Информатика и её применения

Том 11 Выпуск 3 Год 2017

СОДЕРЖАНИЕ

Технический редактор *Л. Кокушкина* Художественный редактор *М. Седакова*
Сдано в набор 23.08.17. Подписано в печать 28.09.17. Формат 60 x 84 / 8
Бумага офсетная. Печать цифровая. Усл.-печ. л. 18,5. Уч.-изд. л. 17. Тираж 100 экз.
Заказ № 995.
Издательство «ТОРУС ПРЕСС», Москва 121614, ул. Крылатская, 29-1-43
Отпечатано в НИПКЦ «Восход-А» с готовых файлов
Москва 109052, ул. Смирновская, д. 25, стр. 3

АНАЛОГИ ТЕОРЕМЫ ГЛЕЗЕРА ДЛЯ ОТРИЦАТЕЛЬНЫХ БИНОМИАЛЬНЫХ И ОБОБЩЕННЫХ ГАММА-РАСПРЕДЕЛЕНИЙ И НЕКОТОРЫЕ ИХ ПРИЛОЖЕНИЯ*

В. Ю. Королев[1]

Аннотация: Доказано, что отрицательные биномиальные распределения с параметром формы, меньшим единицы, являются смешанными геометрическими распределениями. Смешивающее распределение выписывается в явном виде. Тем самым на дискретный случай перенесен аналогичный результат Л. Глезера, устанавливающий, что гамма-распределения с параметром формы, меньшим единицы, являются смешанными показательными законами. Также доказан аналог теоремы Глезера для обобщенных гамма-распределений (GG-распределений, Generalized Gamma distributions). Для смешанных биномиальных распределений, связанных с отрицательными биномиальными распределениями с параметром формы, меньшим единицы, рассмотрен случай малой вероятности успеха и доказан аналог теоремы Пуассона. С помощью представления отрицательных биномиальных распределений в виде смешанных геометрических законов доказаны предельные теоремы для отрицательных биномиальных случайных сумм независимых одинаково распределенных случайных величин (с.в.), в частности аналоги закона больших чисел и центральной предельной теоремы. Рассмотрены случаи как легких, так и тяжелых хвостов. Получены выражения для моментов предельных распределений. Полученные альтернативные эквивалентные представления предельных законов в виде смесей позволяют получить лучшее понимание механизмов, формирующих смешанные вероятностные (байесовские) модели.

Ключевые слова: отрицательное биномиальное распределение; смешанное геометрическое распределение; обобщенное гамма-распределение; устойчивое распределение; распределение Лапласа; распределение Миттаг–Леффлера; распределение Линника; смешанное биномиальное распределение; теорема Пуассона; случайная сумма; закон больших чисел; центральная предельная теорема

DOI: 10.14357/19922264170301

1 Введение

1.1 Мотивация

В большинстве работ, посвященных статистическому анализу метеорологических данных, используемые математические модели наблюдаемых статистических закономерностей довольно далеки от того, чтобы считаться адекватными. В частности, принято считать, что продолжительность периода выпадения осадков, измеренная в сутках (т. е. число последовательных дождливых дней), подчиняется геометрическому распределению вероятностей (см., например, [1]), хотя согласие такой модели с реальными данными очень далеко от допустимого.

Возможно, данный предрассудок основан на общепринятой интерпретации геометрического распределения в терминах испытаний Бернулли как распределения числа последовательных дождливых дней («успехов») до первого дня без осадков («неудачи»). Но схема испытаний Бернулли предполагает, что испытания независимы, тогда как результаты статистического анализа метеорологических данных, зарегистрированных в разных географических точках, демонстрируют, что последовательность дождливых и сухих дней не только не обладает свойством независимости, но даже не является марковской.

Таким образом, классическая схема испытаний Бернулли абсолютно не является адекватной при математическом моделировании метеорологических явлений.

Оказалось, что статистические закономерности поведения некоторых характеристик процесса выпадения осадков (в частности, продолжительность дождливых периодов) очень хорошо описываются отрицательным биномиальным распределением

*Работа выполнена при частичной поддержке Программы Президиума РАН № I.33П (проект 063-2016-0015) и Российского фонда фундаментальных исследований (проекты 15-07-04040 и 17-07-00717).

[1]Факультет вычислительной математики и кибернетики Московского государственного университета имени М. В. Ломоносова; Институт проблем информатики Федерального исследовательского центра «Информатика и управление» Российской академии наук; Университет Дианьзи города Ханчжоу, Китай, vkorolev@cs.msu.su

с параметром формы, меньшим единицы. Так, в работе [2] на примере данных, зарегистрированных в таких разных по своим климатическим параметрам пунктах, как Потсдам и Элиста, было показано, что флуктуации продолжительности дождливых периодов, измеренной в сутках, с очень высокой надежностью описывается отрицательным биномиальным распределением с параметром формы $r \approx 0{,}85$. В той же работе была предложена схематическая интерпретация этого феномена с помощью известного свойства отрицательного биномиального распределения, которое является смешанным пуассоновским распределением, где смешивающим служит гамма-распределение. Как известно (см., например, [3, 4]), пуассоновское распределение является наилучшей вероятностной моделью для дискретных хаотических стохастических процессов, адекватность которой обусловлена универсальным принципом неубывания энтропии в замкнутых системах. Тогда можно считать, что смешивающее гамма-распределение «случайного» параметра пуассоновского распределения в отрицательной биномиальной модели [2] описывает статистические закономерности случайных изменений внешних факторов.

В данной работе предпринимается попытка конкретизировать это возможное объяснение адекватности отрицательной биномиальной модели. С этой целью предлагается использовать понятие смешанного геометрического распределения, введенное в работе [5] (см. также [6, 7]). Ниже будет показано, что любое отрицательное биномиальное распределение с параметром формы, меньшим единицы, является смешанным геометрическим распределением. Тем самым будет доказан «дискретный» аналог теоремы Л. Глезера [8], устанавливающей возможность представления гамма-распределения с параметром формы, меньшим единицы, в виде смешанного показательного распределения. Указанное представление отрицательного биномиального распределения в виде смешанного геометрического можно проинтерпретировать в терминах испытаний Бернулли со случайной вероятностью успеха. Сначала в результате «предварительного» эксперимента определяется значение вероятности успеха, а потом рассматриваемая с.в. определяется как число успехов до первой неудачи в последовательности испытаний Бернулли с так определенной случайной вероятностью успеха. Такая интерпретация позволяет привести дополнительные аргументы, объясняющие адекватность отрицательной биномиальной модели для распределения продолжительности дождливых периодов, а именно: можно предположить, что последовательность дождливых и сухих дней не является независимой, но является *условно независимой* при фиксированном значении с.в., определяющей значение вероятности успеха, которое меняется от одного дождливого периода к другому (например, в зависимости от времени года) и определяется факторами, внешними по отношению к исследуемой локальной системе.

1.2 Структура статьи

Статья организована следующим образом.

В подразд. 1.3 введены понятия, используемые в дальнейшем, и приведены необходимые обозначения.

Основные результаты сформулированы и доказаны в разд. 2. Здесь доказано, что отрицательные биномиальные распределения с параметром формы, меньшим единицы, являются смешанными геометрическими распределениями (теорема 1). Смешивающее распределение выписывается в явном виде. Тем самым на дискретный случай перенесен аналогичный результат Л. Глезера, устанавливающий, что гамма-распределения с параметром формы, меньшим единицы, являются смешанными показательными законами. Изучена связь между смешивающими распределениями в теореме Глезера и теореме 1. Здесь же теорема Глезера распространена на GG-распределения.

В разд. 3 для смешанных биномиальных распределений, связанных с отрицательными биномиальными распределениями с параметром формы, меньшим единицы, рассмотрен случай малой вероятности успеха и доказан аналог теоремы Пуассона.

Раздел 4 посвящен предельным теоремам для отрицательных биномиальных сумм. С точки зрения моделирования статистических закономерностей процессов выпадения осадков полученные здесь предельные распределения могут служить асимптотической аппроксимацией распределения суммарного объема осадков, выпавших в течение одного довольно «продолжительного» дождливого периода. С формальной точки зрения приведенные здесь результаты по сути являются теоремами переноса, которые можно сформулировать и доказать традиционными способами (см., например, [9]), однако приводимые здесь формулировки и доказательства, основанные на представлении отрицательных биномиальных распределений в виде смешанных геометрических законов, могут дать дополнительное понимание эффектов, возникающих при исследовании схемы испытаний Бернулли со случайной вероятностью успехов. Более того, получены альтернативные эквивалентные представления предельных законов в виде смесей, позволяющие по-

лучить лучшее понимание механизмов, формирующих смешанные вероятностные модели, и обеспечить более эффективное применение байесовских методов статистического анализа реальных данных за счет более адекватного выбора априорных распределений.

1.3 Определения и обозначения

Отрицательным биномиальным распределением с параметрами $r > 0$ и $p \in (0,1)$ называется набор положительных чисел

$$q_k = \frac{\Gamma(r+k)}{k!\Gamma(r)}\, p^r (1-p)^k\,,\ \ k = 0,1,2,\ldots,$$

где $\Gamma(r)$ — эйлерова гамма-функция,

$$\Gamma(r) = \int\limits_0^\infty x^{r-1} e^{-x}\, dx\,,\ \ r > 0\,.$$

Несложно убедиться, что $\sum_{k=0}^\infty q_k = 1$, так что набор чисел $\{q_k\}_{k=0}^\infty$ задает распределение вероятностей на $\{0\}\bigcup\mathbb{N}$. Частным случаем отрицательного биномиального распределения, соответствующим значению $r = 1$, является *геометрическое распределение*.

В дальнейшем для удобства изложение будет вестись в терминах с.в. с соответствующими распределениями. При этом будет предполагаться, что все с.в. заданы на одном вероятностном пространстве $(\Omega,\, \mathfrak{F},\, \mathsf{P})$.

В статье используются стандартные обозначения. Символы $\stackrel{\mathrm{d}}{=}$ и $\Longrightarrow$ обозначают совпадение распределений и сходимость по распределению. Целую и дробную часть вещественного числа z будем соответственно обозначать $[z]$ и $\{z\}$.

Случайная величина, имеющая гамма-распределение с параметром формы $r > 0$ и параметром масштаба $\lambda > 0$, будет обозначаться $G_{r,\lambda}$,

$$\mathsf{P}(G_{r,\lambda} < x) = \int\limits_0^x g(z; r, \lambda)\, dz\,,\ \ x \geq 0\,,$$

где

$$g(z; r, \lambda) = \frac{\lambda^r}{\Gamma(r)}\, z^{r-1} e^{-\lambda z}\,,\ \ z \geq 0\,.$$

В принятых обозначениях $G_{1,1}$ — с.в. со стандартным показательным распределением: $\mathsf{P}(G_{1,1} < x) = [1 - e^{-x}]\,\mathbf{1}(x \geq 0)$ (здесь и далее $\mathbf{1}(A)$ — это индикаторная функция множества A).

Гамма-распределение является частным представителем класса GG-распределений, которые были впервые описаны как единое семейство в 1962 г. в работе [10] в качестве семейства вероятностных моделей, включающего в себя одновременно гамма-распределения и распределения Вейбулла. Обобщенным гамма-распределением называется распределение, определяемое плотностью вероятностей вида

$$g^*(x; r, \gamma, \lambda) = \frac{|\gamma|\lambda^r}{\Gamma(r)}\, x^{\gamma r - 1} e^{-\lambda x^\gamma}\,,\ \ x \geq 0\,,$$

где $\gamma \in \mathbb{R}$, $\lambda > 0$, $r > 0$. Более подробное описание свойств GG-распределений см. в [10, 11]. В дальнейшем, как правило, нас будут интересовать GG-распределения с $\gamma \in (0, 1]$. Случайная величина с плотностью $g^*(x; r, \gamma, \lambda)$ будет обозначаться $G^*_{r,\gamma,\lambda}$.

Для частного случая с.в. с GG-распределением — распределением Вейбулла–Гнеденко, определяемым плотностью $g^*(x; 1, \gamma, 1)$ и функцией распределения (ф.р.) $\left[1 - e^{-x^\gamma}\right]\mathbf{1}(x \geq 0)$, будет использоваться особое обозначение W_γ. Таким образом, $G^*_{1,1,1} \stackrel{\mathrm{d}}{=} G_{1,1} \stackrel{\mathrm{d}}{=} W_1$.

Случайная величина со стандартной нормальной ф.р. $\Phi(x)$ будет обозначаться X,

$$\mathsf{P}(X < x) = \Phi(x) = \frac{1}{\sqrt{2\pi}} \int\limits_{-\infty}^x e^{-z^2/2}\, dz\,,\ \ x \in \mathbb{R}\,.$$

Случайную величину с ф.р. Лапласа $F^\Lambda(x)$, соответствующей плотности $\ell(x) = (1/2)e^{-|x|}$, $x \in \mathbb{R}$, будем обозначать Λ.

Функция распределения строго устойчивого распределения с характеристическим показателем α и параметром формы θ, определяемого характеристической функцией (х.ф.)

$$\mathfrak{f}_{\alpha,\theta}(t) = \exp\left\{-|t|^\alpha \exp\left\{-\frac{1}{2}\, i\pi\theta\alpha \mathrm{sign} t\right\}\right\}\,,\ \ t \in \mathbb{R}\,,$$

где $0 < \alpha \leq 2$, $|\theta| \leq \min\{1, 2/\alpha - 1\}$, будет обозначаться $F_{\alpha,\theta}(x)$ (см., например, [12]). Любую с.в. с ф.р. $F_{\alpha,\theta}(x)$ будем обозначать $T_{\alpha,\theta}$. Симметричным строго устойчивым распределениям соответствует значение $\theta = 0$ и х.ф. $\mathfrak{f}_{\alpha,0}(t) = e^{-|t|^\alpha}$, $t \in \mathbb{R}$. Отсюда несложно видеть, что $T_{2,0} \stackrel{\mathrm{d}}{=} \sqrt{2}X$.

Односторонним строго устойчивым законам, сосредоточенным на неотрицательной полуоси, соответствуют значения $\theta = 1$ и $0 < \alpha \leq 1$. Пары $\alpha = 1$, $\theta = \pm 1$ отвечают распределениям, вырожденным в ± 1 соответственно. Остальные устойчивые распределения абсолютно непрерывны. Явные выражения устойчивых плотностей в терминах элементарных функций отсутствуют за четырьмя исключениями (нормальный закон ($\alpha = 2$, $\theta = 0$), распределение Коши ($\alpha = 1$, $\theta = 0$), распределение Леви ($\alpha = 1/2$, $\theta = 1$) и распределение,

симметричное к распределению Леви ($\alpha = 1/2$, $\theta = -1$)). Выражения устойчивых плотностей в терминах функций Фокса (обобщенных G-функций Мейера) можно найти в [13, 14].

Согласно «теореме умножения» (см., например, теорему 3.3.1 в [12]) для любой допустимой пары параметров $(\alpha,\, \theta)$ и произвольного $\alpha' \in (0, 1]$ справедливо мультипликативное представление:

$$T_{\alpha\alpha',\theta} \stackrel{\mathrm{d}}{=} T_{\alpha,\theta} T_{\alpha',1}^{1/\alpha}\,,$$

в котором сомножители в правой части независимы. В частности, для любого $\alpha \in (0, 2]$

$$T_{\alpha,0} \stackrel{\mathrm{d}}{=} X\sqrt{2T_{\alpha/2,1}}\,, \tag{1}$$

т. е. любое симметричное строго устойчивое распределение является масштабной смесью нормальных законов.

Хорошо известно, что если $0 < \alpha < 2$, то $\mathsf{E}|T_{\alpha,\theta}|^\beta < \infty$ для любого $\beta \in (0, \alpha)$, но моменты с.в. $T_{\alpha,\theta}$ порядков $\beta \geq \alpha$ не существуют (см., например, [12]). Несмотря на отсутствие явных выражений плотностей устойчивых распределений в терминах элементарных функций, можно показать [15], что для $0 < \beta < \alpha < 2$

$$\mathsf{E}|T_{\alpha,0}|^\beta = \frac{2^\beta}{\sqrt{\pi}}\,\frac{\Gamma\left((\beta+1)/2\right)\Gamma\left((\alpha-\beta)/\alpha\right)}{\Gamma\left((2-\beta)/\beta\right)}$$

и для $0 < \beta < \alpha \leq 1$

$$\mathsf{E}T_{\alpha,1}^\beta = \frac{\Gamma\left((\alpha-\beta)/\alpha\right)}{\Gamma(1-\beta)}\,.$$

Говорят, что распределение с.в. Z принадлежит к области нормального притяжения строго устойчивого закона $F_{\alpha,\theta}$, $\mathcal{L}(Z) \in \mathrm{DNA}\,(F_{\alpha,\theta})$, если существует конечная положительная постоянная c такая, что

$$\frac{c}{n^{1/\alpha}}\sum_{j=1}^{n} Z_j \Longrightarrow T_{\alpha,\theta}\ \ (n \to \infty)\,,$$

где $Z_1, Z_2, \ldots$ — независимые копии с.в. Z. В дальнейшем будем рассматривать случай стандартного масштаба и полагаем $c = 1$. В работе [16] было показано, что если $\mathcal{L}(Z) \in \mathrm{DNA}\,(F_{\alpha,\theta})$, то $\mathsf{E}|Z|^\beta = \infty$ для любого $\beta > \alpha$.

Пусть $\alpha \in (0, 1]$. Распределения с преобразованием Лапласа–Стилтьеса (п. Л.–С.)

$$\mathfrak{m}_\alpha(s) = \frac{1}{1+s^\alpha}\,,\ \ s \geq 0\,, \tag{2}$$

принято называть *распределениями Миттаг–Леффлера*. Происхождение этого названия связано с тем, что плотность, соответствующая п. Л.–С. (2), имеет вид:

$$f_\alpha^M(x) = \frac{1}{x^{1-\alpha}}\sum_{n=0}^{\infty}\frac{(-1)^n x^{\alpha n}}{\Gamma(\alpha n+1)} = -\frac{d}{dx}\,E_\alpha\left(-x^\alpha\right)\,,$$
$$x \geq 0\,, \tag{3}$$

где $E_\alpha(z)$ — функция Миттаг–Леффлера индекса α, определяемая как степенной ряд

$$E_\alpha(z) = \sum_{n=0}^{\infty}\frac{z^n}{\Gamma(\alpha n+1)}\,,\ \ \alpha > 0\,,\ z \in \mathbb{Z}\,.$$

Функция распределения, соответствующая плотности (3), будет обозначаться $F_\alpha^M(x)$. Случайная величина с ф.р. $F_\alpha^M(x)$ будет обозначаться M_α. В [17] приведено интегральное представление плотности распределения Миттаг–Леффлера:

$$f_\alpha^M(x) = \frac{\sin(\pi\alpha)}{\pi}\int\limits_0^\infty \frac{z^\alpha e^{-zx}\,dz}{1+z^{2\alpha}+2z^\alpha\cos(\pi\alpha)}\,,\ \ x > 0\,.$$

При $\alpha = 1$ распределение Миттаг–Леффлера превращается в стандартное показательное распределение: $M_1 \stackrel{\mathrm{d}}{=} W_1$. Но при $\alpha < 1$ плотность (3) имеет хвост, убывающий степенны́м образом: если $0 < \alpha < 1$, то

$$f_\alpha^M(x) \sim \frac{\sin(\alpha\pi)\Gamma(\alpha+1)}{\pi x^{\alpha+1}}$$

при $x \to \infty$ (см., например, [18]).

Хорошо известно, что распределение Миттаг–Леффлера устойчиво по отношению к геометрическому суммированию (или *геометрически устойчиво*). Это означает, что если $X_1, X_2, \ldots$ — независимые одинаково распределенные неотрицательные с.в. и V_p — независимая от $X_1, X_2, \ldots$ с.в. с геометрическим распределением, то существуют положительные константы $a_p > 0$, гарантирующие сходимость $a_p\left(X_1 + \cdots + X_{V_p}\right) \Longrightarrow M_\alpha$ при $p \to 0$ (см., например, [19] или [20]). Более того, еще в 1965 г. И. Н. Коваленко [21] показал, что распределения с п. Л.–С. (2) и только они являются возможными предельными распределениями для надлежащим образом нормированных геометрических сумм вида $a_p\left(X_1 + \cdots + X_{V_p}\right)$ независимых неотрицательных с.в. при $p \to 0$. Доказательства этого результата были воспроизведены в книгах [9, 22, 23], где вместо термина «распределения Миттаг–Леффлера» класс распределений с п. Л.–С. (2) был назван *классом* $\mathcal{K}$ в честь И. Н. Коваленко.

Спустя 25 лет упомянутое предельное свойство распределений с п. Л.–С. (2) было переоткрыто

Р. Пиллаи [24, 25], который предложил для них использовать термин *распределения Миттаг–Леффлера*, ставший общепринятым.

Распределения Миттаг-Леффлера используются при описании аномальной диффузии или эффектов релаксации (см. [26, 27] и дальнейшие ссылки в этих работах).

Пусть $p \in (0,1)$ и V_p — с.в. с геометрическим распределением с параметром p:

$$\mathsf{P}(V_p = k) = p(1-p)^k\,, \quad k = 0, 1, 2, \ldots \tag{4}$$

Это означает, что

$$\mathsf{P}(V_p \geq m) = \sum_{k=m}^{\infty} p(1-p)^k = (1-p)^m$$

для любого $m \in \mathbb{N}$.

Пусть Y — с.в., принимающая значения в интервале $(0,1)$, причем при всех $p \in (0,1)$ с.в. Y и V_p независимы. Положим $N = V_Y$, т. е. будем считать, что

$$\mathsf{P}(N \geq m) = \int\limits_0^1 (1-y)^m \, d\mathsf{P}(Y < y)$$

для любого $m \in \mathbb{N}$. Распределение с.в. N назовем *Y-смешанным геометрическим*.

2 Результаты

В работе [8] было показано, что гамма-распределение с параметром формы, меньшим единицы, является смешанным показательным распределением. Этот результат для удобства будет сформулирован в виде леммы.

Лемма 1 [8]. *Плотность гамма-распределения $g(x; r, \mu)$ при $0 < r < 1$ может быть представлена в виде*:

$$g(x; r, \mu) = \int\limits_0^{\infty} z e^{-zx} p(z; r, \mu)\, dz\,,$$

где

$$p(z; r, \mu) = \frac{\mu^r}{\Gamma(1-r)\Gamma(r)} \, \frac{\mathbf{1}(z \geq \mu)}{(z-\mu)^r z}\,. \tag{5}$$

Более того, если $r > 1$, то гамма-распределение с параметром формы r нельзя представить в виде смешанного показательного распределения.

Хорошо известно, что отрицательное биномиальное распределение является смешанным пуассоновским распределением, в котором смешивание происходит по гамма-распределению [28] (также см. [4]): для любых $r > 0$, $p \in (0,1)$ и $k \in \{0\} \bigcup \mathbb{N}$

$$\frac{\Gamma(r+k)}{k!\Gamma(r)}\, p^r (1-p)^k = \frac{1}{k!} \int\limits_0^{\infty} e^{-\lambda} \lambda^k g(\lambda; r, \mu)\, d\lambda\,, \tag{6}$$

где $\mu = p/(1-p)$. При условии $0 < r < 1$ продолжим (6) с учетом леммы 1:

$$\begin{aligned}
&\frac{\Gamma(r+k)}{k!\Gamma(r)}\, p^r (1-p)^k = \\
&\quad = \frac{1}{k!} \int\limits_0^{\infty} e^{-\lambda} \lambda^k \left(\int\limits_{\mu}^{\infty} z e^{-z\lambda} p(z; r, \mu)\, dz \right) d\lambda = \\
&\quad = \int\limits_{\mu}^{\infty} \left(\frac{1}{k!} \int\limits_0^{\infty} e^{-\lambda(z+1)} \lambda^k z\, d\lambda \right) p(z; r, \mu)\, dz = \\
&\qquad = \int\limits_{\mu}^{\infty} \left(\frac{z}{z+1} \right) \left(1 - \frac{z}{z+1} \right)^k p(z; r, \mu)\, dz\,.
\end{aligned}$$

Сделав в последнем интеграле замену переменных $z \longmapsto y/(1-y)$, получим

$$\frac{\Gamma(r+k)}{k!\Gamma(r)}\, p^r (1-p)^k = \int\limits_p^1 y(1-y)^k h(y; r, p)\, dy\,,$$

где

$$\begin{aligned}
&h(y; r, p) = \\
&\qquad = \frac{p^r}{\Gamma(1-r)\Gamma(r)} \, \frac{(1-y)^{r-1} \mathbf{1}(p < y < 1)}{y(y-p)^r}\,.
\end{aligned} \tag{7}$$

Тем самым доказана

Теорема 1. *Отрицательное биномиальное распределение с параметрами $r \in (0,1)$ и $p \in (0,1)$ является смешанным геометрическим распределением*: *для любого $k \in \{0\} \bigcup \mathbb{N}$*

$$\begin{aligned}
&\frac{\Gamma(r+k)}{k!\Gamma(r)}\, p^r (1-p)^k = \\
&\quad = \int\limits_{\mu}^{\infty} \left(\frac{z}{z+1} \right) \left(1 - \frac{z}{z+1} \right)^k p(z; r, \mu)\, dz = \\
&\qquad = \int\limits_p^1 y(1-y)^k h(y; r, p)\, dy\,,
\end{aligned}$$

где $\mu = p/(1-p)$, а плотности $p(z; r, \mu)$ и $h(y; r, p)$ определены соответственно в (5) *и* (7).

Следствие 1. Пусть $m > 0$, $p \in (0,1)$, $V_p^{(0)}, V_p^{(1)}, \ldots$ — с.в. с одним и тем же геометрическим распределением (4), $Y_{\{m\},p}$ — с.в. с плотностью распределения $h(y; \{m\}, p)$. Предположим, что все введенные с.в.

независимы. Пусть $N_{m,p}$ — с.в. с отрицательным биномиальным распределением (1) при $r = m$. Тогда

$$N_{m,p} \stackrel{\mathrm{d}}{=} V^{(0)}_{Y_{\{m\},p}} + \sum_{n=1}^{[m]} V^{(n)}_p$$

(для определенности считаем, что $\sum_{n=1}^{0} = 0$ и $V^{(0)}_{Y_{0,p}} = 0$).

Утверждение теоремы 1 можно проинтерпретировать в терминах испытаний Бернулли со случайной вероятностью успеха. Сначала в результате «предварительного» эксперимента определяется значение вероятности успеха, т. е. значение с.в. $Y_{r,p}$, имеющей плотность $h(y; r, p)$. Потом с.в. $N_{r,p}$ определяется как число успехов до первой неудачи в последовательности испытаний Бернулли с так определенной вероятностью успеха $Y_{r,p}$.

Такая интерпретация теоремы 1 позволяет привести дополнительные аргументы, объясняющие адекватность отрицательной биномиальной модели для распределения дождливых периодов, а именно: можно предположить, что последовательность дождливых и сухих дней не является независимой, но является *условно независимой* при фиксированном значении с.в. $Y_{r,p}$, которое меняется от одного дождливого периода к другому (например, в зависимости от времени года) и определяется факторами, внешними по отношению к исследуемой локальной системе.

Рассмотрим смешивающие плотности $p(z; r, \mu)$ и $h(y; r, p)$ более подробно.

Теорема 2. *Для $r \in (0,1)$ пусть $G_{r,1}$ и $G_{1-r,1}$ — независимые гамма-распределенные с.в. Пусть $\mu > 0$, $p \in (0,1)$. Тогда*:

(i) *плотность $p(z; r, \mu)$ соответствует с.в.*

$$Z_{r,\mu} = \frac{\mu(G_{r,1} + G_{1-r,1})}{G_{r,1}};$$

(ii) *плотность $h(y; r, p)$ соответствует с.в.*

$$Y_{r,p} = \frac{p(G_{r,1} + G_{1-r,1})}{G_{r,1} + pG_{1-r,1}}.$$

Доказательство. Из доказательства теоремы 1 вытекает, что

$$Y_{r,p} \stackrel{\mathrm{d}}{=} \frac{Z_{r,\mu}}{1 + Z_{r,\mu}}$$

с $\mu = p/(1-p)$. Поэтому достаточно убедиться в справедливости первого утверждения теоремы. С помощью замены переменных $z \longmapsto \mu(1 + {} + x)$ и элементарных выкладок легко проверить, что плотность $p(z; r, \mu)$ соответствует с.в. $\mu(1 + {} + ((1-r)/r)Q_{2(1-r),\,2r})$, где Q_{ν_1,ν_2} — с.в., имеющая распределение Снедекора–Фишера, для $\nu_1 > 0$, $\nu_2 > 0$ определяемое лебеговой плотностью

$$f_{\nu_1,\nu_2}(x) = \\ = \frac{\Gamma((\nu_1+\nu_2)/2)}{\Gamma(\nu_1/2)\Gamma(\nu_2/2)}\, \nu_1^{\nu_1/2} \nu_2^{\nu_2/2} \frac{x^{\nu_1/2-1}}{(\nu_2 + \nu_1 x)^{(\nu_1+\nu_2)/2}}, \quad x \geq 0$$

(в рассматриваемом случае $\nu_1 = 2(1-r)$, $\nu_2 = 2r$, так что $(\nu_1 + \nu_2)/2 = 1$). Но, как известно,

$$Q_{\nu_1,\nu_2} \stackrel{\mathrm{d}}{=} \frac{\nu_2 G_{\nu_1/2,\,1/2}}{\nu_1 G_{\nu_2/2,\,1/2}} \stackrel{\mathrm{d}}{=} \frac{\nu_2 G_{\nu_1/2,\,1}}{\nu_1 G_{\nu_2/2,\,1}},$$

где с.в. $G_{\nu_1/2,\,1}$ и $G_{\nu_2/2,\,1}$ независимы (см., например, [29, с. 32]). Это замечание завершает доказательство теоремы.

Замечание 1. Несложно видеть, что в числителях представлений с.в. $Z_{r,\mu}$ и $Y_{r,p}$, приведенных в формулировке леммы 2, сумма $G_{r,1} + G_{1-r,1}$ имеет стандартное показательное распределение: $G_{r,1} + G_{1-r,1} \stackrel{\mathrm{d}}{=} W_1$. Однако при этом числители и знаменатели указанных представлений не являются независимыми с.в.

Из утверждения (ii) теоремы 2 вытекает, что при $p \to 0$ с.в. $Y_{r,p}$ является величиной порядка p в том смысле, что при $p \to 0$ с.в. $p^{-1}Y_{r,p}$ по распределению сходится к собственной невырожденной с.в. Придадим этому утверждению строгую формулировку.

Следствие 2. Пусть $r \in (0,1)$, $q \in (0,1)$ и $\mu > 0$ произвольны. Тогда

$$nY_{r,\min\{q,\mu/n\}} \Longrightarrow Z_{r,\mu}$$

при $n \to \infty$.

Доказательство. Согласно пункту (ii) теоремы 2 при $n \to \infty$ имеем:

$$nY_{r,\min\{q,\mu/n\}} = \frac{\min\{nq, \mu\}(G_{r,1} + G_{1-r,1})}{G_{r,1} + \min\{q, \mu/n\}G_{1-r,1}} \Longrightarrow \\ \Longrightarrow \frac{\mu(G_{r,1} + G_{1-r,1})}{G_{r,1}}. \quad (8)$$

Но в соответствии с пунктом (i) теоремы 2 правая часть (8) совпадает с $Z_{r,\mu}$. Следствие доказано.

Докажем результат, распространяющий теорему Глезера (лемму 1) на обобщенные гамма-распределения. Для этой цели понадобится следующее представление с.в. с распределением Вейбулла с параметром $\alpha \in (0,1]$, доказанное в [15].

Лемма 2 [15]. *Пусть $\alpha \in (0,1]$. Тогда*

$$W_\alpha \stackrel{\mathrm{d}}{=} \frac{W_1}{T_{\alpha,1}},$$

где с.в. в правой части независимы.

Лемма 3. *Функция распределения $F(x)$ с $F(0) = 0$ соответствует смешанному показательному распределению тогда и только тогда, когда функция $1 - F(x)$ вполне монотонна, т. е. $F \in C_\infty$ и $(-1)^{n+1}F^{(n)}(x) \geq \geq 0$ при всех $x > 0$.*

Это утверждение немедленно вытекает из теоремы С. Н. Бернштейна [30].

Теорема 3. *Пусть $\alpha \in (0, 1]$, $r \in (0, 1)$, $\mu > 0$. Тогда обобщенное гамма-распределение с параметрами r, α и μ является смешанным показательным распределением*:

$$G^*_{r,\alpha,\mu} \stackrel{\mathrm{d}}{=} \frac{W_1}{T_{\alpha,1}\, Z^{1/\alpha}_{r,\mu}}\,,$$

где с.в. в правой части независимы. Более того, обобщенное гамма-распределение с $\alpha r > 1$ не может быть представлено в виде смешанного показательного распределения.

Доказательство. Докажем первое утверждение теоремы. Во-первых, заметим, что согласно лемме 1 для $x \geq 0$ справедливы соотношения:

$$\begin{aligned}\mathsf{P}(G^{1/\alpha}_{r,\mu} > x) &= \mathsf{P}\left(G_{r,\mu} > x^\alpha\right) = \\ &= \mathsf{P}\left(W_1 > Z_{r,\mu}x^\alpha\right) = \int\limits_0^\infty e^{-zx^\alpha} p(z; r, \mu)\, dz = \\ &= \int\limits_0^\infty \mathsf{P}\left(W_\alpha > xz^{1/\alpha}\right) p(z; r, \mu)\, dz\,,\end{aligned}$$

т. е.

$$G^{1/\alpha}_{r,\mu} \stackrel{\mathrm{d}}{=} \frac{W_\alpha}{Z^{1/\alpha}_{r,\mu}}\,. \tag{9}$$

Теперь воспользуемся леммой 2 и, продолжив (9), получим:

$$G^{1/\alpha}_{r,\mu} \stackrel{\mathrm{d}}{=} \frac{W_1}{T_{\alpha,1} Z^{1/\alpha}_{r,\mu}}\,. \tag{10}$$

Во-вторых, несложно убедиться, что

$$G^{1/\alpha}_{r,\mu} \stackrel{\mathrm{d}}{=} G^*_{r,\alpha,\mu} \tag{11}$$

при любых $r > 0$, $\mu > 0$ и $\alpha > 0$. Теперь первое утверждение теоремы вытекает из (10) и (11).

Докажем второе утверждение. Пусть $\alpha r > 1$. Предположим, что с.в. $G^*_{r,\alpha,\mu}$ имеет смешанное показательное распределение. По лемме 3 это означает, что функция $\psi(s) = \mathsf{P}(G^*_{r,\alpha,\mu} > s)$, $s \geq 0$, является вполне монотонной. Но $\psi'(s) = g^*(s; r, \alpha, \mu) \geq 0$ при всех $s \geq 0$, тогда как

$$\begin{aligned}\psi''(s) &= \left(g^*\right)'(s; r, \alpha, \mu) = \\ &= \frac{\alpha\mu^r}{\Gamma(r)}\, s^{\alpha r - 2} e^{-\mu s^\alpha} \left((\alpha r - 1) - \mu\alpha s^\alpha\right) \leq 0\,,\end{aligned}$$

только если $(\alpha r - 1) - \mu\alpha s^\alpha \leq 0$, т. е.

$$s \geq s_0 \equiv \left(\frac{\alpha r - 1}{\mu\alpha}\right)^{1/\alpha} > 0\,,$$

и $\psi''(s) \geq 0$ при $s \in (0, s_0) \neq \varnothing$, что противоречит вполне монотонности функции $\psi(s)$, доказывая второе утверждение теоремы. Теорема доказана.

3 Смешанные биномиальные распределения, связанные с отрицательными биномиальными законами с $r < 1$, и их асимптотическое поведение при $p \to 0$

Рассмотрим еще одну задачу, связанную с описанной выше схемой испытаний Бернулли со случайной вероятностью успеха $Y_{r,p}$ при условии «малости» последней. В рамках этой схемы сначала в результате «предварительного» эксперимента определяется значение с.в. $Y_{r,p} \in (0, 1)$. Это значение принимается в качестве вероятности успеха в испытаниях Бернулли. Затем с.в. M определяется как число успехов в $m \in \mathbb{N}$ испытаниях Бернулли с так определенной вероятностью успеха $Y_{r,p}$. Чтобы описать бесконечную малость вероятности успеха $Y_{r,p}$, снабдим параметр p и (для общности) параметр m, а также, соответственно, с.в. M «бесконечно большим» индексом n, позволяющим проследить сходимость последовательности с.в. Y_{r,p_n} к нулю при $n \to \infty$. В свою очередь, бесконечная малость Y_{r,p_n} означает, что успехи являются редкими событиями в рамках рассматриваемой последовательности испытаний Бернулли со случайной вероятностью успеха.

В рамках схемы испытаний Бернулли со случайной вероятностью успеха, описанной выше, можно сформулировать и доказать «случайный» аналог классической теоремы Пуассона (так называемого «закона малых чисел») для *смешанных биномиальных распределений* со случайной вероятностью успеха и неограниченно возрастающим целочисленным параметром m_n («числом испытаний») [6, 7]. В ранее известных вариантах «случайного» аналога теоремы Пуассона (см., к примеру, [4]), наоборот, случайным считалось число испытаний, а вероятность успеха оставалась неслучайной.

Пусть для каждого $n \in \mathbb{N}$ Y_n — с.в. такая, что $\mathsf{P}(0 < Y_n < 1) = 1$, $m_n \in \mathbb{N}$, $k = 1, 2, \ldots$ Будем гово-

рить, что с.в. M_n имеет Y_n-*смешанное биномиальное распределение* с параметром m_n, если

$$\mathsf{P}(M_n = j) = C_{m_n}^j \int_0^1 z^k (1-z)^{m_n - j}\, d\mathsf{P}(Y_n < z)\,, \quad j = 0, 1, \ldots, m_n\,. \quad (12)$$

Для $x \in \mathbb{R}$ обозначим $B_n(x) = \mathsf{P}(M_n < x)$. Пусть Z — положительная с.в. Смешанная пуассоновская ф.р. со структурной с.в. Z (по терминологии, принятой в [31]) будет обозначаться $\Pi^{(Z)}(x)$:

$$\Pi^{(Z)}(x+0) = \sum_{j=0}^{[x]} \frac{1}{j!} \int_0^\infty e^{-z} z^j\, d\mathsf{P}(Z < z)\,, \quad x \in \mathbb{R}\,.$$

В [7] доказана следующая теорема.

