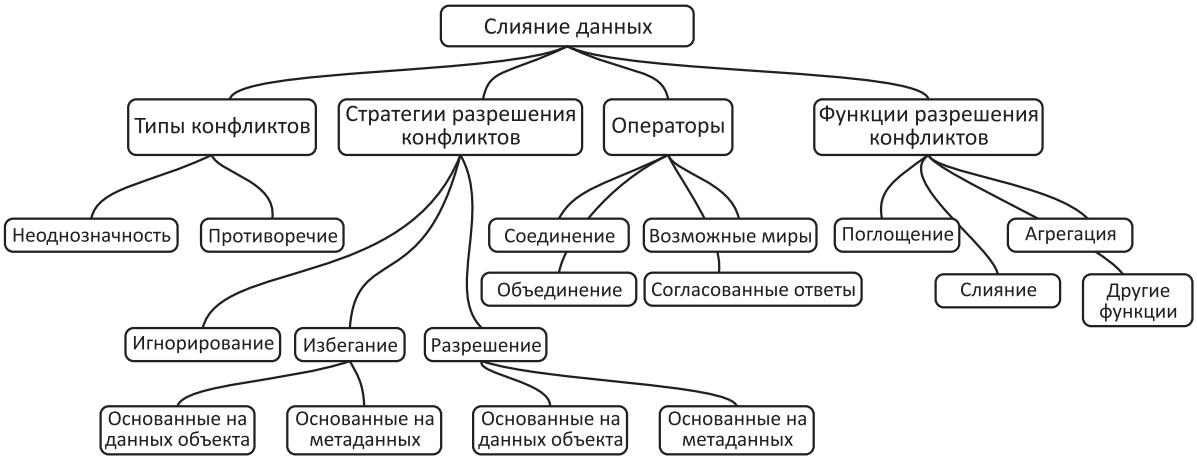


Рис. 1 Различные аспекты проблемы слияния данных

Примеры функций для разрешения конфликтов

Функция	Описание	Стратегия	Пример конфликта
Min, Max, Sum, Count, Avg	Обычная агрегация	Разрешение конфликта	Подсчет зарплаты (средней или максимальной), подсчет возраста или количества детей
Random	Случайное значение	Разрешение конфликта	Размер участка
Longest, Shortest	Самое короткое или длинное значение	Разрешение конфликта	Например, для имен
Choose (source)	Значение из конкретного ресурса	Избежание конфликта	Например, для финансовых данных, если принято решение доверять информации из Yahoo больше, нежели другим ресурсам
Choose Depending (val, col)	Выбирается значение в зависимости от значения в другом атрибуте	Избежание конфликта	Например, если выбран атрибут «город» из одного ресурса, то «почтовый индекс» разумно взять из того же самого ресурса
Vote	Голосование, решение по большинству	Разрешение конфликта	Например, для подсчета рейтинга
Coalesce	Выбор первого ненулевого значения	Избежание конфликта	Например, для имен
Group, Concat	Группировка или конкатенация всех значений	Избежание конфликта	Например, для отзывов о продуктах
MostRecent	Выбор наиболее свежего значения (недавно обновленного)	Разрешение конфликта	Например, если интересует последний адрес местожительства
Escalate	Сохранение всех конфликтующих значений, с тем чтобы пользователь сам решил, какое выбрать	Игнорирование конфликта	Например, для атрибута «пол» сложно придумать объективные причины выбора того или иного значения
...	...	...	...

Например, для строк это может быть обычная конкатенация строк, а пользователь уже сам решает, какие данные верны.

Стратегия избегания конфликтов предполагает выбор данных на основе самих данных (по некоторому алгоритму) или на основе метаданных. Примером функции на основе данных может служить функция coalesce (выбор первого ненулевого значения) или функция выбора самого длинного значения.

Примером функций на основе метаданных может выступать выбор в зависимости от самого источника (например, известно, что один из источников наиболее достоверный). Другим примером является функция, выбирающая значение из того источника, в котором большее число значений было выбрано для других атрибутов.

Стратегии разрешения конфликтов учитывают все значения и выбирают из них «достоверное». Примером подобной функции могут выступать всевозможные функции голосования, функции выбора случайного значения, функции среднего значения, функции наиболее часто встречающегося значения и др.

В таблице представлены примеры функций для разрешения конфликтов.

3.3 Основные функции разрешения конфликтов

Вводится операция outer union [13], результатом которой является объединение двух отношений. Если схемы не совпадают, то результирующая схема является объединением двух исходных схем. Например, пусть даны два отношения: A с набором атрибутов $\{a, b, c, d\}$ и отношение B с набором атрибутов $\{c, d, e, f\}$. Результирующая схема будет содержать набор атрибутов = $\{a, b, c, d, e, f\}$. В результирующие кортежи для недостающих атри-

бутов помещаются нулевые значения. Эта операция не является стандартной и отсутствует в большинстве реляционных систем управления базами данных (СУБД). В реляционной алгебре подобная операция может быть представлена как

```
(SELECT a, b, c, d, NULL as e, NULL as f FROM A)
UNION
(SELECT NULL as a, NULL as b, c, d, e, f FROM B).
```

Вводится функция tuple subsumption [13]. Говорят, что кортеж t1 поглощает другой кортеж t2 (поглощаемый кортеж), если у них

- совпадают схемы;
- в t2 больше неизвестных (null) значений, чем в t1;
- в t2 все известные значения совпадают со значениями в t1.

Например, пусть даны кортежи $\mathrm{t1} = (5, \text{'text'}, \mathrm{null}, 7)$ и $\mathrm{t2} = (5, \mathrm{null}, \mathrm{null}, 7)$. Видно, что каждый атрибут в t2 либо совпадает с аналогичным атрибутом в t1, либо он null. Для этого примера кортеж t1 поглощает кортеж t2.

Вводится функция tuple complementation [13]. Говорят, что кортежи t1 и t2 дополняют друг друга, если

- у них совпадают схемы;
- они не совпадают;
- значения соответствующих атрибутов в t1 и t2 совпадают, либо одно из них не определено, либо оба не определены;
- t1 и t2 имеют как минимум один атрибут, значения которого совпадают.

Например, пусть даны кортежи $\mathrm{t1} = (5, \text{'text'}, \mathrm{null}, \mathrm{null})$ и $\mathrm{t2} = (5, \mathrm{null}, \mathrm{null}, 7)$. Видно, что кортежи дополняют друг друга. Результатом операции дополнения для этих двух кортежей будет новый кортеж $\mathrm{t} = (5, \text{'text'}, \mathrm{null}, 7)$.

3.4 Операторы слияния данных

Различают два основных подхода к слиянию данных. Эти подходы основаны на операции объединения (union-based) или на операции соединения (join-based). Различают следующие основные операции.

Minimum Union [13] (union-based). Операция представляет собой выполнение операции outer union, а затем удаление из результата всех поглощаемых (subsumed [13]) кортежей. Пример операции представлен на рис. 2.

Complemetation Union [13] (union-based). Операция представляет собой выполнение операции outer union, а затем дополнение (complementation) всевозможных кортежей. Пример операции представлен на рис. 3.

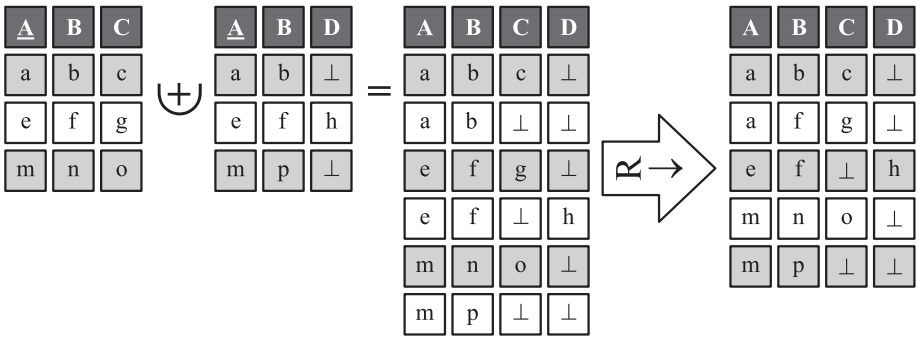


Рис. 2 Пример операции **Minimum Union**

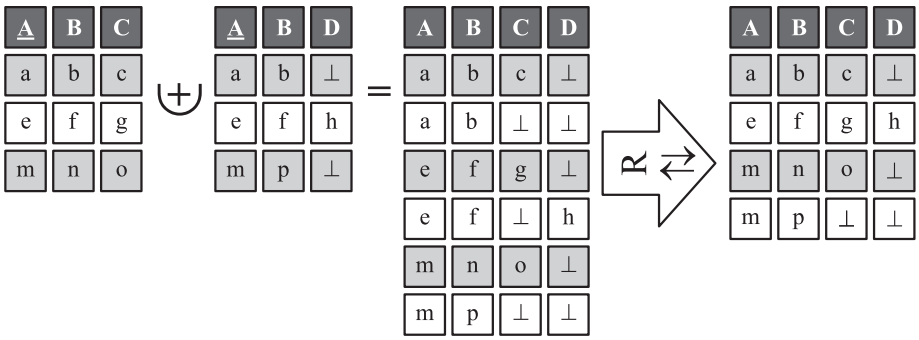


Рис. 3 Пример операции **Complementation Union**

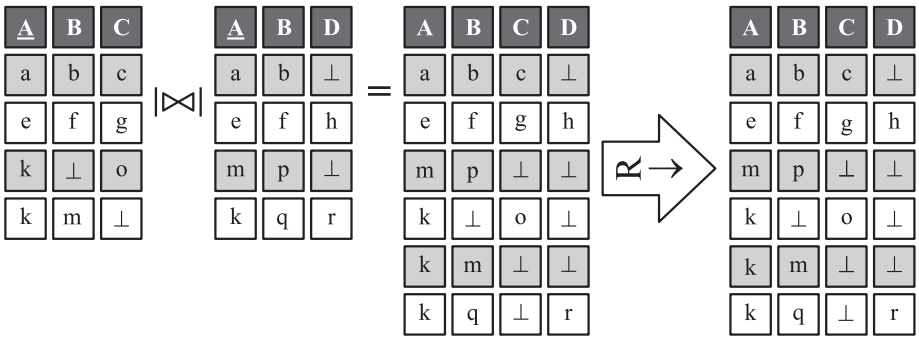


Рис. 4 Пример операции **Full Disjunction**

Grouping and Aggregation [13] (union-based). Операция предполагает выполнение outer union, а затем группировку по общему атрибуту и применение функции агрегации к остальным атрибутам. Пример операции на языке SQL представлен ниже.

```
WITH OU AS (
 (SELECT A, B, C, NULL AS D FROM U1)
   UNION (ALL)
 (SELECT A, B, NULL AS C, D FROM U2)),
SELECT A, MAX(B), MIN(C), SUM(D)
FROM OU
GROUP BY A
```

Full Disjunction [37] (join-based). Операция представляет собой full outer join (стандартную реляционную операцию), после чего применяется subsumption к результату. Пример представлен на рис. 4.

Match Join [13] (union- + join-based). В операции выбираются всевозможные комбинации значений атрибутов, после чего выполняется full outer join. Фактически реализуется outer union двух коллекций. Затем определяется $N-1$ вспомогательных отношений, где N — число атрибутов, а каждое из отношений содержит по два атрибута: один общий и какой-то другой. После чего происходит full outer join $(N-1)$-го отношения. Пример реализации операции на языке SQL представлен ниже.

```
WITH
 OU(A,B,C,D) AS (
 (SELECT A, B, C, NULL AS D FROM U1)
  UNION
 (SELECT A, B, NULL AS C, D FROM U2)),
   // ← Outer Union
 B_V(A,B) AS (SELECT DISTINCT A, B FROM OU),
   // ← 1-е отношение (N = 4)
 C_V(A,C) AS (SELECT DISTINCT A, C FROM OU),
   // ← 2-е отношение (N = 4)
 D_V(A,D) AS (SELECT DISTINCT A, D FROM OU),
   // ← 3-е отношение (N = 4)
SELECT A, B, C, D
FROM B_V FULL OUTER JOIN C_V FULL OUTER
    JOIN D_V // ← Full Outer Join
```

Merge (union- + join-based). Операция объединяет операции соединения и объединения. Для каждого общего атрибута формируются две версии значений, нулевые значения удаляются функцией COALESCE (выбор первого ненулевого значения). Пусть даны два отношения: A с набором атрибутов $\{a, b, c\}$ и B с набором атрибутов $\{a, b, d\}$. Пусть a — конфликтующий атрибут, b — атрибут с нулевыми значениями. Пример реализации операции на языке SQL представлен ниже, а результат показан на рис. 5.

```
(SELECT A.a, COALESCE(A.b, B.b), A.c, B.d
FROM A LEFT OUTER JOIN B ON A.a = B.a)
UNION
(SELECT B.a, COALESCE(B.b, A.b), A.c, B.d
FROM A RIGHT OUTER JOIN B ON A.a = B.a)
```

Grouping and Aggregation (union-based). Операция представляет собой группировку по некоторому атрибуту, а затем использование разнообразных агрегирующих функций. В качестве достоинства данного подхода можно выделить его реализацию в большинстве СУБД и эффективное выполнение. Пример реализации на SQL представлен ниже.

```
WITH OU AS (
 (SELECT A, B, C, NULL AS D FROM U1)
  UNION (ALL)
 (SELECT A, B, NULL AS C, D FROM U2)),
SELECT A, MAX(B), MIN(C), SUM(D)
FROM OU
GROUP BY A
```

Data Fusion оператор [11] — **Fuse By** (union-based). В некоторых системах пошли дальше использования стандартных операций группировки и агрегации. Ключевое слово FUSE BY используется вместо GROUP BY, и семантика у него аналогична. Вместо использования стандартных функций агрегации используется встроенная функция RESOLVE, которой параметром передается само значение и имя

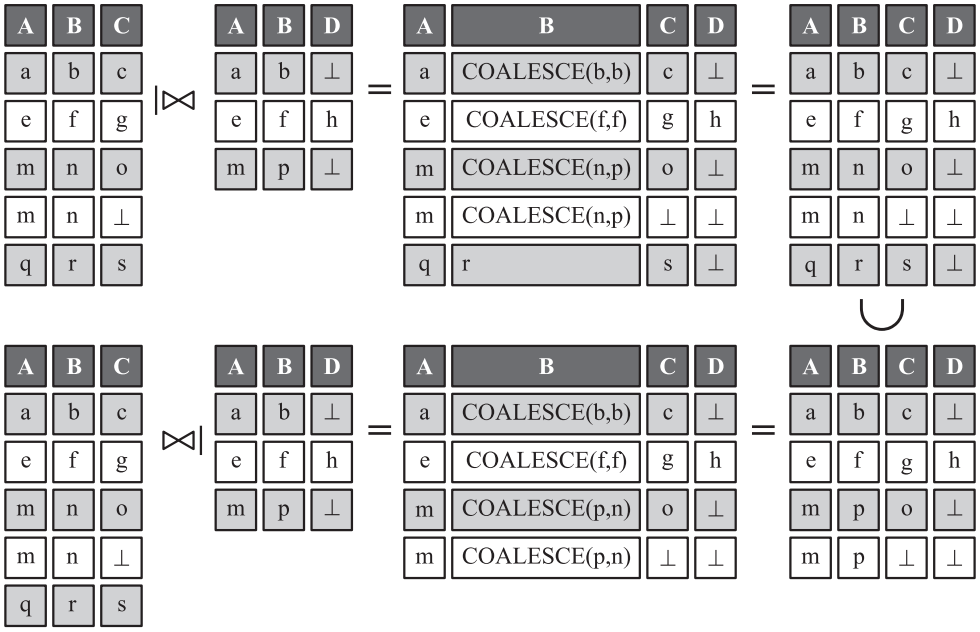


Рис. 5 Пример операции **Merge**

функции разрешения конфликтов. Пример реализации на SQL представлен ниже.

```
SELECT ID,
 RESOLVE(Title, Choose(IMDB)),
 RESOLVE(Year, Max),
 RESOLVE(Director, COALESCE),
 RESOLVE(Rating, COALESCE),
 RESOLVE(Genre, Concat)
FUSE FROM IMDB, Filmdienst
FUSE BY (ID)
ON ORDER Year DESC
```

4 Разрешение сущностей в больших данных

Для манипулирования большими разноструктурированными данными служат Hadoop-инфраструктуры [1, 2], предоставляющие масштабируемое хранилище и обеспечивающие высокую скорость анализа больших данных за счет распределенной их обработки. Для применения методов разрешения сущностей в такой среде нужна *адаптация* алгоритмов для их распределенного выполнения на различных узлах Hadoop-кластера.

В среде Hadoop реализована парадигма распределенного программирования для анализа данных Map-Reduce [38, 39], называемая по именам основных функций. Вначале на всех узлах кластера обрабатываются блоки данных независимо друг от друга (Map). После чего данные группируются по заранее выбранным для алгоритма ключам и поступают на выполнение на один или более узлов в зависимости от алгоритма (Reduce).

Таким образом, для реализации любого алгоритма в Hadoop-инфраструктуре требуется его адаптация к виду Map-Reduce. Другим вариантом является реализация алгоритма на одном из языков высокого уровня, таких как Pig [40], Hive [41], Jaql [42]. Все эти языки автоматически переписывают программы, реализованные на них, в Map-Reduce-приложения для выполнения на Hadoop-кластере.

В случае больших данных и распределенных инфраструктур традиционные подходы требуют доработок. Различают два основных метода разрешения сущностей над большими данными: разбиение данных на блоки (blocking [43, 44]) и распределенный метод разрешения сущностей.

Суть разбиения на блоки заключается в следующем. Пусть имеется 1000 компаний в 1000 городах. Требуется сравнить компании. Алгоритм полного попарного сравнения потребует 10^{12} сравнений. При этом если предположить, что компании из разных городов не могут совпадать, то потребуется 10^9 сравнений. Ключевой проблемой данного подхода является выбор критерия, по которому разбиваются данные.

Различают два основных метода: основанный на хэш-функции [10] и основанный на сходстве соседей [10]. Метод, основанный на хэш-функции, предполагает разбиение на блоки по хэш-ключу.

Основной проблемой алгоритма является выбор хэш-функции. Метод, основанный на сходстве соседей, предполагает, что совпадать могут только объекты, похожие по некоторой мере. Все объекты сортируются по какому-то признаку (ключу — простому или составному, уникальность ключа не требуется). После этого выбирается размер окна, внутри которого объекты сравниваются. Проблемой данного метода является выбор ключа сортировки.

Распределенный метод разрешения сущностей предполагает реализацию традиционных алгоритмов этого семейства в виде Map-Reduce-приложения, что требует зачастую полного пересмотра исходного алгоритма. Другой вариант — реализация алгоритма разрешения сущностей на специализированных языках, чему будет посвящен следующий раздел. Третий вариант — использование специализированных инструментов, направленных на распределенное выполнение методов разрешения сущностей над Hadoop [45].

5 Реализация операций разрешения сущностей и слияния данных в среде Hadoop

Язык HIL [46] — декларативный язык, ориентированный на разрешение и интеграцию сущностей в Hadoop инфраструктуре. HIL компилируется в язык Jaql [43, 44], который, в свою очередь, автоматически переписывается в Map-Reduce, если этого требует алгоритм.

5.1 Реализация методов разрешения сущностей

Пусть даны структуры данных, включающие три атрибута: id, value, name. Тогда простейшее правило разрешения сущностей на языке HIL будет выглядеть следующим образом:

```
declare Duplicated: ?;
declare Generated: ?;
declare Deduplicated: ?;

create link Deduplicated as
select
[gen: [id: g.id, name: g.name, value: g.value],
dup: [id: d.id, name: d.name, value: d.value]]
from Generated g, Duplicated d
match using
 rule_id: g.id = d.id,
 rule_name: g.name = d.name,
 rule_value: g.value = d.value;
```

В этом примере используется простое сопоставление сущностей по совпадению значений. Если требуется ввести какую-то функцию меры для значений, это можно реализовать внешней функцией Jaql:

```
@jaql{
compareValue =
    javaudf("org.ipiran.similarity.ValueSimilarity");
}
```

После этого такую функцию можно вызывать из языка HIL:

```
declare compareValue: function ? to ?;
declare Duplicated: ?;
declare Generated: ?;
declare Deduplicated: ?;

create link Deduplicated as
select
[gen: [id: g.id, name: g.name, value: g.value],
dup: [id: d.id, name: d.name, value: d.value]]
from Generated g, Duplicated d
match using
rule_id:
 compareValue(g.id, d.id) > 0.7,
rule_name:
 compareValue(g.name, d.name) > 0.7,
rule_value:
 compareValue(g.value, d.value) > 0.7;
```

Можно также ввести меру для сравнения объектов целиком.

Пусть описана функция compareObject, которая принимает на вход объекты. Тогда правило на языке HIL изменится, так как в этом случае используется другой вид правил:

```
insert into Deduplicated
select
[gen: [id: g.id, name: g.name, value: g.value],
dup: [id: d.id, name: d.name, value: d.value],
value: compareObject(g,d)]
from Generated g, Duplicated d
where compareObject(g, d) > 0.7;
```

Во всех этих случаях происходит сравнение всех объектов со всеми, сложность подобного сравнения $O(n^2)$. Несмотря на то что сравнения будут выполняться независимо и распределены на всех узлах кластера (так как HIL переписывается в Jaql, а тот, в свою очередь, в Map-Reduce), время их выполнения может быть довольно большим. Для уменьшения числа сравнений, как было описано в разд. 4, можно разбивать данные на блоки.

Пусть имеется функция calcHash, которая вычисляет хэш для объектов. В результате функция может выдавать столько уникальных значений, на сколько блоков требуется разбить данные. Тогда, объединив правила, рассмотренные выше, выбрав

вначале те объекты, что совпадают по хэш-функции, а далее, вычислив общую меру, можно получить результат за более короткое время:

```
declare calcHash: function ? to ?;
insert into GeneratedHash
select [$.*, hash: calcHash($.*)]
from Generated;
insert into DuplicatedHash
select [$.*,hash: calcHash($.*)]
from Duplicated;

create link Deduplicated as
select [
 gen: [id: g.id, name: g.name, value: g.value],
 dup: [id: d.id, name: d.name, value: d.value]]
from GeneratedHash g, DuplicatedHash d
match using
 rule_id: g.hash = d.hash;
insert into Measured
select [gen: dd.gen, dup: dd.dup, value:
  compareObject(dd.gen, dd.dup)]
from Deduplicated dd
where compareObject(dd.gen, dd.dup) > 0.8
```

5.2 Реализация методов слияния данных

Будем считать, что этап разрешения сущностей уже пройден и дана некоторая коллекция Deduplicated, где уже установлены соответствия одним из вышеперечисленных способов. Например, пусть имеются две коллекции: A (id, a, b, c) и B (id, a, b, d). Атрибуты a, b, c, d могут содержать null-значения, атрибуты id совпадают. Ниже дан пример подобных данных для коллекции A в формате JSON:

```
[{"a":null,"b":null,"c":"wmqhxfgmac",
"id":919132322},
{"a":null,"b":null,"c":"wmqhxfgmac",
"id":919132322}]
```

Тогда коллекция разрешенных сущностей может быть получена следующим образом:

```
create link Deduplicated as
select
[gen: [id: a.id, a:a.a, b:a.b, c:a.c],
dup: [id: b.id, a:b.a, b:b.b, d:b.d]]
from A a, B b
match using
 rule1: a.id = b.id;
```

Рассмотрим теперь реализацию Minimum Union и оператор Fusion [11] на языке HIL.

Как было определено в разд. 3, **Minimum Union** — это последовательное применение операций outer union и subsumption [13]. Outer Union фактически реализуется с помощью индекса FusionIndex. Использование индекса оправдано, так как существует несколько записей, описывающих одну сущность. Ключом является атрибут id. Ниже представлена реализация операции Outer Union:

```
insert into FusionIndex![id: f.gen.id] select [a: f.gen.a,
b: f.gen.b, c: f.gen.c] from Deduplicated f;

insert into FusionIndex![id: f.dup.id] select [a: f.dup.a,
b: f.dup.b, d: f.dup.d] from Deduplicated f;
```

Далее для реализации subsumption требуется удалить все ненужные кортежи. Это делается на языке Jaql. Для этого нужна функция, которая бы определяла, поглощается ли один кортеж другим. К сожалению, в языке Jaql нет возможностей написания общих (generic) методов, универсальных для всех коллекций, поэтому функцию сравнения можно реализовать на Java и подключить к языку Jaql подобно тому, как демонстрировалось в подразд. 5.1 на примере функций вычисления меры. Либо же можно реализовать функцию для сравнения конкретных коллекций на языке Jaql, как показано ниже:

```
is_subsumed = fn(i,j) ((
isnull(j.a) or (i.a == j.a) ) and (isnull(j.b) or
 (i.b == j.b)) and (isnull(j.c) or (i.c == j.c)) and
 (snull(j.d) or (i.d == j.d)) and (i! = j));
```

Функция **is_subsumed(i,j)** проверяет, поглощает ли один кортеж другой при помощи попарного сравнения атрибутов или проверки на null.

```
removeSubsumed = fn (a) (b = a,
subs = for (i0 in b) [a → filter is_subsumed(i0,$)],
    s = subs → expand,
a → filter not $ in s);
```

Функция **removeSubsumed** удаляет все поглощенные записи из кортежа. Здесь реализован наивный алгоритм, который попарно для каждого кортежа находит все поглощенные им и удаляет их.

```
minUnion = fn(id,a) ( {id:id, minunion:
 removeSubsumed(a)});
```

Функция **minUnion** нужна для построения результирующих кортежей при реализации Minimum Union. С ее помощью операцию Minumum Union можно описать следующим образом на языке HIL:

```
insert into MinimumUnion
select minUnion(i.dup.id,
       FusionIndex![id : i.dup.id])
from Deduplicated i;
```

Для каждого id достаются все соответствующие записи и удаляются те, которые ими поглощаются.

Оператор **Data Fusion** [11] представляет собой особый вид функции, использующий группировку для преодоления конфликтов. Основная идея заключается в группировке различных представлений одной и той же сущности по общему атрибуту, а затем в применении функций разрешения конфликтов для всех остальных атрибутов, сливая данные

в одну сущность. Различают два вида стратегии для функций разрешения конфликтов:

(1) deciding-стратегия заключается в выборе какого-то одного значения каким-то способом (минимум, максимум, случайное значение);

(2) mediating-стратегия заключается в агрегации всех значений (среднее значение, сумма).

Пусть имеются две коллекции: A (id, name, age) и B (id, name, info), пример которых дан ниже:

A

```
[{"id":760046903,"name":null,"age":null},
{"id":15009544, "name": "zvqcsxkzxk",
"age":938781652}]
```

B

```
[{"id":15009544, "name":null, "info":null},
{"id":760046903, "name":"pjltaghyug","info":null}]
```

Пусть для них пройден этап разрешения сущностей и построена коллекция Deduplicated, как описано выше в этом разделе. Пусть также для этих данных построен индекс FusionIndex, как показано выше для операции Minimum Union. Тогда оператор Data Fusion на языке HIL может быть описан следующим образом:

```
@jaql{
average = fn($a) avg($a[*].age);
any = fn($a) any($a[*].name);
concat = fn ($a) strJoin($a[*].info,"_");
}
insert into Fused
select [
id: i.dup.id,
age:
 average(FusionIndex![id: i.dup.id]),
name:

any(FusionIndex![id: i.dup.id]),
info:
 concat(FusionIndex![id: i.dup.id])]
from Deduplicated i;
```

Функции вычисления среднего, выбора случайного ненулевого значения, а также конкатенации реализованы на Jaql. Данное правило образует коллекцию **Fused**, причем для атрибута age будет подсчитано среднее значение, для имени name выбрано любое ненулевое значение, а для атрибута info будет получена конкатенация всех доступных значений. Таким образом, в данном примере показана реализация обеих стратегий для функций разрешения конфликтов в операторе Data Fusion.

6 Заключение

Рассмотренные методы и операции извлечения и интеграции информации о сущностях реального мира, представленной сырыми разноструктурированными коллекциями данных, позволяют программировать интеграционные потоки вида ETL для образования интегрированных структурированных данных, которые могут быть использованы в приложениях для дальнейшего анализа и обработки. В статье рассмотрены методы разрешения сущностей и слияния данных. В статье показаны способы программирования методов и операций извлечения и интеграции информации о сущностях реального мира, включая методы слияния данных, на декларативном языке HIL.

Литература

1. *White T.* Hadoop: The definitive guide. — 3rd ed. — O'Reilly Media, 2012. 688 p.
2. Apache Hadoop 2.5.1. http://hadoop.apache.org.
3. *Naumann F., Herschel M.* An introduction to duplicate detection. Synthesis lectures on data management. — Morgan & Claypool, 2010. Lecture No. 3. 87 p.
4. *Christen P.* Data matching — concepts and techniques for record linkage, entity resolution, and duplicate detection. Data-centric systems and applications ser. — Springer, 2012. 272 p.
5. *Fan W., Geerts F.* Foundations of data quality management. Synthesis lectures on data management. — Morgan & Claypool, 2012. Lecture No. 29. 217 p.
6. *Bleiholder J., Naumann F.* Data fusion // ACM Computing Surveys (CSUR), 2009. Vol. 41. Iss. 1. Article No. 1. doi: 10.1145/1456650.1456651.
7. *Köpcke H., Thor A., Rahm E.* Evaluation of entity resolution approaches on real-world match problems // Proc. VLDB Endowment, 2010. Vol. 3. Iss. 1-2. P. 484–493.
8. *Köpcke H., Rahm E.* Frameworks for entity matching: A comparison // Data Knowledge Engineering, 2010. Vol. 69. Iss. 2. P. 197–210. doi: 10.1016/j.datak.2009.10.003.
9. *Ganti V., Das Sarma A.* Data cleaning, a practical perspective. Synthesis lectures on data management. — Morgan & Claypool, 2013. Lecture No. 36. 85 p.
10. *Getoor L., Machanavajjhala A.* Entity resolution for big data // KDD'13: 19th ACM SIGKDD Conference on Knowledge Discovery and Data Mining Proceedings, 2013. P. 1527–1527.
11. *Bleiholder J., Naumann F.* Declarative data fusion — syntax, semantics, and implementation // East European Conference on Advances in Databases and Information Systems (ADBIS) Proceedings, 2005. P. 58–73.
12. *Luna Dong X., Naumann F.* Data fusion — resolving data conflicts in integration // Proc. VLDB Endowment, 2009. Vol. 2. Iss. 2. P. 1654–1655.

13. *Bleiholder J.* Data fusion and conflict resolution in integrated information systems. — Potsdam: Hasso-Plattner-Institut, 2010. D.Sc. Diss. 184 p.
14. *Winkler W. E.* Overview of record linkage and current research directions. Research report ser. (Statistics # 2006-2). — Washington, DC: Statistical Research Division, U.S. Census Bureau, 2006. http://www.census.gov/srd/papers/pdf/rrs2006-02.pdf.
15. *Adamic L. A., Adar E.* Friends and neighbors on the Web // Social networks, 2003. Vol. 25. No. 3. P. 211–230.
16. *Bilenko M., Mooney R., Cohen W., Ravikumar P., Fienberg S.* Adaptive name matching in information integration // IEEE Intell. Syst., 2003. Vol. 18. No. 5. P. 16–23.
17. Monge–Elkan distance function. http://www.gabormelli.com/RKB/Monge-Elkan_Distance_Function.
18. *Cochinwala M., Kurienb V., Lalka G., Shasha D.* Efficient data reconciliation // Inform. Sci. Int. J., 2001. Vol. 137. Iss. 1-4. P. 1–15.
19. *Bilenko M., Mooney R.* Adaptve duplicate detecton using learnable string similarity measures // KDD'03: 9th ACM SIGKDD Conference (International) on Knowledge Discovery and Data Mining Proceedings, 2003. P. 39–48.
20. *Christen P.* Automatic record linkage using seeded nearest neighbour and support vector machine classification // KDD'08: 14th ACM SIGKDD Conference (International) on Knowledge Discovery and Data Mining Proceedings, 2008. P. 151–159.
21. *Chen Z., Kalashnikov D. V., Mehrotra S.* Exploiting context analysis for combining multiple entity resolution systems // SIGMOD'09: 2009 ACM SIGMOD Conference (International) on Management of Data Proceedings, 2009. P. 207–218.
22. *Gupta R., Sarawagi S.* Answering table augmentaton queries from unstructured lists on the Web // Proc. VLDB Endowment, 2009. Vol. 2. Iss. 1. P. 289–300.
23. *Ravikumar P., Cohen W.* A hierarchical graphical model for record linkage // UAI'04: 20th Conference on Uncertainty in Artificial Intelligence Proceedings, 2004. P. 454–461.
24. *Tejada S., Knoblock C. A., Minton S.* Learning object identification rules for information integration // Inform. Syst. Data Extraction Cleaning Reconciliation, 2001. Vol. 26. Iss. 8. P. 607–633.
25. *Sarawagi S., Bhamidipaty A.* Interactive deduplication using active learning // KDD'02: 8th ACM SIGKDD Conference (International) on Knowledge Discovery and Data Mining Proceedings, 2002. P. 269–278.
26. *Arasu A., Götz M., Kaushik R.* On active learning of record matching packages // SIGMOD'10: 2010 ACM SIGMOD Conference (International) on Management of Data Proceedings, 2010. P. 783–794.
27. *Bellare K., Iyengar S, Parameswaran A. G., Rastogi V.* Active sampling for entity matching // KDD'12: 18th ACM SIGKDD Conference (International) on Knowledge Discovery and Data Mining Proceedings, 2012. P. 1131–1139.
28. *Adam K., Wu E., Karger D., Madden S., Miller R.* Human-powered sorts and joins // Proc. VLDB Endowment, 2011. Vol. 5. Iss. 1. P. 13–24.
29. *Wang J., Kraska T., Franklin M. J., Feng J.* CrowdER: Crowdsourcing Entity Resolution // Proc. VLDB Endowment, 2012. Vol. 5. Iss. 11. P. 1483–1494.
30. *Ananthakrishna R., Chaudhuri S., Ganti V.* Eliminating fuzzy duplicates in data warehouses // VLDB'02: 28th Conference (International) on Very Large Data Bases Proceedings, 2002. P. 586–597.
31. *Fan W., Geerts F., Jia X., Kementsietsidis A.* Conditional functional dependencies for Data cleaning // ICDE'07: 23rd IEEE Conference (International) on Data Engineering Proceedings, 2007. P. 746–755.
32. *Fan W.* Dependencies revisited for improving data quality // PODS'08: 27th ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems Proceedings, 2008. P. 159–170.
33. *Benjelloun O., Garcia-Molina H., Menestrina D., Su Q., Whang S. E., Widom J.* Swoosh: A generic approach to Entity Resolution // VLDB Int. J., 2009. Vol. 18. Iss. 1. P. 255–276.
34. *Bhattacharya I., Getoor L.* Collective Entity Resolution in relational data // ACM Transactions on Knowledge Discovery from Data (TKDD), 2007. Vol. 1. Iss. 1. Article No. 5. doi: 10.1145/1217299.1217304.
35. *Bhattacharya I., Getoor L.* A latent Dirichlet model for unsupervised Entity Resolution // 6th SIAM Conference (International) on Data Mining Proceedings, 2007. P. 47–58.
36. *Broecheler M., Getoor L.* Probabilistic similarity logic // UAI'10: 26th Conference on Uncertainty in Artificial Intelligence Proceedings, 2010. P. 73–82.
37. *Rajaraman A., Ullman J. D.* Integrating information by outerjoins and full disjunctions // PODS'96: 15th ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems Proceedings, 1996. P. 238–248.
38. *Dean J., Ghemawat S.* MapReduce: Simplified data processing on large clusters // Comm. ACM, 2008. Vol. 51. Iss. 1. P. 107–113.
39. MapReduce Tutorial. http://hadoop.apache.org/docs/r1.2.1/mapred_tutorial.html.
40. Apache Pig Project. http://pig.apache.org.
41. The Apache Hive data warehouse. http://hive.apache.org.
42. IBM InfoSphere BigInsights Version 3.0, Jaql reference. — 2014. http://www-01.ibm.com/support/knowledgecenter/SSPT3X_3.0.0/com.ibm.swg.im.infosphere.biginsights.jaql.doc/doc/c 0057749.html.
43. *Das Sarma A., Jain A., Machanavajjhala A., Bohannon P.* An automatic blocking mechanism for large-scale deduplication tasks // CIKM'12: 21st ACM Conference (International) on Information and Knowledge Management Proceedings, 2012. P. 1055–1064.
44. *Papadakis G., Ioannou E., Niederée C., Palpanas T., Nejdl W.* Beyond 100 million entities: Large-scale blocking-based resolution for heterogeneous data //

WSDM'12: 5th ACM Conference (International) on Web Search and Data Mining Proceedings, 2012. P. 53–62.

45. *Kolb L., Thor A., Rahm E.* Dedoop: Efficient deduplication with Hadoop // Proceedings of the VLDB Endowment, 2012. Vol. 5. Iss. 12. P. 1878–1881.

46. *Hernández M., Koutrika G., Krishnamurthy R., Popa L., Wisnesky R.* HIL: A high-level scripting language for entity integration // EDBT'13: 16th Conference (International) on Extending Database Technology Proceedings, 2013. P. 549–560.

Поступила в редакцию 09.11.14

METHODS OF ENTITY RESOLUTION AND DATA FUSION IN THE ETL-PROCESS AND THEIR IMPLEMENTATION IN THE HADOOP ENVIRONMENT

A. E. Vovchenko, L. A. Kalinichenko, and D. Yu. Kovalev

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

[2]Faculty of Computational Mathematics and Cybernetics, M. V. Lomonosov Moscow State University, 1-52 Leninskiye Gory, GSP-1, Moscow 119991, Russian Federation

Abstract: Entities extraction, their transformation and loading in the integrated repository are the main problem of data integration. These actions are part of the ETL-process (extract–transform–loading). An entity is a digital representation of a real world object (for example, information about a person). Entity resolution takes care of duplicate detection, deduplication, record linkage, object identification, reference matching, and other ETL-related tasks. After the entity resolution step, entities should be merged into the one reference entity (containing information from all related entities). Data fusion is the final step in the data integration process. The paper gives an overview of the entity resolution and data fusion methods. Also, the paper presents the techniques for programming the entity resolution and data fusion methods for implementing the ETL-process in the Hadoop environment. High-Level Integration Language (HIL), a declarative language that focuses on resolution and fusion of entities in the Hadoop-infrastructure, is used in this part of the paper.

Keywords: data integration; ETL; entity resolution; data fusion; big data; Hadoop; Jaql; HIL

DOI: 10.14357/19922264140412

Acknowledgments

This work was supported by the Russian Foundation for Basic Research (projects 13-07-00579 and 14-07-00548), Institute of informatics Problems of the Russian Academy of Sciences (IPI RAN) (theme 38.25 "Specification and problem solving of data analysis in conceptual terms of subject areas with intensive use of data" of the state task for IPI RAN), and the Presidium of the Russian Academy of Sciences (Basic Research Program No. 16 "Fundamental problems of system programming").

References

1. White, T. 2012. Hadoop: The definitive guide. 3rd ed. O'Reilly Media. 688 p.
2. Apache Hadoop 2.5.1. Available at: http://hadoop.apache.org/ (accessed November 01, 2014).
3. Naumann, F., and M. Herschel. 2010. *An introduction to duplicate detection*. Synthesis lectures on data management. Morgan & Claypool. Lecture No. 3. 87 p.
4. Christen, P. 2012. *Data matching — concepts and techniques for record linkage, entity resolution, and duplicate detection*. Data-centric systems and applications ser. Springer. 272 p.
5. Fan, W., and F. Geerts. 2012. *Foundations of data quality management*. Synthesis lectures on data management. Morgan & Claypool. Lecture No. 29. 217 p.
6. Bleiholder, J., and F. Naumann. 2009. Data fusion. *ACM Computing Surveys (CSUR)* 41(1). Article No. 1. doi: 10.1145/1456650.1456651.
7. Köpcke, H., A. Thor, and E. Rahm. 2010. Evaluation of entity resolution approaches on real-world match problems. *Proc. VLDB Endowment* 3(1-2):484–493.
8. Köpcke, H., and E. Rahm. 2010. Frameworks for entity matching: A comparison. *Data Knowledge Engineering* 69(2):197–210. doi: 10.1016/j.datak.2009.10.003.