Теорема 4 [7]. *Пусть $\{m_n\}_{n\geq 1}$ — неограниченно возрастающая последовательность натуральных чисел. Пусть при каждом $n \in \mathbb{N}$ M_n — с.в., имеющая Y_n-смешанное биномиальное распределение* (12) *с целочисленным параметром m_n и ф.р. $B_n(x)$. Предположим, что в* (12) *с.в. Y_n бесконечно малы в том смысле, что существует с.в. Z такая, что $\mathsf{P}(0 < Z < \infty) = 1$ и выполнено условие*

$$m_n Y_n \Longrightarrow Z \quad (13)$$

при $n \to \infty$. Тогда

$$B_n(x) \Longrightarrow \Pi^{(Z)}(x) \ \ (n \to \infty)\,.$$

Пусть числа $r \in (0,1)$, $q \in (0,1)$ и $\mu > 0$ произвольны, $\{m_n\}_{n\geq 1}$ — неограниченно возрастающая последовательность натуральных чисел. Положим

$$Y_n = Y_{r,\min\{q,\mu/m_n\}}\,, \quad n \in \mathbb{N}\,.$$

Тогда из следствия 2 вытекает, что

$$m_n Y_n \Longrightarrow Z_{r,\mu}$$

при $n \to \infty$, т. е. условие (13) выполнено с $Z = Z_{r,\mu}$. При этом из теоремы 4 немедленно получается следующий аналог теоремы Пуассона.

Следствие 3. Пусть числа $r \in (0,1)$, $q \in (0,1)$ и $\mu > 0$ произвольны, $\{m_n\}_{n\geq 1}$ — неограниченно возрастающая последовательность натуральных чисел. Пусть при каждом $n \in \mathbb{N}$ M_n – с.в., имеющая $Y_{r,\min\{q,\mu/m_n\}}$-смешанное биномиальное распределение с параметром m_n. Тогда

$$\mathsf{P}(M_n < x) \Longrightarrow \Pi^{(Z_{r,\mu})}(x) \ \ (n \to \infty)\,.$$

Если $P_{r,\mu}$ — с.в. с распределением $\Pi^{(Z_{r,\mu})}(x)$, то для любого $k = 0, 1, \ldots$

$$\mathsf{P}(P_{r,\mu} = k) = \frac{1}{k!} \int_0^\infty z^k e^{-z} p(z; r, \mu)\, dz = \frac{1}{k!}\, \mathsf{E}\left(Z_{r,\mu}^k \exp\{-Z_{r,\mu}\}\right)\,.$$

4 Предельные теоремы для отрицательных биномиальных случайных сумм с $r < 1$

4.1 Аналог закона больших чисел для неотрицательных слагаемых. Обобщенная теорема Реньи

Пусть $X_1, X_2, \ldots$ — независимые одинаково распределенные неотрицательные с.в. Пусть при каждом $n \in \mathbb{N}$ с.в. N_{r,p_n} имеет отрицательное биномиальное распределение с $r \in (0,1)$ и $p_n \in (0,1)$. Предположим, что при каждом $n \in \mathbb{N}$ с.в. N_{r,p_n} независима от последовательности $X_1, X_2, \ldots$ Обозначим

$$S_k = X_1 + \cdots + X_k\,, \quad k \in \mathbb{N}\,.$$

Начнем с рассмотрения ситуации, в которой существует математическое ожидание $\mathsf{E}X_1 \equiv a$. Согласно усиленному закону больших чисел Колмогорова это условие в определенном смысле необходимо и достаточно для того, чтобы с вероятностью единица

$$\frac{1}{n} \sum_{j=1}^n X_j \longrightarrow a \quad (14)$$

при $n \to \infty$. Поставим целью изучить асимптотическое поведение с.в. $S_{N_{r,p_n}}$ при $p_n \to 0$ при условии (14) и получить аналог закона больших чисел для отрицательных биномиальных случайных сумм при $r < 1$.

В работе [5] доказано следующее утверждение.

Теорема 5 [5]. *Предположим, что неотрицательные с.в. $X_1, X_2, \ldots$ удовлетворяют условию* (14). *Пусть при каждом $n \in \mathbb{N}$ с.в. N_n имеют Y_n-смешанное геометрическое распределение и независимы от $X_1, X_2, \ldots$ Предположим, что существует с.в. Z такая, что $\mathsf{P}(0 < Z < \infty) = 1$ и при $n \to \infty$ выполнено условие*:

$$nY_n \Longrightarrow Z\,. \quad (15)$$

Тогда

$$\lim_{n\to\infty} \sup_{x\geq 0} \left| \mathsf{P}\left(\frac{S_{N_n}}{n} > x\right) - \int_0^\infty e^{-xz/a}\, d\mathsf{P}(Z < z) \right| = 0\,.$$

Теорема 5 является развитием классической теоремы Реньи об асимптотическом поведении прореживаемых процессов восстановления (см., например, [9]). Классическую теорему Реньи можно считать законом больших чисел для геометрических случайных сумм. Эта теорема устанавливает, что однородный точечный процесс с конечным математическим ожиданием длин интервалов между соседними точками, подвергнутый операции простейшего прореживания, при которой каждая точка удаляется с вероятностью $1-p$ и оставляется на своем месте с вероятностью p, сопровождающейся надлежащей компрессией времени с целью обеспечить нетривиальность предельного процесса, сходится к пуассоновскому процессу. Как известно, пуассоновский процесс характеризуется в классе процессов восстановления тем, что длины интервалов времени между последовательными восстановлениями имеют показательное распределение. Теорема 5 обобщает теорему Реньи на отрицательные биномиальные случайные суммы. При этом обобщение сводится к тому, что рассматривается «дважды стохастическое» прореживание, при котором одинаковая для всех точек исходного процесса вероятность p определяется заранее как результат некоторого предварительного случайного эксперимента. Это приводит к тому, что предельный процесс оказывается смешанным пуассоновским, что хорошо согласуется с утверждением следствия 3.

Из теоремы 5, следствия 2 и леммы 1 непосредственно вытекает следующее утверждение.

Следствие 4. Предположим, что независимые одинаково распределенные неотрицательные с.в. $X_1, X_2, \ldots$ удовлетворяют условию (14). Пусть числа $r \in (0,1)$, $q \in (0,1)$ и $\mu > 0$ произвольны. Пусть при каждом $n \in \mathbb{N}$ с.в. N_{r,p_n} имеют отрицательные биномиальные распределения с параметрами r и $p_n = \min\{q, \mu/n\}$ и независимы от $X_1, X_2, \ldots$ Тогда

$$\frac{S_{N_{r,p_n}}}{n} \Longrightarrow aG_{r,\mu} \stackrel{\mathrm{d}}{=} \frac{aW_1}{Z_{r,\mu}} \tag{16}$$

при $n \to \infty$.

Чтобы убедиться в справедливости следствия 4, достаточно заметить, что в соответствии с теоремой 5 и следствием 2 для любого $x \in \mathbb{R}$

$$\mathsf{P}\left(\frac{S_{N_{r,p_n}}}{n} < x\right) \Longrightarrow 1 - \int\limits_0^\infty e^{-xz/a} p(z;\, r, \mu)\, dz \tag{17}$$

при $n \to \infty$, но по лемме 1 смешанная показательная ф.р. в правой части (17) совпадает с функцией гамма-распределения с параметрами r и μ в точке x/a.

С учетом абсолютной непрерывности предельного гамма-распределения в следствии 4 можно заключить, что на самом деле в условии (16) речь идет о равномерной сходимости ф.р.:

$$\lim_{n\to\infty} \sup_{x \geq 0} \left| \mathsf{P}\left(\frac{S_{N_{r,p_n}}}{n} > x\right) - \int\limits_x^\infty g(z/a;\, r, \mu)\, dz \right| = 0\,.$$

В терминах статистических закономерностей процессов выпадения осадков предельное гамма-распределение в следствии 4 может служить асимптотической аппроксимацией распределения суммарного объема осадков, выпавших в течение одного «продолжительного» ($p_n \to 0$) дождливого периода, если средние арифметические ежедневных осадков относительно стабильны.

Замечание 2. Сходимость $n^{-1} S_{N_{r,p_n}} \Longrightarrow aG_{r,\mu}$ имеет место и для $r \geq 1$.

4.2 Суммы неотрицательных слагаемых. Случай тяжелых хвостов

В этом подразделе будет рассмотрена ситуация, когда условие (14) не выполнено, т. е. хвосты распределений слагаемых $X_1, X_2, \ldots$ столь тяжелы, что математическое ожидание отсутствует. Вместо условия (14) здесь будем предполагать, что $\mathcal{L}(X_1) \in \in \mathrm{DNA}\,(F_{\alpha,1})$ при некотором $\alpha \in (0,1)$.

В работе [5] доказано следующее утверждение.

Теорема 6 [5]. *Предположим, что независимые одинаково распределенные неотрицательные с.в. $X_1, X_2, \ldots$ таковы, что $\mathcal{L}(X_1) \in \mathrm{DNA}\,(F_{\alpha,1})$ при некотором $\alpha \in \in (0,1)$. Пусть при каждом $n \in \mathbb{N}$ с.в. N_n имеют Y_n-смешанное геометрическое распределение и независимы от $X_1, X_2, \ldots$ Предположим, что существует с.в. Z такая, что $\mathsf{P}(0 < Z < \infty) = 1$ и при $n \to \infty$ выполнено условие* (15). *Тогда*

$$\frac{S_{N_n}}{n^{1/\alpha}} \Longrightarrow T_{\alpha,1} \left(\frac{W_1}{Z}\right)^{1/\alpha} \quad (n \to \infty)\,, \tag{18}$$

причем с.в. в правой части (18) *независимы.*

Из теоремы 6, следствия 2 и леммы 1 непосредственно вытекает следующее утверждение.

Следствие 5. Предположим, что независимые одинаково распределенные неотрицательные с.в. $X_1, X_2, \ldots$ таковы, что $\mathcal{L}(X_1) \in \mathrm{DNA}\,(F_{\alpha,1})$ при некотором $\alpha \in (0,1)$. Пусть числа $r \in (0,1)$, $q \in (0,1)$ и $\mu > 0$ произвольны. Пусть при каждом $n \in \in \mathbb{N}$ с.в. N_{r,p_n} имеют отрицательные биномиальные распределения с параметрами r и $p_n = \min\{q, \mu/n\}$ и независимы от $X_1, X_2, \ldots$ Тогда

$$\frac{S_{N_{r,p_n}}}{n^{1/\alpha}} \Longrightarrow T_{\alpha,1} G_{r,\mu}^{1/\alpha} \tag{19}$$

при $n \to \infty$.

Чтобы убедиться в справедливости следствия 5, достаточно заметить, что в соответствии с теоремой 6 и следствием 2 для любого $x \in \mathbb{R}$

$$\frac{S_{N_{r,p_n}}}{n^{1/\alpha}} \Longrightarrow T_{\alpha,1}\left(\frac{W_1}{Z_{r,\mu}}\right)^{1/\alpha}$$

при $n \to \infty$, но по лемме 1

$$\frac{W_1}{Z_{r,\mu}} \stackrel{\mathrm{d}}{=} G_{r,\mu}\,.$$

С учетом абсолютной непрерывности предельного распределения в следствии 5 можно заключить, что на самом деле в условии (19) речь идет о равномерной сходимости ф.р.:

$$\lim_{n\to\infty}\sup_{x\ge 0}\left|\mathsf{P}\left(\frac{S_{N_{r,p_n}}}{n^{1/\alpha}} < x\right) - \right.$$
$$\left. - \int\limits_0^x F_{\alpha,1}\left(\frac{x}{z^{1/\alpha}}\right) g(z;\,r,\mu)\,dz\right| = 0\,.$$

С учетом теоремы 3 предельное распределение в следствии 5 можно записать в альтернативных эквивалентных формах. Для этого понадобятся еще два вспомогательных утверждения.

Лемма 4 [17, 32, 33]. *Пусть $\alpha \in (0,1)$. Предположим, что неотрицательные с.в. $T_{\alpha,1}$ и $T'_{\alpha,1}$ независимы и имеют одно и то же строго устойчивое распределение. Тогда плотность $v_\alpha(x)$ с.в. $R_\alpha = T_{\alpha,1}/T'_{\alpha,1}$ имеет вид*:

$$v_\alpha(x) = \frac{\sin(\pi\alpha)x^{\alpha-1}}{\pi[1 + x^{2\alpha} + 2x^\alpha\cos(\pi\alpha)]}\,,\quad x > 0\,.$$

Лемма 5 [17, 32, 33]. *Пусть $\alpha \in (0,1)$, M_α — с.в. с распределением Миттаг–Леффлера с параметром α. Тогда*

$$M_\alpha \stackrel{\mathrm{d}}{=} W_1 R_\alpha\,,$$

где с.в. в правой части независимы.

Первое альтернативное представление распределения в правой части (19) довольно очевидно. С учетом соотношения $G^{1/\alpha}_{r,\mu} \stackrel{\mathrm{d}}{=} G^*_{r,\alpha,\mu}$ (см. доказательство теоремы 3) вместо (19) можно записать:

$$\frac{S_{N_{r,p_n}}}{n^{1/\alpha}} \Longrightarrow T_{\alpha,1}G^*_{r,\alpha,\mu}\,. \tag{20}$$

К сожалению, ни соотношением (19), ни соотношением (20) нельзя пользоваться при статистическом анализе на основе функции правдоподобия, так как плотность с.в. $T_{\alpha,1}$ нельзя выписать в явном виде в терминах элементарных функций за исключением случая $\alpha = 1/2$. В отличие от представлений (19) и (20), два следующих представления вполне пригодны для применения в статистическом анализе.

Чтобы получить второе альтернативное представление распределения в правой части (19), воспользуемся теоремой 3, соотношением (20) и вместо (19) получим:

$$\frac{S_{N_{r,p_n}}}{n^{1/\alpha}} \Longrightarrow \frac{T_{\alpha,1}}{T'_{\alpha,1}}\,\frac{W_1}{Z^{1/\alpha}_{r,\mu}} \stackrel{\mathrm{d}}{=} R_\alpha\, t\frac{W_1}{Z^{1/\alpha}_{r,\mu}},$$

где в каждом выражении с.в. независимы. При этом явный вид плотности с.в. R_α указан в лемме 3, а плотность $w_\alpha(x;r,\mu)$ с.в. $W_1 Z^{-1/\alpha}_{r,\mu}$ выписывается легко:

$$w_\alpha(x;r,\mu) = \frac{\alpha\mu^r}{\Gamma(1-r)\Gamma(r)}\int\limits_{\mu^{1/\alpha}}^{\infty}\frac{e^{-zx}\,dz}{(z^\alpha-\mu)^r}\,,\quad x \ge 0\,.$$

Наконец, чтобы получить третье альтернативное представление распределения в правой части (19), воспользуемся соотношением (20) и леммой 5. При этом получаем:

$$\frac{S_{N_{r,p_n}}}{n^{1/\alpha}} \Longrightarrow \frac{T_{\alpha,1}}{T'_{\alpha,1}}\,\frac{W_1}{Z^{1/\alpha}_{r,\mu}} \stackrel{\mathrm{d}}{=} \frac{M_\alpha}{Z^{1/\alpha}_{r,\mu}}\,. \tag{21}$$

Из соотношения (21) вытекает, что предельные распределения для смешанных геометрических сумм неотрицательных независимых с.в. являются масштабными смесями распределений Миттаг–Леффлера (как уже отмечалось, предельных для «обычных» геометрических случайных сумм неотрицательных с.в.), в которых смешивающей является плотность

$$p_\alpha(x;r,\mu) = \frac{\alpha\mu^r}{\Gamma(1-r)\Gamma(r)}\,\frac{1}{(x^\alpha-\mu)^r x}\,,\quad x \ge \mu^{1/\alpha}\,,$$

с.в. $Z^{1/\alpha}_{r,\mu}$. Интегральное представление плотности распределения Миттаг–Леффлера приведено во введении. Преобразование Лапласа–Стилтьеса с.в. в правой части (21) имеет вид:

$$\mathsf{E}\exp\{-sM_\alpha Z^{-1/\alpha}_{r,\mu}\} = \int\limits_0^\infty \frac{z^{1/\alpha}p(z;r,\mu)\,dz}{z^{1/\alpha}+s^\alpha} =$$
$$= \int\limits_0^\infty \frac{zp_\alpha(z;r,\mu)\,dz}{z+s^\alpha}\,,\quad s > 0\,.$$

В терминах статистических закономерностей процессов выпадения осадков предельное гамма-распределение в следствии 4 может служить асим-

птотической аппроксимацией распределения суммарного объема осадков, выпавших в течение одного «продолжительного» ($p_n \to 0$) дождливого периода, если средние арифметические ежедневных осадков относительно нестабильны.

Замечание 3. Сходимость $n^{-1/\alpha} S_{N_{r,p_n}} \Longrightarrow T_{\alpha,1} \cdot G_{r,\mu}^{1/\alpha}$ имеет место и для $r \geq 1$.

Замечание 4. Использовав формулу для моментов строго устойчивых распределений, сосредоточенных на неотрицательной полуоси (см. подразд. 1.3), для моментов порядков $\beta < \alpha$ с.в., предельных в следствии 5, получим представление:

$$\mathsf{E}\left(T_{\alpha,1} G_{r,\mu}^{1/\alpha}\right)^{\beta} = \frac{\Gamma\left((\beta+\alpha r)/\alpha\right)\Gamma\left((\alpha-\beta)/\alpha\right)}{\mu^{\beta/\alpha}\Gamma(1-\beta)\Gamma(r)}\,.$$

4.3 Суммы асимптотически симметричных слагаемых. Случай тяжелых хвостов

В отличие от ситуации, рассмотренной в предыдущем подразделе, здесь будет предполагаться, что слагаемые в суммах могут принимать значения обоих знаков и, более того, являются асимптотически симметричными в том смысле, что $\mathcal{L}(X_1) \in \mathrm{DNA}\,(F_{\alpha,0})$. В работе [5] доказано следующее утверждение.

Теорема 7 [5]. *Предположим, что независимые одинаково распределенные неотрицательные с.в. $X_1, X_2, \ldots$ таковы, что $\mathcal{L}(X_1) \in \mathrm{DNA}\,(F_{\alpha,0})$ при некотором $\alpha \in (0,2]$. Пусть при каждом $n \in \mathbb{N}$ с.в. N_n имеют Y_n-смешанное геометрическое распределение и независимы от $X_1, X_2, \ldots$ Предположим, что существует с.в. Z такая, что $\mathsf{P}(0 < Z < \infty) = 1$ и при $n \to \infty$ выполнено условие* (15). *Тогда*

$$\frac{S_{N_n}}{n^{1/\alpha}} \Longrightarrow T_{\alpha,0}\left(\frac{W_1}{Z}\right)^{1/\alpha} \quad (n \to \infty)\,, \qquad (22)$$

причем с.в. в правой части (22) *независимы.*

Легко видеть, что х.ф. с.в. $T_{\alpha,0} W_1^{1/\alpha}$ имеет вид:

$$\mathsf{E}\exp\{isT_{\alpha,0}W_1^{1/\alpha}\} = \mathsf{E}\mathsf{E}\left(\exp\{isT_{\alpha,0}W_1^{1/\alpha}\}\big|W_1\right) =$$

$$= \int_0^\infty e^{i(|s|z^{1/\alpha})^\alpha} e^{-z}\,dz = \int_0^\infty e^{iz(|s|^\alpha+1)}\,dz = \frac{1}{1+|s|^\alpha}\,, \quad s \in \mathbb{R}\,. \qquad (23)$$

Распределения с х.ф. (23) и $0 < \alpha \leq 2$ принято называть *распределениями Линника*. Они были введены Ю. В. Линником в 1953 г. [34]. При $\alpha = 2$ распределение Линника превращается в распределение Лапласа. Случайная величина, имеющая распределение Линника с параметром α, ее ф.р. и плотность будут соответственно обозначаться L_α, F_α^L и f_α^L. При этом $F_2^L(x) \equiv F^\Lambda(x)$, $x \in \mathbb{R}$.

Распределения Линника обладают многими интересными свойствами. Прежде всего, как и распределения Миттаг–Леффлера, они являются геометрически устойчивыми, т. е. если $X_1, X_2, \ldots$ — независимые одинаково распределенные с.в., причем $\mathcal{L}(X_1) \in \mathrm{DNA}\,(G_{\alpha,0})$, то при надлежащем выборе положительных постоянных a_p распределения нормированных геометрических случайных сумм $a_p(X_1+\cdots+X_{V_p})$ сходятся к распределению Линника с параметром α. Распределения Линника унимодальны [35], безгранично делимы [36], имеют бесконечный пик плотности в нуле при $\alpha \leq 1$ [36] и т. п. Аналитические и асимптотические свойства распределения Линника рассмотрены в [17, 33, 37, 38]. В частности, в работах [17, 32] установлена интересная связь между распределениями Линника, Лапласа и Миттаг–Леффлера и показано, что

$$L_\alpha \stackrel{\mathrm{d}}{=} X\sqrt{2M_{\alpha/2}} \stackrel{\mathrm{d}}{=} \Lambda\sqrt{R_{\alpha/2}}\,, \qquad (24)$$

где все сомножители независимы, а с.в. $R_{\alpha/2}$ определена в лемме 4.

Из теоремы 7, следствия 2 и леммы 1 непосредственно вытекает следующее утверждение.

Следствие 6. Предположим, что независимые одинаково распределенные неотрицательные с.в. $X_1, X_2, \ldots$ таковы, что $\mathcal{L}(X_1) \in \mathrm{DNA}\,(F_{\alpha,0})$ при некотором $\alpha \in (0,2)$. Пусть числа $r \in (0,1)$, $q \in (0,1)$ и $\mu > 0$ произвольны. Пусть при каждом $n \in \mathbb{N}$ с.в. N_{r,p_n} имеют отрицательные биномиальные распределения с параметрами r и $p_n = \min\{q, \mu/n\}$ и независимы от $X_1, X_2, \ldots$ Тогда

$$\frac{S_{N_{r,p_n}}}{n^{1/\alpha}} \Longrightarrow T_{\alpha,0} \cdot G_{r,\mu}^{1/\alpha} \qquad (25)$$

при $n \to \infty$.

С учетом (1) и (24) с.в., предельная в следствии 6, может быть записана в разных эквивалентных формах:

$$T_{\alpha,0} G_{r,\mu}^{1/\alpha} \stackrel{\mathrm{d}}{=} T_{\alpha,0} G_{r,\alpha,\mu}^{*} \stackrel{\mathrm{d}}{=} X\sqrt{2T_{\alpha/2,1} G_{r,\mu}^{2/\alpha}} \stackrel{\mathrm{d}}{=} X\sqrt{2T_{\alpha/1,1} G_{r,\alpha/2,\mu}^{*}} \stackrel{\mathrm{d}}{=} X\frac{\sqrt{2M_{\alpha/2}}}{Z_{r,\mu}^{1/\alpha}} \stackrel{\mathrm{d}}{=} \Lambda\frac{\sqrt{R_{\alpha/2}}}{Z_{r,\mu}^{1/\alpha}} \stackrel{\mathrm{d}}{=} L_\alpha Z_{r,\mu}^{-1/\alpha},$$

т. е. распределение, предельное в следствии 6, допускает представления в виде масштабной смеси

как симметричного строго устойчивого распределения, так и распределения Линника, или нормального закона, или распределения Лапласа.

С учетом абсолютной непрерывности предельного распределения в следствии 6 можно заключить, что на самом деле в условии (25) речь идет о равномерной сходимости ф.р.:

$$\lim_{n\to\infty}\sup_{x\geq 0}\left|\mathsf{P}\left(\frac{S_{N_{r,p_n}}}{n^{1/\alpha}}<x\right)-\int\limits_0^x F_{\alpha,0}\left(\frac{x}{z^{1/\alpha}}\right)g(z;\,r,\mu)\,dz\right|=0\,.$$

Замечание 5. Сходимость

$$n^{-1/\alpha}S_{N_{r,p_n}}\Longrightarrow T_{\alpha,0}G_{r,\mu}^{1/\alpha}\stackrel{\mathrm{d}}{=}T_{\alpha,0}G_{r,\alpha,\mu}^{*}$$

имеет место и для $r\geq 1$.

Замечание 6. Использовав формулу для абсолютных моментов симметричных строго устойчивых распределений (см. подразд. 1.3), для моментов порядков $\beta<\alpha$ с.в., предельных в следствии 6, получим представление:

$$\mathsf{E}\left(|T_{\alpha,0}|G_{r,\mu}^{1/\alpha}\right)^{\beta}=$$
$$=\frac{2^{\beta}\Gamma\left((\beta+\alpha r)/\alpha\right)\Gamma\left((\alpha-\beta)/\alpha\right)\Gamma\left((\beta+1)/2\right)}{\sqrt{\pi}\,\mu^{\beta/\alpha}\Gamma(r)\Gamma\left((2-\beta)/\beta\right)}\,.$$

4.4 Центральная предельная теорема для отрицательных биномиальных случайных сумм независимых одинаково распределенных случайных величин

В приведенных выше рассуждениях особый интерес представляет случай $\alpha=2$. Хотя соответствующий вариант следствия 6 можно получить простой заменой α на 2 в его формулировке, здесь этот случай будет рассмотрен особо, поскольку соответствующее утверждение можно трактовать как центральную предельную теорему для отрицательных биномиальных случайных сумм $S_{N_{r,p_n}}$, когда $p_n\to 0$ при $n\to\infty$.

Итак, пусть $X_1,X_2,\ldots$ — независимые одинаково распределенные с.в. с $\mathsf{E}X_1=0$ и $\mathsf{E}X_1^2=1$. Известно, что в общем случае в аналогах центральной предельной теоремы для отрицательных биномиальных случайных сумм в качестве предельных законов возникают так называемые VG-распределения (Variance Gamma distributions) — специальные дисперсионно-сдвиговые смеси нормальных законов, в которых смешивающими являются гамма-распределения (см., например, [39, 40]). В рассматриваемом здесь частном случае $r<1$ можно предложить еще одну версию условий и еще одну форму записи предельных VG-распределений в виде смеси распределений Лапласа, что позволяет при статистическом оценивании параметров предельных законов использовать медианные версии EM (expectation-maximization) алгоритма, более устойчивые к исходным данным (см., например, [41]).

Из теоремы 7 с $\alpha=2$, следствия 2 и леммы 1 непосредственно вытекает следующее утверждение.

Следствие 7. Предположим, что независимые одинаково распределенные с.в. $X_1,X_2,\ldots$ таковы, что $\mathsf{E}X_1=0$ и $\mathsf{E}X_1^2=1$. Пусть числа $r\in(0,1)$, $q\in(0,1)$ и $\mu>0$ произвольны. Пусть при каждом $n\in\in\mathbb{N}$ с.в. N_{r,p_n} имеют отрицательные биномиальные распределения с параметрами r и $p_n=\min\{q,\mu/n\}$ и независимы от $X_1,X_2,\ldots$ Тогда

$$\frac{S_{N_{r,p_n}}}{\sqrt{n}}\Longrightarrow X\sqrt{G_{r,\mu/2}}\stackrel{\mathrm{d}}{=}XG_{r,2,\mu/2}^{*}\stackrel{\mathrm{d}}{=}\frac{\Lambda}{\sqrt{Z_{r,\mu}}}\quad(26)$$

при $n\to\infty$. Более того, сходимость ф.р. с.в., участвующих в (26), является равномерной:

$$\lim_{n\to\infty}\sup_{x}\left|\mathsf{P}\left(\frac{S_{N_{r,p_n}}}{\sqrt{n}}<x\right)-\int\limits_0^{\infty}\Phi\left(\frac{x}{\sqrt{z}}\right)g(z;r,\mu/2)\,dz\right|=$$
$$=\lim_{n\to\infty}\sup_{x}\left|\mathsf{P}\left(\frac{S_{N_{r,p_n}}}{\sqrt{n}}<x\right)-\int\limits_0^{\infty}F^{\Lambda}\left(x\sqrt{z}\right)p(z;r,\mu)\,dz\right|=0\,.$$

Другими словами, предельное VG-распределение в данном случае является масштабной смесью распределений Лапласа, в которой смешивающим служит распределение с.в. $Z_{r,\mu}$.

Замечание 7. Сходимость

$$n^{-1/2}S_{N_{r,p_n}}\Longrightarrow X\sqrt{G_{r,\mu/2}}\stackrel{\mathrm{d}}{=}XG_{r,2,\mu/2}^{*}$$

имеет место и для $r\geq 1$.

Литература

1. *Zolina O., Simmer C., Belyaev K., Gulev S., Koltermann P.* Changes in the duration of European wet and dry spells during the last 60 years // J. Climate, 2013. Vol. 26. P. 2022–2047.

2. *Korolev V. Yu., Gorshenin A. K., Gulev S. K., Belyaev K. P., Grusho A. A.* Statistical analysis of precipitation events // AIP Conf. Proc., 2017. Vol. 1863. Iss. 1. doi: 10.1063/1.4992276.
3. *Kingman J. F. C.* Poisson processes. — Oxford: Clarendon Press, 1993. 104 p.
4. *Королев В. Ю., Бенинг В. Е., Шоргин С. Я.* Математические основы теории риска. — 2-е изд. — М.: Физматлит, 2011. 591 с.
5. *Королев В. Ю.* Предельные распределения для дважды стохастически прореженных процессов восстановления и их свойства // Теория вероятностей и ее применения, 2016. Т. 61. Вып. 4. С. 753–773.
6. *Королев В. Ю., Корчагин А. Ю., Зейфман А. И.* Теорема Пуассона для схемы испытаний Бернулли со случайной вероятностью успеха и дискретный аналог распределения Вейбулла // Информатика и её применения, 2016. Т. 10. Вып. 4. С. 11–20.
7. *Korolev V. Yu., Korchagin A. Yu., Zeifman A. I.* On doubly stochastic rarefaction of renewal processes // AIP Conf. Proc., 2017. Vol. 1863. Iss. 1. doi: 10.1063/1.4992275.
8. *Gleser L. J.* The gamma distribution as a mixture of exponential distributions // Am. Stat., 1989. Vol. 43. P. 115–117.
9. *Gnedenko B. V., Korolev V. Yu.* Random summation: Limit theorems and applications. — Boca Raton: CRC Press, 1996. 267 p.
10. *Stacy E. W.* A generalization of the gamma distribution // Ann. Math. Stat., 1962. Vol. 33. P. 1187–1192.
11. *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.
12. *Золотарев В. М.* Одномерные устойчивые распределения. — М.: Наука, 1983. 304 с.
13. *Schneider W. R.* Stable distributions: Fox function representationand generalization // Stochastic processes in classical and quantum systems / Eds. S. Albeverio, G. Casati, D. Merlini. — Berlin: Springer, 1986. P. 497–511.
14. *Uchaikin V. V., Zolotarev V. M.* Chance and stability. — Utrecht: VSP, 1999. 570 p.
15. *Korolev V. M.* Product representations for random variables with Weibull distributions and their applications // J. Math. Sci., 2016. Vol. 218. No. 3. P. 298–313.
16. *Tucker H.* On moments of distribution functions attracted to stable laws // Houston J. Math., 1975. Vol. 1. No. 1. P. 149–152.
17. *Korolev V. Yu., Zeifman A. I.* Convergence of statistics constructed from samples with random sizes to the Linnik and Mittag–Leffler distributions and their generalizations // J. Korean Stat. Soc., 2017. Vol. 46. P. 161–181.
18. *Gorenflo R., Kilbas A. A., Mainardi F., Rogosin S. V.* Mittag–Leffler functions, related topics and applications. — Berlin – New York: Springer, 2014. 443 p.
19. *Bunge J.* Compositions semigroups and random stability // Ann. Probab., 1996. Vol. 24. P. 1476–1489.
20. *Klebanov L. B., Rachev S. T.* Sums of a random number of random variables and their approximations with ε-accompanying infinitely divisible laws // Serdica, 1996. Vol. 22. P. 471–498.
21. *Коваленко И. Н.* О классе предельных распределений для редеющих потоков однородных событий // Литовский математический сборник, 1965. Т. 5. Вып. 4. С. 569–573.
22. *Gnedenko B. V., Kovalenko I. N.* Introduction to queueing theory. — Jerusalem: Israel Program for Scientific Translations, 1968. 281 p.
23. *Gnedenko B. V., Kovalenko I. N.* Introduction to queueing theory. — 2nd ed. — Boston: Birkhauser, 1989. 314 p.
24. *Pillai R. N.* Harmonic mixtures and geometric infinite divisibility // J. Indian Stat. Assoc., 1990. Vol. 28. P. 87–98.
25. *Pillai R. N.* On Mittag–Leffler functions and related distributions // Ann. Stat. Math., 1990. Vol. 42. P. 157–161.
26. *Weron K., Kotulski M.* On the Cole-Cole relaxation function and related Mittag–Leffler distributions // Physica A, 1996. Vol. 232. P. 180–188.
27. *Gorenflo R., Mainardi F.* Continuous time random walk, Mittag–Leffler waiting time and fractional diffusion: Mathematical aspects. Ch. 4. // Anomalous transport: Foundations and applications / Eds. R. Klages, G. Radons, I. M. Sokolov. — Weinheim, Germany: Wiley-VCH, 2008. P. 93–127. http://arxiv.org/abs/0705.0797.
28. *Greenwood M., Yule G. U.* An inquiry into the nature of frequency-distributions of multiple happenings, etc. // J. R. Stat. Soc., 1920. Vol. 83. P. 255–279.
29. *Большев Л. Н., Смирнов Н. В.* Таблицы математической статистики. — 3-е изд. — М.: Наука, 1983. 416 с.
30. *Bernstein S. N.* Sur les fonctions absolument monotones // Acta Math., 1929. Vol. 52. Iss. 1. P. 1–66.
31. *Grandell J.* Mixed Poisson processes. — London: Chapman and Hall, 1997. 268 p.
32. *Kotz S., Ostrovskii I. V.* A mixture representation of the Linnik distribution // Stat. Probabil. Lett., 1996. Vol. 26. P. 61–64.
33. *Korolev V. Yu., Zeifman A. I.* A note on mixture representations for the Linnik and Mittag–Leffler distributions and their applications // J. Math. Sci., 2017. Vol. 218. No. 3. P. 314–327.
34. *Линник Ю. В.* Линейные формы и статистические критерии. I, II // Украинский математический журнал, 1953. Т. 5. Вып. 2. С. 207–243; Вып. 3. С. 247–290.
35. *Laha R. G.* On a class of unimodal distributions // P. Am. Math. Soc., 1961. Vol. 12. P. 181–184.
36. *Devroye L.* A note on Linnik's distribution // Stat. Probabil. Lett., 1990. Vol. 9. P. 305–306.
37. *Kotz S., Ostrovskii I. V., Hayfavi A.* Analytic and asymptotic properties of Linnik's probability densities, I // J. Mathematical Analysis Appl., 1995. Vol. 193. P. 353–371.

38. *Kotz S., Ostrovskii I. V., Hayfavi A.* Analytic and asymptotic properties of Linnik's probability densities, II // J. Math. Anal. Appl., 1995. Vol. 193. P. 497–521.
39. *Carr P. P., Madan D. B., Chang E. C.* The variance gamma process and option pricing // Eur. Financ. Rev., 1998. Vol. 2. P. 79–105.
40. *Королев В. Ю.* Обобщенные гиперболические законы как предельные распределения для случайных сумм // Теория вероятностей и ее применения, 2013. Т. 58. Вып. 1. С. 117–132.
41. *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 индексов // Информатика и её применения, 2008. Т. 2. Вып. 4. С. 12–47.

Поступила в редакцию 11.05.17

ANALOGS OF GLESER'S THEOREM FOR NEGATIVE BINOMIAL AND GENERALIZED GAMMA DISTRIBUTIONS AND SOME OF THEIR APPLICATIONS

V. Yu. Korolev[1,2,3]

[1]Faculty of Computational Mathematics and Cybernetics, M. V. Lomonosov Moscow State University, 1-52 Leninskiye Gory, Moscow 119991, GSP-1, Russian Federation
[2]Institute of Informatics Problems, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, 44-2 Vavilov Str., Moscow 119333, Russian Federation
[3]Hangzhou Dianzi University, Higher Education Zone, Hangzhou 310018, China

Abstract: It is proved that the negative binomial distributions with the shape parameter less than one are mixed geometric distributions. The mixing distribution is written out explicitly. Thus, the similar result of L. Gleser, stating that the gamma distributions with the shape parameter less than one are mixed exponential distributions, is transferred to the discrete case. An analog of Gleser's theorem is also proved for generalized gamma distributions. For mixed binomial distributions related to the negative binomial laws with the shape parameter less than one, the case of a small probability of success is considered and an analog of the Poisson theorem is proved. The representation of the negative binomial distributions as mixed geometric laws is used to prove limit theorems for negative binomial random sums of independent identically distributed random variables, in particular, analogs of the law of large numbers and the central limit theorem. Both cases of light and heavy tails are considered. The expressions for the moments of limit distributions are obtained. The obtained alternative equivalent mixture representations of the limit laws provide better understanding of how mixed probability (Bayesian) models are formed.

Keywords: negative binomial distribution; mixed geometric distribution; generalized gamma distribution; stable distribution; Laplace distribution; Mittag–Leffler distribution; Linnik distribution; mixed binomial distribution; Poisson theorem; random sum; law of large numbers; central limit theorem

DOI: 10.14357/19922264170301

Acknowledgments

The research was partially supported by the RAS Presidium Program No. I.33P (project 0063-2016-0015) and by the Russian Foundation for Basic Research (projects Nos. 15-07-04040 and 17-07-00717).