9. Ganti, V., and A. Das Sarma. 2013. Data cleaning: A practical perspective. Synthesis lectures on data management. Morgan & Claypool. Lecture No. 36. 85 p.
10. Getoor, L., and A. Machanavajjhala. 2013. Entity resolution for big data. *19th ACM SIGKDD Conference (International) on Knowledge Discovery and Data Mining (KDD'13) Proceedings*. Chicago. 1527–1527.
11. Bleiholder, J., and F. Naumann. 2005. Declarative data fusion — syntax, semantics, and implementation. *East European Conference on Advances in Databases and Information Systems (ADBIS) Proceedings*. Tallinn. 58–73.
12. Dong, L. X., and F. Naumann. 2009. Data fusion — resolving data conflicts in Integration. *Proc. VLDB Endowment* 2(2):1654–1655.
13. Bleiholder, J. 2010. Data fusion and conflict resolution in integrated information systems. Potsdam. D.Sc. Diss. 184 p.
14. Winkler, W. E. 2006. Overview of record linkage and current research directions. Research report ser. No. 2006-2. Washington, DC: Statistical Research Division, U.S. Census Bureau. 44 p. Available at: http:// www.census.gov/srd/papers/pdf/rrs2006-02.pdf (accessed November 01, 2014).
15. Adamic, L. A., and E. Adar. 2003. Friends and neighbors on the Web. *Social Networks* 25:211–230.
16. Bilenko, M., R. Mooney, W. Cohen, P. Ravikumar, and S. Fienberg. 2003. Adaptive name matching in information integration. *IEEE Intell. Syst.* 18(5):16–23.
17. Monge–Elkan distance function. Available at: http:// www.gabormelli.com/RKB/Monge-Elkan_Distance_ Function (accessed November 01, 2014).
18. Cochinwala, M., V. Kurienb, G. Lalka, and D. Shasha. 2001. Efficient data reconciliation. *Inform. Sci. Int. J.* 137(1-4):1–15.
19. Bilenko, M., and R. Mooney. 2003. Adaptve duplicate detecton using learnable string similarity measures. *9th ACM SIGKDD Conference (International) on Knowledge Discovery and Data Mining (SIGKDD 2003) Proceedings*. Washington. 39–48.
20. Christen, P. 2008. Automatic record linkage using seeded nearest neighbour and support vector machine classification. *14th ACM SIGKDD Conference (International) on Knowledge Discovery and Data Mining (KDD'2008) Proceedings*. Las Vegas. 151–159.
21. Chen, Z., D. V. Kalashnikov, and S. Mehrotra. 2009. Exploiting context analysis for combining multiple entity resolution systems. *2009 ACM SIGMOD Conference (International) on Management of Bata (SIGMOD 2009) Proceedings*. Providence. 207–218.
22. Gupta, R., and S. Sarawagi. 2009. Answering table augmentaton queries from unstructured lists on the Web. *Proc. VLDB Endowment* 2(1):289–300.
23. Ravikumar, P., and W. Cohen. 2004. A hierarchical graphical model for record linkage. *20th Conference on Uncertainty in Artificial Intelligence (UAI 2004) Proceedings*. Virginia. 454–461.
24. Tejada, S., C. A. Knoblock, and S. Minton. 2001. Learning object identification rules for information integration. *Inform. Syst. Data Extraction Cleaning Reconciliation* 26(8):607–633.
25. Sarawagi, S., and A. Bhamidipaty. 2002. Interactive deduplication using active learning. *8th ACM SIGKDD Conference (International) on Knowledge Discovery and Data Mining (KDD 2002) Proceedings*. Edmonton. 269–278.
26. Arasu, A., M. Götz, and R. Kaushik. 2010. On active learning of record matching packages. *2010 ACM SIGMOD Conference (International) on Management of Data Proceedings*. Indianapolis. 783–794.
27. Bellare, K., S. Iyengar, A. G. Parameswaran, and V. Rastogi. 2012. Active sampling for entity matching. *18th ACM SIGKDD Conference (International) on Knowledge Discovery and Data Mining (KDD 2012) Proceedings*. Beijing. 1131–1139.
28. Adam, K., E. Wu, D. Karger, S. Madden, and R. Miller. 2011. Human-powered sorts and joins. *Proc. VLDB Endowment* 5(1):13–24.
29. Wang, J., T. Kraska, M. J. Franklin, and J. Feng. 2012. CrowdER: Crowdsourcing Entity Resolution. *Proc. VLDB Endowment* 5(11):1483–1494.
30. Ananthakrishna, R., S. Chaudhuri, and V. Ganti. 2002. Eliminating fuzzy duplicates in data warehouses. *28th Conference (International) on Very Large Data Bases (VLDB 2002) Proceedings*. Hong Kong. 586–597.
31. Fan, W., F. Geerts, X. Jia, and A. Kementsietsidis. 2007. Conditional functional dependencies for data cleaning. *2007 IEEE 23rd Conference (International) on Data Engineering Proceeding*. Istanbul. 746–755.
32. Fan, W. 2008. Dependencies revisited for improving data quality. *27th ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems (PODS 2008) Proceedings*. Vancouver. 159–170.
33. Benjelloun, O., H. Garcia-Molina, D. Menestrina, Q. Su, S. E. Whang, and J. Widom. 2009. Swoosh: A generic approach to Entity Resolution. *VLDB Int. J.* 18(1):255–276.
34. Bhattacharya, I., and L. Getoor. 2007. Collective Entity Resolution in relational data. *ACM Trans. Knowledge Discovery Data (TKDD)* 1(1). Article No. 5. doi: 10.1145/1217299.1217304.
35. Bhattacharya, I., and L. Getoor. 2007. A latent Dirichlet model for unsupervised Entity Resolution. *6th SIAM Conference (International) on Data Mining Proceedings*. Maryland. 47–58.
36. Broecheler, M., and L. Getoor. 2010. Probabilistic similarity logic. *26th Conference on Uncertainty in Artificial Intelligence Proceedings*. Corvallis. 73–82.
37. Rajaraman, A., and J. D. Ullman. 1996. Integrating information by outerjoins and full disjunctions. *15th ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems (PODS1996) Proceedings*. Montreal. 238–248.
38. Dean, J., and S. Ghemawat. 2008. MapReduce: Simplified data processing on large clusters. *Comm. ACM* 51(1):107–113.

39. MapReduce tutorial. Available at: http://hadoop.apache.org/docs/r1.2.1/mapred_tutorial.html (accessed November 01, 2014).
40. Apache Pig Project. Available at: http://pig.apache.org/ (accessed November 01, 2014).
41. The Apache Hive data warehouse. Available at: http://hive.apache.org/ (accessed November 01, 2014).
42. IBM InfoSphere BigInsights Version 3.0, Jaql reference. Available at: http://www-01.ibm.com/support/knowledgecenter/SSPT3X_3.0.0/com.ibm.swg.im.infosphere.biginsights.jaql.doc/doc/c0057749.html (accessed November 01, 2014).
43. Sarma, D.A., A. Jain, A. Machanavajjhala, and P. Bohannon. 2012. An automatic blocking mechanism for large-scale de-duplication tasks. *21st ACM Conference (International) on Information and Knowledge Management Proceedings*. Maui. 1055–1064.
44. Papadakis, G., E. Ioannou, C. Niederée, T. Palpanas, and W. Nejdl. 2012. Beyond 100 million entities: Large-scale blocking-based resolution for heterogenous data. *5th ACM Conference (International) on Web Search and Data Mining Proceedings*. Seattle. 53–62.
45. Kolb, L., A. Thor, and E. Rahm. 2012. Dedoop: Efficient deduplication with Hadoop. *Proceedings of the VLDB Endowment* 5(12):1878–1881.
46. Hernández, M., G. Koutrika, R. Krishnamurthy, L. Popa, and R. Wisnesky. 2013. HIL: A high-level scripting language for entity integration. *16th Conference (International) on Extending Database Technology (EDBT'13) Proceedings*. Genoa. 549–560.

Received November 9, 2014

Contributors

Vovchenko Alexey E. (b. 1984) — Candidate of Science (PhD) in technology, senior researcher, Institute of Informatics Problems, Russian Academy of Sciences; 44-2 Vavilov Str., Moscow 119333, Russian Federation; alexey.vovchenko@gmail.com

Kalinichenko Leonid A. (b. 1937) — Doctor of Science in physics and mathematics, professor; Head of Laboratory, Institute of Informatics Problems, Russian Academy of Sciences; 44-2 Vavilov Str., Moscow 119333, Russian Federation; professor, Faculty of Computational Mathematics and Cybernetics, M.V. Lomonosov Moscow State University, 1-52 Leninskiye Gory, GSP-1, Moscow 119991, Russian Federation; leonidk@synth.ipi.ac.ru

Kovalev Dmitry Yu. (b. 1988) — junior researcher, Institute of Informatics Problems, Russian Academy of Sciences, 44-2 Vavilov Str., Moscow 119333, Russian Federation; dm.kovalev@gmail.com

INFORMATICS AND APPLICATIONS 2014. Vol. 8. Iss. 4. P. 110–124

CONCEPTUAL MODELING OF MULTIDIALECT WORKFLOWS

L. Kalinichenko[1,2], S. Stupnikov[1], A. Vovchenko[1], and D. Kovalev[1]

Abstract: This paper contributes to the techniques for conceptual representation of data analysis algorithms and data integration facilities as well as processes to specify data and behavior semantics in one paradigm. An investigation of a novel approach for applying a combination of semantically different platform-independent rule-based languages (dialects) for interoperable conceptual specifications over various rule-based systems (RSs) relying on the rule-based program transformation technique recommended by the W3C Rule Interchange Format (RIF) is extended here. Such approach is combined with the facilities aimed at the semantic rule-based mediation intended for the heterogeneous data base integration. This paper extends a previous research of the authors in the direction of workflow modeling for definition of compositions of algorithmic modules in a process structure. A capability of the multidialect workflow support specifying the tasks in semantically different languages mostly suited to the task orientation is presented. A practical workflow use case, the interoperating tasks of which are specified in several rule-based languages (RIF-CASPD, RIF-BLD, RIF-PRD), is introduced. In addition, OWL 2 is used for the conceptual schema definition, RIF-PRD is used also for the workflow orchestration. The use case implementation infrastructure includes a production rule-based system (IBM ILOG), a logic rule-based system (DLV), and a mediation system.

Keywords: conceptual specification; workflow; RIF; production rule languages; database integration; mediators; PRD; multidialect infrastructure

DOI: 10.14357/19922264140413

1 Introduction

This work keeps on the intention of developing the facilities for conceptual declarative problem specification and solving in data intensive domains (DID). In [1] it was claimed that conceptual data semantics alone (e. g., formalized in ontology languages based on description logic) are insufficient, so that conceptual representation of data analysis algorithms as well as processes for problem solving are required to specify data and behavior semantics in one paradigm.

The results presented in this paper[3] extend the research [1] aimed at the definition and implementation of the facilities for conceptually-driven problems specification and solving in DID aiming at ensuring eventually the following capabilities for expressing the specifications:

(1) an ability to provide complete and precise specification of the abstract structure and behavior of the domain entities, their consistency, relationship, and interaction;

(2) well-grounded diversity of semantics of the modeling facilities providing for the best attainable expressiveness, compactness, and precision of the definition of the problem solving algorithm specifications;

(3) arrangements for the extensions of the modeling facilities satisfying the changing technological and practical needs;

(4) specification independence from implementation platforms (languages, systems);

(5) specification independence from concrete information resources (IRs) (databases, services, ontologies, etc.) combined with facilities for their semantic integration and interoperability; and

(6) built-in methodologies for creation of unifying specification languages providing for construction of semantics-preserving mappings of conceptual specifications into their implementations in specific platforms.

The research reported in [1] investigated the conceptual modeling facilities for DID applying rule-based declarative logic languages possessing different, complementary semantics and capabilities combined with the methods and languages for heterogeneous data mediation and integration. Two fundamental techniques were combined: (*i*) constructing of the unifying extensible language providing for semantics-preserving mapping

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

[2]Faculty of Computational Mathematics and Cybernetics, M. V. Lomonosov Moscow State University, 1-52 Leninskiye Gory, GSP-1, Moscow 119991, Russian Federation

[3]This paper is an extended for the journal version of the results presented in the "Multidialect Workflows" report at the ADBIS'2014 Conference.

into it of various IR specification languages (e. g., such as data definition (DDL) and data manipulation (DML) languages for databases); and (*ii*) creation of the unified extensible family of rule-based languages (dialects) and a model of interoperability of the programs expressed in such dialects with different semantics.

The first technique is based on the experience obtained in course of the SYNTHESIS language development [2]. The kernel of the SYNTHESIS language is based on the object-frame data model used together with the declarative rule-based facilities in the logic language similar to a stratified Datalog with functions and negation. The extensions of the kernel are constructed in such a way that each extension together with the kernel is a result of semantic preserving mapping of some IR language into the SYNTHESIS [2]. The canonical information model is constructed as a union of the kernel with such extensions defined for various resource languages. Canonical model is used for development of *mediators* positioned between the users, conceptually formulating problems in terms of the mediators, and distributed resources. A schema of a subject mediator for a class of problems includes the specification of the domain concepts defined by the respective ontologies.

Another, multidialect technique for rule-based programs interoperability applied is based on the RIF standard [3] of W3C. The RIF standard introduces a unified family of rule-based languages together with a methodology for constructing of semantic preserving mappings of specific languages used in various RSs into RIF dialects. Examples of RSs include SILK, OntoBroker, DLV, IBM Websphere ILOG JRules, RIF4J + IRIS, and others (more examples can be found at http://www.w3.org/2005/rules/wiki/Implementations). From the RIF point of view, an IR is a program developed in a specific language of some RS.

In [1], the first results obtained were presented including the description of an approach and an infrastructure supporting:

- application domain conceptual specification and problem solving algorithms definitions based on the combination of the heterogeneous database mediation technique and the rule-based multidialect facilities;
- interoperability of distributed multidialect rule-based programs and mediators integrating heterogeneous databases; and
- rule delegation approach for the peer interactions in the multidialect environment.

The proof-of-concept prototype of the infrastructure based on the SYNTHESIS environment and RIF standards has been implemented. The approach for multidialect conceptualization of a problem domain, rule delegation, rule-based programs, and mediators interoperability were explained in detail and illustrated on an use-case in the finance domain [1]. For the conceptual definition of the use-case problem, the OWL was used for the domain concepts definition and two RIF logic dialects RIF-BLD [4] and RIF-CASPD [5] were used and mapped for implementation into the SYNTHESIS formula language and the ASP (answer set programming) based DLV [6] language, respectively.

The results obtained so far are quite encouraging for future work: they show that the mentioned in the beginning capabilities (1)–(6) sought for conceptual modeling become feasible. This paper reports the results of extending the research in the direction of modeling of the processes for the problem solving following the approach briefly outlined above. These results include extensions of the infrastructure and specification languages considered in [1] to the workflow level keeping the same approach and paradigm as well as aiming at the capabilities of the conceptualization (1)–(6) that were stated in [1] and mentioned in the beginning of the introduction.

For investigation of such extension with respect to the choice of rule-based languages, it was decided not to go outside the limits of the existing set of the published RIF dialects. Such decision would allow to retain well-defined semantics of the conceptual rule-based languages with a possibility to check preservation of their semantics by various languages of the implementing systems.

The production rule dialect RIF PRD [7] has been chosen as the language for the workflow modeling in such a way that the tasks of the workflow can have multidialect rule-based representation (as defined in [1]). This paper reporting the results of such investigation is structured as follows. To make the paper self-contained, the next section provides a brief overview of the infrastructure supporting multidialect programming defined in details in [1]. Here, it is stressed that this infrastructure is suitable for the workflow tasks specification. Workflow-oriented extension of the multidialect infrastructureis considered in section 3. Use case implementation in the proof-of-concept prototype is given in section 4. Related works are reviewed in section 5. Concluding Remarks summarize contributions of the research.

2 Basic Principles of the Workflow Tasks Representation in the Multidialect Infrastructure

Each workflow task (besides those that for pragmatic reasons are defined as externally specified functions) is assumed to be represented in the novel infrastructure

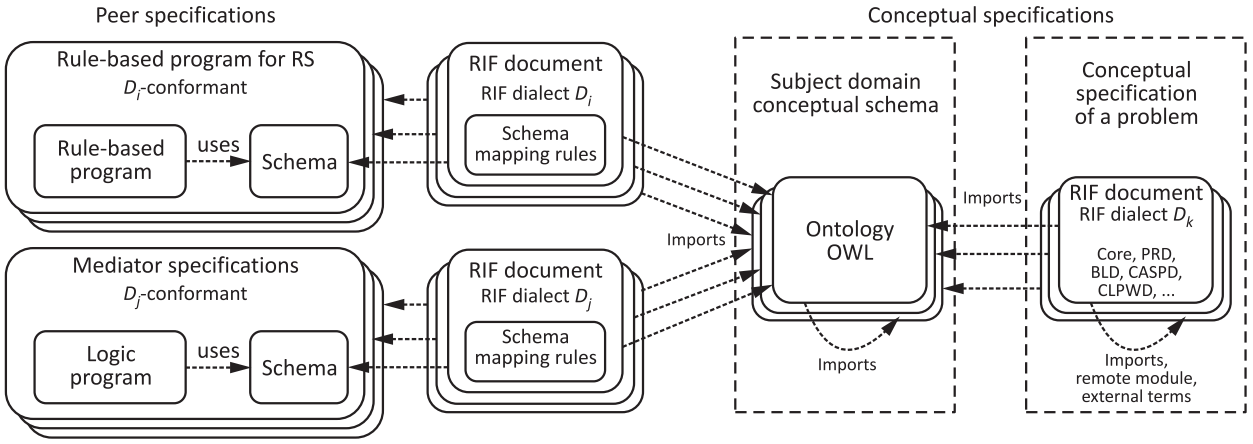


Figure 1 Conceptual schema and peer specifications

defined in details in [1]. Conceptual programming of tasks is performed using the RIF dialects (now not only logic but also PRDs can be used).

Conceptual tasks are implemented by their transformation into the rule-based programs of the respective RSs and mediation systems (MSs). *Conceptual specification of a task* is defined in the context of a subject domain and consists of a set of RIF-documents (document is a specification unit of RIF). The *conceptual schema* of the domain is defined using OWL 2 [8] ontologies. Such usage of ontology is analogous to [9]; however, it is specifically important in the multidialect environment due to the formally defined compatibility between RIF and OWL. The ontologies contain entities of the domain and their relationships (Fig. 1, right-hand part). Conceptual specification of a task is defined over conceptual schema. Ontologies are imported into the RIF-documents specifying an import profile, for instance, OWL Direct. Documents *import* other documents having the same semantics (the *Import* directive), *link* documents defined using other dialects and having different semantics (remote module directive *Module*) or *refer* to entities contained in other documents using *external terms*.

Semantics of a conceptual task definition in such setting becomes a multidialect one. The specification modules of a task are treated as peers. Mediation modules are assumed to be defined in RIF-BLD for representation of the mediator rules (to be interpreted in SYNTHESIS) supporting schema mapping and semantic integration of the IRs. Multidialect task is implemented by means of transformation of conceptual specifications into modular, component-based peer-to-peer (P2P) program represented in the languages of the MSs and RSs with the respective semantics. Interoperability of logic rule components of such distributed program is carried out by means of the delegation technique [1, section 3.3]. Production rule components are considered as external functions, interoperability is achieved through the mechanism of external terms.

A schema S_R of a peer R is a set of entities (classes or relations and their attributes) corresponding to extensional and intensional predicates of the resource implementing the peer R.

The RS or the MS of each peer R should be a conformant D_R consumer where D_R is the respective RIF dialect (Fig. 1, left-hand part). Conformance is formally defined using formula entailment and language mappings [3].

The peer R is relevant to a RIF-document d of a conceptual specification of a problem (Fig. 1, right-hand part) if (i) D_R is a subdialect of the document d dialect (subdialect is a language obtained from some dialect by removing certain syntactic constructsand imposing respective restrictions on its semantics [4]; each program that conforms with the subdialect also conforms with the dialect) and (ii) entities of the peer schema S_R (if they exist) are *ontologically relevant* to entities of the conceptual schema the names of which are used in d for extensional predicates.

The schema of a relevant peer is mapped into the conceptual schema. The mapping establishes the correspondence of the conceptual entities referred in the document d to their expressions in terms of entities of the schema S_R using rules of the D_R dialect. These schema mapping rules constitute separate RIF-document (Fig. 1, middle part).

Peers communicate using a technique for distributed execution of the rule-based programs. The basic notion of the technique is delegation-transferring facts and rules from one peer to another. A peer is installed on a node of the multidialect infrastructure. A node is a combination of a wrapper, an RS or an MS, and a peer (for the details, refer [1, Fig. 3]). A wrapper transforms programs and facts from the specific RIF dialect into the language of

the RS or MS and *vice versa*. A wrapper also implements the delegation mechanism. Transferring facts and rules among peers is performed in the RIF dialects.

A special component (*Supervisor*) of the architecture defined in [1] stores shared information of the environment, i. e., conceptual specifications related to the domain and to the problem, a list of the relevant resources, RIF-documents combining rules for the conceptual specification and a resource schema mapping.

Implementation of the conceptual specification includes the following steps:

(1) rewriting of the conceptual documents into the RIF-programs of the peers performed by the *Supervisor*. The rewriting includes also (i) replacing the document identifiers (used to mark predicates) by peer identifiers and (ii) adding schema mapping rules to programs (Fig. 1, middle part);

(2) a transfer of the rewritten programs to nodes containing peers relevant to the respective conceptual documents. The transfer is performed by the *Supervisor* by calling the method *loadRules* of the respective node wrappers;

(3) a transformation of the RIF-programs into the concrete RS or MS languages. The transformation is performed by the *NodeWrapper* or by the RS or MS itself (if the RS or MS supports the respective RIF dialect); and

(4) an execution of the produced programs in P2P environment.

During the process of rewriting of the conceptual schema into the resource programs, the relationships between RIF-documents of the conceptual schema defined by remote or imported terms are replaced by relationships between peers also defined by remote or imported terms. To implement remote and imported terms, a *rule delegation* mechanism is used to transfer facts and rules from one peer to another. The details of rule delegation approach including description of the related algorithms are provided in [1].

3 Workflow-Oriented Extension of the Multidialect Infrastructure

The aim of the infrastructure proposed is a conceptual programming of problems in the RIF-dialects and an implementation of conceptual specifications using rule-based languages of the RSs and MSs. One of the objectives of this particular paper is to introduce an extension of the existing multidialect infrastructure [1] aiming at the conceptual specification of rule-based workflows.