References

1. Zolina, O., C. Simmer, K. Belyaev, S. Gulev, and P. Koltermann. 2013. Changes in the duration of European wet and dry spells during the last 60 years. *J. Climate* 26:2022–2047.
2. Korolev, V. Yu., A. K. Gorshenin, S. K. Gulev, K. P. Belyaev, and A. A. Grusho. 2017. Statistical analysis of precipitation events. *AIP Conf. Proc.* 1863(1). doi: 10.1063/1.4992276.
3. Kingman, J. F. C. 1993. *Poisson processes*. Oxford: Clarendon Press. 104 p.
4. Korolev, V. Yu., V. E. Bening, and S. Ya. Shorgin. 2011. *Matematicheskie osnovy teorii riska* [Mathematical fundamentals of risk theory]. 2nd ed. Moscow: Fizmatlit. 591 p.

5. Korolev, V. Yu. 2016. Predel'nye raspredeleniya dlya dvazhdy stohasticheski prorezhennykh protsessov vosstanovleniya i ikh svoystva [Limit distributions for doubly stochastically rarefied renewal processes and their properties]. *Teoriya veroyatnostey i ee primeneniya* [Theor. Probab. Appl.] 61(4):753–773.
6. Korolev, V. Yu., A. Yu. Korchagin, and A. I. Zeifman. 2016. Teorema Puassona dlya skhemy ispytaniy Bernulli so sluchaynoy veroyatnost'yu uspekha i diskretnyy analog raspredeleniya Veybulla [The Poisson theorem for the scheme of Bernoulli trials with a random probability of success and a discrete analog of the Weibull distribution]. *Informatika i ee Primeneniya — Inform. Appl.* 10(4):11–20.
7. Korolev, V. Yu., A. Yu. Korchagin, and A. I. Zeifman. 2017. On doubly stochastic rarefaction of renewal processes. *AIP Conf. Proc.* 1863(1). doi: 10.1063/1.4992275.
8. Gleser, L. J. 1989. The gamma distribution as a mixture of exponential distributions. Am. Stat. 43:115–117.
9. Gnedenko, B. V., and V. Yu. Korolev. 1996. *Random summation: Limit theorems and applications*. Boca Raton: CRC Press. 267 p.
10. Stacy, E. W. 1962. A generalization of the gamma distribution. *Ann. Math. Stat.* 33:1187–1192.
11. Zaks, L. M., and V. Yu. Korolev. 2013. Obobshchennye dispersionnye gamma-raspredeleniya kak predel'nye dlya sluchaynykh summ [Generalized variance gamma distributions as limit laws for random sums]. *Informatika i ee Primeneniya — Inform. Appl.* 7(1):105–115.
12. Zolotarev, V. M. 1983. *Odnomernye ustoychivye raspredeleniya* [One-dimensional stable distributions]. Moscow: Nauka. 304 p.
13. Schneider, W. R. 1986. Stable distributions: Fox function representation and generalization. *Stochastic processes in classical and quantum systems*. Eds. S. Albeverio, G. Casati, and D. Merlini. Berlin: Springer. P. 497–511.
14. Uchaikin, V. V., and V. M. Zolotarev. 1999. *Chance and stability*. Utrecht: VSP. 570 p.
15. Korolev, V. Yu. 2016. Product representations for random variables with Weibull distributions and their applications. *J. Math. Sci.* 218(3):298–313.
16. Tucker, H. 1975. On moments of distribution functions attracted to stable laws. *Houston J. Math.* 1(1):149–152.
17. Korolev, V. Yu., and A. I. Zeifman. 2017. Convergence of statistics constructed from samples with random sizes to the Linnik and Mittag–Leffler distributions and their generalizations. *J. Korean Stat. Soc.* 46:161–181.
18. Gorenflo, R., A. A. Kilbas, F. Mainardi, and S. V. Rogosin. 2014. *Mittag–Leffler functions, related topics and applications*. Berlin – New York: Springer. 443 p.
19. Bunge, J. 1996. Compositions semigroups and random stability. *Ann. Probab.* 24:1476–1489.
20. Klebanov, L. B., and S. T. Rachev. 1996. Sums of a random number of random variables and their approximations with ε-accompanying infinitely divisible laws. *Serdica* 22:471–498.
21. Kovalenko, I. N. 1965. O klasse predel'nykh raspredeleniy dlya redeyushchikh potokov odnorodnykh sobytiy [On the class of limit distributions for rarefying flows of homogeneous events]. *Litovskiy matematicheskiy sbornik* [Lithuanian Math. J.] 5(4):569–573.
22. Gnedenko, B. V., and I. N. Kovalenko. 1968. *Introduction to queueing theory*. Jerusalem: Israel Program for Scientific Translations. 281 p.
23. Gnedenko, B. V., and I. N. Kovalenko. 1989. *Introduction to queueing theory*. 2nd ed. Boston: Birkhauser. 314 p.
24. Pillai, R. N. 1990. Harmonic mixtures and geometric infinite divisibility. *J. Indian Stat. Assoc.* 28:87–98.
25. Pillai, R. N. 1990. On Mittag–Leffler functions and related distributions. *Ann. Stat. Math.* 42:157–161.
26. Weron, K., and M. Kotulski. 1996. On the Cole-Cole relaxation function and related Mittag–Leffler distributions. *Physica A* 232:180–188.
27. Gorenflo, R., and F. Mainardi. 2008. Continuous time random walk, Mittag–Leffler waiting time and fractional diffusion: Mathematical aspects. Ch. 4. *Anomalous transport: Foundations and applications*. Eds. R. Klages, G. Radons, and I. M. Sokolov. Weinheim, Germany: Wiley-VCH. 93–127. Available at: http://arxiv.org/abs/0705.0797 (accessed August 25, 2017).
28. Greenwood, M., and G. U. Yule. 1920. An inquiry into the nature of frequency-distributions of multiple happenings, etc. *J. R. Stat. Soc.* 83:255–279.
29. Bol'shev, L. N., and N. V. Smirnov. 1983. *Tablitsy matematicheskoy statistiki* [Tables of mathematical statistics]. 3rd ed. Moscow: Nauka. 416 p.
30. Bernstein, S. N. 1929. Sur les fonctions absolument monotones. *Acta Math.* 52(1):1–66.
31. Grandell, J. 1997. *Mixed Poisson processes*. London: Chapman and Hall. 268 p.
32. Kotz, S., and I. V. Ostrovskii. 1996. A mixture representation of the Linnik distribution. *Stat. Probabil. Lett.* 26:61–64.
33. Korolev, V. Yu., and A. I. Zeifman. 2017. A note on mixture representations for the Linnik and Mittag–Leffler distributions and their applications. *J. Math. Sci.* 218(3):314–327.
34. Linnik, Yu. V. 1953. Lineynye formy i statisticheskie kriterii. I, II [Linear forms and statistical tests, I, II]. *Ukrainskiy matematicheskiy zh.* [Ukrainian Math. J.] 5(2):207–243; 5(3):247–290.
35. Laha, R. G. 1961. On a class of unimodal distributions. *P. Am. Math. Soc.* 12:181–184.
36. Devroye, L. 1990. A note on Linnik's distribution. *Stat. Probabil. Lett.* 9:305–306.
37. Kotz, S., I. V. Ostrovskii, and A. Hayfavi. 1995. Analytic and asymptotic properties of Linnik's probability densities, I. *J. Math. Anal. Appl.* 193:353–371.

38. Kotz, S., I. V. Ostrovskii, and A. Hayfavi. 1995. Analytic and asymptotic properties of Linnik's probability densities, II. *J. Math. Anal. Appl.* 193:497–521.
39. Carr, P. P., D. B. Madan, and E. C. Chang. 1998. The Variance Gamma process and option pricing. *Eur. Financ. Rev.* 2:79–105.
40. Korolev, V. Yu. 2014. Generalized hyperbolic laws as limit distributions for random sums. *Theor. Probab. Appl.* 58(1):63–75.
41. Gorshenin, A. K., V. Yu. Korolev, and A. M. Tursunbaev. 2008. Mediannye modifikatsii EM- i SEM-algoritmov dlya razdeleniya smesey veroyatnostnykh raspredeleniy i ikh primenenie k dekompozitsii volatil'nosti finansovykh indeksov [Median modifications of the EM- and SEM-algorithms for the separation of mixtures of probability distributions and their application to the decomposition of volatility of financial indexes]. *Informatika i ee Primeneniya — Inform. Appl.* 2(4):12–47.

Received May 11, 2017

Contributor

Korolev Victor Yu. (b. 1954) — Doctor of Science in physics and mathematics, professor, Head of the Department of Mathematical Statistics, Faculty of Computational Mathematics and Cybernetics, 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, Federal Research Center "Computer Science and Control"' of the Russian Academy of Sciences, 44-2 Vavilov Str., Moscow 119333, Russian Federation; professor, Hangzhou Dianzi University, Xiasha Higher Education Zone, Hangzhou 310018, China; vkorolev@cs.msu.su

ИНФОРМАТИКА И ЕЁ ПРИМЕНЕНИЯ, 2017. Т. 11. Вып. 3. С. 18–26

СЕГМЕНТИРОВАНИЕ НЕСТАЦИОНАРНЫХ СИГНАЛОВ НА ОСНОВЕ ВЕРОЯТНОСТНЫХ СВОЙСТВ ОКОННОЙ ДИСПЕРСИИ

М. А. Драницына[1], Т. В. Захарова[2]

Аннотация: Выделение фрагментов регистрируемого сигнала, т. е. его сегментация, является актуальной задачей, в частности для биомедицинской отрасли. Сегментация как этап обработки сигналов, зачастую обязательный, может способствовать интерпретации и классификации регистрируемых данных. Особенно сложно сегментировать нестационарные сигналы с малым отношением сигнал/шум. В рамках данной работы основное внимание уделяется изучению шумовой компоненты оконной дисперсии как случайной величины в рассматриваемых моделях. Авторами предложены модели для представления мультикомпонентных сигналов, а также исследованы некоторые вероятностные характеристики шумовой компоненты оконной дисперсии сигналов как случайного процесса в представленных моделях. Результаты работы согласуются с установленными эмпирически свойствами шумовой компоненты оконной дисперсии (для миограммы). Полученные результаты планируется использовать в практических задачах сегментирования сигналов и выделения интервалов с преобладанием тех или иных компонент процесса, а также для прогнозирования поведения сигналов.

Ключевые слова: оконная дисперсия; модель сигнала

DOI: 10.14357/19922264170302

1 Введение

Выделение фрагментов регистрируемого сигнала с различными характеристиками, т. е. его сегментация, является актуальной задачей, в частности для биомедицинской отрасли (например, при анализе электроэнцефалограмм [1, 2], данных различных мониторирующих состояние здоровья устройств [3] и других моно- и мультикомпонентных сигналов [4]). Сегментация сигналов как этап их обработки, зачастую обязательный [3, 4], может способствовать интерпретации и классификации регистрируемых данных.

Будем рассматривать некоторый нестационарный сигнал. Этот сигнал может быть представлен в виде временного ряда, который образуют результаты измерения сигнала в точках τ_k, $k = 1, 2, \ldots, r$. При этом сигнал образован составляющими его процессами $A_1, A_2, \cdots, A_m$, каждый из которых может быть преобладающим на том или ином временно́м интервале регистрации сигнала.

Примером такого сигнала может служить фармакокинетическая кривая, т. е. кривая, отражающая зависимость концентрации вещества, чаще всего в крови, от времени. Для действующего вещества лекарственного препарата, представляющего собой таблетку или капсулу, профиль такой кривой определяется скоростью абсорбции вещества из просвета тонкой кишки, являющейся наиболее частым абсорбирующим органом, и количеством уже абсорбированного вещества, скоростью распределения вещества из крови в периферические ткани (с достижением динамического равновесия) и выведением его из организма как за счет метаболизма, так и выделения соответствующими органами.

Ассоциированная с подлежащими процессами шумовая компонента предполагается случайной величиной. Изменение вероятностных характеристик шумовой компоненты оконной дисперсии будет основанием для сегментирования регистрируемого сигнала в дальнейшем на практике. Оконная дисперсия для выделения определенных участков сигналов была использована, например, для анализа магнитоэнцефалограмм [4] с целью определения момента начала движения. Для выделения опорных точек на миограмме, регистрируемой параллельно с магнитоэнцефалограммой, в работе [5] была предложена методология, которая использована и обоб-

[1]Московский государственный университет имени М. В. Ломоносова, факультет вычислительной математики и кибернетики; margarita13april@mail.ru

[2]Московский государственный университет имени М. В. Ломоносова, факультет вычислительной математики и кибернетики; Институт проблем информатики Федерального исследовательского центра «Информатика и управление» Российской академии наук, lsa@cs.msu.ru

щена для модели с общей шумовой компонентой в настоящей работе. Кроме того, распределение шумовой компоненты оконной дисперсии миограммы было охарактеризовано эмпирически в работе [6].

В рамках данной работы рассматриваются модели мультикомпонентных сигналов, при этом основное внимание уделяется исследованию изменения вероятностных характеристик оконной дисперсии сигнала на разных временны́х интервалах.

2 Модель с общей шумовой компонентой

2.1 Общее представление модели

Пусть регистрируется некий сигнал. Представим в каждой точке t наблюдаемое значение $C(t)$ в виде суммы истинных значений процессов, формирующих результирующий сигнал, $A_1(t), A_2(t), \ldots, A_m(t)$ и белого гауссовского шума $\xi(t)$, характеризующегося нормальным распределением с нулевым математическим ожиданием и дисперсией σ^2. Таким образом, наблюдаемый сигнал $C(t)$ представим в виде:

$$C(t) = \sum_{l=1}^{m} A_l(t) + \xi(t)\,. \tag{1}$$

Пусть n — ширина окна, т. е. число точек, используемых при расчете скользящего среднего $\bar{C}_n$ сигнала $C(t)$. Через C_i и ξ_i для $i = 0, 1, \ldots, n-1$ обозначим соответственно значение сигнала и шума в i-й точке окна.

Во введенных обозначениях для $\bar{C}_n$ справедливо следующее представление:

$$\bar{C}_n = \frac{1}{n}\sum_{i=0}^{n-1} C_i = \frac{1}{n}\sum_{i=0}^{n-1}\left[\sum_{l=1}^{m} A_{l,i} + \xi_i\right] = \sum_{l=1}^{m} \bar{A}_{l,n} + \bar{\xi}_n\,, \tag{2}$$

где $\bar{A}_{1,n}, \bar{A}_{2,n}, \ldots, \bar{A}_{m,n}$, — скользящее среднее истинных составляющих регистрируемого сигнала, а $\bar{\xi}_n$ — скользящее среднее шума.

Оконная дисперсия по определению имеет вид:

$$W_n = \frac{1}{n}\sum_{i=0}^{n-1}\left(C_i - \bar{C}_n\right)^2\,. \tag{3}$$

Исследуем свойства оконной дисперсии сигнала и оконной дисперсии шума в рассматриваемой модели.

Лемма. *Для оконной дисперсии W_n справедливо следующее представление*:

$$W_n = \sum_{l=1}^{m} W_n^{A_l} + \sum_{s\neq l;s,l=1}^{m} W_n^{A_l,s} + W_n^{\xi} + \sum_{l=1}^{m} W_n^{A_l\xi}\,, \tag{4}$$

где

$$W_n^{A_l} = \frac{1}{n}\sum_{i=0}^{n-1} A_{l,i}^2 + \bar{A}_{l,n}^2\,,\ \ l \in \{1, 2, \ldots, m\}\,;$$

$$W_n^{A_l,s} = \frac{2}{n}\sum_{i=0}^{n-1} A_{l,i}\left(A_{s,i} - \bar{A}_{s,n}\right)\,,\quad l, s \in \{1, 2, \ldots, m\}\,, l \neq s\,;$$

$$W_n^{\xi} = \frac{1}{n}\sum_{i=0}^{n-1} \xi_i^2 + \bar{\xi}_n^2\,;$$

$$W_n^{A_l\xi} = \frac{2}{n}\sum_{i=0}^{n-1} \xi_i\left(A_{l,i} - \bar{A}_{l,n}\right)\,,\ \ l \in \{1, 2, \ldots, m\}\,.$$

Доказательство. Подставим (1) и (2) в уравнение (3) и раскроем скобки, тогда

$$W_n = \frac{1}{n}\sum_{i=0}^{n-1}\left(\sum_{l=1}^{m}\left(A_{l,i} - \bar{A}_{l,n}\right) + \left(\xi_i - \bar{\xi}_n\right)\right)^2 =$$

$$= \frac{1}{n}\sum_{i=0}^{n-1}\left(A_{1,i} - \bar{A}_{1,n}\right)^2 + \cdots + \frac{1}{n}\sum_{i=0}^{n-1}\left(A_{l,i} - \bar{A}_{l,n}\right)^2 + \frac{1}{n}\sum_{i=0}^{n-1}\left(\xi_i - \bar{\xi}_n\right)^2 +$$

$$+ \sum_{s\neq l;s,l=1}^{m}\left[\frac{2}{n}\sum_{i=0}^{n-1}\left(A_{l,i} - \bar{A}_{l,n}\right)\left(A_{s,i} - \bar{A}_{s,n}\right)\right] + \sum_{l=1}^{m}\left[\frac{2}{n}\sum_{i=0}^{n-1}\left(A_{l,i} - \bar{A}_{l,n}\right)\left(\xi_i - \bar{\xi}_n\right)\right] =$$

$$= \sum_{l=1}^{m} W_n^{A_l} + \sum_{s\neq l;s,l=1}^{m} W_n^{A_l,s} + W_n^{\xi} + \sum_{l=1}^{m} W_n^{A_l\xi}\,.$$

Последнее равенство совпадает с утверждением леммы.

Такое представление (4) оконной дисперсии может быть интерпретировано следующим образом:

- компоненты $W_n^{A_l}, l \in \{1, 2, \ldots, m\}$, характеризуют тренд, обусловленный истинными компонентами регистрируемого сигнала в отсутствие шума;
- компоненты $W_n^{A_l,s}, s, l \in \{1, 2, \ldots, m\}, s \neq l$, характеризуют суперпозицию истинных компонент регистрируемого сигнала;

– компонента W_n^{ξ} характеризует дисперсию случайной компоненты регистрируемого сигнала — оконная дисперсия шума;

– компоненты $W_n^{A_l\xi}$ характеризуют суперпозицию истинных компонент и шума.

Таким образом, оконная дисперсия рассматриваемого сигнала состоит из суммы компонент, обусловленных изменением истинного сигнала во времени, и компонент, ассоциированных с шумом.

Раскроем скобки для компонент $W_n^{A_l}$, $l \in \{1, 2, \ldots, m\}$, и получим:

$$W_n^{A_l} = \frac{1}{n}\sum_{i=0}^{n-1}\left(A_{l,i} - \bar{A}_{l,n}\right)^2 = \frac{1}{n}\sum_{i=0}^{n-1}\left(A_{l,i}^2 + \bar{A}_{l,n}^2 - 2A_{l,i}\bar{A}_{l,n}\right) = \frac{1}{n}\left[\sum_{i=0}^{n-1}\left(A_{l,i}^2 - 2A_{l,i}\bar{A}_{l,n}\right) + n\bar{A}_{l,n}^2\right] = \frac{1}{n}\sum_{i=0}^{n-1}A_{l,i}^2 - \bar{A}_{l,n}^2 .$$

Для оконной дисперсии шума W_n^{ξ} справедливы аналогичные преобразования и представление:

$$W_n^{\xi} = \frac{1}{n}\sum_{i=0}^{n-1}\left(\xi_i - \bar{\xi}_n\right)^2 = \frac{1}{n}\sum_{i=0}^{n-1}\left(\xi_i^2 + \bar{\xi}_n^2 - 2\xi_i\bar{\xi}_n\right) = \frac{1}{n}\left[\sum_{i=0}^{n-1}\left(\xi_i^2 - 2\xi_i\bar{\xi}_n\right) + n\bar{\xi}_n^2\right] = \frac{1}{n}\sum_{i=0}^{n-1}\xi_i^2 - \bar{\xi}_n^2 .$$

Рассмотрим компоненты, представляющие собой суперпозиции истинных компонент сигнала $W_n^{A_l,s}$ для $s, l \in \{1, 2, \ldots, m\}$, $s \neq l$, тогда

$$W_n^{A_l,s} = \frac{2}{n}\sum_{i=0}^{n-1}\left(A_{l,i} - \bar{A}_{l,n}\right)\left(A_{s,i} - \bar{A}_{s,n}\right) = \frac{2}{n}\sum_{i=0}^{n-1}A_{s,i}\left(A_{l,i} - \bar{A}_{l,n}\right) - \frac{2\bar{A}_{s,n}}{n}\sum_{i=0}^{n-1}\left(A_{l,i} - \bar{A}_{l,n}\right) = \frac{2}{n}\sum_{i=0}^{n-1}A_{s,i}\left(A_{l,i} - \bar{A}_{l,n}\right) .$$

Проведем аналогичные преобразования для компонент $W_n^{A_l\xi}$, $l \in \{1, 2, \ldots, m\}$, и получим:

$$W_n^{A_l\xi} = \frac{2}{n}\sum_{i=0}^{n-1}\left(A_{l,i} - \bar{A}_{l,n}\right)\left(\xi_i - \bar{\xi}_n\right) = \frac{2}{n}\sum_{i=0}^{n-1}\xi_i\left(A_{l,i} - \bar{A}_{l,n}\right) - \frac{2\bar{\xi}_n}{n}\sum_{i=0}^{n-1}\left(A_{l,i} - \bar{A}_{l,n}\right) = \frac{2}{n}\sum_{i=0}^{n-1}\xi_i\left(A_{l,i} - \bar{A}_{l,n}\right) .$$

Подставив полученные выражения в уравнение (4), получим утверждение леммы.

2.2 Свойства шумовой компоненты оконной дисперсии в модели с общей шумовой компонентой

Обозначим через W_n^{Ξ} шумовую компоненту оконной дисперсии регистрируемого сигнала, которая представляет собой сумму оконной дисперсии шума и суперпозицию шума и истинных компонент:

$$W_n^{\Xi} = W_n^{\xi} + \sum_{l=1}^{m} W_n^{A_l\xi} . \qquad (5)$$

Теорема. *В каждой точке наблюдения τ_k шумовая компонента оконной дисперсии W_n^{Ξ} представима в виде*:

$$W_n^{\Xi} = W_n^{\xi} ,$$

если $\left(A_{l,i} - \bar{A}_{l,n}\right) = 0 \;\forall\; l \in \{1, 2, \ldots, m\}$, $\forall\; i \in \{0, 1, \ldots, n-1\}$, где W_n^{ξ} имеет распределение:

$$W_n^{\xi} \sim \Gamma\left(\frac{n}{2\sigma^2}, \frac{n-1}{2}\right) ,$$

иначе

$$W_n^{\Xi} = W_n^{\xi} + \sum_{l:\left(A_{l,i} - \bar{A}_{l,n}\right) \neq 0}^{m} W_n^{A_l\xi} ,$$

если $\exists\, l, i : \left(A_{l,i} - \bar{A}_{l,n}\right) \neq 0$, при этом $W_n^{A_l\xi}$ имеет распределение:

$$W_n^{A_l\xi} \sim N\left(0, \frac{4\sigma^2}{n}W_n^{A_l}\right) .$$

Доказательство. По определению случайные величины ξ_k являются независимыми одинаково распределенными с нулевым математическим ожиданием и дисперсией σ^2. Поэтому распределение компоненты nW_n^{ξ}/σ^2 как оконной дисперсии случайной величины со стандартным нормальным распределением является хи-квадрат-распределением вида:

$$\frac{nW_n^{\xi}}{\sigma^2} \sim \chi_{n-1}^2 = \Gamma\left(\frac{1}{2}, \frac{n-1}{2}\right) .$$

Теперь по свойствам гамма-распределения получаем:

$$W_n^{\xi} \sim \chi^2_{n-1} = \Gamma\left(\frac{n}{2\sigma^2}, \frac{n-1}{2}\right).$$

Компоненты, характеризующие суперпозицию шума и истинных компонент сигнала, $W_n^{A_l\xi}$, $l \in \in \{1, 2, \ldots, m\}$, представляют собой суммы независимых нормально распределенных случайных величин:

$$W_n^{A_l\xi} = \frac{2}{n}\sum_{i=0}^{n-1} \xi_i \left(A_{l,i} - \bar{A}_{l,n}\right),$$

при этом

$$\frac{2}{n}\xi_i \left(A_{l,i} - \bar{A}_{l,n}\right) \sim N\left(0, \left[\frac{2\sigma}{n}\left(A_{l,i} - \bar{A}_{l,n}\right)\right]^2\right).$$

Вследствие усиленной воспроизводимости нормального распределения сумма таких величин по i будет иметь нормальное распределение вида:

$$W_n^{A_l\xi} \sim N\left(0, \frac{4\sigma^2}{n^2}\sum_{i=0}^{n-1}\left(A_{l,i} - \bar{A}_{l,n}\right)^2\right) = N\left(0, \frac{4\sigma^2}{n} W_n^{A_l}\right).$$

Итак, доказано, что в тех случаях, когда истинные компоненты, формирующие сигнал, не изменяются, т. е. $A_{l,i} - \bar{A}_{l,n} = 0$, оконная дисперсия шума характеризуется гамма-распределением с параметрами формы и масштаба $(n-1)/2$ и $n/(2\sigma^2)$ соответственно. Если же для каких-либо из истинных компонент $A_{l,i} - \bar{A}_{l,n} \neq 0$, тогда шумовая компонента оконной дисперсии W_n^{Ξ} представляет собой сумму зависимых случайных величин с гамма-распределением и нормально распределенных.

3 Модель с несколькими различными шумовыми компонентами

3.1 Общее представление модели

Обратимся теперь к несколько иному представлению модели, а именно: представим для каждой точки τ_k значение сигнала C в виде суммы независимых в физическом смысле истинных значений нескольких процессов, формирующих сигнал A_l, $l = 1, 2, \ldots, m$, и соответствующего каждой такой истинной компоненте независимого в статистическом смысле шума $\xi_1, \xi_2, \ldots, \xi_m$, при этом случайная величина $\xi_{l,k}$ характеризуется нормальным распределением с нулевым математическим ожиданием и дисперсией $\sigma^2_{l,k}$, $l = 1, 2, \ldots, m$, причем для всех точек τ_k для фиксированной истинной компоненты A_l изучаемого сигнала случайные величины $\xi_{k,l}$ — независимые одинаково распределенные случайные величины. Тогда для любого τ_k сигнал C представим в виде:

$$C = \sum_{l=1}^{m} C_l = \sum_{l=1}^{m} (A_l + \xi_l). \qquad (6)$$

Обозначим

$$\bar{C}_n = \sum_{l=1}^{m} \bar{C}_{l,n}; \quad W_n = \sum_{l=1}^{m} W_{l,n},$$

где $\bar{C}_n$ — скользящее среднее регистрируемого сигнала C, а W_n — оконная дисперсия сигнала.

Далее рассмотрим отдельно компоненты этой суммы (6) $C_l = A_l + \xi_l$. Для каждой компоненты C_l в соответствии с леммой справедливо разложение (4), поэтому

$$\bar{C}_{l,n} = \frac{1}{n}\sum_{i=0}^{n-1} C_{l,i} = \frac{1}{n}\sum_{i=0}^{n-1}(A_{l,i} + \xi_{l,i});$$

$$W_{l,n} = \frac{1}{n}\sum_{i=0}^{n-1}\left(C_{l,i} - \bar{C}_{l,n}\right)^2 = \frac{1}{n}\sum_{i=0}^{n-1} A^2_{l,i} - \bar{A}^2_{l,n} + \frac{1}{n}\sum_{i=0}^{n-1}\xi^2_{l,i} - \bar{\xi}^2_{l,n} + \frac{2}{n}\sum_{i=0}^{n-1}\xi_{l,i}\left(A_{l,i} - \bar{A}_{l,n}\right).$$

Зафиксируем l и далее для наглядности изложения опустим этот индекс, т. е. будем рассматривать лишь одну истинную компоненту и соответствующую шумовую компоненту.

Рассмотрим шумовую компоненту оконной дисперсии и представим ее в специальном виде:

$$W_n^{\Xi} = \frac{1}{n}\sum_{i=0}^{n-1}\xi_i^2 - \bar{\xi}_n^2 + \frac{2}{n}\sum_{i=0}^{n-1}\xi_i\left(A_i - \bar{A}_n\right) = \frac{1}{n}\sum_{i=0}^{n-1}\left(\xi_i^2 + \xi_i\left(2A_i - 2\bar{A}_n\right)\right) - \bar{\xi}_n^2.$$

Случайная величина $\bar{\xi}_n^2$ сходится по вероятности к нулю, а свойства слагаемых $\xi_i^2 + \xi_i\left(2A_i - 2\bar{A}_n\right)$ будут описаны в подразделе ниже.

Частный случай $l = 1$ соответствует сигналу с единственной истинной компонентой и соответствующим шумом, пример такого сигнала рассмотрен в работах [4, 5].

3.2 Свойства случайной величины вида $(\xi^2 + a\xi)$

Рассмотрим свойства одного слагаемого шумовой компоненты $\xi_i^2 + \xi_i\left(2A_i - 2\bar{A}_n\right)$. Зафиксировав i и введя обозначение $a = 2\left(A - \bar{A}\right)$, получим случайную величину вида $\xi^2 + a\xi$. Ее свойства отражены в следующей лемме.

Лемма. *Пусть случайная величина ξ распределена по нормальному закону $N\left(0, \sigma^2\right)$, тогда $\xi^2 + a\xi$ имеет распределение, соответствующее характеристической функции вида*:

$$\varphi_{\xi^2+a\xi}(t) = \frac{1}{\sqrt{1-2it\sigma^2}}\, e^{a^2t^2/\left(4\left(it-1/\left(2\sigma^2\right)\right)\right)}.$$

Доказательство. Вычислим характеристическую функцию для $\xi^2 + a\xi$.

Запишем определение:

$$\varphi_{\xi^2+a\xi}(t) = Ee^{it\left(\xi^2+a\xi\right)} = \int\limits_{-\infty}^{\infty} e^{it\left(x^2+ax\right)} dF(x) =$$

$$= \int\limits_{-\infty}^{\infty} e^{it\left(x^2+ax\right)} \frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-x^2/\left(2\sigma^2\right)} dx =$$

$$= \frac{1}{\sqrt{2\pi\sigma^2}} \int\limits_{-\infty}^{\infty} e^{it\left(x^2+ax\right)} e^{-x^2/\left(2\sigma^2\right)}\, dx\,,\quad t \in R\,.$$

Сначала подробно рассмотрим частный случай $a = 1$, $\sigma^2 = 1$, который затем обобщим для произвольных параметров a и σ^2.

Проведем элементарные преобразования, выделяя полный квадрат и проводя замену переменных, получим:

$$\varphi_{\xi^2+\xi}(t) = \frac{1}{\sqrt{2\pi}} \int\limits_{-\infty}^{\infty} e^{it\left(x^2+x\right)} e^{-x^2/2}\, dx =$$

$$= \frac{1}{\sqrt{2\pi}} \int\limits_{-\infty}^{\infty} e^{-(1/2-it)(x+it/(2(it-1/2)))^2} \times$$

$$\times\, e^{-(it-1/2)(it/(2(it-1/2)))^2}\, dx =$$

$$= \frac{1}{\sqrt{2\pi\left(1/2-it\right)}}\, e^{-(it-1/2)(it/(2(it-1/2)))^2} \times$$

$$\times \int\limits_{-\infty}^{\infty} e^{-\left(\sqrt{1/2-it}(x+it/(2(it-1/2)))\right)^2} \times$$

$$\times\, d\left(\sqrt{\frac{1}{2}-it}\,\left(x + \frac{it}{2\left(it-1/2\right)}\right)\right) =$$

$$= \frac{\sqrt{\pi}}{\sqrt{2\pi\left(1/2-it\right)}}\, e^{-(it)^2(it-1/2)/\left(4(it-1/2)^2\right)} =$$

$$= \frac{1}{\sqrt{1-2it}}\, e^{t^2/(4(it-1/2))}\,.$$

Теперь проведем аналогичные преобразования для любых a и σ^2:

$$\varphi_{\xi^2+a\xi}(t) = \frac{1}{\sqrt{2\pi\sigma^2}} \int\limits_{-\infty}^{\infty} e^{it\left(x^2+ax\right)} e^{-x^2/\left(2\sigma^2\right)}\, dx =$$

$$= \frac{1}{\sigma\sqrt{1/\sigma^2-2it}}\, e^{a^2t^2/\left(4\left(it-1/\left(2\sigma^2\right)\right)\right)} =$$

$$= \frac{1}{\sqrt{1-2it\sigma^2}}\, e^{a^2t^2/\left(4\left(it-1/\left(2\sigma^2\right)\right)\right)}.$$

Лемма доказана.

Согласно теореме единственности полученная характеристическая функция однозначным образом определяет функцию распределения случайной величины. Рассчитаем первые моменты рассматриваемой случайной величины. Результаты расчета приведены в нижеследующей лемме.

Лемма. Случайная величина $\xi^2 + a\xi$ имеет следующие математическое ожидание и дисперсию:

$$\mathsf{E}\left(\xi^2 + a\xi\right) = \sigma^2\,;\ \ \mathsf{D}\left(\xi^2 + a\xi\right) = \sigma^6 + 3\sigma^5 + 4a^2\sigma^3.$$

Доказательство. Для вывода формул используем следующее свойство характеристических функций:

$$\mathsf{E}\left(\left(\xi^2 + a\xi\right)^n\right) = \frac{\varphi^{(n)}(0)}{i^n}\,.$$

Первая и вторая производные по t для функции $\varphi(t)$ имеют вид:

$$\frac{d}{dt}\,\varphi(t) = \frac{d}{dt}\,\frac{e^{a^2t^2/\left(4\left(-1/\left(2\sigma^2\right)+it\right)\right)}}{\sqrt{1-2it\sigma^2}} =$$

$$= \frac{ie^{a^2t^2/\left(4\left(-1/\left(2\sigma^2\right)+it\right)\right)}}{\sigma\left(1/\sigma^2-2it\right)^{3/2}} +$$

$$+ e^{a^2t^2/\left(4\left(-1/\left(2\sigma^2\right)+it\right)\right)} \left(\frac{a^2t}{2\left(-1/\left(2\sigma^2\right)+it\right)} -\right.$$

$$\left. - \frac{ia^2t^2}{4\left(-1/\left(2\sigma^2\right)+it\right)^2}\right) \Bigg/ \left(\sigma\sqrt{\frac{1}{\sigma^2}-2it}\right);$$

$$\frac{d^2}{dt^2}\,\varphi(t) = -\frac{3e^{a^2t^2/\left(4\left(-1/\left(2\sigma^2\right)+it\right)\right)}}{\sigma\left(1/\sigma^2-2it\right)^{5/2}} +$$

$$+ e^{a^2t^2/\left(4\left(-1/\left(2\sigma^2\right)+it\right)\right)} \left(\frac{a^2}{2\left(-1/\left(2\sigma^2\right)+it\right)} -\right.$$

$$- \frac{ia^2t}{\left(-1/\left(2\sigma^2\right)+it\right)^2} - \\ - \frac{a^2t^2}{2\left(-1/\left(2\sigma^2\right)+it\right)^3}\Bigg) \Bigg/ \left(\sigma\sqrt{\frac{1}{\sigma^2}-2it}\right) + \\ + 2ie^{a^2t^2/\left(4\left(-1/\left(2\sigma^2\right)+it\right)\right)} \left(\frac{a^2t}{2\left(-1/\left(2\sigma^2\right)+it\right)} - \\ - \frac{ia^2t^2}{4\left(-1/\left(2\sigma^2\right)+it\right)^2}\right) \Bigg/ \left(\sigma\left(\frac{1}{\sigma^2}-2it\right)^{3/2}\right) + \\ + e^{a^2t^2/\left(4\left(-1/\left(2\sigma^2\right)+it\right)\right)} \left(\frac{a^2t}{2\left(-1/\left(2\sigma^2\right)+it\right)} - \\ - \frac{ia^2t^2}{4\left(-1/\left(2\sigma^2\right)+it\right)^2}\right)^2 \Bigg/ \left(\sigma\sqrt{\frac{1}{\sigma^2}-2it}\right).$$

Поэтому

$$iE\left(\xi^2+a\xi\right) = \frac{i}{\left(1/\sigma^2\right)^{3/2}\sigma} = i\sigma^2$$

и

$$i^2E\left(\xi^2+a\xi\right)^2 = \\ = -\frac{3}{\left(1/\sigma^2\right)^{5/2}\sigma} - \frac{a^2\sigma}{\sqrt{1/\sigma^2}} = i^2\left(3\sigma^4+a^2\sigma^2\right).$$

С помощью начальных моментов найдем дисперсию:

$$D\left(\xi^2+a\xi\right) = E\left(\xi^2+a\xi\right)^2 - \left(E\left(\xi^2+a\xi\right)\right)^2 = \\ = 3\sigma^4+a^2\sigma^2-\left(\sigma^2\right)^2 = 2\sigma^4+a^2\sigma^2.$$

Таким образом, утверждения леммы доказаны.

3.3 Свойства шумовой компоненты оконной дисперсии в модели с несколькими шумовыми компонентами

Заменим обратно a на $2A-2\bar{A}$ и сформулируем теорему о свойствах шумовой компоненты оконной дисперсии.

Теорема. *Для регистрируемого сигнала $C(t)$ в каждой точке τ_k, $k=\{1,2,\ldots,r\}$, шумовая компонента оконной дисперсии W_n^{Ξ} представляет собой*:

(i) *случайную величину*

$$\sum_{l=1}^{m} W_{l,n}^{\xi} \sim \Gamma\left(\frac{n}{2\sigma_l^2}, \frac{n-1}{2}\right),$$

если

$$A_{l,j} - \bar{A}_{l,n} = \emptyset \; \forall\, l \in \{1,2,\ldots,m\}, \\ \forall\, j \in \{0,1,\ldots,n-1\};$$

(ii) *сумму случайных величин* $\sum_{l=1}^{m}\sum_{j=0}^{n-1}(1/n)\times \\ \times\left(\xi_{l,j}^2+\xi_{l,j}\left(2A_{l,j}-2\bar{A}_{l,n}\right)\right) - \sum_{l=1}^{m}\bar{\xi}_{l,n}^2$, *если* $\exists\, l \in \{1,2,\ldots,m\}$ *и* $j \in \{0,1,\ldots,n-1\}$: $A_{l,j}-\bar{A}_n \neq 0$,
при этом

$$\bar{\xi}_{l,n}^2 \sim \Gamma\left(\frac{n}{2\sigma_l^2}, \frac{1}{2}\right),$$

а характеристическая функция случайной величины $\sum_{l=1}^{m}(1/n)\sum_{j=0}^{n-1}\Big(\xi_{l,j}^2+\xi_{l,j}\left(2A_{l,j}- \\ -2\bar{A}_{l,n}\right)\Big)$ *имеет вид*:

$$\varphi(t) = \prod_{l=1}^{m}\left(1-\frac{2it\sigma_l^2}{n}\right)^{-n/2} \times \\ \times \prod_{j=0}^{n-1} e^{\left(A_{l,j}-\bar{A}_{l,n}\right)^2t^2\sigma_l^2/\left(itn\sigma_l^2-n^2/2\right)}.$$

Доказательство. Сначала будем рассматривать случай для фиксированного l или, иными словами, сигнал, сформированный единственной истинной компонентой и соответствующим шумом.

В этом случае (см. подразд. 3.1) шумовая компонента для оконной дисперсии сигнала имеет вид:

$$W_n^{\Xi} = \frac{1}{n}\sum_{j=0}^{n-1}\left(\xi_j^2+\xi_j\left(2A_j-2\bar{A}_n\right)\right) - \bar{\xi}_n^2.$$

Так как

$$\frac{1}{n}\sum_{j=0}^{n-1}\left(\xi_j^2+\xi_j\left(2A_j-2\bar{A}_n\right)\right) = \\ = \sum_{j=0}^{n-1}\frac{1}{n}\left(\xi_j^2+\xi_j\left(2A_j-2\bar{A}_n\right)\right),$$

то, используя свойства характеристической функции, получим характеристическую функцию для случайной величины $(1/n)\left(\xi_j^2+\xi_j\left(2A_j-2\bar{A}_n\right)\right)$ при каждом фиксированном j:

$$\varphi_{(1/n)\left(\xi_j^2+\xi_j\left(2A_j-2\bar{A}_n\right)\right)}(t) = \\ = \varphi_{\xi_j^2+\xi_j\left(2A_j-2\bar{A}_n\right)}\left(\frac{t}{n}\right) = \\ = \frac{1}{\sqrt{1-2it\sigma^2/n}}\, e^{\left(2A_j-2\bar{A}_n\right)^2t^2/\left(4\left(it\,n-n^2/\left(2\sigma^2\right)\right)\right)}.$$

Тогда характеристическая функция для случайной величины $(1/n)\sum_{j=0}^{n-1}\left(\xi_j^2+\xi_j\left(2A_j-2\bar{A}_n\right)\right)$ может быть записана в виде:

$$\varphi_{(1/n)\sum_{j=0}^{n-1}\left(\xi_j^2+\xi_j\left(2A_j-2\bar{A}_n\right)\right)}(t)=$$
$$=\prod_{j=0}^{n-1}\varphi_{(1/n)\left(\xi_j^2+\xi_j\left(2A_j-2\bar{A}_n\right)\right)}=\frac{1}{\left(1-2it\sigma^2/n\right)^{n/2}}\times$$
$$\times\prod_{j=0}^{n-1}e^{\left(2A_j-2\bar{A}_n\right)^2t^2/\left(4\left(it\,n-n^2/\left(2\sigma^2\right)\right)\right)}=$$
$$=\left(1-\frac{2it\sigma^2}{n}\right)^{-n/2}\times$$
$$\times\prod_{j=0}^{n-1}e^{\left(2A_j-2\bar{A}_n\right)^2t^2/\left(4\left(it\,n-n^2/\left(2\sigma^2\right)\right)\right)}=$$
$$=\left(1-\frac{2it\sigma^2}{n}\right)^{-n/2}\prod_{j=0}^{n-1}e^{\left(A_j-\bar{A}_n\right)^2t^2\sigma^2/\left(itn\sigma^2-n^2/2\right)}.$$

Рассмотрим теперь компоненту шумовой составляющей сигнала $\bar{\xi}_n^2$, которая представляет собой квадрат нормально распределенной случайной величины $\bar{\xi}_n$ с математическим ожиданием 0 и дисперсией σ^2/n. Поскольку

$$\frac{\sqrt{n}}{\sigma}\,\bar{\xi}\sim N\left(0,1\right),$$

то

$$\frac{n}{\sigma^2}\,\bar{\xi}_n^2\sim\chi_1^2=\Gamma\left(\frac{1}{2},\frac{1}{2}\right)$$

и

$$\bar{\xi}_n^2\sim\Gamma\left(\frac{n}{2\sigma^2},\frac{1}{2}\right).$$

Отметим, что математическое ожидание $\bar{\xi}_n^2$ есть σ^2/n, а дисперсия равна $2\sigma^4/n^2$ и эти характеристики зависят от числа точек расчета оконной дисперсии n, убывая как n и n^2 соответственно. Отсюда следует, что можно выбрать такое n, что среднее и дисперсия $\bar{\xi}_n^2$ будут меньше заданной точности измерений.

Принимая во внимание результаты подразд. 2.2, для каждого фиксированного l шумовую компоненту сигнала можно представить в виде:

(i) случайной величины

$$W_n^{\xi}=\frac{1}{n}\sum_{j=0}^{n-1}\xi_j^2-\bar{\xi}_n^2\sim\Gamma\left(\frac{n}{2\sigma^2},\frac{n-1}{2}\right),$$

если $A_j-\bar{A}_n=0\;\forall\,j\in\{0,1,\ldots,n-1\}$;

(ii) суммы случайных величин $(1/n)\sum_{j=0}^{n-1}\left(\xi_j^2+\right.$ $\left.+\,\xi_j\left(2A_j-2\bar{A}_n\right)\right)-\bar{\xi}_n^2$, если $\exists\;j\in$ $\in\{0,1,\ldots,n-1\}:\left(A_j-\bar{A}_n\right)\neq0$, при этом

$$\bar{\xi}_n^2\sim\Gamma\left(\frac{n}{2\sigma^2},\,\frac{1}{2}\right);$$

$$\varphi_{(1/n)\sum_{j=0}^{n-1}\left(\xi_j^2+\xi_j\left(2A_j-2\bar{A}_n\right)\right)}(t)=$$
$$=\left(1-\frac{2it\sigma^2}{n}\right)^{-n/2}\prod_{j=0}^{n-1}e^{\left(A_j-\bar{A}_n\right)^2t^2\sigma^2/\left(itn\sigma^2-n^2/2\right)}.$$

Регистрируемый сигнал C представляет собой сумму независимых истинных компонент и соответствующих независимых шумовых составляющих, поэтому

$$W_{l,n}^{\Xi}=$$
$$=\sum_{l=1}^{m}\left[\frac{1}{n}\sum_{j=0}^{n-1}\left(\xi_{l,j}^2+\xi_{l,j}\left(2A_{l,j}-2\bar{A}_{l,n}\right)\right)-\bar{\xi}_{l,n}^2\right]=$$
$$=\sum_{l=1}^{m}\sum_{j=0}^{n-1}\frac{1}{n}\left(\xi_{l,j}^2+\xi_{l,j}\left(2A_{l,j}-2\bar{A}_{l,n}\right)\right)-\sum_{l=1}^{m}\bar{\xi}_{l,n}^2.$$

Рассмотрим случай, когда $\left(A_{l,j}-\bar{A}_{l,n}\right)=0\;\forall\,l\in$ $\in\{1,2,\ldots,m\},\;\forall\,j\in\{0,1,\ldots,n-1\}$, т. е. истинные компоненты сигнала не меняются на фиксированном окне. Тогда

$$W_{l,n}^{\Xi}=\sum_{l=1}^{m}\sum_{j=0}^{n-1}\frac{1}{n}\xi_{l,j}^2-\sum_{l=1}^{m}\bar{\xi}_{l,n}^2=\sum_{l=1}^{m}W_{l,n}^{\xi}.$$

Таким образом, оконная дисперсия шумовой составляющей сигнала в данном случае представляет собой сумму гамма-распределенных случайных величин

$$W_{l,n}^{\xi}\sim\Gamma\left(\frac{n}{2\sigma_l^2},\frac{n-1}{2}\right).$$

Рассмотрим случай, когда существуют $l\in$ $\in\{1,2,\ldots,m\}$ и $j\in\{0,1,\ldots,n-1\}$, при которых $A_{l,j}-\bar{A}_{l,n}\neq0$. Тогда оконная дисперсия шума представляет разность зависимых сумм случайных величин. При этом функция распределения случайной величины $\sum_{l=1}^{m}\sum_{j=0}^{n-1}(1/n)\left(\xi_{l,j}^2+\xi_{l,j}\left(2A_{l,j}-2\bar{A}_{l,n}\right)\right)$ соответствует характеристической функции

$$\varphi(t)=\prod_{l=1}^{m}\varphi_{(1/n)\sum_{j=0}^{n-1}\left(\xi_{l,j}^2+\xi_{l,j}\left(2A_{l,j}-2\bar{A}_{l,n}\right)\right)}(t)=$$
$$=\prod_{l=1}^{m}\left(1-\frac{2it\sigma_l^2}{n}\right)^{-n/2}\times$$
$$\times\prod_{j=0}^{n-1}e^{\left(A_{l,j}-\bar{A}_{l,n}\right)^2t^2\sigma_l^2/\left(itn\sigma_l^2-n^2/2\right)}.$$

Очевидно, что случайная величина $\sum_{l=1}^{m} \bar{\xi}_{l,n}^2$ представляет собой сумму независимых гамма-распределенных величин с параметрами формы $1/2$ и, вообще говоря, различными параметрами масштаба $n/(2\sigma_l^2)$.

Теорема доказана.

Отметим также, что так как случайная величина $\bar{\xi}_{l,n}^2$ неотрицательна, то на практике величина $\sum_{l=1}^{m} \sum_{i=0}^{n-1} (1/n) \left(\xi_{l,i}^2 + \xi_{l,i} \left(2A_{l,i} - 2\bar{A}_{l,n} \right) \right)$ может служить верхней оценкой шумовой компоненты оконной дисперсии $W_{l,n}^{\Xi}$ регистрируемого сигнала C.

4 Заключение

В рамках работы предложены модели для представления сигналов в виде суммы нескольких подлежащих процессов, а также исследованы некоторые вероятностные характеристики оконной дисперсии сигналов как случайных процессов в представленных моделях. Результаты работы согласуются с установленными эмпирически свойствами шумовой компоненты оконной дисперсии миограммы [6]. В работе продемонстрировано, что на миограмме в период покоя, т. е. в отсутствие полезных компонент сигнала, оконная дисперсия характеризуется гамма-распределением.

Полученные результаты планируется использовать в практических задачах сегментирования сигналов и выделения интервалов с преобладанием тех или иных подлежащих процессов. Кроме того, вероятностные характеристики шумовой компоненты могут использоваться для прогнозирования поведения сигнала. В частности, предполагается применить эти теоретические результаты для анализа фармакокинетических данных.

Литература

1. *Kosar K., Lhotská L., Krajca V.* Classification of long-term EEG recordings // Biological and medical data analysis. — Lecture notes in computer science ser. — Springer, 2004. Vol. 3337. P. 322–332. doi: 10.1007/978-3-540-30547-7_33.
2. *Azami H., Mohammadi K., Hassanpour H.* A hybrid evolutionary approach to segmentation of nonstationary signals // Digit. Signal Process., 2013. Vol. 23. No. 4. P. 1103–1114. doi: 10.1016/j.dsp.2013.02.019.
3. *Kalantarian H., Sarrafzadeh M.* Probabilistic time-series segmentation // Pervasive Mob. Comput., 2017. doi: 10.1016/j.pmcj.2017.03.005.
4. *Захарова Т. В., Никифоров С. Ю., Гончаренко М. Б., Драницына М. А., Климов Г. А., Хазиахметов М. Ш., Чаянов Н. В.* Методы обработки сигналов для локализации невосполнимых областей головного мозга // Системы и средства информатики, 2012. Т. 22. № 2. С. 157–175.
5. *Хазиахметов М. Ш.* Свойства оконной дисперсии миограммы как случайного процесса // Системы и средства информатики, 2014. Т. 24. № 3. С. 110–120.
6. *Allakhverdieva V. M., Chshenyavskaya E. V., Dranitsyna M. A., Karpov P. I., Zakharova T. V.* An approach to the inverse problem of brain functional mapping under the assumption of gamma distributed myogram noise within rest intervals using the independent component analysis // J. Math. Sci., 2016. Vol. 214. No. 1. P. 3–11. doi: 10.1007/s10958-016-2753-x.

Поступила в редакцию 19.04.17

SEGMENTATION OF NONSTATIONARY SIGNALS USING STOCHASTIC CHARACTERISTICS OF THE WINDOW VARIANCE

M. A. Dranitsyna[1] and T. V. Zakharova[1,2]

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

[2]Institute of Informatics Problems, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, 44-2 Vavilov Str., Moscow 119333, Russian Federation

Abstract: Signal or response partitioning (i. e., signal segmentation) is of great interest, e. g., for biomedical research. Signal segmentation, being an essential part of signal processing, may serve as a tool for advanced signal interpretation and data classification. Segmentation of nonstationary signals with a small signal-to-noise ratio is a particulary complicated task. The paper is mainly devoted to exploration of the window variance noise component as a random variable for the proposed signal models. Some stochastic characteristics of the window variance noise

components are investigated in accordance with the models. Theoretical findings are consistent with the previously obtained empirical characteristics of the window variance noise component and are supposed to be of potential use for signal segmentation and prediction.

Keywords: window variance; signal model

DOI: 10.14357/19922264170302

References

1. Kosar, K., L. Lhotska, and V. Krajca. 2004. Classification of long-term EEG recordings. *Biological and medical data analysis*. Lecture notes in computer science ser. 3337:322-332. doi: 10.1007/978-3-540-30547-7_33.
2. Azami, H., K. Mohammadi, and H. Hassanpour. 2013. A hybrid evolutionary approach to segmentation of nonstationary signals. *Digit. Signal Process.* 23(4):1103-1114. doi: 10.1016/j.dsp.2013.02.019.
3. Kalantarian, H., and M. Sarrafzadeh. 2017. Probabilistic time-series segmentation. *Pervasive Mob. Comput.* doi: 10.1016/j.pmcj.2017.03.005.
4. Zakharova, T. V., S. Yu. Nikiforov, M. B. Goncharenko, M. A. Dranitsyna, G. A. Klimov, M. Sh. Khaziakhmetov, and N. V. Chayanov. 2012. Metody obrabotki signalov dlya lokalizatsii nevospolnimykh oblastey golovnogo mozga [Signal processing methods for localization of nonrenewable brain regions]. *Sistemy i Sredstva Informatiki — Systems and Means of Informatics* 22(2):157–175.
5. Khaziakhmetov, M. Sh. 2014. Svoystva okonnoy dispersii miogrammy kak sluchaynogo protsessa [Properties of window dispersion of myogram as a stochastic process]. *Sistemy i Sredstva Informatiki — Systems and Means of Informatics* 24(3):110–120.
6. Allakhverdieva, V. M., E. V. Chshenyavskaya, M. A. Dranitsyna, P. I. Karpov, and T. V. Zakharova. 2016. An approach to the inverse problem of brain functional mapping under the assumption of gamma distributed myogram noise within rest intervals using the independent component analysis. *J. Math Sci.* 214(1):3–11. doi: 10.1007/s10958-016-2753-x.

Received April 19, 2017

Contributors

Dranitsyna Margarita A. (b. 1983) — PhD student, Department of Mathematical Statistics, Faculty of Computational Mathematics and Cybernetics, M. V. Lomonosov Moscow State University, 1-52 Leninskiye Gory, GSP-1, Moscow 119991, Russian Federation; margarita13april@mail.ru

Zakharova Tatiana V. (b. 1962) — Candidate of Science (PhD) in physics and mathematics, associate professor, Department of Mathematical Statistics, Faculty of Computational Mathematics and Cybernetics, M. V. Lomonosov Moscow State University, 1-52 Leninskiye Gory, GSP-1, Moscow 119991, Russian Federation; senior scientist, Institute of Informatics Problems, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, 44-2 Vavilov Str., Moscow 119333, Russian Federation; lsa@cs.msu.ru

ИНФОРМАТИКА И ЕЁ ПРИМЕНЕНИЯ, 2017. Т. 11. Вып. 3. С. 27–33

ОБУЧАЕМАЯ КЛАССИФИКАЦИЯ НЕПОЛНЫХ КЛИНИЧЕСКИХ ДАННЫХ

М. П. Кривенко[1]

Аннотация: Рассматриваются вопросы эффективности методов классификации неполных клинических данных. Обучение байесовского классификатора проводится методом максимального правдоподобия (МП) для модели смеси нормальных распределений. Строгий вывод формул, обеспечивающих реализацию шагов EM (expectation-maximization) алгоритма, позволил корректно применять итерационный процесс получения оценок параметров смеси. Для неполных данных предлагаются приемы выбора начальных значений и коррекции вырождающихся ковариационных матриц элементов смеси. Экспериментальная часть работы заключалась в анализе зависимости качества классификации от степени пропуска отдельных значений, для этого использовались данные о ферментах, полученные для пациентов с заболеваниями печени. Обработка реальных данных продемонстрировала практически идентичные ошибки классификации при применении простых и сложных методов обработки пропусков в случае невысокой степени случайного пропуска отдельных значений.

Ключевые слова: пропущенные данные; EM-алгоритм; смеси нормальных распределений

DOI: 10.14357/19922264170303

1 Введение. Общие принципы обработки неполных данных

В клинических исследованиях отсутствующие данные — это данные, которые планировались к фиксации, но оказались неотраженными в базе данных. Понимание причин отсутствия данных важно для правильной обработки оставшихся данных. Если значения отсутствуют совершенно случайно, выборка данных, скорее всего, останется репрезентативной для популяции. Но если значения отсутствуют систематически, анализ может стать предвзятым. Из-за этих проблем необходимо планировать исследования так, чтобы минимизировать появление отсутствующих значений либо внимательным образом разбираться с причинами появления отсутствующих наблюдений. Поскольку нет уверенности, что принятые предположения о пропусках верны, а сами данные недоступны, необходим дополнительный анализ чувствительности процедур анализа результатов клинических исследований для оценки достоверности проведенных исследований.

Независимо от того, насколько хорошо спроектированы и проведены испытания, появление отдельных недостающих данных является ожидаемым. Подобное явление может быть совершенно не связано с состоянием здоровья пациента и характером лечения: в частности, данные могут быть неполными из-за проблем планирования и реализации отдельных исследований (например, для них не созданы условия проведения), из-за человеческой ошибки при записи данных. С другой стороны, данные могут отсутствовать по причинам, связанным со здоровьем субъекта и экспериментальным лечением, которому он подвергается (например, субъекты могут отказаться от определенных клинических исследований из-за состояния своего здоровья или в силу каких-либо предубеждений). Помимо недостающих данных из-за пропущенных посещений пропуски могут возникать просто из-за способа измерения или характера заболевания (например, данные будут отсутствовать, когда нет смысла в их получении).

Далее рассматриваются способы обработки неполных данных результатов клинических обследований; при этом не затрагивается специфика задач клинических испытаний (оценивание эффективности и безопасности лекарственного препарата или метода лечения и диагностики).

В методологии отсутствующих данных используются два термина: недостающее значение и механизм отсутствия. Механизм отсутствия указывает на распределение вероятностей двоичного события отсутствия информации.

Классификации механизмов пропуска данных, введенные в [1–3], представляют собой формальную структуру, описывающую вероятностные характеристики данных и пропусков их значений; она

[1]Институт проблем информатики Федерального исследовательского центра «Информатика и управление» Российской академии наук, mkrivenko@ipiran.ru

позволяет получить представление о том, как механизм отсутствия может влиять на выводы о клиническом результате. Различают полностью случайный пропуск (Missing Completely at Random, MCAR), случайный пропуск (Missing at Random, MAR) и механизм с отсутствием случайности в пропусках (Missing Not at Random, MNAR). Когда данные пропущены по организационно-административным причинам, механизм отсутствия может быть MCAR, потому что природа отсутствия не имеет ничего общего с моделью получения результатов. Может оказаться востребованной и модель MAR. Важно только отметить, что MAR не является неотъемлемой характеристикой самих данных или механизма пропусков, а тесно связан с моделью анализа: если включить в используемую модель все факторы, от которых зависят пропуски в этой модели, то будем работать в рамках MAR; в противном случае анализ не будет соответствовать предположениям MAR. Самым жестким и наименее реалистичным является MCAR, в то время как MNAR является наименее ограничительным. Тем не менее наличие очень разнообразных предположений, возможных в рамках MNAR, может рассматриваться как проблема (согласно [4]); в силу этого трудно предварительно определить один окончательный вариант анализа MNAR. Перечисленные типы пропусков можно формально определить, используя обозначения, аналогичные обозначениям [5].

При некоторых предположениях пропуски могут быть признаны игнорируемыми. Пропуск классифицируется как игнорируемый, если можно найти действительную оценку результата без учета механизма отсутствия. В своей первой работе, посвященной проблеме недостающих данных, автор [1] показал, что при использовании байесовских процедур или правдоподобия для оценки любого параметра Θ, связанного с клиническим исходом, недостающие данные игнорируются, когда механизм отсутствия является MAR, а Θ является «отличным от» параметра механизма пропуска (кавычки используются, потому что условие отличимости является весьма специфичным).

С учетом специфики данной статьи (обработка матрицы признак–объект, построение классификатора при наличии обучающей выборки) могут быть выделены следующие группы методов анализа неполных данных:

– отбрасывание (игнорирование) пропусков с последующим применением обычных процедур;

– заполнение пропусков (вменение, приписывание недостающих значений) с последующим применением обычных процедур;

– обработка наблюденных и пропущенных данных в совокупности с необходимостью разработки оригинальных процедур.

При отбрасывании пропусков, т. е. обработке комплектных наблюдений (complete-case method в англоязычной литературе), используются только объекты с полными данными. Очевидным преимуществом такого анализа является простота реализации. Кроме того, он дает достоверные результаты в случае MCAR. Тем не менее у подхода, исключающего пациентов с неполными данными, существует ряд недостатков: для относительно небольшого объема данных большой размерности процесс отбрасывания может привести к тому, что нечего будет обрабатывать; игнорирование части данных влечет снижение эффективности получающихся решений; если механизм не является MCAR, то анализ может повлечь необъективное сравнение лечений (см. примеры в [6]).

Проблему пропущенных данных может решить, как кажется на первый взгляд, метод обработки доступных наблюдений (available-case method), когда за счет декомпозиции размерности обрабатываемых данных становится больше комплектных значений (например, если при нахождении ковариационной матрицы отдельно проводить оценивание для пар признаков). Но здесь при «сборке» итоговых характеристик из полученных фрагментов могут возникнуть свои трудности: перестанут выполняться требуемые свойства. В качестве решения можно предложить корректировку получаемых оценок, только она носит индивидуальный характер для каждой решаемой задачи.

Следующий класс методов обработки неполных данных включает процедуры заполнения (imputing). Заполнение — это любой метод, при котором пропущенные значения в наборе данных заполняются правдоподобными оценками. Цель любого метода заполнения заключается в создании полного набора данных, который затем может быть проанализирован с использованием стандартных статистических методов.

Заполнение может осуществляться на основании частных или условных характеристик совокупности признаков. Наиболее простой, широко используемой и постоянно критикуемой является подстановка в качестве пропущенных данных оценки безусловного среднего (или, например, медианы), полученной по доступным данным. Более перспективным выглядит заполнение пропусков с помощью распределений, условных по отношению к признакам, содержащим наблюденные значения. Это могут быть в случае некоторого объекта как простейшие характеристики распределения

(среднее или опять же медиана), так и смоделированные значения из условного распределения.

Существуют и другие методы единичного заполнения, в частности метод hot-deck (обзор дан в [7]). С ним связаны не всегда реализованные надежды на методы заполнения (в частности, см. работу [8] с многообещающим названием).

В последнее время демонстрирует повышенное к себе внимание в литературе так называемое множественное заполнение. Его идея в том, чтобы использовать для замещения пропусков более одного значения. В результате его применения возникают два или более полных набора данных. В зависимости от характера задачи результаты обработки этих наборов данных либо усредняются, либо из них выбираются экстремальные. Каждый из упомянутых методов дает достоверные оценки, когда данные являются MAR.

Обработка наблюденных и пропущенных данных в совокупности основывается на построении модели данных об исследуемых характеристиках и особенностях порождения пропусков. Выводы в этом случае получают с помощью функции правдоподобия при условии, что действует игнорирующий механизм пропуска. В случае неигнорируемого отсева значений данный метод может приводить к необъективным результатам (см., например, [9]).

Понятно, что существует еще одна категория, которая определяется как «другие» методы анализа неполных данных, с их примерами можно познакомиться в [6]. Они обычно ориентированы на специфические задачи и не работают в задаче обучаемой классификации данных с пропусками.

2 Метод максимального правдоподобия

Рассмотрим задачу обучаемой классификации на основе модели смеси нормальных многомерных распределений в условиях, когда отдельные значения признаков могут отсутствовать. Последнее имеет место как для обучающей выборки, так и при классификации нового объекта. Принимая во внимание, что для структуры классов и для каждого класса принята определенная вероятностная модель, можно строить байесовский классификатор. Для этого в первую очередь необходимо найти оценки параметров смеси, описывающей отдельный класс.

Оценивание по методу МП для полных и неполных данных принципиально не отличается, но проблемы реализации все-таки возникают. Фундаментальную роль в их решении играют базовая работа [10] и ее развернутое изложение в [3]. Во-первых, для эффективного применения метода МП требуется лишь MAR, а не более жесткое условие MCAR. Во-вторых, оценка МП неизвестных параметров модели должна находиться путем максимизации $L(\Theta|\mathbf{x}_0)$, причем предполагается, что имеется возможность максимизировать $L(\Theta|\mathbf{x}_0, \mathbf{x}_m)$, в частности с помощью EM-алгоритма; здесь Θ — совокупность параметров, описывающих модель данных; $\mathbf{x}_0$ и $\mathbf{x}_m$ — наблюденные и пропущенные значения соответственно. И наконец, в-третьих, наиболее важным моментом становится нахождение математического ожидания правдоподобия по распределению пропущенных данных.

Смесь нормальных распределений в качестве распределения данных вносит существенные трудности как в вывод базовых соотношений, так и в реализацию итерационных шагов EM-алгоритма. Воспользуемся стандартными обозначениями: если плотность d-мерного нормального распределения обозначить как $\varphi(\mathbf{u}, \vartheta)$, где $\vartheta = (\boldsymbol{\mu}, \boldsymbol{\Sigma})$, то плотность смеси нормальных распределений есть

$$f(\mathbf{u}, \Theta) = \sum_{k=1}^{K} \pi_k \varphi(\mathbf{u}, \vartheta_k)\,,$$

где $\Theta = (\pi_1, \ldots, \pi_K, \vartheta_1, \ldots, \vartheta_K)$.

Исходные данные представляют собой последовательность объектов $\{\mathbf{x}_i,\ i = 1, \ldots, N\}$, причем для каждого i определен набор индексов O_i, указывающих на наблюденные признаки. Дополняет каждый O_i набор M_i, соответствующий пропущенным данным. Эти наборы будут использоваться в качестве верхних индексов у обобщенных параметров и матриц (векторов), служить для отсылки к элементам соответствующих строк и столбцов (строк).

Если предполагается наличие пропущенных данных, то логарифм правдоподобия принимает вид

$$\ln L(\Theta) = \sum_{i=1}^{N} \ln \left(\sum\nolimits_{k=1}^{K} \pi_K \varphi \left(\mathbf{x}_i^{O_i}, \vartheta_k^{O_i} \right) \right)\,,$$

где $\varphi(\mathbf{x}_i^{O_i}, \vartheta_k^{O_i})$ — маргинальная плотность нормального распределения для наблюденных значений $\mathbf{x}_i^{O_i}$.

Вектор значений признаков $\mathbf{x}_i$ для i-го объекта запишем в форме $(\mathbf{x}_{O,i}, \mathbf{x}_{m,i})$, где $\mathbf{x}_{O,i}$ и $\mathbf{x}_{m,i}$ соответственно обозначают наблюденные и пропущенные значения признаков. Это формальная запись, не означающая переупорядочения в соответствии со структурой пропусков. При подборе модели смеси распределений в присутствии пропущенных значений признаков возникают два типа пропусков: один, концептуальный (привнесенный), связанный с ненаблюдаемым индика-

тором z_{ik} принадлежности некоторого объекта $\mathbf{x}_i$ к элементу смеси с номером k, и другой, ненамеренный (непредусмотренный), ему соответствует обозначение $\mathbf{x}_{m,i}$.

ЕМ-алгоритм на $(t+1)$-м шаге итерации требует вычисления

$$Q\left(\Theta|\Theta^{(t)}\right) = \mathsf{E}\left\{L(\Theta|\mathbf{x}_0,\Theta^{(t)}\right\}.$$

Для нахождения $Q(\Theta|\Theta^{(t)})$ понадобится

$$\hat{z}_{ik} = \hat{z}_{ik}^{(t)} = \mathsf{E}\left\{z_{ik}|\mathbf{x}_{O,i};\Theta^{(t)}\right\} = \\ = \frac{\pi_k\varphi_k\left(\mathbf{x}_{O,i};\vartheta_k^{(t)}\right)}{\sum_{k=1}^{K}\pi_k\varphi_k\left(\mathbf{x}_{O,i};\vartheta_k^{(t)}\right)}.$$

Для достаточных статистик каждого элемента смеси сначала получаем:

$$\mathsf{E}\left\{z_{ik}x_{ij}|\mathbf{x}_{O,i};\vartheta_k^{(t)}\right\} = \\ = \begin{cases}\hat{z}_{ik}x_{ij}\,, & \text{если } x_{ij} \text{ наблюдено;}\\ \hat{z}_{ik}\mathsf{E}\left\{x_{ij}|\mathbf{x}_{O,i};\vartheta_k^{(t)}\right\}, & \text{если } x_{ij} \text{ пропущено.}\end{cases}$$

Далее подобным же образом можно выписать выражения для $\mathsf{E}\left\{z_{ik}x_{ij}^2|\mathbf{x}_{O,i};\vartheta_k^{(t)}\right\}$ и $\mathsf{E}\left\{z_{ik}x_{ij}x_{ij'}|\mathbf{x}_{O,i};\vartheta_k^{(t)}\right\}$, $i,j=1,\ldots,N$, $k=1,\ldots,K$, $j\neq j'$. Заметим, если одно значение некоторого признака пропущено, то оно просто заменяется на $E\left\{x_{ij}|\mathbf{x}_{O,i};\vartheta_k^{(t)}\right\}$. Далее для теперь уже полных данных М-шаг ЕМ-алгоритма приводит к оценкам параметров смеси общего вида. Для того чтобы сделать конкретными полученные представления, необходимо воспользоваться видом условного нормального распределения.

Завершим рассуждения итоговыми представлениями. Пусть для каждого элемента смеси заданы π_k, $\boldsymbol{\mu}_k$ и $\boldsymbol{\Sigma}_k$ (ссылка на шаг итерации опущена). Тогда для $i=1,\ldots,N$ и $k=1,\ldots,K$ результаты очередного Е-шага следующие:

$$\hat{z}_{ik} = \frac{\pi_k\varphi_k\left(\mathbf{x}_i^{O_i},\vartheta_k^{O_i}\right)}{\sum_{k=1}^{K}\pi_k\varphi_k\left(\mathbf{x}_i^{O_i},\vartheta_k^{O_i}\right)};$$

$$\tilde{\boldsymbol{\mu}}_{ik}^{M_i} = \mathsf{E}\left\{\mathbf{x}_i^{M_i}|\mathbf{x}_i^{O_i}\right\} = \\ = \boldsymbol{\mu}_k^{M_i} + \boldsymbol{\Sigma}_k^{M_iO_i}\left(\boldsymbol{\Sigma}_k^{O_iO_i}\right)^{-1}\left(\mathbf{x}_i^{O_i}-\boldsymbol{\mu}_k^{O_i}\right);$$

$$\tilde{\mathbf{x}}_{ik} = \begin{pmatrix}\mathbf{x}_i^{O_i}\\ \tilde{\mu}_{ik}^{M_i}\end{pmatrix};$$

$$\tilde{\boldsymbol{\Sigma}}_{ik}^{M_iM_i} = \boldsymbol{\Sigma}_k^{M_iM_i} - \boldsymbol{\Sigma}_k^{M_iO_i}\left(\boldsymbol{\Sigma}_k^{O_iO_i}\right)^{-1}\boldsymbol{\Sigma}_k^{O_iM_i};$$

$$\tilde{\boldsymbol{\Sigma}}_{ik} = \begin{pmatrix}\mathbf{0}^{O_iO_i} & \mathbf{0}^{O_iM_i}\\ \mathbf{0}^{M_iO_i} & \tilde{\boldsymbol{\Sigma}}_{ik}^{M_iM_i}\end{pmatrix}.$$

Далее М-шаг приводит к новым значениям оценок $\hat{\pi}_k$, $\hat{\boldsymbol{\mu}}_k$ и $\hat{\boldsymbol{\Sigma}}_k$:

$$\hat{\pi}_k = \frac{1}{N}\sum_{i=1}^{N}\hat{z}_{ik}\,;$$

$$\hat{\boldsymbol{\mu}}_k = \frac{\sum_{i=1}^{N}\hat{z}_{ik}\tilde{\mathbf{x}}_{ik}}{\sum_{i=1}^{N}\hat{z}_{ik}};$$

$$\hat{\boldsymbol{\Sigma}}_k = \frac{\sum_{i=1}^{N}\hat{z}_{ik}[\tilde{\mathbf{x}}_{ik}-\hat{\boldsymbol{\mu}}_k)(\tilde{\mathbf{x}}_{ik}-\hat{\boldsymbol{\mu}}_k)^{\mathrm{T}}+\tilde{\boldsymbol{\Sigma}}_{ik}]}{\sum_{i=1}^{N}\hat{z}_{ik}} = \\ = \frac{\sum_{i=1}^{N}\hat{z}_{ik}(\tilde{\mathbf{x}}_{ik}-\hat{\boldsymbol{\mu}}_k)(\tilde{\mathbf{x}}_{ik}-\hat{\boldsymbol{\mu}}_k)^{\mathrm{T}}}{\sum_{i=1}^{N}\hat{z}_{ik}} + \frac{\sum_{i=1}^{N}\hat{z}_{ik}\tilde{\boldsymbol{\Sigma}}_{ik}}{\sum_{i=1}^{N}\hat{z}_{ik}}.$$

Остановимся на формулах байесовской классификации (в случае единичной матрицы потерь) наблюдений с пропущенными значениями отдельных признаков. Пусть s-й класс описывается смесью нормальных распределений

$$f_s(\mathbf{u}) = \sum_{k=1}^{K}\pi_{sk}\varphi(\mathbf{u},\vartheta_{sk}), \quad s=1,\ldots,S,$$

а вероятность его появления равна q_s. Тогда предпочтение для некоторого наблюденного вектора $\mathbf{y}_0$ будет отдаваться тому классу, для которого $q_sf_s(\mathbf{y}_0)$ достигает по s своего максимума.

3 Эксперименты

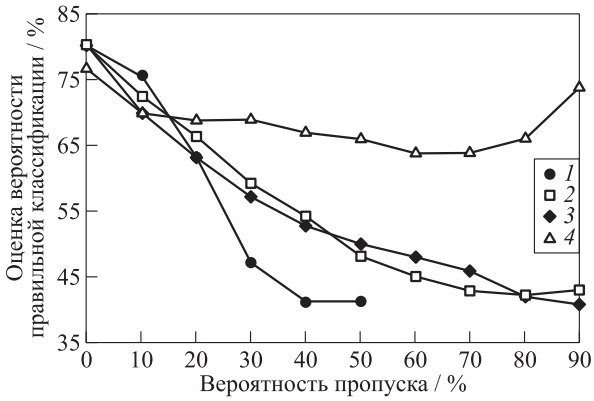
Экспериментальная часть работы заключалась в анализе зависимости качества классификации от степени пропуска отдельных значений. Для этого использовались результаты реальных обследований, приведенных в [11, разд. 5]. В этой работе применялась дискриминация на основе нормального распределения данных о ферментах, полученных для 218 пациентов с заболеваниями печени. Рассматривались четыре заболевания: острый вирусный гепатит (далее для краткости D_1, 57 пациентов); хронический персистирующий гепатит (D_2, 44 пациента); агрессивный хронический гепатит (D_3, 40 пациентов); постнекротический цирроз (D_4, 77 пациентов). Диагноз острого вирусного гепатита основывался на классических клинико-биохимических признаках. Все остальные пациенты были диагностированы на основе результатов лапароскопии и биопсии. Лабораторные исследования, на основе результатов которых должна

проводиться классификация, заключались в измерениях для четырех ферментов печени: аспартатаминотрансфераза (AST), аланинаминотрансфераза (ALT), глутаматдегидрогеназа (GLDH) и орнитинкарбонилтрансфераза (OCT).