Conceptual specification of a problem (class of problems) is defined in the context of a subject domain and consists of a set of RIF-documents. Besides the documents expressed in the logic dialects of RIF, the documents expressed in the production rule dialect (RIF-PRD) also can be a part of conceptual specification of a problem. In particular, these documents are aimed to express a process of solving the problem as the production rule-based workflow.

3.1 Specification of workflow orchestration

A workflow consists of a set of tasks orchestrated by specific constructs (*workflow patterns* [10], for instance, *sequence*, *split*, *join*) defining the order of tasks execution. The specification of such orchestration is called here a *workflow skeleton*. A skeleton is defined using RIF-PRD production rules. Workflows and workflow patterns can be represented using production rules in various ways, e. g., as in [10, 11]. The approach applied in this paper to represent workflows requires the extension of RIF-PRD dialect by several built-in predicates (they are considered to be a part of *wkfl* namespace referenced by http://www.w3.org/2014/rif-workflow-predicate# URI similarly to *func* and *pred* namespaces defined in [12] for built-in functions and predicates of RIF):

– predicate wkfl:end-of-task(?arg) where ?arg is an identifier of a task. The value space of ?arg is the XML-Schema built-in data type xsd:Name representing XML names. The predicate turns into true if a task *?arg* has been completed;

– predicate wkfl:variable-definition(?arg1 ?arg2) where ?arg1 is the identifier of a variable and ?arg2 is the identifier of a type of the variable. The value space for both arguments is xsd:Name. Turning the predicate into true means that a variable ?arg1 of type ?arg2 is defined in the context of a workflow;

– predicate wkfl:variable-value(?arg1 ?arg2) where ?arg1 is the identifier of a variable and ?arg2 is the value of the variable. The value space for the first argument is xsd:Name, the value space for the second argument is the union of value spaces of all RIF built-in datatypes. Turning the predicate into true means that a variable ?arg1 has the value ?arg2;

– predicate wkfl:parameter-definition(?arg1 ?arg2 ?arg3) where ?arg1 is the identifier of a workflow parameter; ?arg2 is the identifier of a type of the parameter; and ?arg3 is the direction of the parameter. The value space for the first and for the second arguments is xsd:Name. The value space for the third argument is {IN, OUT, IN_OUT} (*input*, *output*, or *input–output* parameter). Turning the predicate into true means that a parameter ?arg1 of type ?arg2, and direction ?arg3 is defined for a workflow; and

– predicate wkfl:parameter-value(?arg1 ?arg2) defines values of workflow parameters in the same way as wkfl:variable-value defines values of workflow variables.

Predicates wkfl:variable-definition and wkfl:variable-value allow to specify workflow variables and their values and thus to organize the data flow within a workflow. Predicates wkfl:parameter-definition and wkfl:parameter-value allow to specify workflow parameters and their values and thus to define the interface of a workflow in terms of input and output parameters. Using of workflow parameters and variables is illustrated in the Appendix.

The predicate wkfl:end-of-task(?arg) allows to orchestrate the order of execution of workflows tasks using conditions and actions of production rules. In this section, the template rules intended for representation of several basic workflow patterns (Fig. 2) are provided.

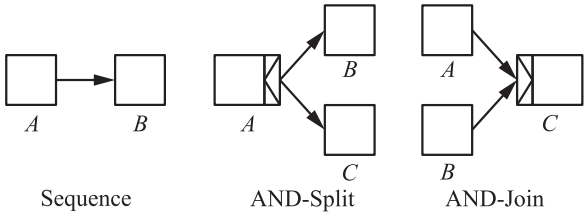


Figure 2 Basic workflow patterns

Three well-known workflow patterns are considered below: Sequence, AND-Split, and AND-Join.

The *AND-Split*[1] workflow pattern is represented in RIF-PRD by the following production ruletemplate using wkfl:end-of-task predicate:

```
If Not(External(wkfl:end-of-task(A)))
Then Do (Act(A)
 Assert(External(wkfl:end-of-task(A))))
If And(Not(External(wkfl:end-of-task(B)))
 External(wkfl:end-of-task(A)))
Then Do (Act(B)
 Assert(External(wkfl:end-of-task(B))))
If And(Not(External(wkfl:end-of-task(C)))
 External(wkfl:end-of-task(A)))
Then Do (Act(C)
 Assert(External(wkfl:end-of-task(C))))
```

The template includes three rules for tasks A, B, and C, respectively. Act(A), Act(B), and Act(C) denote *actions* associated with tasks A, B, and C. Orchestration (tasks B and C are executed concurrently right after task A is completed) is specified using wkfl:end-of-task predicate in conditions and Assert actions of rules.

Similarly, the AND-Split pattern is represented in RIF-PRD by the following production rule template:

```
If Not(External(wkfl:end-of-task(A)))
Then Do (Act(A)
 Assert(External(wkfl:end-of-task(A))))
If And(Not(External(wkfl:end-of-task(B)))
 External(wkfl:end-of-task(A)))
Then Do (Act(B)
 Assert(External(wkfl:end-of-task(B))))
If And(Not(External(wkfl:end-of-task(C)))
 External(wkfl:end-of-task(A)))
Then Do (Act(C)
 Assert(External(wkfl:end-of-task(C))))
```

The Sequence pattern is represented in RIF-PRD by the following production rule template:

```
If Not(External(wkfl:end-of-task(A)))
Then Do (Act(A)
 Assert(External(wkfl:end-of-task(A))))
If And(Not(External(wkfl:end-of-task(B)))
 External(wkfl:end-of-task(A)))
Then Do (Act(B)
 Assert(External(wkfl:end-of-task(B))))
```

More complicated patterns like OR-, XOR- splits and joins, structured loops, subflows, and others are represented in RIF-PRD similarly.

3.2 Workflow tasks specification

Workflow taskscan be specified as:

– separate RIF-documents in various logic RIF-dialects (this is the way how multidialect infrastructure [1] is extended with workflow capabilities);
– separate RIF-documents in the RIF-PRD dialect;
– set of production rules embedded into the workflow skeleton; and
– external functions treated as "black boxes."

Semantics of tasks specified as multidialect logic programs are defined in accordance with the RIF-FLD [3] standard and standards for the respective RIF-dialects (BLD, CASPD, etc.). Semantics of tasks specified as production rule programs are defined in accordance with the RIF-PRD standard. Semantics of external functions "are assumed to be specified externally in some document" [3].

All kinds of tasks (except those that are embedded into a workflow skeleton) are referenced in the workflow skeleton as *external terms* [3] like External (t) where term t is defined by an external resource identified by internationalized resource identifier (IRI) [3].

3.3 Workflow implementation infrastructure

Workflows defined in the conceptual specification are implemented in the environment shown in Fig. 3. Peer-

[1] In this paper, the simplified *presentation syntax* [7] is used.

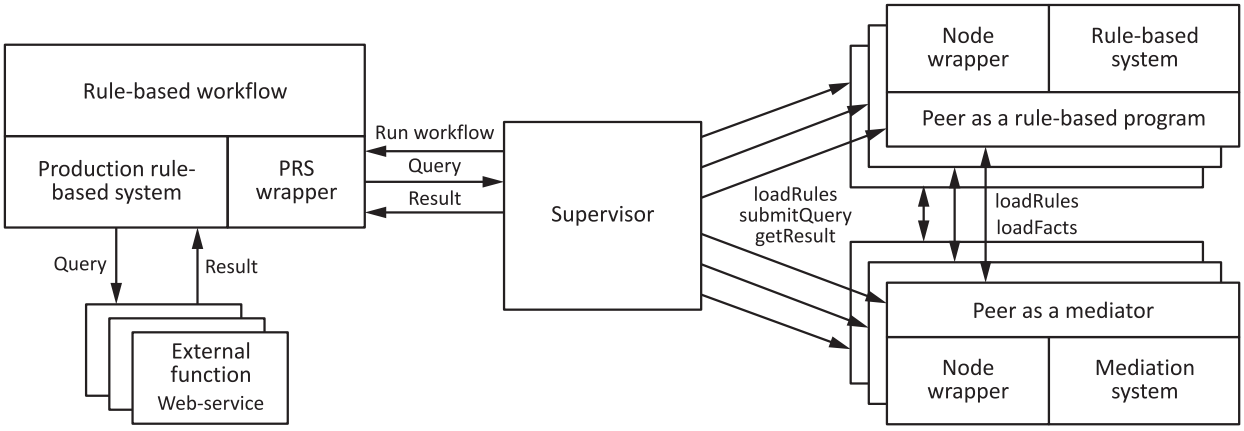


Figure 3 Extended multidialect infrastructure

to-peer environment [1] intended to implement logic programs is extended with a production rule-based system (PRS) (for instance, a production system compliant with the OMG Production Rule Representation [13]) and with external functions, implemented as web-services. Implementation of the conceptual specification includes the following steps:

(1) transfer of the conceptual RIF-documents constituting a workflow skeleton to the production rule-based system node (performed by the *Supervisor* component);

(2) transformation of the conceptual RIF-documents constituting a workflow skeleton into the language of the production rule-based system (performed by the PRS Wrapper component);

(3) transferring RIF logic programs related to tasks to the relevant nodes of the environment and transformation of the RIF-programs into the concrete RS or MS languages [1]; and

(4) execution of the workflow.

The interface of the *Supervisor* includes methods for submitting and executing a workflow represented as a set of RIF-documents, and for getting the result of the workflow execution.

To provide a proof of the multidialect infrastructure concept, a use case in the financial domain has been implemented. The problem to be solved in the use case is called the *investment portfolio diversification problem*. The detailed description of the use case is included in the Appendix.

4 Related Work

Two types of workflow models, namely, abstract and concrete, were identified [14]. In the abstract model, a workflow is described in an abstract form, without referring to specific resources. In this paper, workflow representation in abstract and platform-independent form is suggested.

A classification model for scientific workflow characteristics [10] contributes to better understanding of scientific workflow requirements. The list of structural patterns discovered during this analysis (including sequential, parallel, parallel-split, parallel-merge, and mesh) influenced the choice of the required workflow patterns.

The OMG standard [13] reflects an attitude to production rules from the industrial side providing an OMG MDA (model-driven architecture) platform-independent model (PIM) with a high probability of support at the PSM (platform-specific model) level from the rule engine vendors. Similar capabilities though formally defined are used as the basis for the RIF-PRD [7].

Some vendors of such production rule engines have extended their languages with the workflow specification capabilities. IBM has extended ILOG to provide the ruleflow capability. Microsoft supports Windows Workflow Foundation as a platform providing the workflow and rules capabilities. The examples of specific formalisms for PIM rule-based process specifications are also provided in [11].

Comparing to the known variants of the PIM production rule representations, selection of the RIF-PRD is considered to be well grounded:

(1) the RIF-PRD is formally defined;

(2) RIF ensures support of interoperability of modules written in different rule-based dialects with different semantics;

(3) RIF provides foundations for PIM to PSM semantic preserving transformation; and

(4) RIF also provides ability for specification of the concepts in application domain terms combining rule-based specifications with the OWL ontologies.

Importance of providing the interdialect interoperation is advocated in [15] for combining the functionalities of production systems and logic programs for abductive logic programming (ALP). The ALP framework gives a model-theoretic semantics to both kinds of rules and provides them with powerful proof procedures, combining backward and forward reasoning.

Papers related to RIF-PRD experimentations are focused mainly on the issue of the PRD programs transformation to an implementation system. In [16], a case study of bridging the ILOG Rule Language (IRL) to RIF-PRD and vice versa is considered. In [17], implementation of RIF-PRD in three different paradigms: Answer Set Programming, Production Rules, and Logic Programming (XSB) is investigated.

The contribution of this paper with regard to previous works of the authors [1] consists in extensions of the infrastructure and specification languages considered in [1] to the workflow level.

5 Concluding Remarks

Progress in the investigation of the infrastructure [1] for the conceptual multidialect interoperable programming in the abstract, rule-based, platform-independent notations is reported. An extension of the coherent combination of the multidialect rule-based programming technique recommended by the W3C RIF with the approach for unifying modeling of heterogeneous data bases for their semantic mediation is presented. The extension of the infrastructure and specification languages considered in [1] in the direction of the workflow modeling is presented.

Sticking to the limits of the existing set of the published RIF dialects, a capability of the multidialect workflow support is presented with the tasks specified in semantically different languages mostly suited to the task orientation. Also, a realistic problem solving use case containing the interoperating tasks specified in several platform-independent rule-based languages: RIF-CASPD, RIF-BLD, RIF-PRD, is presented. In addition, OWL 2 is used for the conceptual schema definition, RIF-PRD is applied for the workflow orchestration. The platforms selected for implementation of the tasks include: DLV, SYNTHESIS, IBM ILOG. Such approach retains well-defined semantics of the platform-independent rule-based languages with a possibility to check preservation of their semantics by various languages of the implementing systems. The principle of independence of tasks from the specific IRs is carried out by the heterogeneous database mediation facilitates contributing to the reuse of tasks and workflows. Alongside with the further extension of the approach, in the future work, the authors plan to apply the conceptual multidialect programming philosophy for support of the experiments in data intensive sciences. In particular, they plan to investigate modeling hypotheses in astronomy representing them as a set of rules applying the multiplicity of the dialects required.

APPENDIX A

MULTIDIALECT WORKFLOW USE CASE

A.1 Investment portfolio diversification problem extended

Motivation of the use case that illustrates the proposed approach comes from the finance area. The use case extends the *investment portfolio diversification problem* defined in [1, Appendix] by adding workflow orchestration applying the RIF-PRD. The idea of the portfolio diversification problem is as follows. The portfolio is a collection of securities of companies, and its size is the number of securities in the portfolio. The problem is to build a diversified portfolio of maximum size. Diversification means that the prices of the securities in portfolio should be almost independent of each other. If the price of one security falls, it will not significantly affect the prices of others. Thus, the risk of a portfolio sharp decrease is reduced.

The input data for the problem is a set of securities and respective time series of indicators of the security price for each security. Time series for each security is a set of pairs (d, v) where d is a date and v is an indicator of the security price (for instance, closing price). The financial services *Google Finance* (https://www.google.com/finance) and *Yahoo! Finance* (http://finance.yahoo.com/) are considered. They include various indicators of the security price for all trading days of the last decades. For the diversified portfolio, the securities having noncorrelated time series should be used. Noncorrelation of the time series means that their correlation is less than some predetermined price correlation value. The output data for the problem is a set of subsets of securities of the maximum size, for which the pair wise correlation will be less than the predetermined one.

The maximum satisfying subset of securities is calculated in the following way. Let G be a graph where the vertices are the securities. An edge between two securities exists if absolute value of their correlation is less than a specified number. So, any two securities connected by an edge are considered as noncorrelated. In such case, the problem of finding the portfolio of the maximum size is exactly the problem of finding a maximum clique in an undirected graph. A maximal clique is a maximal portfolio. Note that several different maximal portfolios can be found.

The conceptual specification of the use case [1] used two RIF-dialects: RIF-BLD and RIF-CASPD. The use case was implemented in the environment containing a mediation system used as a platform for RIF-BLD [4] and ASP-based DLV system [6] — a platform for RIF-CASPD. The RIF-BLD was used to specify the problem of data integration, and RIF-CASPD — the problem of finding a maximum clique in an undirected graph.

In this work, the portfolio use case is extended in the following way. The goal is not only to build a set of diversified portfolios, but also to choose the "best" of them according to some criteria. There are several approaches to choose the most appropriate portfolio.

The most recognized one is based on the Markovitz portfolio theory [18]. The idea is to choose the portfolio, which has the maximum risk/return ratio. The most well-known metric to operate with risk/return is Sharpe-ratio [19]: $(r_p - r_f)/\sigma^2$. Here, r_p denotes the expected return of the portfolio; r_f denotes the risk free rate; and σ^2 denotes the portfolio standard deviation (risk). The more the Sharpe-ratio is, the better the investment is.

Another approach is based on an idea that with the advent of social networks, it became possible to monitor ideas, sentiments, actions of people and lots of available information has to do with the markets and investments. In [20], Bollen *et al.* draw the connection between the mood of investor tweets and the move of Dow Jones Index, stating that correlation between them is more than 80%. The idea of using tweets to assess market movements has been implemented in several hedge funds.

Combining these two strategies could provide benefits of both of them, which leads to the following problem statement: having S&P500 (a stock market index maintained by the Standard & Poor's, comprising 500 large-cap American companies) list of companies, compute the diversified portfolio of maximum size with the best risk/return and sentiment ratios.

A.2 Conceptual specification of the application domain and the problem

Conceptual schema (ontology) of the application domain of historical prices of securities is written in the simplified OWL functional syntax [8] (Declaration keyword is omitted; property, domain, and range declarations are combined).

```
Ontology(<http://synthesis.ipi.ac.ru/portfolio/
    ontology>
 Class(Portfolio)
  ObjectProperty(securities domain(Portfolio)
   range(Portfolio))
  DataProperty(expected_return domain(Portfolio)
   range(xsd:double))
  DataExactCardinality(1 expected_return
   Portfolio)
  DataProperty(std_dev domain(Portfolio)
   range(xsd:double))
  DataExactCardinality(1 std_dev Portfolio)
  DataProperty(sharpe_ratio domain(Portfolio)
   range(xsd:double))
  DataExactCardinality(1 sharpe_ratio Portfolio)
  DataProperty(twitter_positive_ratio
   domain(Portfolio) range(xsd:double))
  DataExactCardinality(1 twitter_positive_ratio
   Portfolio)
  DataProperty(risk_free_rate domain(Portfolio)
   range(xsd:double))
  DataExactCardinality(1 risk_free_rate
   Portfolio)
  DataProperty(recommended domain(Portfolio)
   range(xsd:boolean))
  DataExactCardinality(1 recommended Portfolio)

 Class(Security)
  DataProperty(ticker  domain(Security)
   range(xsd:string))
  DataExactCardinality(1 ticker Security)
  DataProperty(rates  domain(Security)
   range(StockRate))
  DataProperty(positive_tweets domain(Security)
   range(xsd:double))
  DataExactCardinality(1 positive_tweets
   Security)
  DataProperty(sec_expected_return
   domain(Security) range(xsd:double))
  DataExactCardinality(1 sec_expected_return
   Security)
  DataProperty(sec_std_dev domain(Security)
   range(xsd:double))
  DataExactCardinality(1 sec_std_dev Security)

 Class(StockRate)
  DataProperty(date domain(StockRate)
   range(xsd:date))
  DataExactCardinality(1 date StockRate)
  DataProperty(price domain(StockRate)
   range(xsd:double))
  DataExactCardinality(1 price StockRate)
)
```

A portfolio (the Portfolio class) is characterized by a set of securities (securities attribute) contained in the portfolio, by several metrics: expected return (expected_return attribute), standard deviation (std_dev attribute), Sharpe ratio (sharpe_ratio attribute), risk free rate (risk_free_rate attribute), and ratio of positive tweets mentioning securities of the portfolio (twitter_positive_ratio attribute).

A security (the Security class) is characterized by identifier (ticker attribute), time series of historical prices (attribute rates), ratio of positive tweets mentioning the security (positive_tweets attribute), expected return (sec_expected_return attribute), and standard deviation (sec_std_dev attribute).

The workflow of the extended portfolio problem is demonstrated in Fig. 4. The workflow contains six tasks[1]:

(1) getPortfolios. A set of diversified portfolio candidates is computed. The multidialect task specification consists of two RIF-documents in BLD and CASPD dialects [1, Appendix]. Portfolios received as a result contain only security tickers, they have to be augmented by financial and sentiments ratios;

(2) getPositiveTweetRatio. This task is responsible for computing a sentiment ratio of tweets for every security. Every

[1]To save space, specifications are provided only for getPortfolios, getPositiveTweetRatio, and computePortfolioTwitterMetrics tasks.

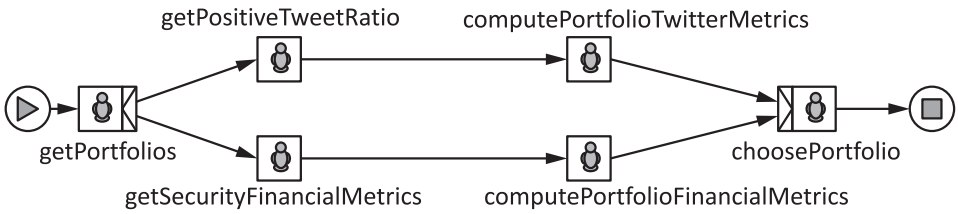


Figure 4 Portfolio workflow

tweet is assessed to be positive, negative, or neutral. The task is specified as a call of external function;

(3) computePortfolioTwitterMetrics. The portfolio sentiment ratio is computed as the average of its securities sentiment ratio. The task is specified using RIF-PRD;

(4) getSecurityFinancialMetrics. For every security in a portfolio the financial rates (the expected return and the standard deviation) are calculated on the basis of historical rates of securities specified as an OWL 2 class of the ontology of the application domain. The task is specified using RIF-BLD dialect;

(5) computePortfolioFinancialMetrics. The computation of the portfolio expected return, risk, and Sharpe-ratio is done within this task. The task is specified using RIF-PRD dialect; and

(6) choosePortfolio. The best portfolio is chosen according to maximizing the (*Sharpe ratio * sentiment ratio*) coefficient. The task is specified using RIF-PRD dialect.