Для представления элементов обучающей выборки для каждого из классов D_1–D_4 использовалась смесь нормальных распределений из пяти элементов. Значение $K = 5$ было выбрано как компромисс после применения информационных критериев AIC (Akaike Information Criterion), BIC (Bayesian Information Criterion), AWE (Approximate Weight of Evidence) [12, разд. 2.3.3], которые показали достаточно четко, что $K > 1$ (модель нормального распределения не подходит), и весьма расплывчато, что $K \in [2, 10]$. Возникновение пропусков моделировалось с помощью индикаторной переменной $r_{ij} = 1$ с вероятностью $1 - p_m$ и $r_{ij} = 0$ с вероятностью p_m.

Для каждого выбранного значения p_m генерировалась индикаторная матрица, затем на ее основе с помощью одного из методов обработки данных с пропусками строились оценки смесей, описывающих классы D_1–D_4, и тем самым формировался эмпирический байесовский классификатор; далее методом перепроверки оценивалась вероятность p_c^* правильного решения подобного классификатора. Подобная процедура повторялась N_{exp} раз, и оценивались выборочные характеристики для p_c^*: среднее, стандартное отклонение, минимальное и максимальное значения. В ходе экспериментов при больших значениях p_m возникла проблема с появлением «пустых» (с полностью пропущенными значениями) столбцов и строк матрицы признак–объект для исходных данных. В рамках данного исследования она решалась путем отбрасывания подобных столбцов (для некоторого объекта нет ни одного наблюденного признака) и строк (для некоторого признака и хотя бы одного класса полностью отсутствуют наблюдения). Это привело к фактическому снижению размерности признакового пространства d и объема обучающей выборки N и необходимости в отказе от классификации в случае полного вырождения признакового пространства и снижения объема части обучающей выборки для некоторого класса ниже критического (в данной работе это $d + 1$).

Результаты моделирования при $N_{\text{exp}} = 100$ для методов комплектных данных (CC), заполнения частным средним (PM), заполнения условным средним (CM), МП (ML) приведены на рисунке. Для метода CC проявляется эффект отказа от классификации: если при $p_m = 20\%$ относительная частота составляла 0%, то при $p_m = 30\%$ — 4%, при $p_m = 40\%$ — 64%, при $p_m = 50\%$ — 100%.



Зависимость оценки вероятности правильной классификации p_c^* от вероятности пропуска p_m для методов CC (*1*), PM (*2*), CM (*3*) и ML (*4*)

Из рисунка можно сделать выводы:

– в диапазоне $p_m = 0\%$–20% результативность всех методов практически одинакова;

– в диапазоне всех представленных значений p_m методы PM и CM фактически совпадают;

– метод ML обладает очевидными преимуществами.

Если к полученным результатам добавить информацию о временно́й сложности анализируемых методов, то можно заключить, что при малых значениях вероятности пропуска ($p_m = 0\%$–20%) следует пользоваться простыми методами комплектных данных и заполнения частным средним, при больших значениях вероятности пропуска предпочтение следует отдать методу МП.

Целесообразность усложнения модели данных за счет перехода от просто нормального распределения к смеси таковых можно проиллюстрировать сравнением поведения эффективности методов PM и ML при $K = 5$ и 1, что дает приблизительно 10%-ное снижение значений p_c^* в последнем случае.

4 Заключение

В ходе исследований пришлось заново провести строгий вывод формул, обеспечивающих реализацию шагов EM-алгоритма. Дело в том, что в большинстве доступных работ [13–16] лишь декларировались базовые соотношения, иногда с неточностями. Самое важное, встречались сомнительные переходы при получении результатов (например, не было ясно, является ли фактическое заполнение пропущенных значений на M-шаге следствием удобства или формальным результатом,

следующим из общих оптимизационных принципов), вообще не упоминались условия и приемы регуляризации оценок вторых моментов.

Полученные результаты позволили корректно применять итерационный процесс получения оценок параметров смеси. В качестве начального шага итерационного EM-алгоритма было предложено использовать метод заполнения пропущенных данных частными средними и случайный перебор матриц апостериорных вероятностей.

Экспериментальная часть работы заключалась в анализе зависимости качества классификации от степени пропуска отдельных значений на примере данных о ферментах, полученных для пациентов с заболеваниями печени. Обработка реальных данных продемонстрировала практически идентичные ошибки классификации при применении простых и сложных методов обработки пропусков в случае невысокой степени случайного пропуска отдельных значений. Рассмотренный метод заполнения пропусков условными средними является вариацией на тему использования метода hot-deck, его очевидная «непрактичность» (времена обработки по методам CC, PM, CM, ML относятся как 1 : 1 : 92 : 2) при фактически той же эффективности по сравнению с более простыми методами позволяют усомниться в выводах [8].

Метод МП совместно со смесью нормальных распределений в качестве модели данных обладает безусловными преимуществами, хотя и оказывается достаточно сложным при реализации. Последнее стимулирует исследования по анализу и разработке соответствующих алгоритмов обработки данных, в частности с использованием идеи sweep-оператора [3] или приемов древовидной организации вычислений, упомянутых в [15].

Литература

1. *Rubin D. B.* Inference and missing data // Biometrika, 1976. Vol. 63. P. 581–592.
2. *Rubin D. B.* Multiple imputation for nonresponse in surveys. — New York, NY, USA: John Wiley & Sons, 1987. 256 p.
3. *Little R. J. A., Rubin D. B.* Statistical analysis with missing data. — 2nd ed. — New York, NY, USA: John Wiley & Sons, 2002. 408 p.
4. *Mallinckrodt C. H., Lane P. W., Schnell D., Peng Y., Mancuso J.* Recommendation for the primary analysis of continuous endpoints in longitudinal clinical trials // Drug Inf. J., 2008. Vol. 42. P. 303–319.
5. *Molenberghs G., Kenward M. G.* Missing data in clinical studies. — West Sussex: John Wiley & Sons, 2007. 526 p.
6. *Myers W. R.* Handling missing data in clinical trials: An overview // Drug Inf. J., 2000. Vol. 34. P. 525–533.
7. *Andridge R. R., Little R. J. A.* A review of hot deck imputation for survey non-response // Int. Stat. Rev., 2010. Vol. 78. No. 1. P. 40–64.
8. *Myers T. A.* Goodbye, listwise deletion: Presenting hot deck imputation as an easy and effective tool for handling missing data // Commun. Meth. Measures, 2011. Vol. 5. No. 4. P. 297–310.
9. *Little R. J. A.* Modeling the drop-out mechanism in repeated-measures studies // J. Am. Stat. Assoc., 1995. Vol. 90. No. 431. P. 1112–1121.
10. *Dempster A. P., Laird N. M., Rubin D. B.* Maximum likelihood from incomplete data via EM algorithm // J. Roy. Stat. Soc. B Met., 1977. Vol. 39. No. 1. P. 1–38.
11. *Alber A.* Multivariate interpretation of clinical laboratory data. — New York, NY, USA: CRC Press, 1987. 386 p.
12. *Кривенко М. П.* Статистические методы представления и статистической предварительной обработки референсных значений. — М.: ФИЦ ИУ РАН, 2016. 160 с.
13. *Ghahramani Z., Jordan M. I.* Learning from incomplete data. — MIT AI, 1994. A.I. Memo No. 1509. C.B.C.L. Paper No. 108. https://dspace.mit.edu/handle/1721.1/7202.
14. *Hunt L., Jorgensen M.* Mixture model clustering for mixed data with missing information // Comput. Stat. Data An., 2003. Vol. 41. P. 429–440.
15. *Delalleau O., Courville A., Bengio Y.* Efficient EM training of Gaussian mixtures with missing data. arXiv.org, 2012. https://arxiv.org/abs/1209.0521.
16. *Eirola E., Lendasse A., Vandewalle V., Biernacki C.* Mixture of Gaussians for distance estimation with missing data // Neurocomputing, 2014. Vol. 131. P. 32–42.

Поступила в редакцию 14.06.17

SUPERVISED LEARNING CLASSIFICATION OF INCOMPLETE CLINICAL DATA

M. P. Krivenko

Institute of Informatics Problems, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, 44-2 Vavilov Str., Moscow 119333, Russian Federation

Abstract: The article examines the effectiveness of classification methods for incomplete clinical data. Training Bayesian classifier is carried out by the maximum likelihood method for the model of a mixture of normal distributions. Rigorous derivation of formulas ensuring the realization of the steps of the EM algorithm allowed correctly applying the iterative process of obtaining estimates of the parameters of the mixture. For incomplete data, methods for selecting initial values and correcting degenerate covariance matrices for the elements of the mixture are proposed. The experimental part of the work consisted in analyzing the dependence of the quality of classification on the number of missing individual values, using data on enzymes obtained for patients with liver diseases. The real data treatment has demonstrated almost identical classification errors when applying simple and complex methods of processing of missing values in the case of low number of randomly missing individual values.

Keywords: missing data; EM algorithm; mixtures of normal distributions

DOI: 10.14357/19922264170303

References

1. Rubin, D. B. 1976. Inference and missing data. *Biometrika* 63:581–592.
2. Rubin, D. B. 1987. *Multiple imputation for nonresponse in surveys*. New York, NY: John Wiley & Sons. 256 p.
3. Little, R. J. A, and D. B. Rubin. 2002. *Statistical analysis with missing data*. 2nd ed. New York, NY: John Wiley & Sons. 408 p.
4. Mallinckrodt, C. H., P. W. Lane, D. Schnell, Y. Peng, and J. Mancuso. 2008. Recommendation for the primary analysis of continuous endpoints in longitudinal clinical trials. *Drug Inf. J.* 42:303–319.
5. Molenberghs, G., and M. G. Kenward. 2007. *Missing data in clinical studies*. West Sussex: John Wiley & Sons. 526 p.
6. Myers, W. R. 2000. Handling missing data in clinical trials: An overview. *Drug Inf. J.* 34:525–533.
7. Andridge, R. R., and R. J. A. Little. 2010. A review of hot deck imputation for survey non-response. *Int. Stat. Rev.* 78(1):40–64.
8. Myers, T. A. 2011. Goodbye, listwise deletion: presenting hot deck imputation as an easy and effective tool for handling missing data. *Commun. Meth. Measures* 5(4):297–310.
9. Little, R. J. A. 1995. Modeling the drop-out mechanism in repeated-measures studies. *J. Am. Stat. Assoc.* 90(431):1112–1121.
10. Dempster, A. P., N. M. Laird, and D. B. Rubin. 1977. Maximum likelihood from incomplete data via EM algorithm. *J. Roy. Stat. Soc. B Met.* 39(1):1–38.
11. Alber, A. 1987. *Multivariate interpretation of clinical laboratory data*. New York, NY: CRC Press. 386 p.
12. Krivenko, M. P. 2016. *Statisticheskie metody predstavleniya i statisticheskoy predvaritel'noy obrabotki referensnykh znacheniy* [Statistical methods for representation and pretreatment of reference values]. Moscow: FRC CSC RAS. 160 p.
13. Ghahramani, Z., and M. I. Jordan. 1995. Learning from incomplete data. MIT AI. A.I. Memo No. 1509. C.B.C.L. Paper No. 108. Available at: https://dspace.mit.edu/handle/1721.1/7202 (accessed June 14, 2017).
14. Hunt, L., and M. Jorgensen. 2003. Mixture model clustering for mixed data with missing information. *Comput. Stat. Data An.* 41:429–440.
15. Delalleau, O., A. Courville, and Y. Bengio. 2012. Efficient EM training of Gaussian mixtures with missing data. Available at: https://arxiv.org/abs/1209.0521 (accessed June 14, 2017).
16. Eirola, E., A. Lendasse, V. Vandewalle, and C. Biernacki. 2014. Mixture of Gaussians for distance estimation with missing data. *Neurocomputing* 131:32–42.

Received June 14, 2017

Contributor

Krivenko Michail P. (b. 1946) — Doctor of Science in technology, professor, leading scientist, Institute of Informatics Problems, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, 44-2 Vavilov Str., Moscow 119333, Russian Federation; mkrivenko@ipiran.ru

КОМПЬЮТЕРНАЯ МОДЕЛЬ СИНЕРГИИ КОЛЛЕКТИВНОГО ПРИНЯТИЯ РЕШЕНИЙ

И. А. Кириков[1], А. В. Колесников[2], С. В. Листопад[3]

Аннотация: Задачи управления сложными социально-техническими системами характеризуются множеством НЕ-факторов (в смысле А. С. Нариньяни), затрудняющих решение. Традиционно к подобным задачам привлекаются коллективы экспертов под руководством лица, принимающего решения (ЛПР), что позволяет справиться с разнородностью информации и динамичностью проблемной ситуации. По этой же причине для автоматизированного решения сложных задач актуально моделирование коллективных процессов в системах поддержки принятия решений. В статье рассматриваются вопросы моделирования коллективного решения сложных задач и возникающего при этом эффекта синергии, когда интегрированное решение лучше любого решения экспертов, работающих индивидуально.

Ключевые слова: малый коллектив экспертов; синергия; гибридная интеллектуальная многоагентная система

DOI: 10.14357/19922264170304

1 Введение

Традиция решения сложных задач коллективами экспертов под руководством ЛПР, имеет давние корни: это военные советы, коллегии министерств, всевозможные совещания, планерки, консилиумы, аналитические центры и т. д. [1]. Актуальность коллективного решения сложных задач обусловлена преимуществами перед индивидуальной работой управленца: повышение качества принимаемых решений за счет учета многообразия мнений и интеграции знаний различных специалистов; повышение доверия всех членов коллектива к результатам его работы и мотивации к реализации таких решений; соблюдение этических норм. Они во многом определяются процессами и эффектами взаимодействия экспертов, изучаемых социальной психологией и социологией со второй четверти XX в. [2–4]. Часть этих эффектов: социальная фасилитация, адаптация, самоорганизация, синергия — положительно влияют на решения сложных задач, другие, например социальная ингибация, эффект Рингельмана, групсинк и конформизм, — отрицательно.

Опытные ЛПР обеспечивают условия возникновения положительных групповых эффектов и минимизируют отрицательные, перестраивая состав и структуру системы управления, адаптируясь к изменениям во внешней среде.

Проблема в том, что бóльшая часть современных компьютерных технологий — среда реализации методов, а не инструментальное средство их синтеза. Отсюда аналогично экспертным системам, рассуждающим «не хуже» одного человека, актуальны информационные технологи для управления в условиях сложных задач не хуже коллектива специалистов.

В работе рассматривается компьютерное моделирование эффекта синергии методами гибридных интеллектуальных многоагентных систем (ГиИМАС) [5], когда коллективное решение лучше любого решения экспертов, работающих индивидуально, не взаимодействуя.

2 Понятие синергии

Синергия в широком смысле относится к «кооперативным», коллективным эффектам — в буквальном смысле это эффекты, вызываемые сущностями, которые «работают вместе» (части, элементы или отдельные лица). Этот термин часто ассоциируется с принципом «целое больше, чем сумма его частей» метафизики Аристотеля. Однако это узкое и вводящее в заблуждение понимание многогранной концепции: эффекты, вызываемые целым, отличаются от того, что детали могут производить по отдельности.

[1] Калининградский филиал Федерального исследовательского центра «Информатика и управление» Российской академии наук, baltbipiran@mail.ru

[2] Балтийский федеральный университет им. И. Канта; Калининградский филиал Федерального исследовательского центра «Информатика и управление» Российской академии наук, avkolesnikov@yandex.ru

[3] Калининградский филиал Федерального исследовательского центра «Информатика и управление» Российской академии наук, ser-list-post@yandex.ru

Таблица 1 Синергия в научных дисциплинах

Научная дисциплина	Характерный пример	Связанные термины
Физика	Квантовая когерентность Теория хаоса Фазовые переходы	Холизм, упорядочение Эмерджентность, аттрактор, порядок Кооперативные эффекты, нарушение симметрии
Термодинамика	Диссипативные структуры	Порядок/хаос, низкая энтропия, отрицательная энтропия
Биофизика	Гиперциклы	Сотрудничество, взаимодействие, координация, эмерджентность
Химия	Молекулярные макроструктуры	Симметрия, коллективная стабильность, порядок
Биохимия	Супрамолекулы	Взаимодействие, функциональная интеграция, координация
Нейробиология	Синаптическая передача	Кооперативность, пороговые эффекты, эмерджентность
Экология и поведенческая биология	Коэволюция Симбиоз Социобиология	Взаимность, паразитизм Взаимность, сотрудничество Взаимность, взаимный альтруизм, эмерджентность, сотрудничество
Экономика	Финансовая синергия Производственная синергия	Конгломерат, поглощение, слияние Взаимодействие, координация, эффект от масштаба

Известно много видов кооперативных/синергетических эффектов. Некоторые возникают из линейных или аддитивных явлений. Агрегация большого числа однородных сущностей может обеспечить преимущество коллективу. Например, колония хищных миксобактерий (*лат*. Myxococcus xanthus) способна поглотить гораздо более крупную добычу, чем одна или несколько бактерий. Колония способна коллективно вырабатывать пищеварительные ферменты в больших концентрациях, которые рассеивались бы в окружающей среде, если бы вырабатывались одной или несколькими бактериями [6]. В экономике известна синергия:

(1) финансовая, возникающая при объединении организаций разного профиля деятельности в результате снижения рисков и, соответственно, повышения доступности кредитных средств;

(2) предпринимательская, обусловленная лучшими инвестиционными возможностями для объединенной организации;

(3) расширения от совместного использования ресурсов;

(4) рыночная как следствие перекрестного субсидирования;

(5) оперативная как результат обмена знаниями [7].

В настоящее время нет согласованной теории возникновения эффекта синергии в различных системах. В табл. 1 сведены известные и хорошо описанные примеры проявления синергетических эффектов с указанием дисциплин, в которых они изучаются.

Выделяются основные особенности эффекта синергии [7]: системный характер; согласуемость с теорией гештальта М. Вертхаймеера, В. Кёлера и К. Кофки, согласно которой «целое больше, чем сумма его частей»; невоспроизводимость результатов работы системы как целого при работе всех ее элементов по отдельности.

3 Эффект синергии в малых коллективах экспертов, решающих сложные задачи

Преимущества малого коллектива экспертов (МКЭ) ориентированы на реализацию идей, не выполнимых при индивидуальном принятии решений из-за того, что у конкретного ЛПР нет возможности выйти за рамки его непосредственной деятельности. Профессиональные обязанности в МКЭ распределяются в соответствии со способностями и компетентностями исполнителей в зависимости от сложности деятельности. Синергетический эффект в МКЭ достигается «групповой компенсацией индивидуальных неспособностей». Внутрикомандное взаимодействие, партнерство и сотрудничество при решении задач «повышают эффективность не менее чем на 10%» [8], что и порождает синергетический эффект в МКЭ, когда неумения одного компенсируются навыками и способностями другого.

Отсюда актуальны методы выработки единого решения МКЭ на основе частных рекомендаций экспертов [9] (табл. 2): Дельфи [10, 11], анализа иерархий [12, 13], мозгового штурма [10], синектики [11], пула мозговой записи [10] и др.

Таблица 2 Преимущества и недостатки методов экспертного оценивания

Метод	Преимущества	Недостатки
Дельфи	Анонимное обсуждение, исключение групсинка; заочное участие экспертов	Длительность принятия решений; мнение большинства не всегда правильное; многократный пересмотр мнения экспертом
Анализа иерархий	Высокая скорость выработки решений; качественный (а не количественный, как в методе Дельфи) анализ альтернатив	Возможен групсинк; плохо подходит к условиям неопределенности; не использует коллективное творчество для выработки решений
Мозгового штурма	Генерирует инновационные варианты, синтезируя идеи; высокая скорость выработки решений; ответственность участников коллектива за принятое решение	Высокие интеллектуальные усилия участников; исключает «управление мышлением»
Синектики	Генерирует инновационные варианты, синтезируя идеи, используя аналогии, метафоры и т. п.; высокая скорость выработки решений; ответственность участников коллектива за принятое решение	Высокие интеллектуальные усилия участников; требуется обученный коллектив, иначе возрастает его критичность и снижается продуктивность; коллектив решает аналог задачи
Пула мозговой записи	Анонимное обсуждение, исключение групсинка; генерирует инновационные варианты, синтезируя идеи; высокая скорость выработки решений	Высокие интеллектуальные усилия участников

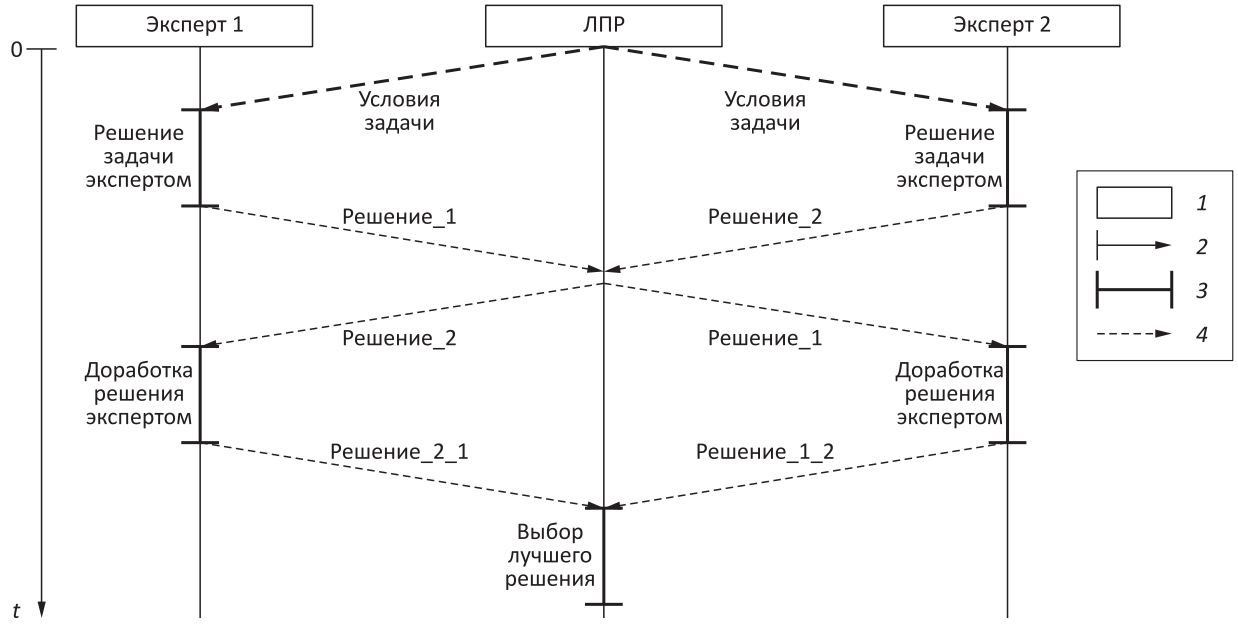


Рис. 1 Схема работы МКЭ методом пула мозговой записи: *1* — участники коллектива; *2* — шкала времени; *3* — действие; *4* — процесс передачи информации между участниками коллектива; t — модельное время

Анализ релевантности методов из табл. 2 компьютерному моделированию работы МКЭ позволяет сделать выбор в пользу пула мозговой атаки, когда функционирование МКЭ имитируется последовательностью действий: (1) ЛПР передает экспертам условия задачи; (2) каждый эксперт на основе собственной модели и оценок вырабатывает вариант ее решения и передает его ЛПР; (3) ЛПР полученные варианты конфиденциально сообщает всем экспертам, кроме источника, для доработки и улучшения; (4) эксперты улучшают варианты и возвращают их ЛПР: третий и четвертый этапы

повторяются, пока каждый из экспертов не обработает хотя бы один раз каждый вариант решения; (5) ЛПР оценивает все варианты решения, в том числе и промежуточные, и выбирает лучший на основе собственной модели задачи.

Схематично функционирование МКЭ из ЛПР и двух экспертов показано на рис. 1, где обмен решениями чередуется с индивидуальной работой экспертов по поиску новых или доработке имеющихся решений через призму своей модели внешнего мира.

В результате сложная задача редуцируется в подзадачи со специфическими для эксперта оценками, при этом эксперт использует один из множества методов ее решения. Процессы обмена мнениями касательно решения задачи указывают на случайный характер взаимодействия экспертов.

4 Моделирование эффекта синергии в гибридных интеллектуальных многоагентных системах

Моделирование МКЭ и возникающего в них эффекта синергии предлагается реализовывать с использованием ГиИМАС, которые представляют собой гибридные интеллектуальные системы (ГиИС), практикующие многоагентный подход [14]. Элементы таких ГиИС реализуются в виде агентов, обладающих свойством автономности [15]. Как и многоагентные системы (МАС), они моделируют взаимодействия автономных агентов между собой и с внешней средой, в результате которых архитектура системы может динамически перестраиваться в соответствии с конкретными функциями (ролями) агентов и установившимися отношениями между ними. В результате ГиИМАС сочетают в себе положительные стороны ГиИС и МАС: благодаря сочетанию нескольких методов искусственного интеллекта они релевантны задачам с высокой сложностью моделирования [14]; за счет имитации взаимодействия экспертов и возникающих при этом коллективных процессов они способны менять свою архитектуру для достижения синергетического эффекта.

Для компьютерной реализации модели МКЭ разработана функциональная структура ГиИМАС (рис. 2). Она может применяться при проектировании ГиИМАС для широкого круга неоднородных задач, поскольку: (1) использована общая многоагентная модель действительности; (2) перечень агентов-решателей охватывает пять классов методов из шести, используемых в ГиИС [1]; (3) порядок взаимодействия агентов определяется моделью предметной области.

Рассмотрим назначение ее агентов:

(1) интерфейсный агент запрашивает входные данные и выдает результат;

(2) агент, принимающий решения (АПР), рассылает агентам поиска решения условия задачи, определяет порядок их взаимодействия. Когда

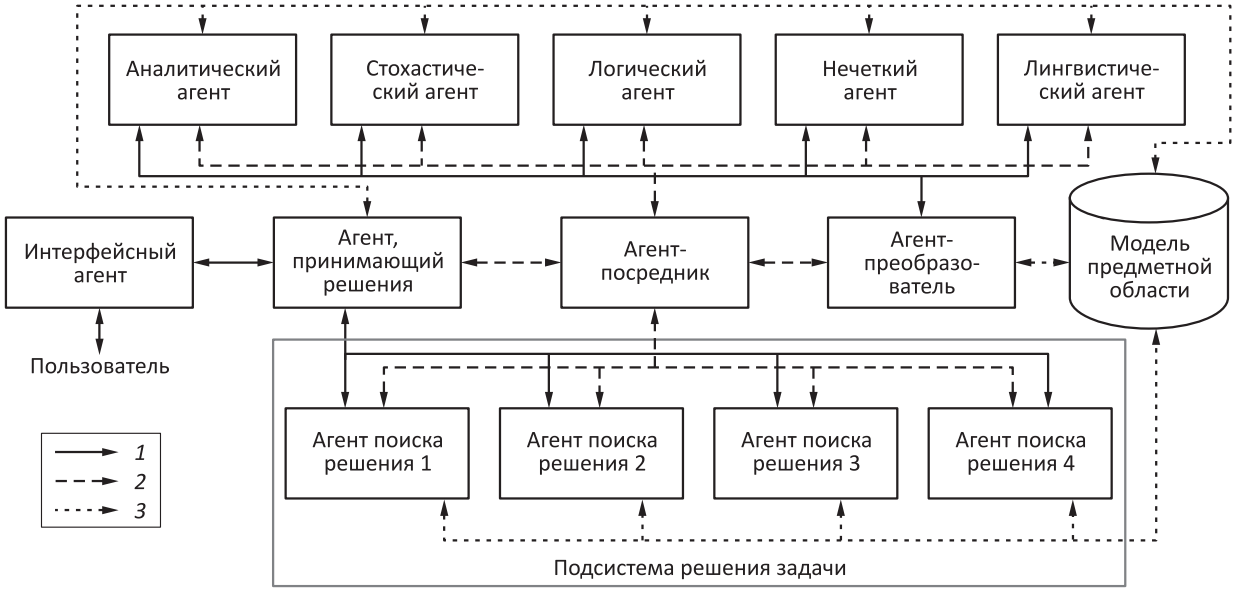


Рис. 2 Функциональная структура ГиИМАС: *1* — взаимоотношения агентов (запросы информации, передача результатов их решения); *2* — взаимоотношения агентов (запросы помощи в решении подзадач); *3* — взаимодействие (получение сведений из модели, обновление модели) агентов с моделью предметной области

последние решили задачу, он выбирает альтернативу и передает интерфейсному агенту или запускает новую итерацию, рассылая решение остальным агентам поиска;

(3) агенты поиска решения имеют знания о предметной области и выполняют генерацию и оценку решений каждый по своему критерию. Для решения подзадач тестовой сложной транспортно-логистической задачи (СТЛЗ) эти агенты используют муравьиный алгоритм;

(4) агент-посредник отслеживает имена, модели и возможности зарегистрированных агентов интеллектуальных технологий (решателей). Агенты обращаются к нему, чтобы узнать, какой из решателей может помочь в поставленной перед ними подзадаче;

(5) решатели в верхней части рис. 2 вместе с агентом-преобразователем реализуют гибридную составляющую ГиИМАС, комбинируя разнородные знания, и предоставляют «услуги» агентам с использованием моделей и алгоритмов: алгебраических уравнений для описания причинно-следственных связей концептов предметной области; метода Монте Карло; продукционной экспертной системы с рассуждениями в прямом направлении; нечеткого вывода Мамдани;

(6) модель предметной области — семантическая сеть, основа взаимодействия агентов, построена по концептуальной модели решаемой задачи. Агенты интерпретируют смысл получаемых сообщений на этой модели.

Для оценки влияния эффекта синергии на качество решений ГиИМАС проведены серии экспериментов, в которых требовалось решить СТЛЗ, т. е. найти для нескольких транспортных средств совокупность маршрутов, оптимальную по четырем критериям: суммарная стоимость; общая длительность поездок для всех транспортных средств; вероятность опоздания хотя бы к одному клиенту; надежность (мерой надежности выбрано математическое ожидание увеличения стоимости совокупности маршрута) [1]. Учитывались такие стохастические факторы, как вероятность возникновения дорожных пробок и вероятность опоздания к клиенту, потери от боя груза и др.

Исходные данные:

(1) запросы клиентов на доставку грузов (наименование, количество товара, временной интервал его доставки);

(2) сведения о дорогах к клиентам (протяженность, загруженность, качество);

(3) паспортные данные транспортных средств (расход горючесмазочных материалов, грузоподъемность и т. п.);

(4) сведения о графиках работы и заработной плате персонала (водителей и грузчиков);

(5) информация о грузе (вес, габариты, хрупкость и т. п.).

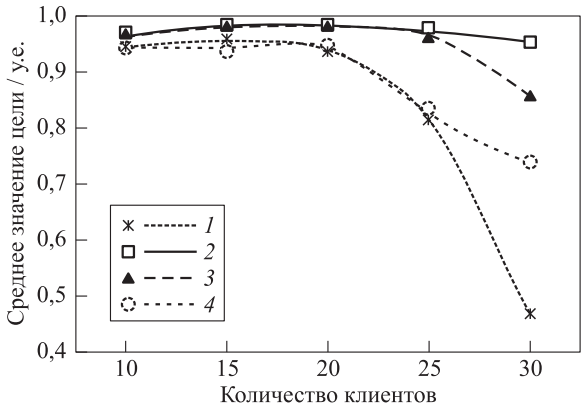
Выходные данные: совокупность маршрутов доставки грузов (по одному на транспортное средство) и их параметры: стоимость, длительность, надежность и вероятность опоздания, сводный критерий качества маршрута. Для тестирования использованы задачи из табл. 3.

Исследовались три архитектуры ГиИМАС, работающие по схеме, представленной на рис. 1: с нейтральными, сотрудничающими и конкурирующими агентами. В ГиИМАС с нейтральными агентами каждый из четырех агентов поиска решения минимизирует значение «своего» критерия оценки решения. В ГиИМАС с сотрудничающими агентами все четыре агента-поисковика минимизируют все четыре критерия оценки решения (аналогично АПР). В ГиИМАС с конкурирующими агентами один агент минимизирует стоимость и максимизирует длительность, второй — максимизирует стоимость и минимизирует длительность, третий — минимизирует вероятность опоздания и максимизирует надежность, а четвертый — максимизирует вероятность опоздания и минимизирует надежность.

Качество тестовых решений оценивалось по объективным показателям и субъективно экспертами. Для пяти задач и каждой архитектуры ГиИМАС

Таблица 3 Количественные параметры тестируемых задач

Задача	Количество клиентов	Количество дорог	Количество водителей	Количество грузчиков	Количество автомобилей
З_10	10	75	3	3	3
З_15	15	240	5	5	5
З_20	20	420	5	5	5
З_25	25	650	9	9	9
З_30	30	377	6	6	6



Архитектура ГиИМАС	Количество клиентов				
	10	15	20	25	30
1 — конкуренция	0,9449	0,9528	0,9348	0,8135	0,4687
2 — нейтралитет	0,9669	0,9802	0,9796	0,9699	0,9504
3 — сотрудничество	0,9661	0,9800	0,9795	0,9639	0,8592
4 — без взаимодействия	0,9439	0,9422	0,9411	0,8248	0,7370

Рис. 3 Среднее значение сводного критерия качества маршрута

проведено по 100 вычислительных экспериментов. По всем задачам и архитектурам ГиИМАС, а также для архитектуры без взаимодействия (агенты-поисковики не обмениваются индивидуальными решениями) построены графические зависимости числа ситуаций, когда коллективное решение лучше любого индивидуального, среднего значения сводного критерия качества маршрута (рис. 3), средних значений стоимости, длительности, надежности, вероятности опоздания для маршрутов, от числа клиентов, анализ которых показал высокое качество маршрутов, рекомендуемых ГиИМАС.

Как видно из рис. 3, в большинстве случаев любая из архитектур ГиИМАС предлагает более качественные решения, чем ГиИМАС без взаимодействия агентов, т. е. проявляется эффект синергии. Качество принимаемых решений ГиИМАС с нейтральными агентами выше, чем ГиИМАС других архитектур. Это прямое следствие того, что в ГиИМАС с нейтральными агентами вероятность возникновения синергетического эффекта выше, но чем меньше размерность задачи, тем меньше его влияние на качество решения.

Для задач с 25 и 30 клиентами эффективность ГиИМАС с конкурирующими агентами резко снижается и она демонстрирует результаты хуже, чем ГиИМАС без взаимодействия, т. е. возникает дисергия, когда коллективное решение не лучше решений индивидуальных агентов. Очевидно, этот эффект обусловлен невозможностью конкурирующих агентов «договориться» на задачах с высокой комбинаторной сложностью.

Таким образом, при правильной организации взаимодействия в ГиИМАС эффект синергии повышает качество принимаемых решений по сравнению с ГиИМАС, в которой он не моделируется. По итогам тестовой эксплуатации программного продукта ТРАНСМАР, реализующего ГиИМАС, на двух объектах средняя суммарная себестоимость доставки грузов в день сократилась на 7,2%, средняя суммарная длительность доставки в день — на 12,13%, среднее время построения маршрутов в день уменьшилось на 23,14%.

5 Заключение

В работе рассмотрено понятие эффекта синергии и один из подходов для его достижения при решении сложных задач МКЭ — организация рассуждений методом пула мозговой записи. Данный метод положен в основу компьютерной модели МКЭ — ГиИМАС, отображающей и комбинирующей на компьютере разнообразие знаний экспертов о проблемной среде, что имитирует полиязыковой характер сложных задач, с одной стороны, и социальный, коллективный характер решений, когда моделируется взаимодействие экспертов друг с другом и с ЛПР — с другой стороны.

Результаты лабораторных экспериментов с системой, а также практического использования программного продукта ТРАНСМАР, реализующего модель ГиИМАС, показали, что моделирование эффекта синергии в ГиИМАС повышает качество

принимаемых решений по сравнению с ГиИМАС, в которой агенты поиска решений не обмениваются решениями и эффект синергии отсутствует.

Литература

1. *Трахтенгерц Э. А., Степин Ю. П., Андреев А. Ф.* Компьютерные методы поддержки принятия управленческих решений в нефтегазовой промышленности. — М.: СИНТЕГ, 2005. 592 с.
2. *Freud S.* Group psychology and the analysis of the ego. — The international psycho-analytical library ser., 1922. Vol. 6. P. 1–134.
3. *Lewin K.* Resolving social conflicts: Selected papers on group dynamics. — New York, NY, USA: Harper & Row, 1948. 230 p.
4. *Кириков И. А., Колесников А. В., Листопад С. В.* Моделирование систем поддержки принятия решений синергетическим искусственным интеллектом // Информатика и её применения, 2013. Т. 7. Вып. 3. С. 62–69.
5. *Колесников А. В.* Гибридные интеллектуальные системы. Теория и технология разработки. — СПб.: СПбГТУ, 2001. 711 с.
6. *Bonner J. T.* The evolution of complexity. — Princeton, NJ, USA: Princeton University Press, 1988. 272 p.
7. *Benecke G., Schurink W., Roodt G.* Towards a substantive theory of synergy // SA J. Human Resource Management, 2007. Vol. 5. No. 2. P. 9–19.
8. *Зимняя И. А.* Педагогическая психология. — Ростов-на-Дону: Феникс, 1997. 480 с.
9. *Орлов А. И.* Теория принятия решений. — М.: Экзамен, 2005. 656 с.
10. *Сладкевич В. П., Чернявский А. Д.* Современный менеджмент (в схемах). — Киев: МАУП, 2003. 152 с.
11. *Колпаков В. М.* Теория и практика принятия управленческих решений. — Киев: МАУП, 2004. 504 с.
12. *Саати Т.* Принятие решений. Метод анализа иерархий / Пер. с англ. — М.: Радио и связь, 1993. 278 с. (*Saaty T. L.* The analytic hierarchy process. — New York, NY, USA: McGraw-Hill, 1980. 296 p.)
13. *Сухарев М. Г.* Методы прогнозирования. — М.: РГУ нефти и газа, 2009. 208 с.
14. *Колесников А. В., Кириков И. А., Листопад С. В., Румовская С. Б., Доманицкий А. А.* Решение сложных задач коммивояжера методами функциональных гибридных интеллектуальных систем / Под ред. А. В. Колесникова. — М.: ИПИ РАН, 2011. 295 с.
15. *Тарасов В. Б.* От многоагентных систем к интеллектуальным организациям: философия, психология, информатика. — М.: Эдиториал УРСС, 2002. 352 с.