Workflow skeleton is specified as a RIF-PRD document importing the ontology of the application domain:

```
Document( Dialect(RIF-PRD)
 Base(<http://synthesis.ipi.ac.ru/portfolio/
  workflow#>)
 Import(<http://synthesis.ipi.ac.ru/portfolio/
  ontology#>
 <http://www.w3.org/ns/entailment/OWL-Direct>)
Prefix(ont<http://synthesis.ipi.ac.ru/portfolio/
 ontology#>)
Prefix(ofws<http://synthesis.ipi.ac.ru/
 synthesis/projects/RuleInt/OpinionFinderWS#>)
Prefix(mws<http://synthesis.ipi.ac.ru/
 synthesis/projects/RuleInt/MediatorWS#>)

Group 2 (
 Do(
  Assert(External(wkfl:parameter-definition(
   startDatexsd:string IN)))
  Assert(External(wkfl:parameter-definition(
   endDatexsd:string IN)))
  Assert(External(wkfl:parameter-definition(
   bestPortfolioont:Portfolio OUT)))
  Assert(External(wkfl:variable-definition(
   ps  List<ont:Portfolio> IN)))
  Assert(External(wkfl:
   variable-value(ps List())))
 )
)
```

```
Group 1 (
 Forall ?sd ?ed such that (
  External(wkfl:parameter-value(startDate ?sd))
  External(wkfl:parameter-value(endDate ?ed))  )
( If Not(External(wkfl:
   end-of-task(getPortfolios)))
  Then
   Do( Modify(External(wkfl:variable-value(ps
    External(mws:getPortfolios(?sd ?ed) )))
   Assert(External(wkfl:
    end-of-task(getPortfolios))) )
 )

 Forall ?ps ?p ?scs ?s ?t such that (
  External(wkfl:variable-value(ps ?ps))
  ?p#?ps  ?p[securities->?scs]
   ?s#?scs ?s[ticker->?t] )
( If And( Not(External(wkfl:
     end-of-task(getTweets)))
   External(wkfl:end-of-task(getPortfolios)))
  Then
  Do( Modify(?s[positive_tweets->
   External(ofws:computeSecPosTweets(?t))] )
   Assert(External(wkfl:
    end-of-task(getTweets))) )
)

Forall ?ps ?p such that (
 External(wkfl:variable-value(ps  ?ps))
  ?p#?ps)
( If And(Not(External(wkfl:
   end-of-task(countTwitterMetrics)))
   External(wkfl:end-of-task(getTweets)) )
  Then Do(
   Modify(?p[twitter_positive_ratio->
    External(func:numeric-divide(
    Sum{?pt | Exists
     ?scs ?s(?p[securities->?scs]
     ?s#?scs  ?s[positive_tweets->?pt])}
    External(func:count(?ps))))])
   Assert(External(wkfl:
    end-of-task(countTwitterMetrics)))
) )) )
```

Production rules of the document are divided into two groups. The first group with priority 2 contains rules defining workflow parameters and variable. Input parameters are *start date* and *end date* of historical rates used for calculation of *portfolio metrics*. Workflow variable ps denotes a set containing *portfolio candidates*.

The second group with priority 1 contains the orchestration rules — workflow skeleton. The only orchestration rule provided in the example above corresponds to the task getPortfolios. The external function getPortfolios encapsulates a multidialect logic program calculating portfolio candidates [1, Appendix]. A Modify action is used to call the function and to put the returned result into the ps variable.

A.3 Revised portfolio problem infrastructure

The implementation structure of the use case is shown in Fig. 5.

The RIF-PRD workflow skeleton was transformed into a program (rule set) in the ILOG [21] language combining production rules and workflow facilities (like fork and sequence). The ILOG program was executed in the IBM Operational Decision Manager tool [22]. In order to execute ILOG programs, the underlying execution model (XOM) [23] was defined as a set of Java classes: Portfolio, Security, and StockRate. The Portfolio class corresponds to a financial portfolio and contains as attributes a set of securities in it, its expected return, standard deviation, Sharpe ratio, and twitter positive ratio. Code of this class is provided below:

```
public class Portfolio {
 private Collection<Security> securities;
 private double expected_return;
 private double std_dev;
 private double sharpe_ratio;
 private double twitter_positive_ratio;
 // as of 05.04.14 US 5-year treasuries
 private static double risk_free_rate = 0.0169;
 private boolean recommended;
}
```

Class Security corresponds to real world financial securities. The class contains as attributes a ticker, ratio of positive tweet number to the sum of positive and negative tweets, a set of stock rates, security's standard deviation, and expected return. These attributes are set as responses to corresponding web services queries:

```
public class Security {
 public String ticker;
 public double positive_tweets;
 public Collection<StockRate> rates;
 public double std_dev;
 public double expected_return;
 public static int number_of_periods = 5;
}
```

StockRate is a simple class and contains just two attributes — price and date:

```
public class StockRate {
 public float price;
 public String date;
}
```

It is easy to see that the one-to-one mapping exists between conceptual schema entities and execution model entities.

Parameters of RIF-PRD workflow skeleton (startDate, endDate, and bestPortfolio) are mapped into the respective parameters of ILOG rule set (Fig. 6).

The variable of RIF-PRD workflow skeleton (ps) is mapped into a local variable of the rule set. Specification of the variable looks as follows:

```
<?xml version="1.0" encoding="UTF-8"?>
<ilog.rules.studio.model.base:VariableSetxmi:
  version="2.0"
xmlns:xmi="http://www.omg.org/XMI"
  xmlns:ilog.rules.studio.model.base =
"http://ilog.rules.studio/model/base.ecore">
 <name>local_vars</name>
 <variables name="ps" type="java.util.ArrayList"
   initialValue=""verbalization="ps"/>
</ilog.rules.studio.model.base:VariableSet>
```

Rules of the RIF-PRD workflow skeleton are mapped into ILOG *ruleflow* [23]:

```
flowtask portfolio$_$flow {
 property mainflowtask = true;
 property ilog.rules.business_name =
```

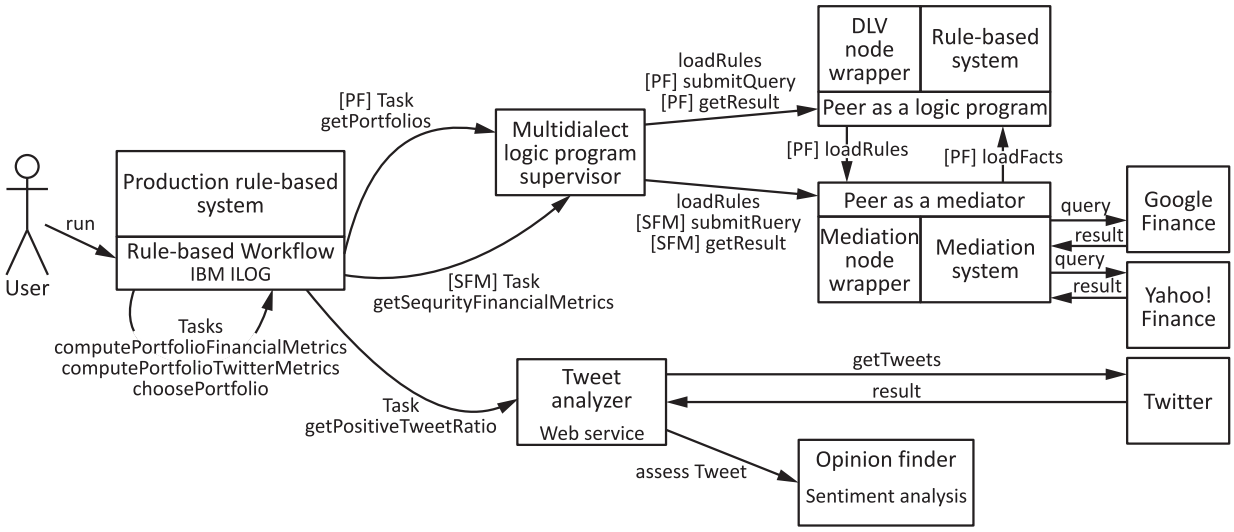


Figure 5 Portfolio problem implementation infrastructure

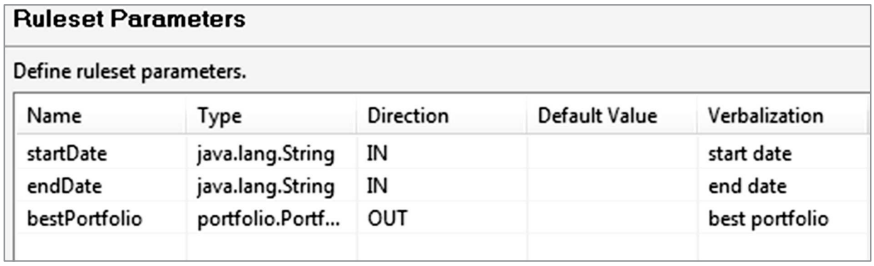


Figure 6 Rule set parameters

```
  "portfolio_flow";
 body {
  portfolio$_$flow#getPortfolios;
  fork {
   portfolio$_$flow#getRates;
   portfolio$_$flow
   #computePortfolioFinancialMetrics;} &&
  { portfolio$_$flow#getTweets;
   portfolio$_$flow#
    computePortfolioTwitterMetrics;}
  portfolio$_$flow#choosePortfolio;
 }
};

ruletask portfolio$_$flow#getPortfolios {
 property ilog.rules.business_name =
   "portfolio_flow>getPortfolios";
 body { getPortfolios.*}
};

ruletask portfolio$_$flow#
   computePortfolioTwitterMetrics {
 propertyilog.rules.business_name =
  "portfolio_flow>
   computePortfolioTwitterMetrics";
 body { computePortfolioTwitterMetrics.* }
};

ruletask portfolio$_$flow#getTweets {
 property ilog.rules.business_name =
  "portfolio_flow>getTweets";
 property ilog.rules.package_name = "";
 body {getTweets.*}
};
```

The computePortfolioTwitterMetrics, computePortfolioFinancialMetrics, and choosePortfolio tasks are implemented as production rules in ILOG:

```
package computePortfolioTwitterMetrics {
 use ps;
 import portfolio.*;

 rule computePortfolioTwitterMetrics {
  property status = "new";
  when { IlrContext() from ?context; }
  then {
   foreach (Portfolio p in ps) {
    double ?twitter_metrics = 0;
    int ?length = 0;
     foreach (Security security
       in p.securities) {
      ?twitter_metrics= ?twitter_metrics +
       security.positive_tweets;
      ?length = ?length + 1; }
     p.twitter_positive_ratio=
      ?twitter_metrics / ?length;
}}}}
```

The getPortfolios and computeSecurityFinancialMetrics tasks are implemented by the following production rules in ILOG:

```
package getPortfolios {
 use ps;
 import portfolio.*;

 rule getPortfolios {
  when { IlrContext() from ?context; }
  then {
   ps = Supervisor.getPortfolios(startDate,
    endDate);
} } }
```

Here, the Supervisor is the Java class wrapping execution of logic programs in multidialect infrastructure including two nodes [1]. The nodes correspond to the mediation system (which integrates *Google Finance* and the *Yahoo! Finance* services) and to a rule-based programming system DLV.

The getSecurityFinancialMetrics task uses the same instance of the mediation system as the getPortfolios task. The reason is that financial metrics are calculated using the historical rates of the securities. This is exactly the information that is extracted by the mediation system from Google Finance and Yahoo! Finance. The difference between two tasks is that the getPortfolios is implemented as a submission of a query to the DLV node, but the getSecurityFinancialMetrics is implemented as a submission of a different query to the Mediation Node.

The getPositiveTweetRatio task is implemented by the following production rule in ILOG:

```
package getTweets {
 use ps;
 import portfolio.*;

 rule getTweets {
  when { IlrContext() from ?context; }
```

Table 1 Metrics for the securities

Security identifier	Expected return	Standard deviation	Positive tweet ratio
COG	0.163	0.201	0.507
DO	0.015	0.019	0.651
EQR	0.150	0.022	0.846
FOSL	0.513	0.030	0.579
SCG	0.050	0.010	0.622

Table 2 Metrics for the portfolio candidates

Portfolio identifier	Expected return	Standard deviation	Risk free rate	Sharpe ratio	Positive tweet ratio	Sharpe ratio × Positive tweet ratio
1	0.111	0.008	0.0169	11.755	0.660	7.758
2	2.400	0.507	0.0169	4.701	0.508	2.388
3	2.381	0.508	0.0169	4.662	0.557	2.597
4	2.347	0.505	0.0169	4.606	0.708	3.261
5 (best)	0.178	0.011	0.0169	14.227	0.641	9.120
6	0.147	0.008	0.0169	15.577	0.521	8.166

```
  then {
   foreach (Portfolio p in ps) {
    foreach (Security s in p.securities) {
     s.positive_tweets =
      WebServices.computeSecPosTweets(s.ticker);
} } } } }
```

Here, WebServices is the Java-class wrapping invocation of a web-service. The WSDL specification of the web-service can be found at http://synthesis.ipi.ac.ru/synthesis/projects/RuleInt/OpinionFinderWS. The web-service, in its turn, encapsulates a Java-program. The program first collects tweets using the Twitter Streaming API. After that, a sentiment analysis is done by the Polarity Classifier of the OpinionFinder tool [24] which assesses if tweet is positive, negative, or neutral. Finally, the sentiment ratio for every security in a portfolio is calculated and returned as the result.

A.4 Result of the use case workflow execution

The results obtained by one of the use case runs are as follows. The task getPortfolios computes portfolio candidates on the basis of historical rates of daily closing prices of securities from S&P500 list for the 2011–2013. Six portfolios of size 5 were calculated. Each portfolio is a set of identifiers (tickers) of companies:

```
Candidate 1: { ALXN, BF.B, EW, POM, VNO }
Candidate 2: { BMC, JBL, LUK, MNST, POM }
Candidate 3: { AVP, BMC, JPL, MNST, POM }
Candidate 4: { ALTR, BF.B, BMC, DGX, PEG }
Candidate 5: { COG, DO, EQR, FOSL, SCG }
Candidate 6: { ADSK, GILD, INTC, POM, TJX }
```

The task getSecurityFinancialMetrics computes the expected return and the standard deviation for every security mentioned in portfolio candidates. The task getPositiveTweetRatio computes positive sentiment ratios for every security mentioned in portfolio candidates (500 latest tweets for every security were used for the computation). Financial and twitter metrics for several securities are provided in Table 1.

The task computePortfolioFinancialMetrics computes financial metrics for every portfolio candidate on the basis of respective metrics for securities in a portfolio. The task computePortfolioTwitterMetrics computes sentiment metrics for every portfolio candidate on the basis of sentiment metrics for securities in a portfolio. Financial and twitter metrics for portfolio candidates are provided in Table 2. The task choosePortfolio identifies the best portfolio by maximum value of the products of Sharpe ratio and positive tweet ratio obtained for every portfolio (see Table 2).

Acknowledgments

This research has been done under the support of the RFBR (projects13-07-00579, 14-07-00548) and the Program for Basic Research of the Presidium of RAS.

References

1. Kalinichenko, L. A., S. A. Stupnikov, A. E. Vovchenko, and D. Y. Kovalev. 2013. Conceptual declarative problem specification and solving in data intensive domains. *Informatics and Applications — Inform Appl.* 7(4):112–139. Available at: http://synthesis.ipi.ac.ru/synthesis/publications/13ia-multidialect (accessed December 9, 2014).
2. Kalinichenko, L. A., S. A. Stupnikov, and D. O. Martynov. 2007. *SYNTHESIS: A language for canonical information modeling and mediator definition for problem solving in heterogeneous information resource environments*. Moscow: IPIRAN. 171 p.
3. Boley, H., and M. Kifer, eds. 2013. RIF framework for logic dialects. W3C recommendation. 2nd ed. Available at: http://www.w3.org/TR/2013/REC-rif-fld-20130205/ (accessed December 9, 2014).

4. Boley, H., and M. Kifer, eds. 2013. RIF basic logic dialect. W3C Recommendation. 2nd ed. Available at: http://www.w3.org/TR/2013/REC-rif-bld-20130205/ (accessed December 9, 2014).
5. Heymans, S., and M. Kifer, eds. 2009. RIF core answer set programming dialect. Available at: http://ruleml.org/rif/RIF-CASPD.html (accessed November 5, 2014).
6. Leone, N., G. Pfeifer, W. Faber, T. Eiter, G. Gottlob, S. Perri, and F. Scarcello. 2006. The DLV system for knowledge representation and reasoning. *ACM Trans. Comput. Logic* 7(3):499–562.
7. DeSante, M. C., G. Hallmark, and A. Paschke, eds. 2013. RIF production rule dialect. W3C Recommendation. 2nd ed. Available at: http://www.w3.org/TR/2013/REC-rif-prd-20130205/ (accessed December 9, 2014).
8. Motik, B., P. F. Patel-Schneider, and B. Parsia, eds. 2012. OWL 2 Web Ontology Language structural specification and functional-style syntax. W3C Recommendation. 2nd ed. Available at: http://www.w3.org/TR/owl2-syntax/ (accessed November 5, 2014).
9. Calvanese, D., G. De Giacomo, D. Lembo, M. Lenzerini, A. Poggi, and R. Rosati. 2007. Ontology-based database access. *15th Italian Symposium on Advanced Database Systems Proceedings*. 324–331.
10. Ramakrishnan, L., and B. Plale. 2010. A multi-dimensional classification model for scientific workflow characteristics. *1st Workshop (International) on Workflow Approaches to New Data-Centric Science Proceedings*. New York: ACM. Article No. 4. 12 p. Available at: http://dl.acm.org/citation.cfm?id=1833402 (accessed December 9, 2014).
11. Boukhebouze, M, Y. Amghar, A.-N. Benharkat, and Z. Maamar. 2011. A rule-based approach to model and verify flexible business processes. *Int. J. Business Process Integration Management* 5(4):287–307.
12. Polleres, A., H. Boley, and M. Kifer, eds. 2013. RIF datatypes and Built-Ins 1.0 W3C Recommendation. 2nd ed. Available at: http://www.w3.org/TR/2013/REC-rif-dtb-20130205/ (accessed December 9, 2014).
13. Production Rule Representation (PRR), Version 1.0. OMG Document Number: formal/2009-12-01. Available at: http://www.omg.org/spec/PRR/1.0 (accessed November 5, 2014).
14. Yu, J., and R. Buyya. 2005. A taxonomy of scientific workflow systems for grid computing. *ACM SIGMOD Records* 34(3):44–49.
15. Kowalski, R., and F. Sadri. 2009. Integrating logic programming and production systems in abductive logic programming agents. *Web reasoning and rule systems*. Eds. A. Polleres and T. Swift. Lecture notes in computer science ser. Springer. 5837:1–23.
16. Cosentino, V., M. D. Del Fabro, and A. El Ghali. 2012. A model driven approach for bridging ILOG rule language and RIF. *6th Symposium (International) on Rules RuleML Proceedings*. CEUR-WS.org. 874:96–102.
17. Veiga, F. D. J. 2011. Implementation of the RIF-PRD. Universidade Nova de Lisboa. Master Thesis. Available at: http://run.unl.pt/bitstream/10362/6310/1/Veiga_2011.pdf (accessed November 5, 2014).
18. Markowitz, H. M. 1991. *Portfolio selection: Efficient diversification of investments*. Wiley. 402 p.
19. Sharpe, W. F. 1966. Mutual fund performance. *J. Business* 39(S1):119–138.
20. Bollen, J., H. Maoa, and X. Zeng. 2011. Twitter mood predicts the stockmarket. *J. Comput. Sci.* 2(1):1–8.
21. IBM WebSphere ILOG JRules Version 7.0. Online documentation. Available at: http://pic.dhe.ibm.com/infocenter/brjrules/v7r0/index.jsp (accessed November 5, 2014).
22. IBM Operational Decision Manager. Available at: http://www-03.ibm.com/software/products/en/odm (accessed November 5, 2014).
23. IBM Operational Decision Manager Version 8.5 Information Center. Available at: http://pic.dhe.ibm.com/infocenter/dmanager/v8r5/index.jsp (accessed November 5, 2014).
24. Wilson, T., J. Wiebe, and P. Hoffmann. 2005. Recognizing contextual polarity in phrase-level sentiment Analysis. *Conference on Human Language Technology and Empirical Methods in Natural Language Processing Proceedings*. Stroudsburg: Association for Computational Linguistics. 347–354.