Поступила в редакцию 16.07.17

COMPUTER MODEL OF SYNERGY OF TEAM DECISION-MAKING

I. A. Kirikov[1], A. V. Kolesnikov[1,2], and S. V. Listopad[1]

[1]Kaliningrad Branch of the Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, 5 Gostinaya Str., Kaliningrad 236000, Russian Federation
[2]Immanuel Kant Baltic Federal University, 14 A. Nevskogo Str., Kaliningrad 236041, Russian Federation

Abstract: The problems of the practice of complex sociotechnical systems management are characterized by a variety of NOT-factors (following the terminology suggested by A. S. Narinyani) that hamper their solution. Traditionally, teams of experts under the leadership of a decision-maker are involved in such problems to deal with the heterogeneity of information and the dynamic nature of the problem. For the same reason, the modeling of team processes in decision support systems is important for the automated solving of complex problems. The article deals with the issues of modeling the process of collective solving of complex problems and the resulting synergy effect, when an integrated solution is better than any decision of experts working individually.

Keywords: small team of experts; synergy; hybrid intelligent multiagent system

DOI: 10.14357/19922264170304

References

1. Trakhtengerts, E. A, Yu. P. Stepin, and A. F. Andreev. 2005. *Komp'yuternye metody podderzhki prinyatiya upravlencheskikh resheniy v neftegazovoy promyshlennosti* [Computer methods for management decision making support in the oil and gas industry]. Moscow: SINTEG. 592 p.
2. Freud, S. 1922. *Group psychology and the analysis of the ego*. The international psycho-analytical library ser. 6:1–134.
3. Lewin, K. 1948. *Resolving social conflicts: Selected papers on group dynamics*. New York, NY: Harper & Row. 230 p.
4. Kirikov, I. A., A. V. Kolesnikov, and S. V. Listopad. 2013. Modelirovanie sistem podderzhki prinyatiya resheniy sinergeticheskim iskusstvennym intellektom [Decision support systems modeling with synergetic artificial intelligence]. *Informatika i ee Primeneniya — Inform. Appl.* 7(3):62–69.

5. Kolesnikov, A. V. 2001. *Gibridnye intellektual'nye sistemy. Teoriya i tekhnologiya razrabotki* [Hybrid intelligent systems: Theory and technology of development]. St. Petersburg: SPbGTU Publ. 711 p.
6. Bonner, J. T. 1988. *The evolution of complexity*. Princeton, NJ: Princeton University Press. 272 p.
7. Benecke, G., W. Schurink, and G. Roodt. 2007. Towards a substantive theory of synergy. *SA J. Human Resource Management* 5(2):9–19.
8. Zimnyaya, I. A. 1997. *Pedagogicheskaya psikhologiya* [Pedagogical psychology]. Rostov-on-Don: Phoenix. 480 p.
9. Orlov, A. I. 2005. *Teoriya prinyatiya resheniy* [Decision theory]. Moscow: Examen Publ. 656 p.
10. Sladkevich, V. P., and A. D. Chernyavskiy. 2003. *Sovremennyy menedzhment (v skhemakh)* [Modern management (in diagrams)]. Kiev: Interregional Academy of Personnel Management. 152 p.
11. Kolpakov, V. M. 2004. *Teoriya i praktika prinyatiya upravlencheskikh resheniy* [Theory and practice of management decision-making]. Kiev: Interregional Academy of Personnel Management. 504 p.
12. Saaty, T. L. 1980. *The analytic hierarchy process*. New York, NY: McGraw-Hill. 296 p.
13. Sukharev, M. G. 2009. *Metody prognozirovaniya* [Forecasting methods]. Moscow: Russian State University of Oil and Gas. 208 p.
14. Kolesnikov, A. V., I. A. Kirikov, S. V. Listopad, S. B. Rumovskaya, and A. A. Domanitskiy. 2011. *Reshenie slozhnykh zadach kommivoyazhera metodami funktsional'nykh gibridnykh intellektual'nykh system* [Complex travelling salesman problems solving by the methods of the functional hybrid intelligent systems]. Moscow: IPI RAN. 295 p.
15. Tarasov, V. B. 2002. Ot mnogoagentnykh sistem k intellektual'nym organizatsiyam: Filosofiya, psikhologiya, informatika [From multiagent systems to intelligent organizations: Philosophy, psychology, and informatics]. Moscow: Editorial URSS. 352 p.

Received July 16, 2017

Contributors

Kirikov Igor A. (b. 1955) — Candidate of Science (PhD) in technology; director, Kaliningrad Branch of the Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, 5 Gostinaya Str., Kaliningrad 236000, Russian Federation; baltbipiran@mail.ru

Kolesnikov Alexander V. (b. 1948) — Doctor of Science in technology; professor, Department of Telecommunications, Immanuel Kant Baltic Federal University, 14 A. Nevskogo Str., Kaliningrad 236041, Russian Federation; senior scientist, Kaliningrad Branch of the Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, 5 Gostinaya Str., Kaliningrad 236000, Russian Federation; avkolesnikov@yandex.ru

Listopad Sergey V. (b. 1984) — Candidate of Science (PhD) in technology; senior scientist, Kaliningrad Branch of the Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, 5 Gostinaya Str., Kaliningrad 236000, Russian Federation, ser-list-post@yandex.ru

ИНФОРМАТИКА И ЕЁ ПРИМЕНЕНИЯ, 2017. Т. 11. Вып. 3. С. 42–50

МЕТОДЫ ТЕОРИИ КАТЕГОРИЙ В МОДЕЛЬНО-ОРИЕНТИРОВАННОЙ СИСТЕМНОЙ ИНЖЕНЕРИИ

С. П. Ковалёв[1]

Аннотация: Предложен математический аппарат на базе теории категорий, который позволяет формально описывать и строго исследовать процедуры применения моделей в инженерной деятельности, составляющие сущность модельно-ориентированной системной инженерии (Model-Based Systems Engineering, MBSE). В основе аппарата лежит математическое представление сборочных чертежей (мегамоделей систем) диаграммами в категориях, объектами которых служат модели, а морфизмы представляют действия по сборке моделей систем из моделей компонентов. Адекватность аппарата обоснована исходя из требований стандартов, регламентирующих описание структуры систем, в том числе IEC 81346. Предложены и исследованы теоретико-категорные методы решения ряда практических задач сборки систем. Приведены примеры решения таких задач в категориях, представляющих две ключевые области применения MBSE: геометрическое моделирование изделий сложной формы и дискретно-событийное имитационное моделирование поведения технических систем.

Ключевые слова: модельно-ориентированная системная инженерия; мегамодель; теория категорий; копредел

DOI: 10.14357/19922264170305

1 Введение

Модельно-ориентированная системная инженерия состоит в формализованном применении моделирования в поддержке жизненного цикла систем, включая сбор требований, проектирование, проверку и приемку, другие стадии [1]. Модели, разрабатываемые в ходе процедур MBSE, пригодны к автоматической обработке на компьютерах. Это позволяет сначала задавать, верифицировать и оптимизировать проектные решения на моделях «в цифре и только потом воплощать «в железе», снижая затраты на организацию жизненного цикла изделий и сокращая сроки выполнения работ [2].

И все же внедрение технологий MBSE в инженерную деятельность происходит медленно. Это связано во многом с нехваткой единой концептуальной базы инженерного моделирования: предлагается много частных языков и технологий, слабо совместимых друг с другом и плохо приспособленных для совместной разработки моделей большими мультидисциплинарными коллективами [3]. Тем самым затрудняется переход от набора электронных чертежей к полноценному электронно-цифровому макету (digital mock-up) промышленного изделия.

Естественный, хотя и «трудный», подход к получению результатов общего характера, унифицирующих разнородные технологии, состоит в том, чтобы как можно более строго формализовать процедуры моделирования. Формализация позволит совершенствовать процедуры MBSE и передавать их на исполнение компьютеру без пробелов и искажений. Самый высокий уровень строгости достигается при привлечении математического аппарата, поскольку математика позволяет надежно доказывать или опровергать утверждения, характеризующие корректность и эффективность процедур.

В настоящей работе предложен аппарат, основанный на математическом представлении сборочных чертежей («мегамоделей» систем) ориентированными графами (диаграммами). Узлы такого графа помечаются обозначениями моделей частей, а ребра помечаются обозначениями действий (activities), посредством которых части собираются в систему. Представление структуры систем графами регламентируется, в частности, стандартом IEC 81346 [4]. Естественным источником математических методов конструирования и анализа мегамоделей служит теория категорий (см., например, [5, 6]). Модели рассматриваются как объекты подходящих категорий, а действия формально описываются морфизмами. Строятся и исследуются теоретико-категорные конструкции, описывающие процедуры MBSE на абстрактном концептуальном уровне. Определенный опыт такого исследования был накоплен в инженерии программного обеспечения [7] и теперь может быть обобщен для системной инженерии в целом. На-

[1] Институт проблем управления им. В. А. Трапезникова Российской академии наук, kovalyov@nm.ru

пример, сборке системы согласно некоторой мегамодели отвечает построение копределa диаграммы — универсальной конструкции [5].

Статья построена следующим образом. В разд. 2 приведен обзор принципов описания структуры систем согласно стандарту IEC 81346. Раздел 3 посвящен практическим проблемам мегамоделирования и сборке систем. В разд. 4 вводятся конструкции теории категорий, позволяющие формально решать задачи мегамоделирования. В заключении приводятся выводы и намечаются направления дальнейших исследований.

2 Структура систем и стандарт IEC 81346

Важной проблемой MBSE, отмеченной во введении, является слабая совместимость языков и инструментов моделирования от разных поставщиков. Основным подходом к достижению совместимости является стандартизация — принятие обязывающих документов, устанавливающих требования и принципы взаимозаменяемости инструментов. Многие стандарты определяют конкретные форматы машиночитаемой записи моделей, нейтральные относительно разработчиков инструментов MBSE. Примером служит формат описания твердотельных геометрических моделей STEP, стандартизованный семейством ISO 10303. Однако для формализации MBSE в целом интерес представляют в первую очередь стандарты более общего плана, унифицирующие принципы и методы применения моделей в жизненном цикле систем независимо от способа записи моделей. С этой точки зрения внимания заслуживает международный стандарт IEC 81346-1:2009 «Промышленные системы, установки и оборудование — принципы структурирования и ссылочные обозначения — часть 1: основные правила» («Industrial Systems, Installations and Equipment and Industrial Products — Structuring Principles and Reference Designations — Part 1: Basic Rules») [4]. Стандарт не принят в России, однако ряду его положений в области структуры систем соответствует российский ГОСТ 2.053-2013 «ЕСКД. Электронная структура изделия. Общие положения».

В стандарте IEC 81346 рассматривается ряд вопросов моделирования структуры систем и идентификации отдельных единиц в составе систем. Системная единица названа в стандарте объектом, причем принципиально не проводится различие между объектами реального мира, составляющими реально существующие системы, и объектами мыслительной деятельности — моделями единиц, составляющими модели систем. Таким образом, стандарт выходит за рамки MBSE и рассматривает ряд вопросов системной инженерии вообще. Иерархическая структура системы (холархия [3]) изображается деревом, узлы которого помечены обозначениями объектов. Важным достижением стандарта является выявление того факта, что одна и та же система задается не одной, а несколькими в общем случае различными иерархическими структурами, возникающими в результате декомпозиции согласно различным принципам (аспектам). В их числе:

- функциональная (function-oriented) структура, отвечающая разделению системных единиц по выполняемым ими функциям в составе системы;
- продуктовая (product-oriented), или модульная, структура, отражающая сборочную (технологическую) конфигурацию системы;
- структура размещения (location-oriented), в соответствии с которой единицы располагаются в физическом пространстве.

Ясно, что один и тот же объект может входить в несколько структур и при этом находиться на различных уровнях. В то же время в некоторых аспектах объект может никак не проявлять себя и вследствие этого отсутствовать в соответствующих структурах. Полное идентифицирующее ссылочное обозначение объекта (reference designation) конструируется путем последовательного перечисления всех объектов, находящихся на пути от корня дерева рассматриваемой структуры до данного объекта включительно. Наименование каждого объекта в этом перечислении составляется из символьного обозначения аспекта, буквенного обозначения класса (типа), к которому относится объект, и порядкового номера объекта среди экземпляров своего класса. Таким путем обеспечивается уникальность наименования любой единицы в пределах системы. Например, функциональная структура обозначается символом «=», а функциональный класс переключателей потоков ресурсов обозначается буквами QA, так что первая по порядку единица, выполняющая функцию переключения, называется =QA1, а ее полное ссылочное обозначение может выглядеть как =WP1=WC1=QA1. Если объект присутствует в нескольких структурах, то он может иметь несколько ссылочных обозначений, как показано на рис. 1 [4].

С точки зрения практики системной инженерии большой интерес представляет описание эволюции структурного представления системы по ходу жизненного цикла, приведенное в приложении B к стандарту IEC 81346. «Строительный материал»

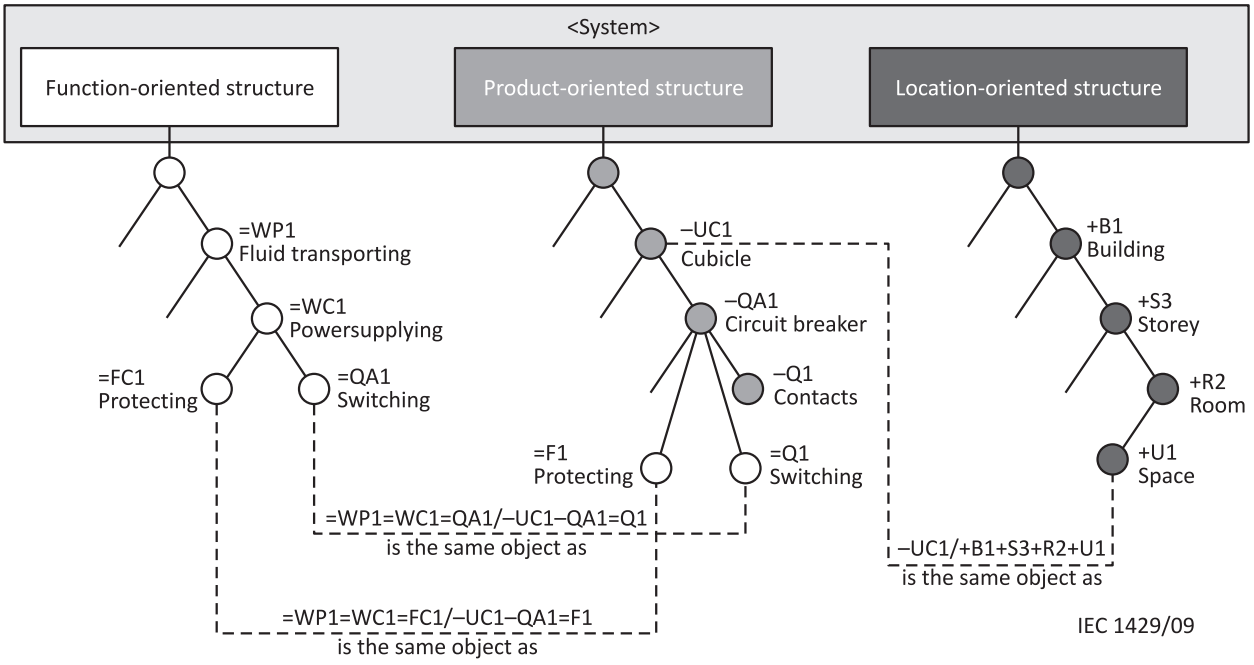


Рис. 1 Пример ссылочных обозначений структурных единиц системы

для структур имеет вид (виртуального) справочника или каталога объектов, из которого выбираются объекты для включения в структуру.

В начале жизненного цикла системы на основе исходных требований к ней конструктор строит ее функциональную структуру. Затем определяется пространственное положение функциональных объектов, в результате чего создается структура размещения. На следующей стадии формируются закупочные спецификации, образующие продуктовую структуру. В ходе последующих стадий жизненного цикла эти структуры могут трансформироваться. На каждой стадии могут происходить замена, слияние и расщепление объектов. Таким образом, объекты разных структур системы связаны отношением вида «многие ко многим», вдоль которого прослеживаются (трассируются) исходные требования.

В то же время стандарт не предусматривает указание способов, какими объекты собраны в системы. Поэтому структуру системы можно рассматривать как эскизный проект, в котором отражены лишь факты вхождения системных единиц более низкого уровня иерархии в единицы более высокого уровня.

Проект такого рода поступает на вход технологу, который определяет конкретные операции сборки каждой единицы каждого уровня иерархии. При необходимости технолог вносит изменения в конструкцию объектов (такие как нарезка резьбы) и добавляет связующие интерфейсные объекты (такие как клей, трансформатор и др.). В результате для каждого составного объекта формируется сборочный чертеж, на котором указаны все составляющие объекты и действия по их соединению в целях получения системы. Технологическая проработка требуется на всех стадиях жизненного цикла, на которых формируется либо изменяется какая-либо из структур системы.

3 Мегамоделирование и сборка систем

В MBSE объекты, образующие структуры систем, описываются формализованными компьютерными моделями различных видов: геометрическими фигурами и телами, численными аппроксимациями дифференциальных уравнений, оснащенными графами и т. д. При этом, как свидетельствуют стандарты типа IEC 81346, для анализа структуры систем и организации сборки необходимо знать не столько внутреннюю структуру моделей, сколько ассортимент их возможностей соединяться с другими моделями в целях формирования моделей составных объектов. Иными словами, модели рассматриваются как «черные ящики» с известным поведением по отношению к другим моделям. Каталог объектов, упоминавшийся в предыдущем разделе, в условиях приме-

нения MBSE составляется из моделей и описаний действий по их соединению.

Структуры систем и сборочные чертежи представляют собой частные случаи мегамоделей (megamodel) — моделей, состоящих из моделей и связей между ними [8]. Мегамодель, в которой связи описывают соединение моделей, образующих некоторую систему, называется конфигурацией этой системы [5]. Существуют и другие виды мегамоделей, предназначенные для описания других процедур MBSE, таких как формирование модели согласно заданной метамодели (instantiating) [9]. Но в настоящей работе сосредоточимся на конфигурациях и сборке систем.

Например, в моделировании механических систем, состоящих из твердых тел, моделями деталей и сборочных единиц служат геометрические тела, которые могут быть представлены для компьютерной обработки различными способами: конструктивным, воксельным, граничным [10]. Объекты, составляющие механические системы, т. е. представления экземпляров тел, получаются из моделей путем аффинных изометрий и растяжений. Так, из набора цилиндров разных размеров составляется модель штанги (спортивного снаряда). В функциональной структуре штанги по IEC 81346 цилиндры представлены разными объектами, поскольку они выполняют разные функции, хотя порождаются одной и той же геометрической моделью. Соответственно, в каталоге моделей содержится тело в форме цилиндра, допускающее несколько разных действий по включению в состав штанги.

В качестве еще одного примера рассмотрим дискретно-событийное имитационное моделирование, поддержка которого относится к числу важнейших достижений MBSE [1]. Здесь модель имеет вид сценария — фрагмента предполагаемой истории поведения моделируемой системы, представленного потоком дискретных событий различных видов. Некоторые события могут вызывать либо запрещать возникновение других событий. Описания действий по сборке сценариев поведения систем отражают вклад сценариев поведения составляющих. Так, сценарий работы цеха составляется из сценариев работы станков, связанных друг с другом согласно маршрутным картам [11].

Сформулируем задачу мегамоделирования сборки систем в общем виде следующим образом. По мегамодели, представляющей конфигурацию некоторой системы, требуется сконструировать модель системы как целого и рассчитать для нее моделируемые параметры, в том числе эмерджентные — не присущие никакой из составляющих единиц в отдельности. Принцип конструирования модели системы легко усмотреть из организации структурного представления: система должна находиться на иерархическом уровне, располагающемся непосредственно над уровнем составляющих ее объектов. Иными словами, модель системы должна включать в себя модели всех составляющих с учетом их конфигурационных связей и в то же время включаться в любые модели, включающие в себя модели всех составляющих конфигурации.

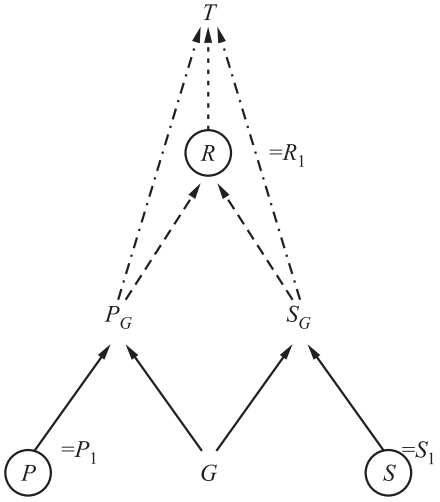


Рис. 2 Схема склеивания

Поясним этот принцип на простом примере. Предположим, что нужно объединить в систему два объекта P и S и что технолог решил сделать это с помощью клея — третьего объекта G, который может быть соединен и с P, и с S. Действие клея описывается конфигурацией следующего вида: объекты G и P порождают в результате соединения известный промежуточный комплексный объект P_G, содержащий их, а объекты G и S порождают объект S_G. Система R, полученная путем склеивания P с S при помощи G, отбирается среди объектов, содержащих P_G и S_G, по следующему структурному критерию: объект R должен содержаться в любом объекте T, содержащем P_G и S_G. Схематически этот критерий изображен на рис. 2.

Если объект R, удовлетворяющий указанному структурному критерию, существует, то он действительно отвечает системе, которая собрана из S и P путем склеивания посредством G (и не содержит ничего «лишнего»). Более того, легко видеть, что такой объект R определяется, по существу, однозначно в том смысле, что любые два объекта R и R', удовлетворяющие структурному критерию, содержатся друг в друге. Если же нужного объекта R не существует, то делается вывод, что технолог ошибся: клей G не способен соединить объекты P и S.

В структурное представление, выполненное по стандарту IEC 81346 либо по ГОСТу 2.053-2013, входят только объекты P, S и R и две композитные стрелки: $P \to R$, проходящая через P_G, и $S \to \to R$, проходящая через S_G (так что мегамодель склеивания — это часть схемы, ограниченная треугольником PSR). Кроме того, стрелки на схеме склеивания, в отличие от структуры, представляют не просто факты включения объектов друг в друга, а конкретные действия по их соединению. При этом соблюдается следующее естественное условие структурной корректности: если из одного объекта можно прийти в другой разными путями по схеме, то эти пути задают одно и то же композитное действие. Например, клей G включается в состав системы R единственным способом, несмотря на наличие двух путей $G \to P_G \to R$ и $G \to S_G \to R$: в действительности не имеет значения, через какой промежуточный объект «прослеживается» включение клея в систему. Таким образом, мегамодель сборки содержит больше информации, чем иерархическая структура системы.

Если модели содержат значения тех или иных параметров, а описание действий по их соединению позволяет выявить правила преобразования значений, то по мегамодели сборки можно вычислить значения параметров для системы. Известны примеры вычислений такого рода в области разработки новых композиционных материалов [12]. Осредненные (эффективные) физические характеристики композитов, такие как модуль Юнга и коэффициент Пуассона, сложным образом зависят от характеристик компонентов и способов изготовления композита из них. При помощи методов теории упругости эти зависимости задаются в форме линеаризованных матричных соотношений, которые приписываются к стрелкам мегамоделей, представляющим включение компонентов в композиты. Появляется возможность рассчитывать на компьютере свойства композитов по базе данных компонентов, без проведения дорогостоящих физических экспериментов.

В заключение раздела отметим, что хотя прямой расчет системы по конфигурации имеет большое значение, в MBSE он играет вспомогательную роль. Согласно стандарту IEC 81346 и практикам системной инженерии, система обычно проектируется сверху вниз — от корня структурной иерархии к составляющим [13]. Это означает, что технолог в основном решает не прямую, а обратную задачу: модель системы, которую нужно собрать, известна, а нужно построить (восстановить) конфигурацию, из которой такая система может быть получена путем сборки, с учетом различных ограничений. Формальные математические постановки и методы решения обратных задач мегамоделирования представляют собой крупную перспективную тему исследований, выходящую за рамки настоящей статьи.

4 Теория категорий в мегамоделировании

Как указывалось во введении, естественным источником математических методов конструирования и анализа мегамоделей служит теория категорий. Категорией называется коллекция абстрактных объектов, попарно связанных морфизмами (стрелками). Точное определение занимает буквально несколько строк [14]: категория C состоит из совокупности объектов $\mathrm{Ob}\, C$ и совокупности морфизмов $\mathrm{Mor}\, C$, на которых заданы следующие операции:

(1) каждому морфизму f сопоставляется два объекта: область $\mathrm{dom}\, f$ и кообласть $\mathrm{codom}\, f$ (соотношения вида $\mathrm{dom}\, f = A$ и $\mathrm{codom}\, f = B$ наглядно записываются в форме стрелки f: $A \to B$, а множество всех морфизмов, удовлетворяющих этим соотношениям, обозначается через $\mathrm{Mor}(A, B)$);

(2) для любой пары морфизмов f, g, удовлетворяющей условию $\mathrm{codom}\, f = \mathrm{dom}\, g$, определена композиция — морфизм $g \circ f : \mathrm{dom}\, f \to \to \mathrm{codom}\, g$, причем она ассоциативна: для любой тройки морфизмов f, g, h, удовлетворяющей условиям $\mathrm{codom}\, f = \mathrm{dom}\, g$ и $\mathrm{codom}\, g = = \mathrm{dom}\, h$, выполняется соотношение $h \circ (g \circ f) = = (h \circ g) \circ f$;

(3) любой объект A обладает тождественным морфизмом $1_A : A \to A$ таким, что для любого морфизма $f : A \to B$ выполняется соотношение $f \circ 1_A = 1_B \circ f = f$.

Классическим примером категории служит **Set**, состоящая из всех множеств и всех их отображений: закон композиции отображений задается стандартной подстановкой, а тождественным морфизмом произвольного множества служит его тождественное отображение на себя.

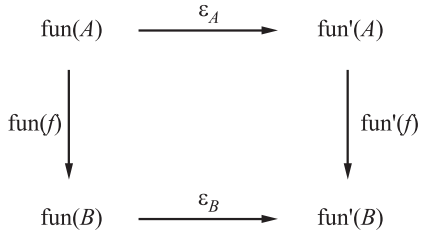
Вместе с категорией вводится понятие функтора — отображения категорий, сохраняющего структуру. Функтор $\mathrm{fun} : C \to D$, действующий из категории C в D, — это пара одноименных отображений $\mathrm{fun} : \mathrm{Ob}\, C \to \mathrm{Ob}\, D$, $\mathrm{fun} : \mathrm{Mor}\, C \to \mathrm{Mor}\, D$, удовлетворяющая следующим условиям (для произвольных C-морфизмов f, g и C-объекта A):

(1) $\mathrm{fun}\,(\mathrm{dom}\, f) = \mathrm{dom}\,\mathrm{fun}\,(f), \mathrm{fun}\,(\mathrm{codom}\, f) = = \mathrm{codom}\,\mathrm{fun}\,(f)$;

(2) $\mathrm{fun}\,(g \circ f) = \mathrm{fun}\,(g) \circ \mathrm{fun}\,(f)$, если композиция $g \circ f$ определена;

(3) $\mathrm{fun}\,(1_A) = 1_{\mathrm{fun}\,(A)}$.

Все категории и все функторы образуют (формальную) категорию **CAT**. Чтобы исследовать взаимосвязь между функторами, вводится следующее понятие: естественным преобразованием ε функтора $\mathrm{fun} : C \to D$ в $\mathrm{fun}' : C \to D$ называется любое семейство D-морфизмов $\varepsilon_A : \mathrm{fun}\,(A) \to \mathrm{fun}'(A)$, $A \in \mathrm{Ob}\,C$, такое что для любого C-морфизма $f : A \to B$ выполняется соотношение $\varepsilon_B \circ \mathrm{fun}\,(f) = \mathrm{fun}'(f) \circ \varepsilon_A$:



Эффективность применения теории категорий в качестве математического аппарата MBSE обусловлена тем, что любой каталог моделей представляет собой не что иное, как категорию. Действительно, любая цепочка действий по соединению моделей порождает композитное действие (процесс) и, кроме того, любая модель допускает пустое действие над самой собою, не подразумевающее никаких изменений (процедура «ничегонеделания»). Например, в твердотельном моделировании механических систем объектами категории моделей выступают тела — подмножества в $\mathbb{R}^3$, которые являются ограниченными, регулярными (совпадают с замыканием своей внутренности) и полуаналитическими (допускают представление конечными булевыми комбинациями множеств вида $\{(x, y, z) | F_i(x, y, z) \leq 0\}$, где $F_i : \mathbb{R}^3 \to \mathbb{R}$ является вещественной аналитической функцией для всех i) [10]. Чтобы было возможно задавать процедуры типа склеивания участков поверхности тел, в категорию геометрических моделей добавляются ограниченные регулярные полуаналитические подмножества в $\mathbb{R}^n$, $0 \leq n \leq 2$, при помощи стандартного вложения $\mathbb{R}^n$ в $\mathbb{R}^3$. Далее выполняется факторизация: отождествляются друг с другом все множества, переходящие друг в друга под действием аффинных изометрий. Морфизмы таких классов эквивалентности, описывающие действия по сборке составных механических систем, порождаются изометрическими вложениями множеств и растяжениями. Получается подкатегория в **Set**, которую будем обозначать через **MBS** (от Multibody Systems).

Для многих известных технологий MBSE формальное описание каталогов поддерживаемых моделей приводит к категориям множеств со структурой — алгебраических систем, топологических пространств, графов и т. д. Морфизмами в таких категориях служат отображения множеств, совместимые со структурой. На любой такой категории действует канонический функтор в **Set**, «забывающий» структуру.

В качестве примера приведем дискретно-событийное моделирование, в котором математической моделью сценария служит множество событий, частично упорядоченное причинно-следственными зависимостями и размеченное видами событий [15]. Действия по сборке сложных сценариев задаются монотонными отображениями, сохраняющими разметку, поскольку ни события, ни зависимости, ни метки не могут быть «потеряны» при соединении сценариев поведения компонентов в сценарии поведения систем. Получается категория **Pomset**, состоящая из всех помеченных частично упорядоченных множеств и всех их монотонных отображений, сохраняющих разметку. Имеется функтор $|-| : \mathbf{Pomset} \to \mathbf{Set} : S \mapsto |S|$, «забывающий» порядок и разметку.

Зафиксируем произвольную категорию C, представляющую некоторый каталог моделей. Как и для любой алгебраической системы, определена конструкция подкатегории в C — это пара, состоящая из подкласса в $\mathrm{Ob}\,C$ и подкласса в $\mathrm{Mor}\,C$, замкнутых относительно унаследованных из C операций. Подкатегория в C называется полной, если любой C-морфизм, область и кообласть которого содержатся в ней, сам содержится в ней. Например, подкатегориями описываются различные аспекты структурного представления систем согласно стандарту IEC 81346. Действительно, композиция двух морфизмов, представляющих действия по формированию некоторого аспекта структуры, также должна входить в этот аспект, поскольку стандарт предписывает строить цепочки для идентификации объектов в структуре системы. Кроме того, если объект присутствует в аспекте, то его тождественный морфизм формально должен быть включен в этот аспект. В то же время подкатегории, описывающие все аспекты, не обязаны образовывать в совокупности разбиение категории C: как показывает рис. 1, возможны как действия, входящие в несколько аспектов одновременно, так и композитные действия с переходом между структурами, не входящие ни в один аспект. Требуется лишь, чтобы объединение классов объектов всех этих подкатегорий совпадало с $\mathrm{Ob}\,C$, поскольку не имеет смысла вводить модели, не входящие ни в одну структуру.

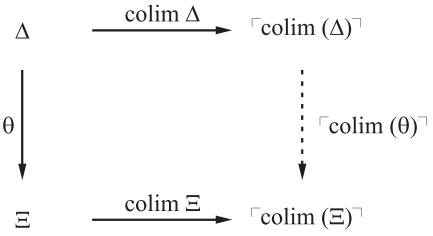
Категории можно получать из графов: любой ориентированный мультиграф порождает категорию, объектами в которой служат все узлы, а морфизмами — все пути. Областью и кообластью морфизма являются соответственно начало и конец пути, композиция морфизмов действует как конкатенация путей, а тождественным морфизмом узла a является пустой путь из a в a, не содержащий ни одного ребра. Отсюда получается фундаментальное понятие C-диаграммы — это функтор вида $\Delta : X \to C$, где X — категория, порожденная некоторым графом и называемая схемой диаграммы. Все C-диаграммы образуют категорию $\mathbf{D}C$ (ковариантная категория «сверхзапятой» [14]), в которой морфизмом диаграммы $\Delta : X \to C$ в $\Xi : Y \to C$ служит любая пара вида $\langle \gamma, fd \rangle$, состоящая из функтора $fd : X \to Y$ и естественного преобразования $\gamma : \Delta \to \Xi \circ fd$; закон композиции морфизмов диаграмм имеет вид:

$$\langle \gamma, fd \rangle \circ \langle \varphi, gd \rangle = \langle \gamma_{gd(-)} \circ \varphi, fd \circ gd \rangle .$$

В теории категорий накоплен богатый арсенал алгебраических методов конструирования и анализа диаграмм.

Любая мегамодель задается C-диаграммой, так что категорное представление каталогов моделей позволяет формально решать задачи мегамоделирования. Морфизмы диаграмм описывают структурные преобразования мегамоделей, выполняемые при помощи инструментов MBSE. Покажем, как решаются средствами теории категорий прямые задачи мегамоделирования. Здесь применяется одна из основных теоретико-категорных конструкций — копредел диаграммы [5], который строится следующим образом. Обозначим через $\mathbf{1}$ категорию, состоящую из одного объекта 0 и одного морфизма 1_0. Из любой категории X имеется в точности один функтор $!_X : X \to \mathbf{1}$, сопоставляющий объект 0 любому X-объекту (иными словами, $\mathbf{1}$ является терминальным $\mathbf{CAT}$-объектом). Имеется вложение (инъективный функтор) $\ulcorner - \urcorner : C \hookrightarrow \mathbf{D}C$, сопоставляющее произвольному C-объекту Q точку — диаграмму $\ulcorner Q \urcorner : \mathbf{1} \to C : 0 \mapsto Q$. Коконусом (cocone) называется $\mathbf{D}C$-морфизм, имеющий точку в качестве кообласти. Можно изобразить коконус $\langle \sigma, !_X \rangle : \Delta \to \ulcorner Q \urcorner$ над диаграммой $\Delta : X \to C$ в виде диаграммы, «пририсовав» к Δ дополнительную вершину, помеченную объектом Q, и набор ребер — стрелок, по одной для каждого узла $I \in \mathrm{Ob}\, X$, направленной из I в вершину и помеченной морфизмом $\sigma_I : \Delta(I) \to Q$. Копределом (colimit) диаграммы Δ называется коконус $\mathrm{colim}\, \Delta : \Delta \to \ulcorner R \urcorner$, универсальный в том смысле, что для любых C-объекта T и коконуса $\delta : \Delta \to \ulcorner T \urcorner$ существует единственный C-морфизм $w : R \to T$ такой, что $\delta = \ulcorner w \urcorner \circ \mathrm{colim}\, \Delta$. Легко видеть, что это условие универсальности представляет собой в точности структурный критерий из разд. 3. Таким образом, конструирование копредела конфигурации Δ описывает на строгом математическом языке сборку системы, которой отвечает вершина R. В категориях типа **MBS** и **Pomset** построение копредела сводится к факторизации раздельных объединений объектов, представляющих компоненты системы, по отношениям эквивалентности, индуцированным моделями клея и других средств сборки.

Копредел любой диаграммы, если он существует, определяется однозначно с точностью до изоморфизма. Более того, можно описать сборку систем из конфигураций в виде функтора. Пусть Cd — некоторый класс C-диаграмм, имеющих копределы. Он порождает полную подкатегорию в $\mathbf{D}C$, из которой в C действует функтор копредела colim, сопоставляя каждой диаграмме из Cd вершину некоторого ее копредела, а каждому $\mathbf{D}C$-морфизму $\theta : \Delta \to \Xi$, где $\Delta, \Xi \in Cd$ — стрелку копредела $\mathrm{colim}\,(\theta)$ такую, что $\mathrm{colim}\, \Xi \circ \theta = \ulcorner \mathrm{colim}\,(\theta) \urcorner \circ \mathrm{colim}\, \Delta$.



Например, в категории **Set** любая диаграмма имеет копредел [14, упражнение 5.1.8], поэтому имеется функтор $\mathrm{colim} : \mathbf{D}(\mathbf{Set}) \to \mathbf{Set}$. Примечательно, что этот функтор является рефлектором: он сопряжен слева с вложением $\ulcorner - \urcorner : \mathbf{Set} \hookrightarrow \mathbf{D}(\mathbf{Set})$, причем единица рефлексии состоит из $\mathbf{D}(\mathbf{Set})$-морфизмов $\mathrm{colim}\, \Delta : \Delta \to \ulcorner \mathrm{colim}\,(\Delta) \urcorner$, $\Delta \in \mathrm{Ob}\, \mathbf{D}(\mathbf{Set})$. Напомним, что единица рефлексии — это естественное преобразование тождественного функтора в композицию рефлектора и вложения (в данном случае, естественное преобразование функтора $1_{\mathbf{D}(\mathbf{Set})}$ в $\ulcorner \mathrm{colim}\,(-) \urcorner$), состоящее из универсальных стрелок [14, разд. 4.3]. И для произвольного класса Cd, содержащего достаточное количество одноточечных диаграмм, функтор colim сопряжен слева с ограничением вложения $\ulcorner - \urcorner$ на подходящую полную подкатегорию в C. А поскольку сопряженный функтор задается однозначно с точностью до изоморфизма [14, разд. 4.1], можно сделать вывод, что сборка систем в некотором смысле «зашифрована» в процедуре построения одноточечных диаграмм — моделей систем как целого без раскрытия структуры.

Так наглядно проявляется двойственность прямых и обратных задач мегамоделирования.

5 Заключение

Аппарат теории категорий обладает большим потенциалом в области повышения полезной отдачи от MBSE, в том числе путем математически строгого решения задач мегамоделирования. Так, базовая процедура системной инженерии — сборка системы из заданной конфигурации взаимосвязанных компонентов — формально описывается теоретико-категорной конструкцией копредела диаграммы. Более сложные конструкции отвечают сложным процедурам сборки, таким как связывание (weaving) общесистемных функций, рассеянных по всем компонентам (crosscutting concerns), например мониторинговых или защитных [16]. Математического представления требуют и другие процедуры MBSE, в частности коллективная модификация мегамоделей и составляющих моделей, восстановление конфигурации заданной системы, оценка взаимозаменяемости компонентов.

Актуальны вопросы внедрения аппарата теории категорий в практику, в том числе путем развития программных инструментов моделирования и мегамоделирования. Здесь открывается широкий спектр направлений для дальнейших исследований.

Литература

1. Modeling and simulation-based systems engineering handbook / Eds. D. Gianni, A. D'Ambrogio, A. Tolk. — London: CRC Press, 2014. 513 p.
2. *Ковалёв С. П., Толок А. В.* Применение модельно-ориентированного подхода в управлении жизненным циклом технических изделий // Информационные технологии в проектировании и производстве, 2015. № 2. С. 3–9.
3. *Левенчук А. И.* Системноинженерное мышление. — М.: ТехИнвестЛаб, 2015. 305 с.
4. IEC 81346-1:2009. Industrial Systems, Installations and Equipment and Industrial Products — Structuring Principles and Reference Designations — Part 1: Basic Rules. — Geneva: ISO, 2009. 168 p.
5. *Ginali S., Goguen J.* A categorical approach to general systems // Conference (International) on Applied General Systems Research Proceedings / Ed. G. J. Klir. — NATO conference series. — New York, NY, USA: Plenum Press, 1978. Vol. 5. P. 257–270.
6. *Mabrok M. A., Ryan M. J.* Category theory as a formal mathematical foundation for model-based systems engineering // Appl. Math. Inform. Sci., 2017. Vol. 11. No. 1. P. 43–51.
7. *Ковалёв С. П.* Теоретико-категорный подход к проектированию программных систем // Фундаментальная и прикладная математика, 2014. Т. 19. Вып. 3. С. 111–170.
8. *Bézivin J., Jouault F., Rosenthal P., Valduriez P.* Modeling in the large and modeling in the small // Model Driven Architecture: European MDA Workshops on Foundations and Applications Proceedings / Eds. U. Aßmann, M. Aksit, A. Rensink. — Lecture notes in computer science ser. — Springer, 2005. Vol. 3599. P. 33–46.
9. *Diskin Z., Kokaly S., Maibaum T.* Mapping-aware megamodeling: Design patterns and laws // Software Language Engineering: 6th Conference (International) Proceedings / Eds. M. Erwig, R. F. Paige, E. Van Wyk. — Lecture notes in computer science ser. — Springer, 2013. Vol. 8225. P. 322–343.
10. *Requicha A. G.* Representations for rigid solids: Theory, methods, and systems // ACM Comput. Surv., 1980. Vol. 12. Iss. 4. P. 437–464.
11. *Kádár B., Pfeiffer A., Monostori L.* Discrete event simulation for supporting production planning and scheduling decisions in digital factories // 37th CIRP Seminar (International) on Manufacturing Systems Proceedings. — Budapest, 2004. P. 444–448.
12. *Giesa T., Spivak D. I., Buehler M. J.* Category theory based solution for the building block replacement problem in materials design // Adv. Eng. Mater., 2012. Vol. 14. Iss. 9. P. 810–817.
13. *Косяков А., Свит У., Сеймур С., Бимер С.* Системная инженерия. Принципы и практика / Пер. с англ. — М.: ДМК-Пресс, 2014. 636 с. (*Kossiakoff A., Sweet W. N., Seymour S., Biemer S. M.* Systems engineering principles and practice. — 2nd ed. — New York, NY, USA: John Wiley, 2011. 560 p.)
14. *Маклейн С.* Категории для работающего математика / Пер. с англ. — М.: Физматлит, 2004. 352 с. (*Mac Lane S.* Categories for the working mathematician. — New York, NY, USA: Springer, 1978. 317 p.)
15. *Pratt V. R.* Modeling concurrency with partial orders // Int. J. Parallel Prog., 1986. Vol. 15. No. 1. P. 33–71.
16. *Ковалёв С. П.* Семантика аспектно-ориентированного моделирования данных и процессов // Информатика и её применения, 2013. Т. 7. Вып. 3. С. 70–80.

Поступила в редакцию 16.01.17

METHODS OF CATEGORY THEORY IN MODEL-BASED SYSTEMS ENGINEERING

S. P. Kovalyov

Institute of Control Sciences, Russian Academy of Sciences, 65 Profsoyuznaya Str., Moscow 117997, Russian Federation

Abstract: A mathematical device based on the category theory is proposed to formally describe and rigorously explore procedures of employing models in engineering that constitute the contents of model-based systems engineering (MBSE). The essence of the device consists in mathematical representation of assembly drawings (megamodels of systems) as diagrams in categories whose objects are models, and morphisms represent actions associated with assembling system models from component models. The soundness of the device is justified on the basis of standards that govern description of the systems' structure such as IEC 81346. Category-theoretical methods for solving a number of practical problems of assembling systems are proposed and explored. Examples of solving such problems are provided in categories that represent two key application areas for MBSE: geometric modeling of complex shapes and discrete-event simulation of the behavior of industrial systems.

Keywords: model-based systems engineering; megamodel; category theory; colimit

DOI: 10.14357/19922264170305

References

1. Gianni, D., A. D'Ambrogio, and A. Tolk, eds. 2014. *Modeling and simulation-based systems engineering handbook*. London: CRC Press. 513 p.
2. Kovalyov, S. P., and A. V. Tolok. 2015. Primenenie model'no-orientirovannogo podkhoda v upravlenii zhiznennym tsiklom tekhnicheskikh izdeliy [Applying model-based approach to product lifecycle management]. *Informatsionnye tekhnologii v proektirovanii i proizvodstve* [Information Technologies in Design and Industry] 2(158):3–9.
3. Levenchuk A. I. 2015. *Sistemnoinzhenernoe myshlenie* [Systems engineering thinking]. Moscow: TechInvestLab. 305 p.
4. IEC 81346-1:2009. 2009. Industrial Systems, Installations and Equipment and Industrial Products — Structuring Principles and Reference Designations — Part 1: Basic Rules. Geneva: ISO. 168 p.
5. Ginali, S., and J. Goguen. 1978. A categorical approach to general systems. *Conference (International) on Applied General Systems Research Proceedings*. Ed. G. J. Klir. NATO conference ser. Plenum Press. 5:257–270.
6. Mabrok, M. A., and M. J. Ryan. 2017. Category theory as a formal mathematical foundation for model-based systems engineering. *Appl. Math. Inform. Sci.* 11(1):43–51.
7. Kovalyov, S. P. 2016. Category-theoretic approach to software systems design. *J. Math. Sci.* 214(6):814–853.
8. Bézivin, J., F. Jouault, P. Rosenthal, and P. Valduriez. 2005. Modeling in the large and modeling in the small. *Model Driven Architecture: European MDA Workshops on Foundations and Applications Proceedings*. Eds. U. Aßmann, M. Aksit, and A. Rensink. Lecture notes in computer science ser. Springer. 3599:33–46.
9. Diskin, Z., S. Kokaly, and T. Maibaum. 2013. Mapping-aware megamodeling: Design patterns and laws. *6th Conference (International) on Software Language Engineering Proceedings*. Eds. M. Erwig, R. F. Paige, and E. Van Wyk. Lecture notes in computer science ser. Springer. 8225:322–343.
10. Requicha, A. G. 1980. Representations for rigid solids: Theory, methods, and systems. *ACM Comput. Surv.* 12(4):437–464.
11. Kádár, B., A. Pfeiffer, and L. Monostori. 2004. Discrete event simulation for supporting production planning and scheduling decisions in digital factories. *37th CIRP Seminar (International) on Manufacturing Systems Proceedings*. Budapest. 444–448.
12. Giesa, T., D. I. Spivak, and M. J. Buehler. 2012. Category theory based solution for the building block replacement problem in materials design. *Adv. Eng. Mater.* 14(9):810–817.
13. Kossiakoff, A., W. N. Sweet, S. Seymour, and S. M. Biemer. 2011. *Systems engineering principles and practice*. 2nd ed. New York, NY: John Wiley. 560 p.
14. Mac Lane, S. 1978. *Categories for the working mathematician*. New York, NY: Springer. 317 p.
15. Pratt, V. R. 1986. Modeling concurrency with partial orders. *Int. J. Parallel Prog.* 15(1):33–71.
16. Kovalyov, S. P. 2013. Semantika aspektno-orientirovannogo modelirovaniya dannykh i protsessov [Semantics of aspect-oriented modeling of data and processes]. *Informatika i ee Primeneniya — Inform. Appl.* 7(3):70–80.

Received January 16, 2017

Contributor

Kovalyov Sergey P. (b. 1972) — Doctor of Science in physics and mathematics, leading scientist, Institute of Control Problems, Russian Academy of Sciences, 65 Profsoyuznaya Str., Moscow 117997, Russian Federation Federation; kovalyov@nm.ru

ОБ ЭФФЕКТИВНОСТИ ИЕРАРХИЧЕСКОГО АЛГОРИТМА ПОИСКА ПРИБЛИЖЕННОГО БЛИЖАЙШЕГО СОСЕДА В ЗАДАННОМ НАБОРЕ ИЗОБРАЖЕНИЙ*

М. М. Ланге[1], С. Н. Ганебных[2], А. М. Ланге[3]

Аннотация: Исследуется эффективность иерархического алгоритма поиска в заданном наборе изображений близкого представителя к предъявляемому изображению с негарантированной погрешностью относительно ближайшего соседа. Алгоритм использует пространство квадропирамидальных представлений изображений и стратегию направленного поиска на последовательных уровнях представления с нарастающим разрешением. Эффективность алгоритма исследуется в терминах эмпирического распределения погрешностей поиска и вычислительной сложности относительно сложности полного перебора. Приводятся эмпирические распределения погрешностей и оценки вычислительной сложности алгоритма для двух приложений: поиска в наборе изображений рукописных цифр из базы данных MNIST и координатной привязки зашумленных изображений к аэрокосмической карте местности из сетевого сервиса Google Maps.

Ключевые слова: изображение; квадропирамидальное представление; цифровая карта; ближайший сосед; приближенный ближайший сосед; погрешность поиска; эмпирическое распределение; вычислительная сложность

DOI: 10.14357/19922264170306

1 Введение

Существует широкий класс прикладных задач, связанных с поиском в большом наборе данных объекта, близкого по заданной мере к предъявляемому объекту. К таким задачам относится поиск в наборах изображений образцов, сходных с предъявляемыми изображениями, координатная привязка наблюдаемого изображения к цифровой карте местности и др. Перечисленные примеры относятся к проблеме извлечения изображений (Image Retrieval) [1]. При больших объемах данных решающий алгоритм должен удовлетворять заданным требованиям к допустимой погрешности (точности) поиска и вычислительной сложности (быстродействию). Как правило, эти характеристики находятся в обратной зависимости: с увеличением быстродействия уменьшается точность (растет погрешность) и наоборот. Поэтому необходимо обеспечить баланс этих требований путем варьирования параметрами решающего алгоритма.

В качестве решающих алгоритмов могут быть использованы алгоритмы поиска приближенного ближайшего соседа [2–6] или их модификации. При фиксированной размерности $d \geq 1$ векторного пространства известные алгоритмы с гарантированной точностью, задаваемой допустимой погрешностью $\varepsilon > 0$, реализуют поиск в наборе из векторов представителя на расстоянии $D \leq (1+\varepsilon)D_{\min}$ от предъявляемого вектора, где $D_{\min} > 0$ — расстояние до ближайшего соседа. Такие алгоритмы используют древовидные структуры данных, которые при фиксированных d и ε позволяют уменьшить порядок роста вычислительной сложности по n по сравнению со сложностью переборного алгоритма. В частности, BBD-алгоритм, использующий решающее Balance Box Decision дерево [5], имеет сложность $O(d\lceil 1 + 6d/\varepsilon\rceil^d \log n)$, а сложность LSH-алгоритма на основе локального хеширования Locality Sensitive Hashing [6] составляет $O(dn^{1/(1+\varepsilon)})$. Для сравнения переборный алгоритм поиска ближайшего соседа ($\varepsilon = 0$) имеет сложность $\Theta(dn)$. Характер зависимости вычислительной сложности указанных алгоритмов от размерности d и допустимой погрешности ограничивает их применение для поиска изображений размера $N \times N$ из-за чрезмерно высокой размерности $d = N^2$ при $N \geq 100$.

В настоящей работе рассматривается альтернативный алгоритм для быстрого поиска в заданном наборе изображений мощности n приближенного

*Работа выполнена при поддержке РФФИ (проекты 15-07-07516 и 15-07-09324).

[1]Федеральный исследовательский центр «Информатика и управление» Российской академии наук, lange_mm@ccas.ru

[2]Федеральный исследовательский центр «Информатика и управление» Российской академии наук, sng@ccas.ru

[3]Федеральный исследовательский центр «Информатика и управление» Российской академии наук, lange_am@mail.ru

ближайшего соседа к предъявляемому изображению. Предлагаемый алгоритм является модификацией алгоритма, рассмотренного в [7], которая использует квадропирамидальное представление изображений с многоуровневым разрешением [8–10]. Алгоритм базируется на параметрической стратегии экспоненциального сужения зоны поиска, аналогичной использованной в работе [11]. Такая стратегия поиска обеспечивает вычислительную сложность $O(n \log N)$, но не гарантирует точности приближенного решения.

Эффективность предложенного алгоритма исследована в терминах эмпирического распределения погрешностей поиска и вычислительного выигрыша относительно алгоритма полного перебора. При различных значениях параметра алгоритма указанные характеристики получены для поиска изображений рукописных цифр из базы данных MNIST [12] и для координатной привязки зашумленных изображений к цифровой карте участка земной поверхности, взятой на интернет-сервисе Google Maps [13].

2 Формальная постановка задачи

Рассматривается множество изображений $\mathbf{X}$ размера $N \times N$, элементы которых принадлежат алфавиту $A = \{0, 1, \ldots, q-1\}$ $(q \geq 2)$. Предполагается, что каждое изображение $\mathbf{x} \in \mathbf{X}$ имеет по крайней мере один ненулевой элемент, а размер изображения является целочисленной степенью числа 2, так что $N = 2^L$, где $L \gg 1$. В общем случае множество $\mathbf{X}$ содержит набор изображений $\hat{\mathbf{X}}$, в котором необходимо найти изображение $\hat{\mathbf{x}} \in \hat{\mathbf{X}}$, ближайшее или достаточно близкое по заданной мере к заданному изображению $\mathbf{x} \in \mathbf{X}$.

Любое изображение $\mathbf{x} \in \mathbf{X}$ с указанными размерами допускает квадропирамидальное представление

$$\mathbf{x}_L = (x_0, \ldots, x_l, \ldots, x_L) \qquad (1)$$

порядка L, которое содержит последовательность описаний изображения $\mathbf{x}$ с уровнями разрешения $l = 0, \ldots, L$ [8]. Пример квадропирамиды порядка $L = 2$ дан на рис. 1.

Описание l-го уровня x_l в (1) является изображением размера $2^l \cdot 2^l$, получаемым из описания x_{l+1} путем усреднений по непересекающимся группам из четырех смежных элементов. Ненулевое значение элемента в вершине пирамиды x_0 позволяет сформировать нормализованное квадропредставление

$$\mathbf{y}_L = (y_1, \ldots, y_l, \ldots, y_L) \qquad (2)$$

путем деления всех элементов пирамиды (1) на значение элемента вершины x_0. Нормализация элементов уменьшает зависимость представления (2) от средней яркости изображения по сравнению с представлением (1).

Пусть $z(k_{l1}, k_{l2}) > 0$ — значение элемента с индексами $(k_{l1}, k_{l2}) = 1, \ldots, 2^l$ в нормализованном описании $y_l \in \mathbf{y}_L$. Для любой пары изображений $\mathbf{x} \in \mathbf{X}$, $\hat{\mathbf{x}} \in \mathbf{X}$, имеющих нормализованные описания $\mathbf{y}_L, \hat{\mathbf{y}}_L$, вводится мера их различия l-го порядка:

$$D_l(\mathbf{x}, \hat{\mathbf{x}}) = \frac{1}{(2^l \cdot 2^l)} \sum_{k_{l1}=1}^{2^l} \sum_{k_{l2}=1}^{2^l} |z(k_{l1}, k_{l2}) - \hat{z}(k_{l1}, k_{l2})|\,, \quad l = 1, \ldots, L\,, \qquad (3)$$

где $z(k_{l1}, k_{l2})$ и $\hat{z}(k_{l1}, k_{l2})$ — элементы описаний $y_l \in \mathbf{y}_L$ и $\hat{y}_l \in \hat{\mathbf{y}}_L$ соответственно для изображений $\mathbf{x}$ и $\hat{\mathbf{x}}$. Суммирование мер $D_t(\mathbf{x}, \hat{\mathbf{x}})$ вида (3) порядка $t = 1, \ldots, l$ $(l \leq L)$ с весами $w_t = (1/2)\log_2(2^t \cdot 2^t) = t$ порождает взвешенную меру различия порядка l:

$$\tilde{D}_l(\mathbf{x}, \hat{\mathbf{x}}) = \sum_{t=1}^{l} w_t D_t(\mathbf{x}, \hat{\mathbf{x}})\,, \quad 1 \leq l \leq L\,. \qquad (4)$$

Точный или приближенный поиск ближайшего соседа для предъявляемого изображения $\mathbf{x} \in \mathbf{X}$ выполняется на подмножестве $\hat{\mathbf{X}} \subset \mathbf{X}$, содержащем n изображений. Алгоритм поиска принимает решение $\hat{\mathbf{x}}^* \in \hat{\mathbf{X}}$ по мере (4) наибольшего порядка L на наборе изображений $\hat{\mathbf{X}}^* \subseteq \hat{\mathbf{X}}$ мощности $n^* \leq n$ в соответствии с решающим правилом

$$\hat{\mathbf{x}}^* = \arg \min_{\hat{\mathbf{x}} \in \hat{\mathbf{X}}^*} \tilde{D}_L(\mathbf{x}, \hat{\mathbf{x}})\,. \qquad (5)$$

Набор $\hat{\mathbf{X}}^*$ в (5) определяется используемой стратегией направленного (иерархического) поиска, которая в случае $\hat{\mathbf{X}}^* = \hat{\mathbf{X}}$ обеспечивает точное решение $\hat{\mathbf{x}}^*_{\mathrm{NN}}$ (Nearest Neighbor), совпадающее

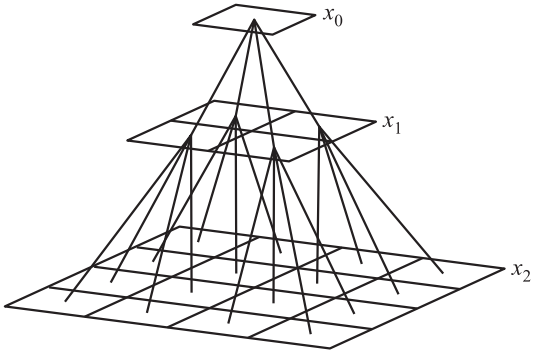


Рис. 1 Квадропирамида $\mathbf{x}_2 = (x_0, x_1, x_2)$ порядка $L = 2$

с ближайшим соседом, а в случае $\hat{\mathbf{X}}^* \subset \hat{\mathbf{X}}$ — приближенное решение $\hat{\mathbf{x}}^*_{\mathrm{AN}}$ (Approximate Nearest), которое может отличаться от ближайшего соседа или совпадать с ним.

Погрешность поиска по правилу (5) приближенного ближайшего изображения в заданном наборе определяется различием значений $\tilde{D}_L(\mathbf{x}, \hat{\mathbf{x}}^*_{\mathrm{AN}})$ и $\tilde{D}_L(\mathbf{x}, \hat{\mathbf{x}}^*_{\mathrm{NN}})$ и в случае $\tilde{D}_L(\mathbf{x}, \hat{\mathbf{x}}^*_{\mathrm{NN}}) > 0$ равна:

$$\varepsilon^s_{n^*}\left(\hat{\mathbf{x}}^*_{\mathrm{AN}}, \hat{\mathbf{x}}^*_{\mathrm{NN}}\right) = \frac{\tilde{D}_L(\mathbf{x}, \hat{\mathbf{x}}^*_{\mathrm{AN}}) - \tilde{D}_L(\mathbf{x}, \hat{\mathbf{x}}^*_{\mathrm{NN}})}{\tilde{D}_L(\mathbf{x}, \hat{\mathbf{x}}^*_{\mathrm{NN}})}. \quad (6)$$

В задаче координатной привязки погрешность определяется отклонением координат (a^*, b^*) найденного по правилу (5) изображения $\hat{\mathbf{x}}^*_{\mathrm{AN}}$ от координат (a, b) предъявляемого изображения $\mathbf{x}$ и в случае изображений размера $N \times N$ равна:

$$\varepsilon^g_{n^*}\left(\hat{\mathbf{x}}^*_{\mathrm{AN}}, \mathbf{x}\right) = \frac{1}{2}\left(\frac{|a - a^*|}{N} + \frac{|b - b^*|}{N}\right). \quad (7)$$

Качество рассматриваемого алгоритма исследуется в терминах распределений

$$\Pr\left\{\varepsilon^s_{n^*}\left(\hat{\mathbf{x}}^*_{\mathrm{AN}}, \hat{\mathbf{x}}^*_{\mathrm{NN}}\right) \le \varepsilon\right\}; \quad (8)$$

$$\Pr\left\{\varepsilon^g_{n^*}\left(\hat{\mathbf{x}}^*_{\mathrm{AN}}, \hat{\mathbf{x}}\right) \le \varepsilon\right\} \quad (9)$$

погрешностей (6) и (7) с параметром n^*: $1 \le n^* \le n$, где $\varepsilon \ge 0$ — допустимая погрешность, принимающая значения с некоторым шагом. При фиксированном значении ε вероятности (8) и (9) соответствуют надежности выполнения точности поиска и привязки.

В разд. 3 дается описание алгоритма поиска и приводятся оценки его вычислительной сложности при больших значениях n и N, связанных соотношением $N^2 \ge \log_q n$. В разд. 4 приводятся параметрическое семейство эмпирических распределений (8) с параметром n^* и численные оценки сложности алгоритма, полученные на множестве изображений рукописных цифр. Аналогичные параметрические семейства эмпирических распределений (9) и оценки вычислительной сложности, полученные для координатной привязки зашумленных изображений к карте участка земной поверхности, приводятся в разд. 5.

3 Алгоритм поиска

Предполагается, что изображения из набора $\hat{\mathbf{X}}$, на котором производится поиск, заданы нормализованными пирамидальными представлениями вида (2) и образуют многоуровневую сеть

$$\hat{\mathbf{Y}}_1, \ldots, \hat{\mathbf{Y}}_l, \ldots, \hat{\mathbf{Y}}_L, \quad (10)$$

в которой $\hat{\mathbf{Y}}_l$ — подмножество представлений всех изображений из $\hat{\mathbf{X}}$, заданных l уровнями нормализованных пирамид. Алгоритм поиска решения (5) использует стратегию последовательного сужения зоны поиска на уровнях $l = 1, \ldots, L$ сети (10). Согласно этой стратегии число анализируемых изображений на l-м уровне определяется экспоненциальной функцией

$$n_l = \left\lfloor n \cdot 4^{-\alpha(l-1)} \right\rfloor, \quad l = 1, \ldots, L, \quad (11)$$

с коэффициентом $\alpha = (L-1)^{-1} \log_4(n/n^*)$, где $n^* = 1, 2, \ldots, n$ — мощность набора $\hat{\mathbf{X}}^* \subseteq \hat{\mathbf{X}}$, на котором принимается решение (5) по представлениям $\hat{\mathbf{Y}}_L$ последнего уровня сети (10).

Алгоритм поиска. Для предъявляемого изображения $\mathbf{x} \in \mathbf{X}$ на последовательных уровнях $l = 1, \ldots, L-1$ сети (10) вычисляются значения меры различия $\tilde{D}_l(\mathbf{x}, \hat{\mathbf{x}})$ вида (4) для n_l изображений из набора $\hat{\mathbf{X}}$ и среди них отбираются n_{l+1} изображений с наименьшими значениями $\tilde{D}_l(\mathbf{x}, \hat{\mathbf{x}})$; на уровне $l = L$ среди $n_L = n^*$ изображений отбирается ближайшее с наименьшим значением $\tilde{D}_L(\mathbf{x}, \hat{\mathbf{x}})$, которое дает решение (5).

В случае $1 \le n^* < n$ параметр $\alpha > 0$ в (11) обеспечивает экспоненциальное сужение зоны иерархического поиска, которое приводит к нахождению приближенного ближайшего соседа на наборе изображений $\hat{\mathbf{X}}^* \subset \hat{\mathbf{X}}$ мощности n^*; в случае $n^* = n$ параметр $\alpha = 0$ приводит к переборному поиску ближайшего соседа на наборе $\hat{\mathbf{X}}^* \equiv \hat{\mathbf{X}}$ мощности n. В обоих случаях вычисление меры различия изображений производится с использованием рекурсии

$$\tilde{D}_l\left(\mathbf{x}, \hat{\mathbf{x}}\right) = \tilde{D}_{l-1}\left(\mathbf{x}, \hat{\mathbf{x}}\right) + w_l D_l\left(\mathbf{x}, \hat{\mathbf{x}}\right), \quad l = 1, \ldots, L, \quad (12)$$

при начальном условии $\tilde{D}_0(\mathbf{x}, \hat{\mathbf{x}}) = 0$.

Вычислительная сложность алгоритма определяется числом элементарных операций, затрачиваемых на вычисление меры на всех уровнях сети (10), и на сортировку значений меры на последовательных уровнях для отбора ближайших изображений согласно (11), включая отбор решения на последнем уровне. В работе [7] дана асимптотическая оценка вычислительной сложности сформулированного иерархического алгоритма поиска при больших значениях мощности n набора изображений $\hat{\mathbf{X}}$, размере изображения N, удовлетворяющем условию $N^2 \ge \log_q n$ (q — размер алфавита), и соотношении $n/n^* \ge N^2/4$, обеспечивающем коэффициент сужения зоны поиска $\alpha \ge 1$. Асимптотика

сложности иерархического алгоритма с указанными параметрами имеет вид $C_{n^*\le 4n/N^2} = O(n \log N)$, вычислительная сложность переборного алгоритма — $C_{n^*=n} = \Omega(nN^2)$. Из приведенных оценок следует, что доля сложности иерархического поиска приближенного решения относительно сложности переборного поиска ближайшего соседа убывает с увеличением размера изображения как $O(N^{-2} \log N)$.

Численные оценки сложности алгоритма получены для суммарного количества элементарных операций, затрачиваемых на вычисление меры и сортировку вставками значений меры со сложностью $m \log_2 m$ на наборе из m элементов [14]. Поэтому при фиксированных n и $N = 2^L$ вычислительная сложность алгоритма равна

$$C_{n^*} = C_{n^*}^{\mathrm{msr}} + C_{n^*}^{\mathrm{srt}}, \qquad (13)$$

где

$$C_{n^*}^{\mathrm{msr}} = \sum_{l=1}^{L} n_l \cdot 4^l \le n \sum_{l=1}^{L} 4^l \left(\frac{n}{n^*}\right)^{-(l-1)/(L-1)}; \qquad (14)$$

$$C_n^{\mathrm{srt}} = (n-1)\,[n^* = n] + \left((n^*-1) + \sum_{l=1}^{L-1} n_l \log_2 n_l\right)[n^* < n] \le$$

$$\le (n-1)\,[n^* = n] + n \log_2 n \left(\sum_{l=1}^{L-1}\left(\frac{n}{n^*}\right)^{-(l-1)/(L-1)} - \right.$$

$$\left. - \sum_{l=1}^{L-1} \frac{l-1}{L-1}\left(\frac{n}{n^*}\right)^{-(l-1)/(L-1)}\right)[n^* < n] +$$

$$+ (n^*-1)\left(a + \right.$$

$$\left. + n \log_2 e \sum_{l=1}^{L-1} \frac{l-1}{L-1}\left(\frac{n}{n^*}\right)^{-(l-1)/(L-1)}\right)[n^* < n] \qquad (15)$$

соответственно затраты на вычисление меры и на сортировку; $[f]$ — индикатор f. В случае $n^* = n$ формулы (13)–(15) дают оценки вычислительной сложности переборного поиска точного решения, а в случае $n^* < n$ — оценки сложности иерархического поиска приближенного решения.

Формулы (12)–(14) использованы для получения численных оценок относительной сложности $C_{n^*\le n}/C_{n^*=n}$ при значениях $\lfloor n^* = n/2^k \rceil$, $k = 0, 1, \ldots$, и параметрах n и $N = 2^L$, с которыми проведены эксперименты по поиску рукописных цифр (см. разд. 4) и по координатной привязке изображений к карте местности (см. разд. 5). Величина, обратная относительной сложности, соответствует вычислительному выигрышу иерархического алгоритма по сравнению с алгоритмом перебора.

4 Оценки качества и сложности поиска рукописных цифр

Экспериментальные характеристики качества алгоритма поиска получены на наборе полутоновых изображений рукописных цифр из базы данных MNIST [12]. Вычислительный эксперимент выполнен с помощью кода, написанного на языке MATLAB [15] . Примеры изображений рукописных цифр даны на рис. 2.

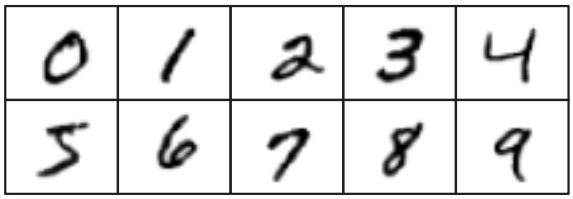


Рис. 2 Примеры рукописных цифр из базы данных MNIST

Цифры на изображениях нормированы по размеру и центрированы в поле изображения. Параметры изображений: $N = 32$; $q = 256$; число уровней представления изображений $L = \log_2 N = 5$; мощность набора данных $\hat{\mathbf{X}}$ равна $n = 60\,000$. Параметры N, q и n удовлетворяют необходимому условию $N^2 > \log_q n$. Набору $\hat{\mathbf{X}}$ предъявлялось 10 000 изображений, не входящих в набор данных $\hat{\mathbf{X}}$, так что $\tilde{D}_L(\mathbf{x}, \hat{\mathbf{x}}^*_{\mathrm{AN}}) \ge \tilde{D}_L(\mathbf{x}, \hat{\mathbf{x}}^*_{\mathrm{NN}}) > 0$. Для каждого $\mathbf{x}$ и соответствующей пары $\hat{\mathbf{x}}^*_{\mathrm{AN}}, \hat{\mathbf{x}}^*_{\mathrm{NN}}$ вычислялась погрешность поиска вида (6) и на 10 000 предъявляемых изображений строилось семейство эмпирических распределений вида (8) с параметром $\lfloor n^* = n/2^k \rceil$, $k = 0, 5, \ldots, 10$, при значениях допустимой погрешности $0 \le \varepsilon \le 0{,}1$. Графики семейства распределений даны на рис. 3.

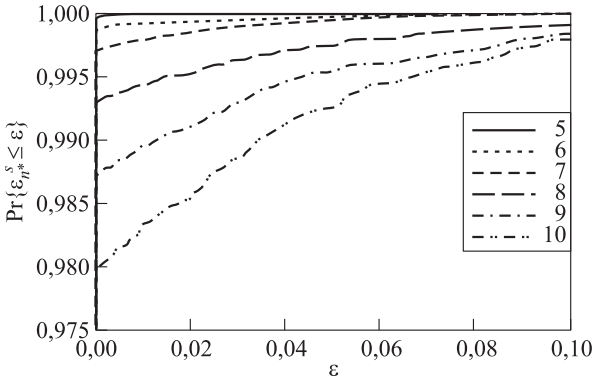


Рис. 3 Эмпирические распределения погрешностей поиска рукописных цифр на наборе изображений с параметрами $N = 32$, $q = 256$, $n = 60\,000$, $k = 5$, 6, 7, 8, 9, 10

Таблица 1 Оценки относительной сложности иерархического алгоритма поиска изображений рукописных цифр ($N = 32$, $q = 256$, $n = 60\,000$)

k	$C^{\mathrm{msr}}_{n^*\le n}/C_{n^*=n}$	$C^{\mathrm{srt}}_{n^*\le n}/C_{n^*=n}$	$C_{n^*\le n}/C_{n^*=n}$
0	0,9993	0,0007	1,0000
1	0,5325	0,0356	0,5681
2	0,2885	0,0285	0,3170
3	0,1597	0,0237	0,1834
4	0,0908	0,0205	0,1113
5	0,0535	0,0182	0,0717
6	0,0329	0,0166	0,0496
7	0,0214	0,0154	0,0368
8	0,0146	0,0146	0,0292
9	0,0107	0,0139	0,0246
10	0,0082	0,0134	0,0216

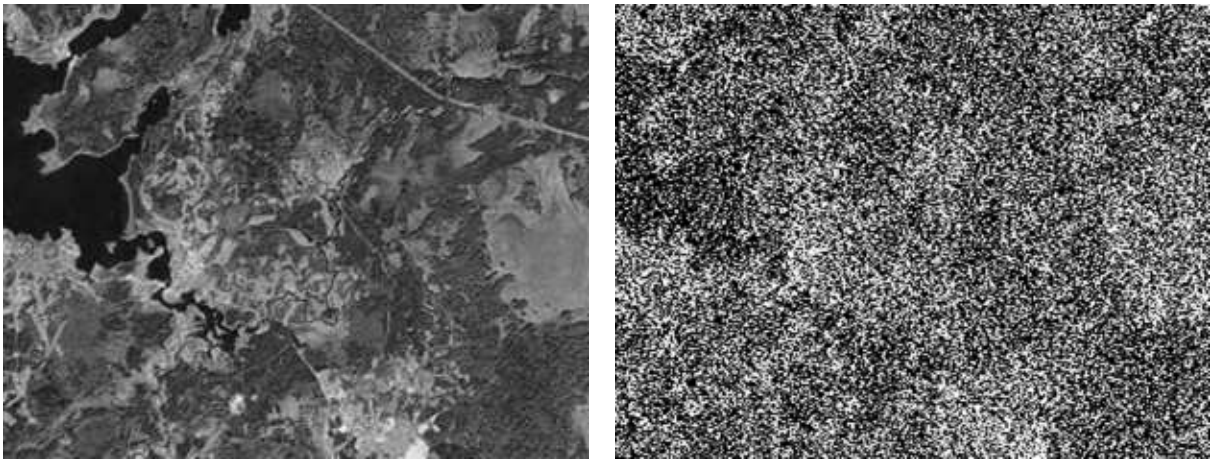
Численные оценки сложности алгоритма поиска представлены значениями $C_{n^*\le n}/C_{n^*=n}$, вычисленными в точках $\lfloor n^* = n/2^k \rceil$, $k = 0, 1, \ldots, 10$, при $N = 32$, $n = 60\,000$. Полученные оценки сложности алгоритма поиска даны в табл. 1.

Из приведенных распределений и оценок сложности следует, что при значениях $8 \le k \le 10$ иерархический алгоритм реализует нулевую погрешность поиска ближайшего соседа ($\varepsilon^s_{n^*} = 0$) с вероятностью 0,980–0,997 и обеспечивает вычислительный выигрыш в 34–46 раз по сравнению с переборным алгоритмом. С ростом допустимой погрешности $\varepsilon > 0$ вероятность нахождения ближайшего соседа с точностью $\varepsilon^s_{n^*} \le \varepsilon$ увеличивается при тех же значениях вычислительного выигрыша.

5 Оценки качества и сложности координатной привязки изображений

Экспериментальные оценки эффективности алгоритма поиска для координатной привязки к карте местности получены на цифровом аэрокосмическом снимке участка земной поверхности [13]. Обработка данных выполнена программным кодом [15].

Размеры аэрокосмического снимка 300×236 (в пикселях), число уровней яркости $q = 256$. Использованный снимок и его зашумленная версия даны на рис. 4.