Received November 3, 2014

Contributors

Kalinichenko Leonid A. (b. 1937) — Doctor of Science in physics and mathematics, professor; Head of Laboratory, Institute of Informatics Problems, 44-2 Vavilov Str., Moscow 119333, Russian Federation; professor, Faculty of Computational Mathematics and Cybernetics, M. V. Lomonosov Moscow State University, 1-52 Leninskiye Gory, GSP-1, Moscow 119991, Russian Federation; leonidandk@gmail.com

Stupnikov Sergey A. (b. 1978) — Candidate of Science (PhD) in technology, senior scientist, Institute of Informatics Problems, Russian Academy of Sciences, 44-2 Vavilov Str., Moscow 119333, Russian Federation; ssa@ipi.ac.ru

Vovchenko Alexey E. (b. 1984) — Candidate of Science (PhD) in technology, senior scientist, Institute of Informatics Problems, Russian Academy of Sciences, 44-2 Vavilov Str., Moscow 119333, Russian Federation; itsnein@gmail.com

Kovalev Dmitry Yu. (b. 1988) — junior scientist, Institute of Informatics Problems, Russian Academy of Sciences, 44-2 Vavilov Str., Moscow 119333, Russian Federation; dm.kovalev@gmail.com

КОНЦЕПТУАЛЬНОЕ МОДЕЛИРОВАНИЕ МУЛЬТИДИАЛЕКТНЫХ ПОТОКОВ РАБОТ*

Л. А. Калиниченко[1,2], С. Ступников[1], А. Вовченко[1], Д. Ковалев[1]

[1]Институт проблем информатики Российской академии наук
[2]Московский государственный университет им. М. В. Ломоносова, факультет вычислительной математики и кибернетики

Аннотация: Рассматриваются методы концептуального представления алгоритмов анализа данных, средств интеграции данных, а также процессов, направленных на спецификацию семантики данных и поведения в единой парадигме. Расширяется новый подход к применению комбинации семантически различных платформонезависимых языков на правилах (диалектов) для создания интероперабельных концептуальных спецификаций над различными системами на правилах. Подход опирается на методику трансформации программ на правилах, рекомендованную стандартом W3C Rule Interchange Format (RIF). Подход, предлагаемый в стандарте RIF, сочетается со технологией семантической интеграции неоднородных баз данных в предметных посредниках. Статья расширяет предыдущие исследования авторов в направлении моделирования потоков работ для определения композиций алгоритмических модулей в процессной структуре. Рассмотрены возможности спецификации задач в мультидиалектных потоках работ с применением семантически различных языков, наиболее подходящих для конкретных задач. Приведен практический пример потока работ, задачи которого специфицированы с использованием нескольких языков на правилах (RIF-CASPD, RIF-BLD, RIF-PRD). Для определения концептуальной схемы использован язык OWL 2, для оркестровки потока работ использован язык RIF-PRD. Инфраструктура реализации примера включает систему на продукционных правилах (IBM ILOG), систему на логических правилах (DLV) и предметный посредник.

Ключевые слова: концептуальные спецификации; потоки работ; RIF; языки продукционных правил; интеграция баз данных; посредники; PRD; мультидиалектная инфраструктура

DOI: 10.14357/19922264140413

Литература

1. *Kalinichenko L. A., Stupnikov S. A.. Vovchenko A. E., Kovalev D. Y.* Conceptual declarative problem specification and solving in data intensive domains // Информатика и её применения, 2013. Т. 7. Вып. 4. С. 112–139. http://synthesis.ipi.ac.ru/synthesis/publications/13ia-multidialect.
2. *Kalinichenko L. A., Stupnikov S. A., Martynov D. O.* SYNTHESIS: A language for canonical information modeling and mediator definition for problem solving in heterogeneous information resource environments. — Moscow: IPI RAN, 2007. 171 p.
3. RIF framework for logic dialects. W3C Recommendation / Eds. H. Boley, M. Kifer. 2nd ed. http://www.w3.org/TR/2013/REC-rif-fld-20130205/.
4. RIF basic logic dialect. W3C Recommendation / Eds. H. Boley, M. Kifer. 2nd ed. http://www.w3.org/TR/2013/REC-rif-bld-20130205/.
5. RIF core answer set programming dialect / Eds. S. Heymans, M. Kifer, 2009. http://ruleml.org/rif/RIF-CASPD.html.
6. *Leone N., Pfeifer G., Faber W., Eiter T., Gottlob G., Perri S., Scarcello F.* The DLV system for knowledge representation and reasoning // ACM Trans. Comput. Logic, 2006. Vol. 7. No. 3. P. 499–562.
7. RIF production rule dialect. W3C Recommendation / Eds. De Sante Marie C., Hallmark G., A. Paschke. 2nd ed. http://www.w3.org/TR/2013/REC-rif-prd-20130205/.
8. OWL 2 Web Ontology Language Structural Specification and Functional-Style Syntax. W3C Recommendation / Eds. B. Motik, P. F. Patel-Schneider, B. Parsia. 2nd ed. http://www.w3.org/TR/owl2-syntax/.
9. *Calvanese, D., De Giacomo G., Lembo D., Lenzerini M., Poggi A., Rosati R.* Ontology-based database access // 15th Italian Symposium on Advanced Database Systems Proceedings, 2007. P. 324–331.
10. *Ramakrishnan L., Plale B.* A multi-dimensional classification model for scientific workflow characteristics // 1st Workshop (International) on Workflow Approaches to New Data-Centric Science Proceedings. New York: ACM, 2010. Aricle No. 4. 12 p. http://dl.acm.org/citation.cfm?id=1833402.

*Работа выполнена при поддержке РФФИ (проекты 13-07-00579, 14-07-00548) и Программы фундаментальных исследований Президиума РАН.

11. *Boukhebouze M., Amghar Y., Benharkat A.-N., Maamar Z.* A rule-based approach to model and verify flexible business processes // Int. J. Business Process Integration Management, 2011. Vol. 5. No. 4. P. 287–307.

12. RIF Datatypes and Built-Ins 1.0. W3C Recommendation / Eds. A. Polleres, H. Boley, M. Kifer. 2nd ed. http://www.w3.org/TR/2013/REC-rif-dtb-20130205/.

13. Production Rule Representation (PRR), Version 1.0. OMG Document Number: formal/2009-12-01. http://www.omg.org/spec/PRR/1.0.

14. *Yu J., Buyya R.* A taxonomy of scientific workflow systems for grid computing // ACM SIGMOD Records, 2005. Vol. 34. No. 3. P. 44–49.

15. *Kowalski R., Sadri F.* Integrating logic programming and production systems in abductive logic programming agents // Web reasoning and rule systems / Eds. A. Polleres, T. Swift. Lecture notes in computer science ser. — Springer, 2009. Vol. 5837. P. 1–23.

16. *Cosentino V., Del Fabro M. D., El Ghali A.* A model driven approach for bridging ILOG rule language and RIF // 6th Symposium (International) on Rules, RuleML 2012 Proceedings. 2012. CEUR-WS.org. Vol. 874. P. 96–102.

17. *Veiga F. D. J.* Implementation of the RIF-PRD. Universidade Nova de Lisboa, 2011. Master Thesis.

18. *Markowitz H. M.* Portfolio selection: Efficient diversification of investments. Wiley, 1991. 402 p.

19. *Sharpe, W. F.* Mutual fund performance // J. Business, 1966. Vol. 39(S1). P. 119–138.

20. *Bollen J., Maoa H., Zeng X.* Twitter mood predicts the stock market // J. Comput. Sci., 2011. Vol. 2. No. 1. P. 1–8.

21. IBM WebSphere ILOG JRules Version 7.0. Online documentation. http://pic.dhe.ibm.com/infocenter/brjrules/v7r0/index.jsp.

22. IBM Operational Decision Manager. http://www-03.ibm.com/software/products/en/odm.

23. IBM Operational Decision Manager Version 8.5 Information Center. http://pic.dhe.ibm.com/infocenter/dmanager/v8r5/index.jsp.

24. *Wilson T., Wiebe J., Hoffmann P.* Recognizing contextual polarity in phrase-level sentiment analysis. *Conference on Human Language Technology and Empirical Methods in Natural Language Processing Proceedinhgs*. Stroudsburg: Association for Computational Linguistics, 2005. P. 347–354.

Поступила в редакцию 03.11.2014

INFORMATICS AND APPLICATIONS 2014. Vol. 8. Iss. 4. P. 125–136

AUTOMATION BEYOND WEB 2.0

A. Sorokin[1]

Abstract: This paper introduces a new approach to the analysis of information systems (IS) evolution based on a range of technological activities. The issue centres on the prospect that Web-driven IS will be expanded from business processes to other domains of activities. The classical approach by which automation eliminates bottlenecks in business processes does not work under these conditions. Current trends in information technologies (IT) increase the capability for Web integration that leads to new types of virtual systems that will create a new Web architecture, conditionally named a Web "spiral." The spiral type of integration on Web supported by integrated cross-industry solutions is more promising and effective in comparison with the "radial" ones. The paper describes this new class of IT systems.

Keywords: automation; business process reengineering; collaborative software; economies of scale; Internet topology; sociotechnical systems; systems of systems; virtual enterprises; Web 2.0

DOI: 10.14357/19922264140414

1 Introduction

Not much time has passed — on a historical scale — since computers came on the technology scene. However, the time and origin of invention and even authorship are still under discussion. Clearly, the improvement in computing is a process that combines a variety of ideas, technologies, and drivers. In Japan, for example, according to some sources, computing after the war had tight links with the emergence of telephone switchboards (http://museum.ipsj.or.jp/en/computer/dawn/0005.html).

In the U.S., one of the motivations for searching new ways of information processing was difficulties in working with clerical card indexes in the first 30 years of the XX century. In the USSR, the first prototypes of computing technology were closely associated with defense programmes and space research and development (R&D). After its birth, computing technology began to show not only rapid development, but also the ability to penetrate virtually all spheres of human activities, going far beyond the range of tasks originally referred to. With the emergence of local area network (LAN) and Internet, computing evolved along with telecommunications originating the term "information and communications technology" (ICT).

Current ICT development reflects the issues that are relevant to the professional roles of specialists involved. Information managers view as dominant the phenomenal data avalanche that has to be overcome with the help of high performance parallel computing. The IT architects are concerned about the complexity of systems integration. Business analysts are trying to cope with the growing number of approaches and notations defining business process modeling.

Narrow professional observations on the issues lead to multiple conclusions in which direction the Web of the Future will go. The center of gravity of such observations, if focused on the technical qualities of new IT, are considered here in comparison with each other, in isolation from the economy and lifecycle (LC) environment in which they operate.

Dominant approaches are looking like a kind of technological Darwinism, most characteristically expressed by Gartner's hypercycle of emerging technologies [1]. Meanwhile, automation has firmly implemented in the overall human environment, the conditions of existence which are regulated by known factors, such as economic crises, resource depletion, population growth, and many others. The IT fashion is changing rapidly. In one to two years, new bright idea is highlighted. One can endlessly study trends whilst more practical question stays in shadow — what artifacts will appear at their crossroads?

A more general conception for describing the evolution of IT systems has been proposed by IBM Almaden Lab. They noted the marked complexity of models and objects, which designers of modern IS face today. The concept is presented in the form of a new Science of Services, Management, and Engineering (SSME) aimed at developing design techniques for highly complex objects. Stakeholders may include large organizations, cities, and even whole states. (http://campustechnology.com/articles/2009/04/13/ibm-and-higher-ed-push-for-a-smarter-planet-with-ssme-curriculum.aspx).

Viewing IT evolution from the SSME perspective, one may come to the conclusion that a classical ap-

[1]IBM EE/A, 10 Presnenskaya Nab., Moscow 123317, Russian Federation

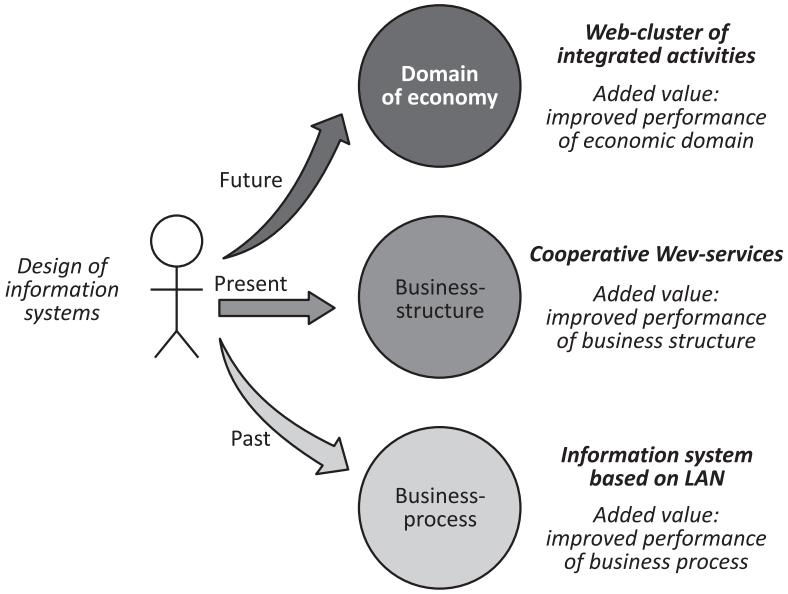


Figure 1 Evolution of automation scale and impact criteria

proach to IT system design, based on business processes analysis, does not work here at all, or must be extensively revised. The reason lies in the multiplicity of such processes that must be simultaneously analyzed and automated. Additionally, advanced complex IT systems automate not one or a few business processes but whole domains of activities in general. In such cases, one is faced with a large number of processes and technologies embedded in living environments; so, each of them needs to be repeatedly changed, removed, or substituted. This leads to the necessity of an initial analysis of the domain itself and, second, the related business processes, implying transition to a meta-design approach, and the modification of basic ICT system design, scaling to *domain* of activity. Under the term 'activity,' the bunch of connected business processes, covering professional areas (domains), supported by groups of technologies with the purpose of improving economic efficiency is considered. Hence, ***automation and virtualization*** in this work are regarded as group technologies for the reproduction of the artificial environment with the purpose of significantly improving business performance.

In this paper, the system and design aspects of future Web evolution are investigated that will undoubtedly affect the architecture and LC of IS. Below, ***the stack of activities*** is introduced which analysis can help to make some predictions on ICT development in the near future. Figure 1 shows automation changes in terms of scale and added-value criteria.

In this paper, the present author will try to find some answers on described above issues considering Web as self-organizing, self-sustaining, and evolving system with multiple direct and feedback connections between domains.

2 Lifecycle Ecosystem

An area of activity (domain) as functionally homogeneous set of business processes and technologies appear in division of labor is considered. Quantity, composition, and relationship between such activities form products LC and environment that represent a model of technological expansion. It uses the principle of procreation, when the original activity generates the following one as a result of internal conflict that limited its capacity (Fig. 2).

First of all — LC realization is impossible without resources. Therefore, the root domain will be 1. "Access to resources" activity. This domain is denoted by number one, and main types of resources that are consumed by the technology community create a set of subdomains. Let list them in a logical sequence — from basic to the more complex ones appearing later:

1.1 Data (information resources) collection.

1.2 Natural resources extraction.

1.3 Finances.

1.1 "Data collection." This is the root activity in number of basic types playing fundamental role in every human's life and society in general. Probably for this reason, the history of automated systems started with the automation of information processes. Since that and going on, automated systems are called as "information." Without this subdomain, the very existence of

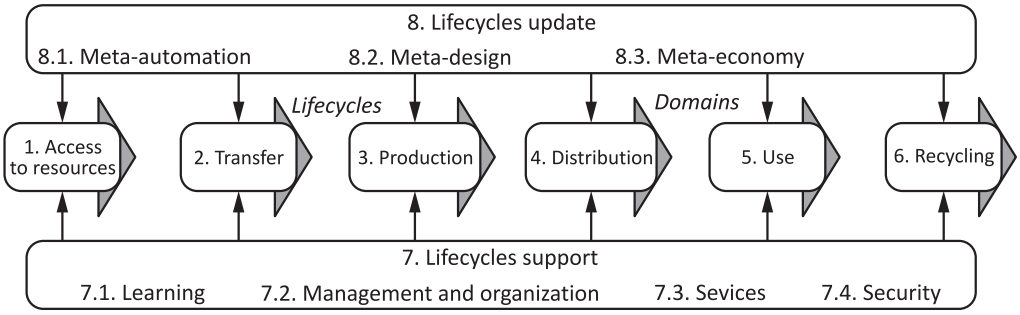


Figure 2 LC Ecosystem

society is impossible. Initial stage here is observation. After that, people are beginning to measure the observed phenomena. Next stage is introduction measures of weight, distance, and time. Then development moves towards the inventions of better tools to collect information. Thus, new tools and correspondently domain states appeared: microscope and telescope, photography, microphone, audio recording, telephone, x-ray, filming, radar, video, electronic microscope, sensors and telemetry, space probe, databases, search engines, scanner, geopositioning, global information systems. Each tool opened new and more powerful opportunities for high quality information activities.

1.2 "Natural resources extraction." To implement new technologies and inventions, natural resources are needed. They determine the place of this activity. Stages evolve in the direction from the exploitation of the biosphere and then — extraction and use of minerals and metals (mining), extraction of hydrocarbons, uranium enrichment. The sequence of stages is moving from easy to more difficult availability and after that — to creation of artificial resources, recycling, and meta-materials [2].

1.3 "Finances." Recourses evolution also came from natural to artificial and led to emergence of money as more general and valuable recourse. Milestones indicating the progress of this type of activity are: introduction into circulation of money substitutes, coinage, emergences of usury and banks, invention of paper money, introduction into circulation of securities to financial markets, globalization and automated support of financial markets.

Intra conflict: resources cannot be used instantly at the place of extraction. This conflict is an origin of new type of activity — 2. "Transfer." Subdomains in this area are also ordered by complexity:

2.1 Information channels.

2.2 Material channels (transport, pipelines, energy lines).

2.1 "Information channels." Stages that mark the domain expansion form a sequence starting of personal contacts between people, and further on — postal service, electrical transmission of discrete information (telegraph, teletype), electromagnetic analog transmission (telephone, radio), digital communications, computer networks, satellite repeaters.

2.2 "Material channels." Transfer of resources absorbs achievements of related domains and is developing toward modern supply chain. They provide movement of a large number of people, transportation of goods, raw materials and energy. Stages marked expansion in this domain are: road construction, use of natural ways (river and sea), the emergence of railways, air transport, container transport, and, finally, modern automated supply chain.

Intra conflict: inability of immediate use resources by end user. To improve this, they must be processed, which generates the next logical domain — 3. "Production." Its subdomains:

3.1 Power production.

3.2 Production of goods and services.

3.1 "Power production." Previous activity of extractive industries makes basis of energy production. On qualitative scale of stages, one may mark use of muscular energy as a starting point and further on, opening of fire, use of wind and water energy, steam energy (in the beginning of the industrial revolution), electricity, nuclear energy, alternative energy sources in a modern, high-tech version, and, finally, thermonuclear (forecast). Evolution of this domain makes it possible mass productive activity in the next one.

3.2 "Production of goods and services." Initial milestone of this activity is served to satisfy primary needs with the help of manual production of food, clothing, and footwear. Next step is providing of services. Then goes the chain of stages: deployment of production of consumer goods on a commercial scale, industrial and residential construction, heavy machinery, entertainment business, high-tech manufacturing of communications and computing equipment.

Intra conflict: manufactured products and services must be delivered to consumer. This function is outside

this domain, which generates the following area 4. "Distribution." Communities cooperate with each other by information, services, and products exchange they produced as objects of trade. So, according to this logic, trade became the following domain of activity in the constructed technologic stack. It has evolved from a simple barter towards the emergences of money trade, after that — major trading houses, laying large trade routes and distribution channels, unification of production, transportation, and sales businesses, global international trade, sales through communication channels and networks (television, mail, catalogs, and over the Internet).

Stages of this area include: barter, monetary trade, wholesale, trading networks, integration of trade and production, global trade, and trade through communication channels.

Intra conflict: Customers are buying goods which quality is entirely determined by the manufacturer. This may not comply what customers are really need. In domain evolution, certain number of compensational tools was born and developed: advertising, marketing, etc. However, these tools are not always fully able to suppress this conflict manifestations. Flexible production is a way to solve it getting feedbacks and experience from products use history.

The fifth domain is "Use." Its subdomains, as well as subdomains of all other activities arranged by the principle "from simple to complex" are taken here from classical Maslow pyramid [3]:

5.1 Satisfaction of physiological needs.

5.2 Safety needs.

5.3 Social needs realization.

5.4 Esteem support.

5.5 Realizing of personal potential.

Intra conflict: produced items with time become morally and physically outdated and must be recycled.

6. "Recycling" — next and the last activity in the row of domains that form LC chain of activities. Its implementation became more difficult with increasing scales of production, using modern synthetic materials, pollution, and many other factors.

Main stages: indiscriminate dumping of waste; disintegration technologies; and recycling and reuse.

Intra conflict: environmental pollution. It can be solved by sending of extracted reusable resources into initial first domain with cross-domain feedback.

Technology development makes LC more complex, and an inherent intra conflicts increase their vulnerability. Compensatory mechanisms bring to life two other large domains 7. "Lifecycles support" and 8. "Lifecycles update."

7. "Lifecycles support" includes business processes and technology that support existing industries and all domains included in LC ecosystem. It has four subdomains:

7.1 "Learning."

7.2 "Management and organization."

7.3 "LC services."

7.4 "Meta-security."

7.1 "Learning." Information shared through communication channels has to be analyzed. Analysis results in the new knowledge. Large number of professions is associated with the knowledge accumulation and sharing. Development is moving from primary artifacts in the form of oral tradition, education, and then, after invention of writing, towards book printing, calculation, research, establishment of educational and research institutions, lecturing by radio and television, knowledge bases/expert systems, computer-aided learning systems, social networks, virtual universities, and, finally, to virtual labs and universities. However, knowledge consumption must be effective and purposely managed. Later stages include knowledge socialization where new role functions appeared: leaders, experts, facilitators, and others that support the process of knowledge exchange in networked social groups. They also need to be managed.

7.2 "Management and organization." Every domain has its specialties in division of labor processes. It increases performance of society in general. On top of this division, management plays special role to improve effectiveness of every domain by better organization and resources utilization.

Initial stages here are: management of row information, communications management, and knowledge consuming management. Then, path of evolution goes to emergence of organizational skills, processes management, organizations structures management, material objects control (tools, machines), asset management, and territories and global structures governing (states, transnational corporations).

Improving social performance by organizational means quickly achieves the limits where performance stops growing. To complete the mission of forth domain, additional instruments are needed. Technology, innovations, and inventions are these tools. Needs in their development are calling for life next area of activity.

7.3 "LC services." This subdomain of meta-activity provides B2B (Business-to-Business) is services that differ than B2C (Business-to-Customer) serviced which are producing by subdomain 3.2. Technology expansion is growing from repair on demand to subscription for services, CALS (Continuous Acquisition and Lifecycle Support), web-services for LC updating, and embedded self-services.

7.4 "Meta-security." Very important subdomain that defends all other activities from various threats. Collapse and damage to any of its elements are fraught with losses and even economic or social disasters. That is why, all other domains contribute to its maintenance, and, in its turn, domain provides a feedback to all other stack's elements. Milestones and specific implementations are: the emergence of the concepts of society, property, including intellectual property, and from here onwards — personal rights and duties, emergence of law and legislation, tools for protection of family and possessions, health, judiciary, police and security guards, armed forces, international law, and cyber defense.

Intra conflict: multiple LC, heterogeneous technologies and manufactured products cannot be effectively organized by supporting tools. Quality of training, maintenance, management, and security measures often follows by incidents that show backlogs off requirements to LC sustainability. This demands permanent improvements and innovations. For such purposes, the ecosystem of LC evolution contends 8th domain — "Lifecycles update" that is dedicated to modernize, renovate, optimize, and integrate LC.

Unlike "Lifecycle support," this domain's technologies replace old LC by the new ones or, at least, introduce new elements and provide LC adjustment and optimization. It also contents three main subdomains:

8.1 "Meta-design."

8.2 "Meta-automation."

8.3 "Meta-economy."

8.1 "Meta-design." In context of this paper, the term is associated with innovative and inventive activity, the result of which is modernization and replacement domains (1–6) with new and modern components. It also means design of LC, their integration and optimization.

Accumulated knowledge as well as managerial and organizational skills allow invent and produce tools of increasingly perfect design on industrial scale. Its creative stages and expansion phenomena in some senses repeat the picture of development in "Management" for current domain's purpose and is also directed to improvement of technological culture. Very important inventions were made in basic domains starting progress with inventions in data collection (computer images that represent designing objects, etc.), communications (mapping, navigation instruments, etc.), and knowledge accumulation (writing, printing, etc.).

Further evolution goes through inventions of organizing technologies, towards specialization of industries and crafts, drawing, technology development and manufacturing processes of these objects, production of technology equipment, automation, and virtualization of manufacturing.

8.2 "Meta-automation." Unlike single IS implementations, "Meta-automation" belongs to a class of "system of systems" for it affects not only separate business processes but the parts of the whole domains. It can be regarded as an instrument for domains efficiency management; however, unlike economy tools, "Meta-automation" makes it with different means.

Initially, automation pasts the simplest form allowing reduction of LC cost by replacing manual labor. Then, it goes in a way of creation an artificial environment that increases business processes performance in all others domains. That is the reason why evolving stages of automation repeat sequence of described human activities expansion. Evolution started with the first milestone that represents the concepts and initial prototypes of IS and databases, which at one time were considered as main computer applications for indefinitely long term. According to some sources, in 1962, American company System Development Corporation first coined the term database (http://en.wikipedia.org/wiki/Database).

Next milestone was circuit switching. ARPANET project where packet switching was first implemented in communications between computers became the third one.

This key technology opened the era of networking. Next is a milestone that marks the period when efforts were focused on management control system (MCS) later evolved in ERP (Enterprise Resource Planning). Boom of CAD/CAM (computer aided design / computer aided manufacturing) evolved in CIM (computer integrated manufacturing) and factory automation made the following states of development. It produced complete systems including not only computer graphics, but robotic workstations, information retrieval systems, plotters, and various versions of LAN. Seventh milestone: noticeable progress in simulation modeling of economic processes and the development of computer models of the economy.

Communication protocol TCP/IP (Transmission Control Protocol / Internet Protocol) opened new era. Speed and scale of Internet as unprecedented in human history monster artificial environment mean a new phase of automation — spread or even absorption of all kinds of human activities by sociotechnosphere. With the implementation of Internet technologies, sociotechnical systems, concepts of which appeared in the 1960s [4], nowadays reached a new level. It leveraged by deep penetration of IS and all kinds of technology in social and organizational structures, as well as the quality of innovations. The influence of this phenomenon will be discussed in the following sections.

8.3 "Meta-economy" is a toolkit of economic regulation not inside but under LC. Activity that determines behavior of industrial units.

Intra conflict: similar to the case of "LC support" domains, the contradictions lie in multiplicity of LC that make it difficult to effectively coordinate all instruments of updating and innovations. Predicted solution is the development of Meta-design technologies to leverage LC modernization.

It should be noted that both domains (7 and 8) have also an external additional conflict between them since stability and modernization are inherently contradictory concepts. For example, implementation of advanced and saving car engines was periodically hampered by oil-producing companies interested in increase consuming of petroleum for internal combustion engines. Nevertheless, economic conditions tightening shifts the equilibrium between the domains in favor of modernization and forces manufacturers to accelerate the transition to new LC design. Measures taken to energy saving in the current crisis stimulated the development of "green" technologies and production of economic components for electronics and lighting.

3 Technologic Stack of Activities

Stack of activities is a model for presenting a set of key areas of lines of business (LOB), or business domains, in logical connection between the individual domains. In definition of "activity" which was introduced above, two entities are presented — the set of business processes and a variety of technologies. For this reason, one can build up two types of stacks — first, the processes stack and, second, the technological stack of activities. This couple gives an overall model for domain description. However, in this paper, we, in the first turn, are interested in IT analysis and, therefore, should focus on building a technological stack. Then, it will be applied to exploration of Web evolution.

Theoretical model of technology stack is N-dimensional. In fact, the stages also have their history, description, and content parameters, which imply the possibility of including additional elements called states. So, three-dimensional (3D) stack will include a description of each stage decomposed by states. For example, for stage "construction industry" in domain "Production of goods and services" (8 → 8.4), additional chain of stage decomposition will appear: 8.4.1 "Use of natural shelters" → 8.4.2 "Construction of stone and wood" → 8.4.3 "Construction of the man made materials" → 8.4.4 "Use of 3D printers."

Each state may be subjected to further decomposition, etc. until the desired degree of granularity is obtained. For studies in local business areas, it is possible to make slices of this model. For the immediate objectives of the present study, the two dimensions of model are enough and will not sufficiently distort the results.

Stack element (domain or milestone) affects others with *direct links* and *feedbacks*. Direct links mean that progress could be measured by emergences of new milestones in higher positioned domains. Feedbacks are measured by the progress observed in previous domains.

When direct links and feedbacks form a *loop*, it gives an enhanced effect of activities interaction often described as a "*revolution*." Thus, feedback from banks and users capital (activity "Finance") to the activity "Production of goods and services" and backward made a loop and gave direct effect in the creation of capitalist industry (the first industrial revolution) and a modern market economy. Direct links from activity 7.1. ("Learning") to 8.2. "Meta-design" led to invention of computer as a mean of information processing. Implementation of computing in lower domains in all subsequent forms provided an effect, often defined as the *second industrial revolution*.

Stages of technology evolution form the space that may be defined as "Space of Technology" (Fig. 3). Technologic stack enables one also to fix the level of social and technology development. If one draws a line across the selected milestones, it will be a line of development level.

The lower level of development crosses the initial stages of domains and forms the foundation of "Space of technologies." Higher stages of domains mark the front of development in technological space.

North-West corner of built technologic space covers basic technologies taken from nature and natural analogs. Large scale automated systems are concentrated in the South-Eastern corner. They leverage technologically closed communities with artificial internal rules of existence.

Direction from the North-West to the South-Eastern corner figuratively plays a role of vector of development, on which the most crucial inventions, technologies, and discoveries are located.

4 Special Role of Automation in "Space of Technologies"

Mentioning this special role of meta-automation in shaping of technology, let make the closer look on evolution of this domain. Web is the space for interaction. Men, machine, and system are the general parts of such interaction between them in IT environment. Let also fix three modern types of Web environment in which interaction is taking place: user centered, machine dominated, and systems integrated.

Computing platforms and algorithms also play very important roles. But in this sense, these roles supporting and computing evolution will go in the direction to satisfy

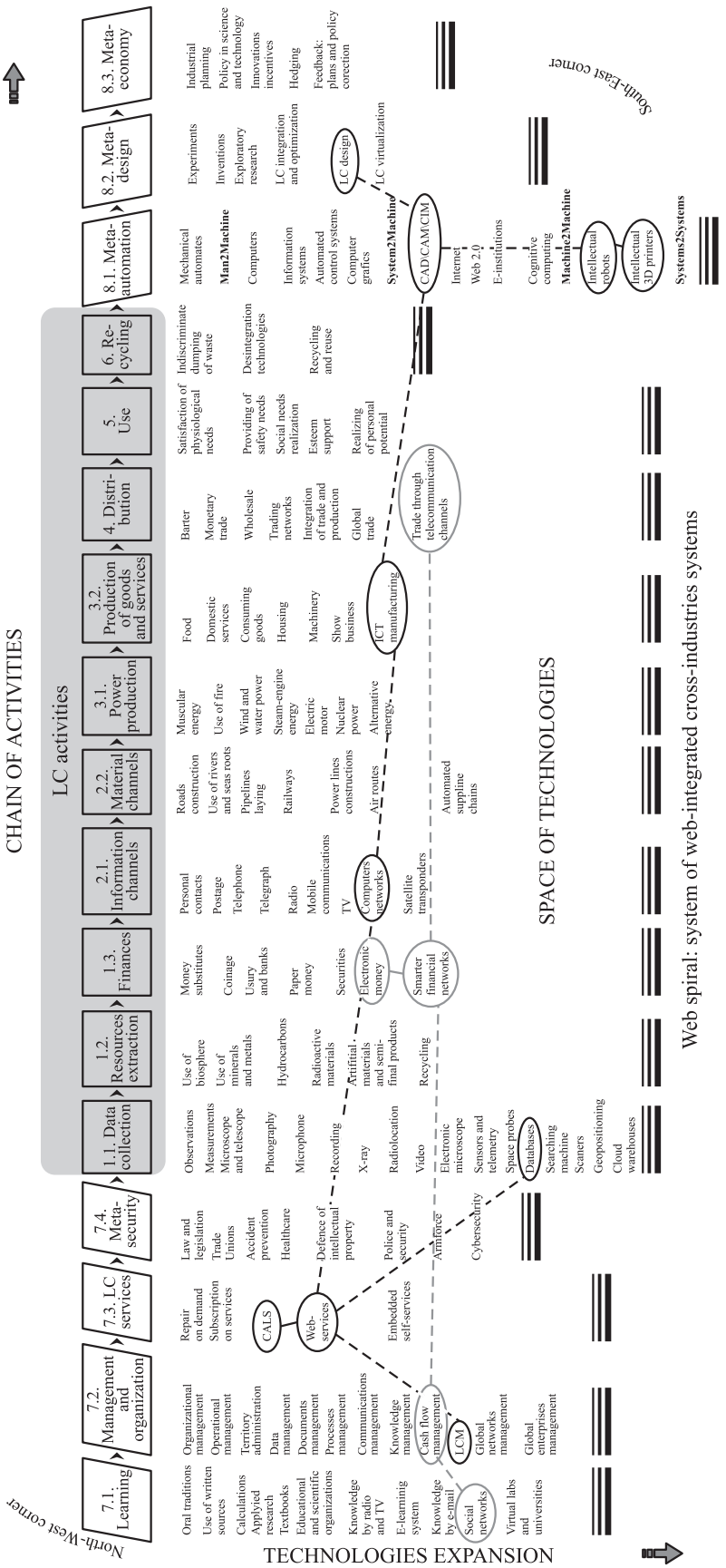


Figure 3 Technologic stack of activities

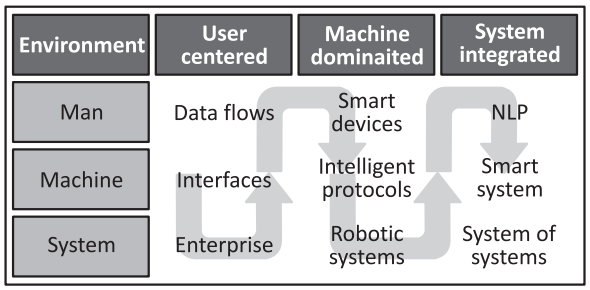


Figure 4 Basic IT classes supporting interaction in Web

the growing demands coming from further development of interaction needs. Hence, let build classification that presents generic classes of automation that support Web interaction using these 6 entities (Fig. 4).

Constructing this classification, the rules introduced early have been followed: classes evolve from simple to sophisticated — from left higher corner to the right low one. And each higher class may absorb and use all technology solutions of lower classes. And the present model is 3D that gives possibilities for classes to provide new artifacts as vertical components of the model. Interaction tools between human and user centered environment set up the initial class. It is *data flow* supporting technology — e-mail, e-messengers, Skype, and so forth.

Second class contents the means for intercommunion among machine and user centered environment. They are *Interface* design technology and examples include standard API (application programming interface), speech-gesture interfaces, Google glasses, and so on.

The next class is *Enterprise*. This type of technology intermediates between work stations (user centered environment) and systems. Commercial ERP systems also satisfy this class definition.

Then, class with *Smart devices* is going. It helps one to operate in machine dominated area: smart houses, machine control with embedded processors, etc.

Evolution of machine–machine interaction is based on intelligent protocol class of technology. The examples are SDN (software defined networking), intelligent navigation software.

Systems managing machine class demands use of robotic conception. Artificial intelligence mechanisms must be embedded in drones (UAV — unmanned aerial vehicle)), industrial and military robots for their navigation, interaction and mission execution.

Next step in Web interaction evolution is emerging of the new technology class that permits human to solve complex tasks in system integrated media by use of natural language (NLP — Natural Language Processing). This class does not include relatively simple voice recognition technology related to interface class. It has a deal with Q&A (questions and answers) system, artificial intelligence for decision-making, advanced training systems, etc. IBM Watson is a good example of commercial system that occupied this position.

Further, let come to *Smart systems* class of interaction technologies permitting any machine to be integrated and interact in systems environment. This class is a child of fast developing cloud technology, social networking — from computing side, and demands to smart consuming of resources — from side of economy. Examples embrace IBM series of Smarter Planet system that will be considered in more details later on.

The last, ninth, and the highest class is *System of systems* technology. The most spectacular example of this type is Web itself. Applying the famous analogy of Internet and real web made by spider — a wonderful example of natural stress-resistant construction — it is possible to answer the following question: What quality of our IT Web that is the greatest artificial system in human history, provides its integrity instead possible fragmentation with increasing complexity?

In the model shown in Fig. 3, one may recognize "radial" "filaments" of WWW (world wide web) representing industrial/LOB directions of automation technologies development. Cross-industries connections of automation milestones shown there represent "spiral" way of integration. Cloud computing, social networks, and M2M (machine-to-machine) technologies open practically unlimited prospects for such integration. Projects of global free wi-fi like Outernet (https://www.outernet.is/) have to accelerate movement toward "spiral" at the humanitarian (healthcare, education, etc.) parts of it. Some assessments show that 15 billion devices will be connected by 2015 (Intel) in Internet of Everything (IoE) that will cover of $14.4 trillion (Cisco) [5]. One may foresee that further development of AHCP (Ad-Hoc Configuration Protocol) inside *intelligent protocol* class will open the possibility of *smarter integration* when IT platforms of "spiral" systems could "negotiate" with each other and install proper ad hoc configuration. Large potential of *intellectual integration* is opened by IBM Watson. With ability to deep intelligent search of absent and necessary system's components based on artificial intelligence, Watson computer will be very useful assistant in design of "spiral" systems.

Mentioned technologies produce powerful synergetic effect for automation development providing WWW as platform for further deployment of extralarge complex systems, predicted above as class 9 (see Fig. 4). Growing in scale and accumulating innovations taken from different domains, such types of systems will reinforce total WWW performance and lead to new web topology which is similar to weaving of web spirals. One may call it "spiral" architecture, employing mentioned parallel. And, correspondingly, such future systems will be called as Web "spiral" systems.

In meta-automation expansion scenario outlined here, the described 9 classes of interaction technologies are used as reference solutions that determine general direction of developments.

Each domain produces technology that tends to expand into others domains "territory." As a result of expansion, communication technologies are observed everywhere, e-learning is everywhere. Power and transport are also penetrating in every part of technological space. It will be shown that automation like each other activity enjoys this ability but beside this, it also possesses some special features that make it meta-activity.

As was noted above, each domain may be decomposed into subdomains. Technological breakthroughs corresponding to each of them can be represented as innovative milestones that determine "vertical" or "radial" directions of technology diffusion. New milestone also provides impacts to neighborhood environment in horizontal directions through direct and feedback connections. As a result, all of the domains in varying degrees contribute to the technological development of each activity. In general, it looks like the acceleration of scientific and technological progress. Automation follows the innovations in all directions but sometimes precedes and accelerates them.

Taking as an example, breakthrough in nanotechnology with the graphene invention (8.2. "Meta-design") leads to implementations into domains 3.1 (rechargeable battery, solar panels), 3.2 (processors), 1.1 and 7.4 (sensors), etc. Collaborative works in these areas performed onto IT integrated platform will be more productive than process of "natural" technology diffusion.

Thus, automation as a special kind of activity, from the one hand, by means of innovations is tightly connected with Meta-design domain and from the second — is a tool for managing of economic performance set by "Meta-economy." Automation fulfills this role in two ways. First — in labor allocation scheme it transfers to machine only those operations that man performs worse than computer. The second way is integration that adds value for multitude of enterprises helping them to manage common data flows and effectively use their huge computing resources. That, in fact, makes it possible to combine stack elements to obtain economic benefits in new models of entrepreneurship.

The innovative quality of automation provides results and artifacts that cannot be achieved without automation. The most famous examples are robots, 3D printers, and calculations of satellites flights trajectories. Cognitive computing, cloud technology, and M2M are taking out automation beyond the role of just service tool that helps to solve different tasks inside LC processes. With latest advent, automation obtains the ability not just redesign and improves LC themselves but to invent new LC for new products. In accordance with such new quality, this activity can be called the "Meta-automation" as a part of "Update of lifecycles" domain, ensuring LC design, renewal, and integration.

5 Steps to "Spiral" Systems

Let consider main properties of the "spiral" systems. First of all:

- "Spiral" is a "system of systems" component of Web architecture and more advanced form in evolution of Web hosting and data centers;
- "Spiral" system belongs to different owners and supports multitenant mode;
- "Spiral" integrates self-organized communities: emerge collaborative projects and maintain and implement their outcomes with the help of rented instruments and resources;
- "Spiral" system projects management is conducted in a virtualized environment (domain "Management and Organization") and also may attract social lending alongside with traditional financial sources;
- "Spiral" integrates, first, larger parts of LC: information and research, economics and finance, and design and production as secured distributed clods; and
- "Spiral" platforms predominantly provide utility computing in the mode of EaaS (Everything as a Service).

Deployment corresponding architecture of Web brings forth the task of new design methods or even new approaches to Internet topology.

The constructed model allows to look at automation with more general civilization positions, defining its place and role in overall progress.

For qualitative assessment of possible value for synergetic effect of automation, it is necessary to involve some economic considerations.

IBM Institute for Business Value published a study model of the world economy representing 11 core systems [6] (Fig. 5). These core systems include: Infrastructure, Electricity, Finance, Education, Food, Communications, Water Resources, Transportation, Entertainment, Fashion, Leisure, Health, Governance, and Security. In Fig. 5, each core is represented by a bubble whose value is proportional to the size of the economy. For scaling, a bubble of 1 trillion dollars is shown in the lower right corner. The model has the fractal properties. Production, business, IT systems, engineering, energy resources, materials, and trade are incorporated in each core system as components. In many cases, the component's name and the name of core system may be the same. The components inside cores are also indicated by scalable bubbles. Arrows,

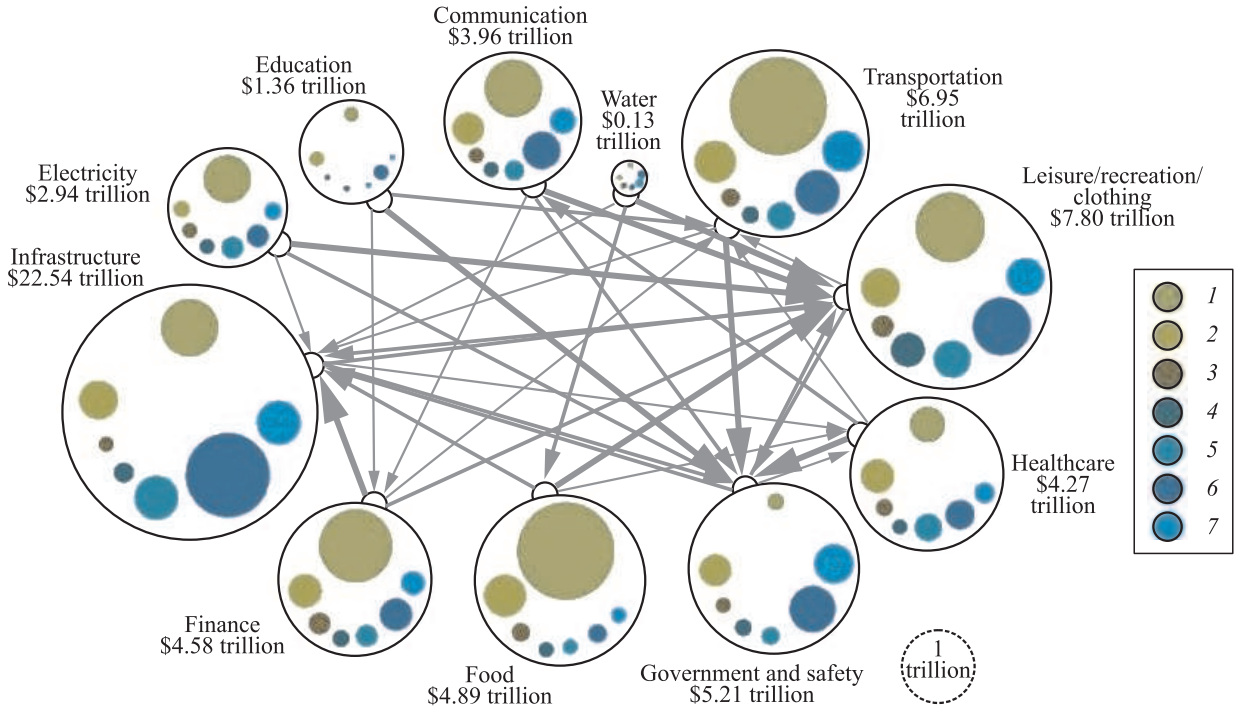


Figure 5 World economy model according IBM Institute for Business Value: *1* — same industry; *2* — business support; *3* — IT systems; *4* — energy resources; *5* — machinery; *6* — materials; and *7* — trade

whose width characterizes the degree of influence of one core system to another, show contribution of all core systems in the operation of each separate one.

IBM analysts based on regression economic models argue that the total loss of world economy is estimated as 15 trillion dollars resulted due to inefficiency. They also found that the loss of at least 4 trillion dollars could be prevented through more rational organization and use of modern information technologies. Certain part of such losses occurs at the level of "system of systems" as poor integration between the cores. Executive Report made by Korsten and Seider was published by IBM in 2010 in the period when global economic crisis has begun [6]. At that time, IBM developed and presented the "Smarter Planet" initiative [7] as a technological response to the crisis. The purpose of this initiative was to increase value delivered to the end users with the help of "smart" IS aimed at the automation of complex activities that approximately correspond to system cores in the economic model described above. Such systems were labeled as "Smarter Grid," "Smarter Healthcare," "Smarter Finance," "Smarter Work," and so on.

Their design should provide cost savings and effective business processes in the target domain (power production, customer services, water resources, etc.) for interdomains integration and building of heterogeneous systems at a very broad scale. They can be considered as the first steps to Web-"spiral" systems of systems (class 8, see Fig. 4).

Platforms of this series are different in structure, but have many common characteristics:

- high degree of security provided by software products of IBM Tivoli family;
- dynamic infrastructure supported by cloud technologies;
- meta-models and tools for business processes management that make possible quick adaption to changes in business environment;
- ability to accumulate knowledge and provide access to cognitive resources with help of social tools;
- high scalability and mobile access to services; and
- good performance records (cost savings and use of "green technologies" reducing total cost of ownership).

Platforms presented in LC ecosystem (see Fig. 2) are: Smarter SCADA for Oil and Gas (1. Access to resources), Smarter Transportation (2. Transfer); IBM Smart Grid and Rational software platform for automotive systems (3. Production); IBM Smarter Commerce (4. Distribution); and IBM i2, IBM Defense Operations Platform, and IBM i2 Defense Solution (7. LC support).

Early examples of complex automation corresponding to "spiral" concept were DiFac project launched in second framework program of the European Commission [8] and BioVLAB [9].

DiFac is a complex sociotechnical system designed to boost both economic efficiency and performance of human labor in production area. Collaborative participants of the project developed methods for industrial control, interaction in the network team, whose members are located in different countries and were connected through 3D virtual reality.

BioVLAB is a cloud environment for microRNA and mRNA (ribonucleic acid) integrated analysis (MMIA) on Amazon EC2. It makes vast amount of microRNA expression profile data publicly available. BioVLAB is positioned by its developers as an easy-to-use computing environment for researchers who plan to perform genome-wide integrated analysis tasks with advanced features:

– readily expanded computational tools;
– easily modifiable by reconfiguring in the workflow;
– on-demand cloud computing resources; and
– distributed orchestration supports complex and long running workflows asynchronously.

Special place in the row is occupied by IBM Intelligent Operations Center (IOC) introduced by IBM as a part of Smarter Planet initiative. IBM IOC had been applied in many target areas of Smarter Planet in a purpose to integrate and use data from multiple sources and present results of their processing in single interface. Covered sources may belong to absolutely different domains of activity and this complex integration permits to monitor and manage their states and support operative decisions. Data processing and decision-making use advanced analytics, asset management, and collaboration tools. Smarter City is one of the most complex and promising platforms introductions in modern Web. Perhaps, IBM IOC is the largest commercial solutions currently distributed at the IT market. It provides the following functions [10]:

– visual workspace;
– events and incident management;
– resource, response, and activity management;
– status monitoring;
– collaboration, instant notification, and messaging;
– reports; and
– semantic model.

System's architecture includes multilevel SOA (service-oriented architecture) structure, power infrastructure based on IBM Tivoli software, including clouds and system security. Key performance indicator managed dashboard uses event management and workflows engine to react on real-world situation and to keep specified policy and performance level.

All above mentioned systems are designed for collaborative works performed by legally independent organizations acting as a single Web alliance. As such, they meet the definition of a virtual enterprise and the 9th class "system of systems" as well. Consequently, "spiral" systems may also be regarded as virtual organization of next generation.

6 Concluding Remarks

1. Further economy development, as it follows from the IBM Institute for Business Value model will be not for intensification of natural resources consumption but for losses reduction. This requires the new type of Web systems and Web architecture with the ability to automate not just business processes but domain of activities.
2. Design of this type of information systems based on classical approach that automation eliminates bottlenecks in business process does not work and to BPM (business process management) must be added AMS (activity management system).
3. Analysis of technologic stack and requirements of the modern economy permits to expect with a high probability that new type of IS for domain automation conditionally defined as "spiral" will evolve in the direction responding the introduced requirements.

References

1. Gartner hype cycle. Available at: http://www.gartner.com/technology/research/methodologies/hype-cycle.jsp (accessed November 21, 2014).
2. News tagged with metamaterials. Available at: http://phys.org/tags/metamaterials/ (accessed November 21, 2014).
3. Maslow, A. 1954. *Motivation and personality*. New York, N.Y.: Harper&Row Publs. Inc. 15–31.
4. Emery, F. E., and E. L. Trist. 1960. Socio-technical systems. *Management science, models and techniques*. Eds. C. W. Churchman and M. Verhurst. London: Pergamon Press. 2:83–97.
5. Internet of things market forecast. Available at: http://postscapes.com/internet-of-things-market-size (accessed November 21, 2014).
6. Korsten, P., and Ch. Seider. 2010. The world's 4 trillion dollar challenge: Using a system-of-systems approach to build a smarter planet. IBM Institute for Business Value. IBM Global Business Services Executive Report. Available at: http://www-05.ibm.com/tr/events/ibmcozumlerzirvesi2011/pdf/GBE03278USEN.PDF (accessed June 17, 2014).
7. IBM Smarter Planet publications. Available at: http://www.ibm.com/smarterplanet/us/en/overview/ideas/index.html?re=sph; http://www.ibm.com/smarterplanet/ru/ru/; http://en.wikipedia.org/wiki/

Smarter_Planet; http://www.ibm.com/smarterplanet/us/en/?ca=v_smarterplanet (accessed June 17, 2014).
8. DiFac success story. http://www.ims.org/wp-content/uploads/2012/03/DiFac-SUCCESS-STORY_100917.pdf (accessed June 17, 2014).
9. Lee, H., Y. Yang, H. Chae, S. Nam, D. Choi, P. Tangchaisin, C. Herath, S. Marru, K. P. Nephew, and S. Kim. 2012. BioVLAB-MMIA: A cloud environment for microRNA and mRNA integrated analysis (MMIA) on Amazon EC2. *IEEE Trans. Nanobiosci.* 11(3):266–272. doi: 10.1109/TNB.2012.2212030.
10. IBM Corp., International Technical Support Organization. November 15, 2012. IBM Intelligent Operations Center for Smarter Cities. IBM Redbooks Solution Guide.

Received June 10, 2014

Contributor

Sorokin Alexander V. (b. 1946) — Candidate of Science (PhD) in technology, University Relations Manager for Russia & CIS, IBM EE/A; asorokin27@gmail.com

АВТОМАТИЗАЦИЯ ЗА ПРЕДЕЛАМИ WEB 2.0

А. Сорокин

IBM EE/A, Пресненская наб. 10, Москва 123317, Россия

Аннотация: Рассматривается новый подход к анализу эволюции информационных систем, основанный на разработанном автором стеке активностей. С помощью введенного подхода исследуются перспективные тенденции построения на платформе Вэб информационных систем, которые начинаются с автоматизации отдельных бизнес-процессов и затем, в результате дальнейшей экспансии информационных технологий (ИТ), охватывают области профессиональной деятельности. В результате классический подход к проектированию информационных систем, базирующийся на устранении посредством автоматизации узких мест бизнес-процессов, перестает работать. Текущие тенденции в развитии ИТ, связанные с новыми возможностями «ортогональной» интеграции систем, делают вероятным появление нового типа больших информационных систем и нового типа их Вэб-архитектуры, условно названного в данной работе «спиралью паутины». По сравнению с «радиальной» интеграцией Вэб в рамках одной профессиональной области такой тип архитектуры является более эффективным.

Ключевые слова: автоматизация; реинжиниринг бизнес-процессов; совместная разработка программных продуктов; экономика масштабирования; интернет-топология; социотехнические системы; системы систем; виртуальные предприятия; Вэб 2.0

DOI: 10.14357/19922264140414

Литература

1. Gartner's hype cycle. http://www.gartner.com/technology/research/methodologies/hype-cycle.jsp.
2. News tagged in metamaterials. http://phys.org/tags/metamaterials/.
3. *Maslow A.* Motivation and personality. New York, N.Y.: Harper&Row Publs. Inc., 1954. P. 15–31.
4. *Emery F. E., Trist E. L.* Socio-technical systems // *Management science, models and techniques* / Eds. C. W. Churchman and M. Verhurst. — London: Pergamon Press, 1960. Vol. 2. P. 83–97.
5. Internet of things market forecast. http://postscapes.com/internet-of-things-market-size.
6. *Korsten P., Seider Ch.* The world's 4 trillion dollar challenge. Using a system-of-systems approach to build a smarter planet. IBM Institute for Business Value, 2010. IBM Global Business Services Executive Report. http://www-05.ibm.com/tr/events/ ibmcozumlerzirvesi2011/pdf/GBE03278USEN.PDF.
7. IBM Smarter Planet publications. http://www.ibm.com/smarterplanet/us/en/overview/ideas/index.html?re=sph; http://www.ibm.com/smarterplanet/ru/ru/; http://en.wikipedia.org/wiki/Smarter_Planet; http://www.ibm.com/smarterplanet/us/en/?ca=v_smarterplanet (accessed June 17, 2014).
8. DiFac Success story. http://www.ims.org/wp-content/uploads/2012/03/DiFac-SUCCESS-STORY_100917.pdf.
9. *Lee H., Yang Y., Chae H., Nam S., Choi D., Tangchaisin P., Herath C., Marru S., Nephew K. P., Kim S.* BioVLAB-MMIA: A cloud environment for microRNA and mRNA integrated analysis (MMIA) on Amazon EC2 // IEEE Trans. Nanobiosci., 2012. Vol. 11. No. 3. P. 266–272. doi: 10.1109/TNB.2012.2212030.
10. IBM Corporation, International Technical Support Organization. IBM Intelligent Operations Center for Smarter Cities. IBM Redbooks Solution Guide. November 15, 2012.

Поступила в редакцию 10.06.2014

ОБ АВТОРАХ

Бронштейн Ефим Михайлович (р. 1946) — доктор физико-математических наук, профессор Уфимского государственного авиационного технического университета

Вовченко Алексей Евгеньевич (р. 1984) — кандидат технических наук, старший научный сотрудник Института проблем информатики Российской академии наук

Галина Ирина Владимировна (р. 1967) — старший научный сотрудник Института проблем информатики Российской академии наук

Горшенин Андрей Константинович (р. 1986) — кандидат физико-математических наук, старший научный сотрудник ИПИ РАН; доцент Московского государственного технического университета радиотехники, электроники и автоматики (МГТУ МИРЭА)

Грушо Александр Александрович (р. 1946) — доктор физико-математических наук, член-корреспондент Академии криптографии РФ; ведущий научный сотрудник Института проблем информатики Российской академии наук; профессор факультета вычислительной математики и кибернетики Московского государственного университета им. М. В. Ломоносова

Грушо Николай Александрович (р. 1982) — кандидат физико-математических наук, старший научный сотрудник Института проблем информатики Российской академии наук

Ерошенко Александр Андреевич (р. 1989) — аспирант факультета вычислительной математики и кибернетики Московского государственного университета им. М. В. Ломоносова

Зацаринный Александр Алексеевич (р. 1951) — доктор технических наук, профессор, заместитель директора Института проблем информатики Российской академии наук

Зелёв Павел Александрович (р. 1988) — аспирант Уфимского государственного авиационного технического университета

Золотарев Олег Васильевич (р. 1959) — кандидат технических наук, доцент ВПО «Российский Новый Университет»

Калиниченко Леонид Андреевич (р. 1937) — доктор физико-математических наук, профессор, заведующий лабораторией Института проблем информатики Российской академии наук; профессор факультета вычислительной математики и кибернетики Московского государственного университета им. М. В. Ломоносова

Ковалев Дмитрий Юрьевич (р. 1988) — младший научный сотрудник Института проблем информатики Российской академии наук

Козеренко Елена Борисовна (р. 1959) — кандидат филологических наук, заведующая лабораторией Института проблем информатики Российской академии наук

Королев Виктор Юрьевич (р. 1954) — доктор физико-математических наук, профессор кафедры математической статистики факультета вычислительной математики и кибернетики Московского государственного университета им. М. В. Ломоносова; ведущий научный сотрудник Института проблем информатики Российской академии наук

Корчагин Александр Юрьевич (р. 1989) — аспирант факультета вычислительной математики и кибернетики Московского государственного университета им. М. В. Ломоносова

Миронов Андрей Михайлович (р. 1966) — кандидат физико-математических наук, старший научный сотрудник Института проблем информатики Российской академии наук

Михеев Михаил Юрьевич (р. 1957) — доктор филологических наук, ведущий научный сотрудник НИВЦ Московского государственного университета им. М. В. Ломоносова; ведущий научный сотрудник Института проблем информатики Российской академии наук

Морозова Юлия Игоревна (р. 1984) — научный сотрудник Института проблем информатики Российской академии наук

Печинкин Александр Владимирович (1946–2014) — доктор физико-математических наук, профессор, главный научный сотрудник ИПИ РАН

Разумчик Ростислав Валерьевич (р. 1984) — кандидат физико-математических наук, старший научный сотрудник ИПИ РАН; доцент Российского университета дружбы народов

Сомин Николай Владимирович (р. 1947) — кандидат физико-математических наук, ведущий научный сотрудник Института проблем информатики Российской академии наук

Сорокин Александр Викторович (р. 1946) — кандидат технических наук, менеджер университетских проектов в России и СНГ, IBM Восточная Европа и Азия

Ступников Сергей Александрович (р. 1978) — кандидат технических наук, старший научный сотрудник Института проблем информатики Российской академии наук

Тимонина Елена Евгеньевна (р. 1952) — доктор технических наук, профессор, ведущий научный сотрудник Института проблем информатики Российской академии наук

Френкель Сергей Лазаревич (р. 1951) — кандидат технических наук, старший научный сотрудник Институт проблем информатики Российской академии наук; доцент Московского государственного технического университета радиотехники, электроники и автоматики (МГТУ МИРЭА)

Черток Андрей Викторович (р. 1987) — младший научный сотрудник факультета вычислительной математики и кибернетики Московского государственного университета им. М. В. Ломоносова; генеральный директор ООО «Эйфория Груп»

Чупраков Константин Григорьевич (р. 1985) — кандидат технических наук, ведущий математик Института проблем информатики Российской академии наук

Шарнин Михаил Михайлович (р. 1959) — кандидат технических наук, старший научный сотрудник Института проблем информатики Российской академии наук

Шестаков Олег Владимирович (р. 1976) — доктор физико-математических наук, доцент факультета вычислительной математики и кибернетики Московского государственного университета им. М. В. Ломоносова; старший научный сотрудник Института проблем информатики Российской академии наук

АВТОРСКИЙ УКАЗАТЕЛЬ ЗА 2014 г.

2014 AUTHOR INDEX

Профессор Александр Владимирович Печинкин

7.10.1946–4.12.2014

Институт проблем информатики Российской академии наук, редакционный совет и редакционная коллегия журнала «Информатика и её применения» с глубоким прискорбием извещают, что 4 декабря 2014 г. после продолжительной и тяжелой болезни скончался Александр Владимирович Печинкин — лауреат премии Правительства РФ в области науки и техники, доктор физико-математических наук, профессор, главный научный сотрудник ИПИ РАН, член редколлегии журнала «Информатика и её применения».

А. В. Печинкин родился в 1946 г. в Москве. Еще до окончания в 1968 г. механико-математического факультета МГУ им. М. В. Ломоносова А. В. Печинкин начал вести научную и педагогическую деятельность, которую затем продолжил в различных научно-исследовательских учреждениях и высших учебных заведениях столицы (НИИ ССУ, МИЭМ, МГТУ им. Н. Э. Баумана, РУДН). С 2000 г. его работа была неразрывно связана с ИПИ РАН. Выдающийся ученый, получивший признание в России и за рубежом, внесший существенный научный вклад в развитие теории массового обслуживания, А. В. Печинкин основал крупную научную школу, из которой вышло большое число молодых ученых. Научные работы А. В. Печинкина главным образом относятся к теории вероятностей и ее приложениям. Он автор свыше 200 фундаментальных трудов по прикладной теории вероятностей и теории массового обслуживания. А. В. Печинкин являлся членом различных диссертационных советов, редколлегий научных журналов, программных комитетов международных научных конференций.

А. В. Печинкин был выдающимся преподавателем. Его педагогический талант нашел свое отражение в учебниках «Теория вероятностей и математическая статистика» и «Теория массового обслуживания», которые были переведены на английский язык и по которым училось несколько поколений студентов-математиков в России и за рубежом. За серию учебников «Математика в техническом университете» он был удостоен премии Правительства РФ в области науки и техники.

Для научного творчества А. В. Печинкина была характерна любовь к конкретно поставленным вопросам любой важности — от занимательных задач для школьников и студентов до сложных вопросов чистой и прикладной математики. Лекции А. В. Печинкина стимулировали в каждом заинтересованном слушателе представления о существовании замечательных связей между разнородными математическими объектами, на первый взгляд совершенно различными. Он с одинаково большим вниманием и участием относился и к своим молодым ученикам, и к уже состоявшимся ученым. А. В. Печинкин обладал большим личным обаянием, имел широкий круг интересов.

Все знавшие А. В. Печинкина всегда будут помнить его как замечательного ученого и прекрасного товарища.

Институт проблем информатики Российской академии наук, редакционный совет и редакционная коллегия журнала «Информатика и её применения» выражают глубокое соболезнование родным и близким покойного.

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;
   - источники финансирования работы (ссылки на гранты, проекты, поддерживающие организации и т. п.).

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;
- Acknowledgments.

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