(*а*) (*б*)

Рис. 4 Аэрокосмический снимок участка земной поверхности: (*а*) без шума; (*б*) с гауссовым шумом с нулевым средним и дисперсией $\sigma^2 = 0{,}25$

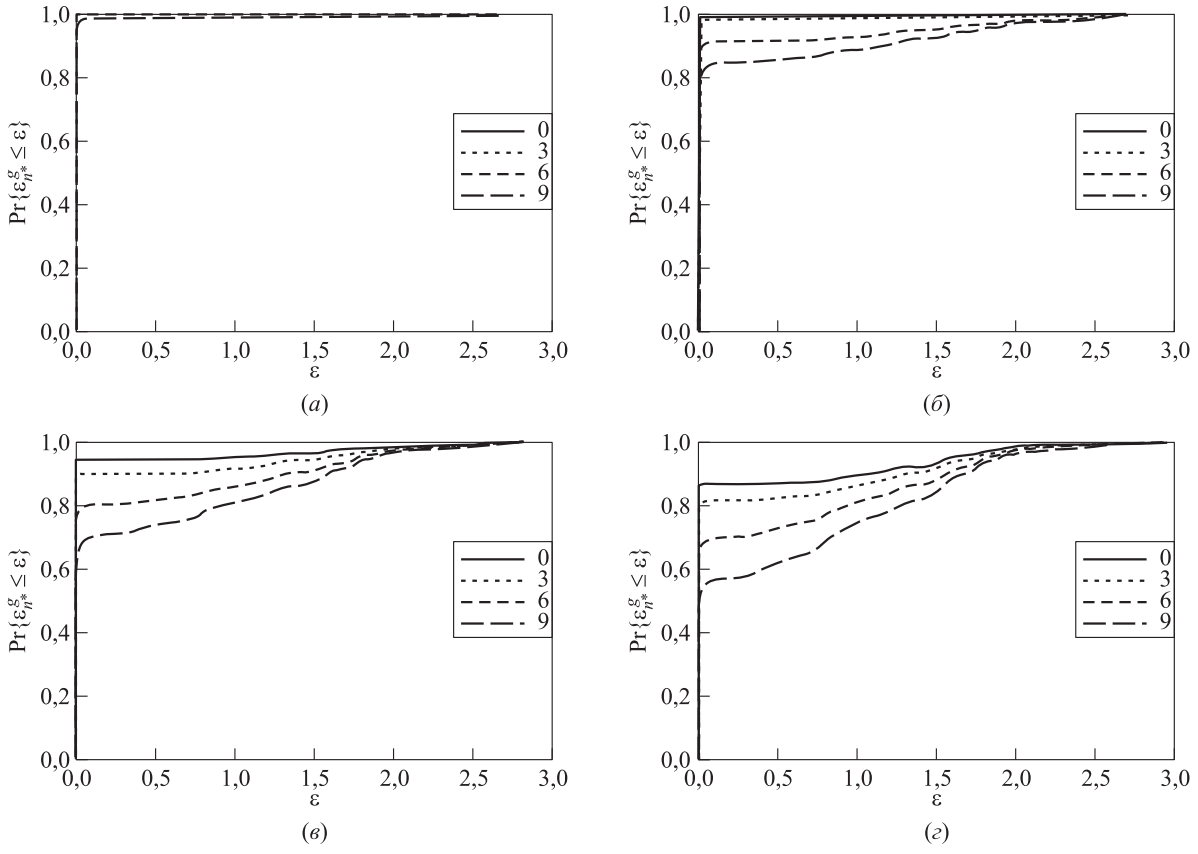


Рис. 5 Эмпирические распределения погрешностей координатной привязки при различных дисперсиях гауссова шума ((*а*) 0,1; (*б*) 0,15; (*в*) 0,2; (*г*) 0,25) и параметрах $N = 64$, $q = 256$, $n = 10\,353$, $k = 0, 3, 6, 9$

Привязка в координатах снимка проводилась по изображению размера 64×64 ($N = 64$). Мощность набора данных $\hat{\mathbf{X}}$ определялась количеством всевозможных изображений указанного размера, взятых на снимке с шагом в два пикселя, и равна $n = 10\,353$. Параметры N, q и n удовлетворяют необходимому условию $N^2 > \log_q n$.

Набору $\hat{\mathbf{X}}$ предъявлялись поочередно все n изображений с аддитивным гауссовым шумом с дисперсией $0 \le \sigma^2 \le 0{,}25$. Для любого предъявляемого изображения $\mathbf{x} \in \hat{\mathbf{X}}$ с координатами (a, b) вычислялись координаты (a^*, b^*) найденного алгоритмом изображения $\hat{\mathbf{x}}^*_{\mathrm{AN}}$ и погрешность привязки вида (7). При различных значениях σ^2 строились семейства эмпирических распределений (9) с параметром n^* в диапазоне значений погрешности $0 \le \varepsilon \le 2{,}5$.

На рис. 5 даны семейства распределений, полученные для четырех значений дисперсии шума при $\lfloor n^* = n/2^k \rceil$, $k = 0, 3, 6, 9$.

В табл. 2 приведены численные оценки относительной сложности $C_{n^* \le n}/C_{n^*=n}$ при значениях $k = 0, 3, \ldots, 10$.

Экспериментально установлено, что при отсутствии шума ($\sigma^2 = 0$) иерархический алгоритм с параметром $n^* \ge 1$ с вероятностью единица дает нулевую погрешность привязки $\varepsilon^g_{n^*} = 0$ и, следовательно, по качеству эквивалентен алгоритму перебора. В случае шума с дисперсией $\sigma^2 \le 0{,}10$ погрешность $\varepsilon^g_{n^*} = 0$ реализуется практически с единичной вероятностью при более чем 100-кратном вычислительном выигрыше иерархического алгоритма относительно переборного алгоритма.

Семейства распределений, полученные при значениях $\sigma^2 = 0{,}15$, $0{,}20$ и $0{,}25$, демонстрируют динамику соотношения показателей качества привязки и быстродействия иерархического алгоритма. При любой фиксированной дисперсии шума σ^2 и заданной допустимой погрешности привязки ε вероятность реализации точности привязки $\varepsilon^g_{n^*} \le \varepsilon$ уменьшается с ростом вычислительного выигрыша алгоритма. При фиксированных значениях ε и n^* вероятность реализации точности $\varepsilon^g_{n^*} \le \varepsilon$ уменьшается с увеличением дисперсии шума.

Необходимо отметить, что представление цифровой карты местности набором всевозможных

Таблица 2 Оценки относительной сложности иерархического алгоритма привязки изображений к карте местности ($N = 64$, $q = 256$, $n = 10\,353$)

k	$C^{\mathrm{msr}}_{n^*\le n}/C_{n^*=n}$	$C^{\mathrm{srt}}_{n^*\le n}/C_{n^*=n}$	$C_{n^*\le n}/C_{n^*=n}$
0	0,9998	0,0002	1,0000
3	0,1504	0,0059	0,1563
4	0,0824	0,0050	0,0874
5	0,0462	0,0044	0,0506
6	0,0265	0,0039	0,0304
7	0,0158	0,0036	0,0194
8	0,0098	0,0034	0,0132
9	0,0063	0,0032	0,0095
10	0,0028	0,0034	0,0062

изображений, взятых с шагом в один пиксель, должно привести к увеличению вероятности требуемой точности при заданных значениях ε, n^* и σ^2.

6 Заключение

Исследована эффективность иерархического алгоритма поиска в заданном наборе изображений приближенного ближайшего соседа к заданному изображению в терминах эмпирического распределения значений погрешности и вычислительной сложности алгоритма. Рассматриваемый алгоритм предназначен для реализации поиска с заданным быстродействием и надежностью выполнения требования по точности.

Эффективность алгоритма продемонстрирована на двух источниках изображений: на наборе изображений рукописных цифр из базы MNIST и на наборе пересекающихся изображений аэрокосмической карты местности, взятой из интернет-сервиса Google Maps. При фиксированной вероятности реализации требуемой точности показана возможность «размена» точности и вычислительной сложности путем варьирования параметром алгоритма.

Полученные экспериментальные результаты показали достаточно высокую надежность требуемой точности поиска в наборе изображений рукописных цифр и требуемой точности координатной привязки зашумленного изображения к цифровой карте местности. Для уменьшения порядка роста вычислительной сложности поиска от мощности набора изображений планируется исследовать модификацию иерархического алгоритма с использованием структуры, которая объединяет представление с многоуровневым разрешением и решающее дерево.

Литература

1. *Datta R., Joshi D., Li J., Wang J.* Image retrieval: Ideas, influences, and trends of the new age // ACM Comput. Surv., 2008. Vol. 40. No. 2. P. 1–60.
2. *Friedman J., Bentley J., Finkel R.* An algorithm for finding best matches in logarithmic expected time // ACM T. Math. Software, 1977. Vol. 3. No. 3. P. 209–226.
3. *Cleary J.* Analysis of an algorithm for finding nearest neighbors in Euclidean space // ACM T. Math. Software, 1979. Vol. 5. No. 2. P. 183–192.
4. *Soleymani M., Morgera S.* An efficient nearest neighbor search method // IEEE T. Commun., 1987. Vol. 35. No. 6. P. 677–679.
5. *Arya S., Mount D., Netanyahu N., Silverman R., Wu A.* An optimal algorithm for approximate nearest neighbor searching in fixed dimensions // J. ACM, 1998. Vol. 45. No. 6. P. 891–923.
6. *Andoni A., Indyk P.* Near-optimal hashing algorithms for approximate nearest neighbor in high dimensions // Commun. ACM, 2008. Vol. 51. No. 1. P. 117–122.
7. *Lange M. M., Ganebnykh S. N., Lange A. M.* Algorithm of approximate search for the nearest digital array in a hierarchical data set // Machine Learning Data Analysis, 2016. Vol. 2. No. 1. P. 6–16.
8. *Rosenfeld A.* Quadtrees and pyramids for pattern recognition and image analysis // 5th Conference (International) on Pattern Recognition Proceedings, 1980. P. 802–811.
9. *Jackins C., Tanimoto S.* Quadtrees, octtrees, and K-trees: A generalized approach to recursive decomposition of Euclidean space // IEEE T. Pattern Anal., 1983. Vol. 5. No. 5. P. 533–539.
10. *Samet H.* The quadtree and related hierarchical data structures // Comput. Surv., 1984. Vol. 16. No. 2. P. 187–260.
11. *Ланге М. М., Новиков Н. А.* Представление данных с многоуровневым разрешением для быстрой координатной привязки изображений // Техническое зрение в системах управления: Сб. тр. науч.-технич. конф. — М.: ИКИ РАН, 2012. С. 242–249.

12. MNIST database. http://yann.lecun.com/exdb/mnist/index.html.
13. Network service Google Maps. http://www.maps.google.com.
14. *Cormen T., Leiserson C., Rivest R., Stein C.* Introduction to algorithms. — 3rd ed. — MIT Press, 2009. 1312 p.
15. Algorithm for searching approximate nearest neighbor. http://sourceforge.net/projects/edivis/files/.

Поступила в редакцию 13.12.16

ON EFFICIENCY OF THE HIERARCHICAL ALGORITHM FOR SEARCHING APPROXIMATE NEAREST NEIGHBOR IN A GIVEN SET OF IMAGES

M. M. Lange, S. N. Ganebnykh, and A. M. Lange

Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, 44-2 Vavilov Str., Moscow 119333, Russian Federation

Abstract: The efficiency of the hierarchical algorithm for searching approximate nearest neighbor in a given set of images subject to an unwarranted error about the nearest image is investigated. The algorithm uses a space of quad pyramidal image representations as well as a guided search strategy in successive representation levels of increasing resolution. The efficiency is studied in terms of both an empirical distribution of search errors and computational complexity of the hierarchical algorithm relative to the exhaustive search. The above characteristics are obtained for two applications, namely, search for approximate nearest image in a set of hand-written digits from the MNIST data base and gridding a given noisy image in an aerospace digital map from the Google maps network service.

Keywords: image; quad pyramidal representation; digital map; nearest neighbor; approximate nearest neighbor; search error; empirical distribution; computational complexity

DOI: 10.14357/19922264170306

Acknowledgments

The research was supported by the Russian Foundation for Basic Research (projects 15-07-07516 and 15-07-09324).

References

1. Datta, R., D. Joshi, J. Li, and J. Wang. 2008. Image retrieval: Ideas, influences, and trends of the new age. *ACM Comput. Surv.* 40(2):1–60.
2. Friedman, J., J. Bentley, and R. Finkel. 1977. An algorithm for finding best matches in logarithmic expected time. *ACM T. Math. Software* 3(3):209–226.
3. Cleary, J. Analysis of an algorithm for finding nearest neighbors in Euclidean space. 1979. *ACM T. Math. Software* 5(2):183–192.
4. Soleymani, M., and S. Morgera. 1987. An efficient nearest neighbor search method. *IEEE T. Commun.* 35(6):677–679.
5. Arya, S., D. Mount, N. Netanyahu, R. Silverman, and A. Wu. 1998. An optimal algorithm for approximate nearest neighbor searching in fixed dimensions. *J. ACM* 45(6):891–923.
6. Andoni, A., and P. Indyk. 2008. Near-optimal hashing algorithms for approximate nearest neighbor in high dimensions. *Commun. ACM* 51(1):117–122.
7. Lange, M. M., S. N. Ganebnykh, and A. M. Lange. 2016. Algorithm of approximate search for the nearest digital array in a hierarchical data set. *Machine Learning Data Analysis* 2(1):6–16.
8. Rosenfeld, A. 1980. Quadtrees and pyramids for pattern recognition and image analysis. *5th Conference (International) on Pattern Recognition Proceedings*. 802–811.
9. Jackins, C., and S. Tanimoto. 1983. Quadtrees, octtrees, and K-trees: A generalized approach to recursive decomposition of Euclidean space. *IEEE T. Pattern Anal.* 5(5):533–539.
10. Samet, H. 1984. The quadtree and related hierarchical data structures. *Comput. Surv.* 16(2):187–260.
11. Lange, M. M., and N. A. Novikov. 2012. Predstavlenie dannykh s mnogourovnevym razresheniem dlya bystroy koordinatnoy privyazki izobrazheniy [Multiresolution data representation for fast image gridding]. *Sb. tr. nauch.-tekhnich. konf. "Tekhnicheskoe zrenie v sistemakh upravleniya* [Scientific-Technical Conference "Computing

Vision in Control Systems" Proceedings]. Moscow: IKI RAN. 242–249.

12. MNIST database. Available at: http://yann.lecun.com/exdb/mnist/index.html (accessed January 10, 2017).
13. Network service Google Maps. Available at: http://www.maps.google.com (accessed January 10, 2017).
14. Cormen, T., C. Leiserson, R. Rivest, and C. Stein. 2009. *Introduction to algorithms*. 3rd ed. MIT Press. 1312 p.
15. Algorithm for searching approximate nearest neighbor. Available at: http://sourceforge.net/projects/edivis/files/ (accessed January 10, 2017).

Received December 13, 2016

Contributors

Lange Mikhail M. (b. 1945) — Candidate of Science (PhD) in technology, leading scientist, Federal Research Center "Computer Sciences and Control" of the Russian Academy of Sciences, 44-2 Vavilov Str., Moscow 119333, Russian Federation; lange_mm@ccas.ru

Ganebnykh Sergey N. (b. 1968) — scientist, Federal Research Center "Computer Sciences and Control" of the Russian Academy of Sciences, 44-2 Vavilov Str., Moscow 119333, Russian Federation; sng@ccas.ru

Lange Andrey M. (b. 1979) - Candidate of Science (PhD) in physics and mathematics, scientist, Federal Research Center "Computer Sciences and Control" of the Russian Academy of Sciences, 44-2 Vavilov Str., Moscow 119333, Russian Federation; lange_am@mail.ru

ПОВЫШЕНИЕ КАЧЕСТВА КЛАССИФИКАЦИИ В ЗАДАЧЕ ОБНАРУЖЕНИЯ ВНУТРЕННЕГО ПЛАГИАТА*

И. О. Молибог[1], А. П. Мотренко[2], В. В. Стрижов[3]

Аннотация: Исследуется задача классификации объектов в многомерных пространствах. Для снижения размерности задачи предлагается модификация алгоритма t-SNE (*англ.* t-distributed Stochastic Neighbor Embedding), в которой при обучении используется информация о разметке, не возникает необходимости заново обучать алгоритм при добавлении новых данных, а также предусмотрена параллельная реализация. Предлагаемый алгоритм решает задачу внутреннего плагиата, в которой признаками являются частотные словесные профили сегментов текста. Показано, что качество классификации после применения алгоритма выше, чем без него или с другими алгоритмами.

Ключевые слова: анализ данных; снижение размерности; нелинейные методы снижения размерности; обучение многообразий; обнаружение внутреннего плагиата

DOI: 10.14357/19922264170307

1 Введение

В работе рассматривается задача классификации объектов в пространствах большой размерности, признаковое описание которых имеет в себе скрытые функциональные зависимости. Предполагается, что объекты содержатся вблизи многообразия много меньшей размерности, чем размерность исходного пространства. Назовем это предположение гипотезой многообразия [1]. Данные ряда практических задач, включая задачи анализа генома, анализа текста и распознавания изображений, не противоречат этой гипотезе [2]. В [3] было дано ее формальное определение и перечислены идеи методов, которыми ее можно проверить. Практической задачей, рассматриваемой в данной работе, является задача обнаружения внутреннего плагиата [4, 5].

Задача обнаружения внутреннего плагиата состоит в поиске заимствованных частей документа без использования внешних источников. При решении задачи исследуемый текст некоторым образом разбивается на сегменты. Каждому сегменту соответствует его вектор признаков. Сегмент считается минимальной единицей заимствования. Он считается либо полностью заимствованным, либо полностью оригинальным. Тогда задача обнаружения внутреннего плагиата является задачей классификации, где объектами являются векторы признаков сегментов, а классами — метки заимствования или оригинальности.

Способы разбиения на сегменты, как и способы вычисления вектора признаков, являются предметом отдельного исследования. Подходы [4, 6–8] продемонстрировали на конкурсе PAN-2011 [9] наилучшее качество решения задачи обнаружения внутреннего плагиата. Они включают разбиение документа на абзацы, предложения, блоки слов или символов. В них используются признаки, основанные на частотных профилях сегментов. Такие признаки имеют размерность, пропорциональную числу слов в документе, сильно разрежены и не всегда информативны.

В данной работе предполагается, что объекты с таким признаковым описанием подчиняются гипотезе многообразия. Это означает, что метрически близкие объекты могут быть геодезически далекими, и дает возможность применить методы снижения размерности для улучшения качества классификации.

В задаче понижения размерности требуется построить гладкое отображение множества $\mathbf{X}$ в пространстве исходных данных в некоторое множество $\mathbf{Z}$ в пространстве меньшей размерности. Будем называть элементы $\mathbf{Z}$ образами элементов $\mathbf{X}$. Пространство образов будем называть результирующим. В конкретных алгоритмах на это отображе-

*Работа выполнена при финансовой поддержке РФФИ (проект 16-07-01155).

[1]Центр энергетических систем, Сколковский институт науки и технологий; Московский физико-технический институт, i.molybog@skoltech.ru

[2]Московский физико-технический институт, anastasiya.motrenko@phystech.edu

[3]Вычислительный центр им. А. А. Дородницына Федерального исследовательского центра «Информатика и управление» Российской академии наук, strijov@phystech.edu

ние накладывают необходимые ограничения, исходя из специфики задачи [10]. Приведем некоторые из них.

Для снижения размерности широко применяются линейные методы, основанные на анализе дисперсии: латентно-семантический анализ [11, 12], анализ главных компонент [13]. Однако они могут не сохранять кластерную структуру исходных данных и потому не применимы для решения задач вложений из нелинейных многообразий.

Для выполнения вложений из нелинейных многообразий были разработаны алгоритмы, использующие изометрические отображения. Алгоритмы ISOMAP (Isometric Mapping) [14] и Laplacian Eigenmap [15] приближают геодезическое расстояние с помощью графа k ближайших соседей. Алгоритмы Local Linear Embedding (LLE) [16] и Hessian-based LLE [17] основаны на предположении, что многообразие аппроксимируется кусочно-линейной функцией. Для каждого объекта исходного пространства строится его линейное приближенное описание через соседние объекты, после чего по этим описаниям строятся образы в результирующем пространстве. Метод [17] использует для описания объектов специальную квадратичную форму, что гарантирует асимптотическую оптимальность метода даже в случае невыпуклых множеств.

Алгоритм Local Tangent Space Alignment Algorithm [18] также использует кусочно-линейную аппроксимацию. Многообразие приближается гиперплоскостью в окрестности каждой точки, после чего полученные приближения сглаживаются между собой. При помощи Semidefinite Embedding [19] можно получить вложение, в котором сохранены точные расстояния между ближайшими объектами. Для этого метод максимизирует след матрицы Грама для образов при ограничениях, накладываемых отношением соседства объектов исходного пространства и их матрицей Грама.

Все перечисленные методы нацелены на наиболее точное сохранение расстояний между объектами при снижении размерности. Это может привести к неустойчивости решения, связанной с тем, что изменения расстояния между далекими и близкими объектами штрафуются одинаково. Кроме того, они не приспособлены для решения задачи классификации, поскольку не учитывают разметку при выполнении вложения, хотя существуют их модификации, обладающие этим свойством. В [20] метод аппроксимации расстояний, используемый в ISOMAP, модифицирован в методе оптимизации целевого функционала. Полученный метод получил название TRIMAP. В нем при обучении используется разметка обучающей выборки.

В данной работе применяется метод t-NSE [2]. Выгодной особенностью метода t-SNE является склонность к локализации изолированных плотных пространственных структур произвольной геометрии. Под изолированной плотной структурой подразумевается множество точек, имеющих близких соседей из той же структуры, но сравнительно удаленных от всех точек не из нее. Такой эффект достигается тем, что близким и далеким объектам назначаются разные приоритеты.

Недостатком метода t-SNE в отношении задачи классификации является то, что в нем не предусмотрено функции вложения объектов, не участвовавших в построении уже существующего вложения. В работе [21] описана параметрическая модификация t-SNE, которая частично избавлена от этой особенности, однако в данной работе она не использовалась.

Дополнительным ограничением применимости метода t-SNE является высокая по сравнению с другими методами вложений вычислительная сложность. Хотя в [22] предлагаются два способа вычисления градиента, при использовании которых сложность непараметрического t-SNE составляет $O(km\log(m))$, где m — размер выборки, а k — размерность результирующего пространства, этого ускорения недостаточно для обеспечения комфортной работы даже с выборками длиной порядка 10^3.

Основным вкладом данной статьи в теорию распознавания образов является предложенная модификация метода t-SNE, позволяющая строить классификаторы в результирующем пространстве. Преимуществом предлагаемого метода является то, что он расширяет границы применимости оригинального метода t-SNE. Разработанная модификация предусматривает вложение тестовых данных без повторного вложения обучающих, а также может учитывать разметку обучающих данных и имеет параллельную реализацию.

2 Постановка задачи

Обозначим $\mathbb{X} \subset \mathbb{R}^n$ множество всех возможных векторов $\mathbf{x}$ признаков изучаемых объектов. Предполагается, что объекты $\mathbb{X}$ подчиняются *гипотезе многообразия*: найдется гладкое отображение $\mathbf{f} : \mathbb{R}^d \to \mathbb{R}^n$ такое, что

$$\text{для } \mathbf{x} \in \mathbf{X} \text{ существует } \mathbf{z}^* \in \mathbb{R}^d : \mathbf{x} = \mathbf{f}(\mathbf{z}^*) + \boldsymbol{\varepsilon}\,,$$

где $\boldsymbol{\varepsilon}$ — случайный вектор с нулевым математическим ожиданием и конечной матрицей корреляций. Будем называть d эффективной размерностью исходного пространства $\mathbb{X}$. Она определяется природой признакового пространства. Поскольку d

заранее не известно, введем понятие результирующего пространства $\mathbb{R}^k$, в котором выполняется поиск решения. В общем случае $k \neq d$. Процесс поиска образов объектов выборки в результирующем пространстве назовем вложением в него.

Рассмотрим выборку из m объектов, заданную матрицей

$$\mathbf{X} = [\mathbf{x}_1 \cdots \mathbf{x}_m]^\mathsf{T}\,, \quad \mathbf{x}_i \in \mathbb{X}\,, \quad i = 1, \ldots, m\,. \tag{1}$$

Пусть $p_{ij} = P(\mathbf{x}_i, \mathbf{x}_j)$ и $q_{ij} = Q(\mathbf{z}_i, \mathbf{z}_j)$ — расстояния между объектами в $\mathbb{R}^n$ и $\mathbb{R}^k$ соответственно:

$$p_{ij} = \frac{p_{j|i} + p_{i|j}}{2m},$$

$$p_{i|j} = \frac{\exp\left(-||\mathbf{x}_i - \mathbf{x}_j||^2 / \left(2\sigma_i^2\right)\right)}{\sum_{k \neq i} \exp\left(-||\mathbf{x}_i - \mathbf{x}_k||^2 / \left(2\sigma_i^2\right)\right)};$$

$$q_{ij} = \frac{\left(1 + ||\mathbf{z}_i - \mathbf{z}_j||^2\right)^{-1}}{\sum_{k \neq i} \left(1 + ||\mathbf{z}_i - \mathbf{z}_k||^2\right)^{-1}}\,, \quad q_{ii} = 0\,,$$

$$i, j \in \{1, \ldots, m\}\,.$$

Параметр σ_i в условном распределении p_{ij} задан для каждого i и зависит от расположения $\mathbf{x}_i$ относительно других объектов в исходном пространстве. Если он расположен в области высокой концентрации исходных данных, то коэффициент σ_i имеет меньшие значения, чем если бы концентрация была низкой.

Расположение

$$\mathbf{Z} = [\mathbf{z}_1 \cdots \mathbf{z}_m]^\mathsf{T} \subset \mathbb{R}^k \tag{2}$$

как образов $\mathbf{X}$ в результирующем пространстве $\mathbb{R}^k$ находится путем минимизации дивергенции Кульбака–Лейблера:

$$\mathbf{Z}_{\min} = \underset{\mathbf{Z} \in \mathbb{R}^{m \times k}}{\operatorname{argmin}}\, C(\mathbf{X}, \mathbf{Z})\,, \tag{3}$$

где

$$C(\mathbf{X}, \mathbf{Z}) = \mathrm{KL}(P||Q) = \sum_{i \neq j} p_{ij} \log \frac{p_{ij}}{q_{ij}}\,. \tag{4}$$

Заметим, что минимизация происходит только по координатам объектов $\mathbf{z}_1, \ldots, \mathbf{z}_m$ как по переменным, а координаты $\mathbf{x}_1, \ldots, \mathbf{x}_m$ считаются известными константами.

Задача решается градиентными методами [2]. Для инициализации начальных точек $\mathbf{Z}^{(0)} = [\mathbf{z}_1^{(0)} \cdots \mathbf{z}_m^{(0)}]^\mathsf{T}$ градиентного спуска в стандартной реализации было предложено [2] два метода: инициализировать случайными точками либо использовать для задания начальной инициализации метод Principal Components Analysis. От качества начальной инициализации, в случае с невыпуклой задачей оптимизации, зависят не только скорость сходимости к оптимуму, но и локальный минимум, к которому будет сходиться градиентный метод.

3 Предлагаемая модификация t-SNE

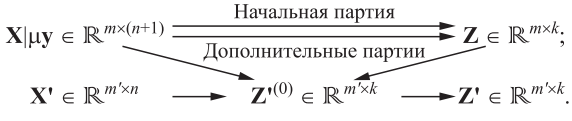
Рассмотрим задачу классификации с обучающей выборкой $\mathbf{X}$ (1) и тестовой выборкой из m' объектов $\mathbf{X}' = [\mathbf{x}_{m+1} \cdots \mathbf{x}_{m+m'}]^\mathsf{T} \subset \mathbb{X}$. Соответственно, метки классов $y_i \in \{0, 1\}$, $i = 1, \ldots, m$, известны, а $\hat{y}_i$, $i = m+1, \ldots, m+m'$, необходимо оценить. Так как на этапе обучения данные $\mathbf{X}'$ могут быть недоступны, метод непараметрического t-SNE не применим для снижения размерности в задачах классификации. Назовем это проблемой непросмотренных объектов (out-of-sample problem). Для ее решения предлагается минимизировать (4) независимо по различным подмножествам объектов.

Для повышения качества классификатора в результирующем пространстве предлагается перед вложением обучающей выборки добавить в ней метки классов в качестве признаков и улучшить таким образом начальное приближение градиентного метода. Идея такого подхода заключается в том, что, поскольку t-SNE сохраняет только локальную структуру схожести между объектами, после проведения процедуры понижения размерности классифицируемые объекты отображаются в кластеры, предварительно разнесенные с учетом меток. При этом используется предположение, что объекты из $\mathbf{X}'$ больше схожи с объектами $\mathbf{X}$ того же класса, чем с объектами противоположного. Таким образом удается увеличить расстояние между образами классифицируемых объектов из различных классов, что упрощает их классификацию. Ниже показаны основные отображения оригинального непараметрического t-SNE

$$\mathbf{X} \in \mathbb{R}^{m \times n} \longrightarrow \mathbf{Z} \in \mathbb{R}^{m \times k}\,;$$

$$\mathbf{X}' \in \mathbb{R}^{m' \times n} \longrightarrow \mathbf{Z}' \in \mathbb{R}^{m' \times k}$$

и предложенной модификации



Использование исходной разметки выборки при вложении для обучения классификатора. Для учета

разметки обучающей выборки признаковая матрица $\mathbf{X}$ расширяется дополнительным столбцом признаков

$$\tilde{\mathbf{X}} = \left(\mathbf{X}\middle|\mu\mathbf{y}\right),$$

где μ — вес меток как признаков. В модифицированном алгоритме на основе расширенной матрицы $\tilde{X}$ выполняется поиск образов $\mathbf{Z}$ (4), на которых обучается классификатор. Таким образом, при построении вложения обучающей выборки решается задача

$$\mathbf{Z}_{\min} = \operatorname*{argmin}_{\mathbf{Z}\in\mathbb{R}^{m\times k}} C\left(\left(\mathbf{X}\middle|\mu\mathbf{y}\right), \mathbf{Z}\right).$$

Вложение новых объектов в пространство со сниженной размерностью для классификации. Обозначим через $\mathbf{Z}' = [\mathbf{z}_{m+1}\cdots\mathbf{z}_{m+m'}]^{\mathsf{T}}$ образы $\mathbf{X}'$ в результирующем пространстве. Аналогично (3) сформулируем задачу поиска $\mathbf{Z}'$ в виде m' задач k-мерной минимизации, которые могут быть решены независимо:

$$\mathbf{z}_i^{\min} = \operatorname*{argmin}_{\mathbf{z}_i\in\mathbb{R}^{m'}} C\left(\begin{bmatrix}\mathbf{X}\\ \hline \mathbf{x}_i^{\mathsf{T}}\end{bmatrix}, \begin{bmatrix}\mathbf{Z}\\ \hline \mathbf{z}_i^{\mathsf{T}}\end{bmatrix}\right), \quad i = m+1,\ldots,m+m',$$

где матрицы $\begin{bmatrix}\mathbf{X}\\ \hline \mathbf{x}_i^{\mathsf{T}}\end{bmatrix}$ и $\begin{bmatrix}\mathbf{Z}\\ \hline \mathbf{z}_i^{\mathsf{T}}\end{bmatrix}$ получены из $\mathbf{X}$ и $\mathbf{Z}$ добавлением строк $\mathbf{x}_i^{\mathsf{T}}$ и $\mathbf{z}_i^{\mathsf{T}}$ соответственно. При использовании такого подхода предполагается, что обучающая выборка $\mathbf{X}$ (1) достаточно репрезентативна.

Для инициализации образов $\mathbf{z}_{i'}$ классифицируемых объектов предлагается использовать метод взвешенного среднего по образам соседей:

$$\mathbf{z}_{i'}^{(0)} = \sum_{i=1}^{m}\mathbf{z}_i w_{ii'}, \quad \sum_{i=1}^{m} w_{ii'} = 1, \quad i' = m+1,\ldots,m+m',$$

где $w_{ii'}$ — веса образов объектов $\mathbf{x}_i$, $i = 1,\ldots,m$. В работе рассмотрены два способа задания весов:

Алгоритм 1: Вложение выборки с известным вектором ответов классификации $\mathbf{y}$

Data: $\mathbf{X}, \mathbf{y}, \mu, S_s, S_b$
Result: $\mathbf{Z}$

```
X̃ = (X|μy)
Инициализировать Z (2) случайно или при помощи PCA(X̃). Положить инициализацию начальной точкой градиентного метода: Z^(0).
if m > S_s then
    Разбить Z на партии: начальная партия Z_0 размером S_s и B = ⌈(m − S_s)/S_b⌉ дополнительных партий Z_1, ..., Z_B размером не больше чем S_b каждая. Оптимизировать (4) по Z_0, зафиксировав координаты остальных объектов из Z, известные из предыдущего шага. for Z_i ∈ {Z_1, ... Z_B} do
        Оптимизировать (4) по Z_i, зафиксировав координаты остальных объектов из Z, известные из предыдущего шага.
    end
else
    Оптимизировать (4)
end
```

Алгоритм 2: Вложение выборки без известного вектора ответов классификации

Data: $\begin{bmatrix}\mathbf{X}\\ \hline \mathbf{X}'\end{bmatrix}, \mathbf{Z}$
Result: $\mathbf{Z}'$

```
Инициализировать Z (2) случайно, при помощи PCA(X̃), либо используя (6) или (5) для расчета W и считать Z'^(0) = Z^T W.
for i ∈ {m+1, ..., m+m'} do
    Оптимизировать (4) по z_i, зафиксировав координаты остальных объектов из [Z / Z'], известные из предыдущего шага.
end
```

$$w_{ii'}^{\text{softmax}} = \frac{\exp(-\|\mathbf{x}_i - \mathbf{x}_{i'}\|)}{\sum_{k=1}^{m} \exp(-\|\mathbf{x}_k - \mathbf{x}_{i'}\|)} \qquad (5)$$

или

$$w_{ii'}^{\text{stud}} = \frac{(1 + \|\mathbf{x}_k - \mathbf{x}_{i'}\|^2)^{-1}}{\sum_{k=1}^{m} (1 + \|\mathbf{x}_k - \mathbf{x}_{i'}\|^2)^{-1}}\,. \qquad (6)$$

Для ускорения процедуры вложения при работе с большими данными предлагается процедура поэтапного вложения объектов блоками, размер которых — S_s для первого по очереди и S_b для всех остальных — много меньше размера m всей выборки.

Псевдокод предложенного метода приведен в алгоритмах 1 и 2.

4 Вычислительный эксперимент

Вычислительный эксперимент состоит из двух частей: исследование разработанного алгоритма на синтетических данных и применение разработанного алгоритма для решения задачи внутреннего плагиата.

Для инициализации вложения тестовых данных использовались четыре различных подхода: случайный — инициализация случайным образом; PCA (Principal Component Analysis) — инициализация образами при снижении размерности методом главных компонент; Softmax и Student — задаваемые по формулам (5) и (6).

На рис. 1–5 представлены результаты экспериментов, проведенных при использовании всех этих способов инициализации.

Все инициализированные таким образом объекты далее преобразуются, минимизируя (4) при фиксированных образах объектов обучающей выборки. После этого происходит классификация полученных образов.

Методы инициализации (5) и (6) были предложены так, чтобы инициализированные данные обладали свойством сохранения локальной структуры исходной выборки. Предполагалось, что это

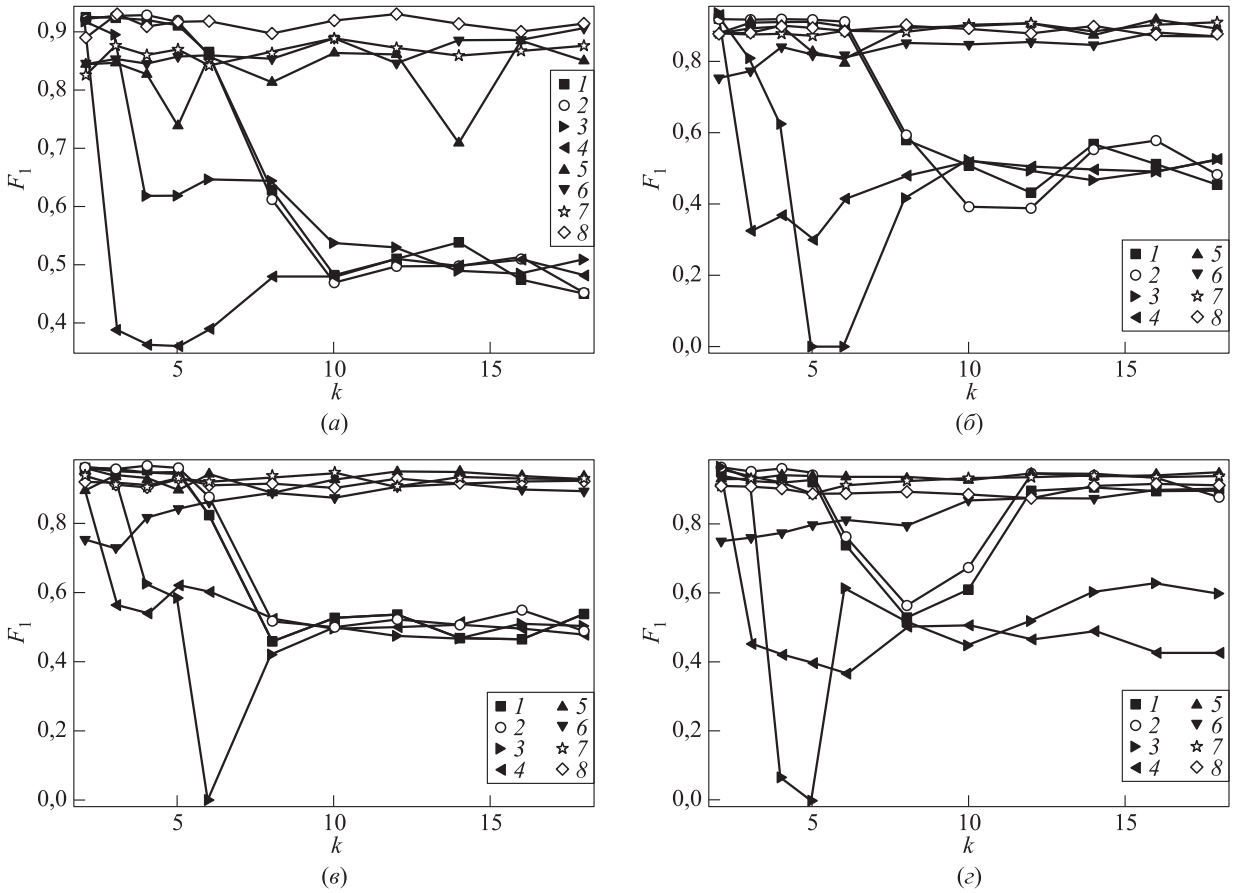


Рис. 1 Зависимость F_1 от k при различных эффективных размерностях выборки $d = 4$ (*а*), 8 (*б*); 12 (*в*) и 16 (*г*) при использовании t-SNE (*1* — Student; *2* — Softmax; *3* — Random; *4* — PCA) и других методов снижения размерности (*5* — LLE; *6* — PCA; *7* — ISOMAP), а также без применения снижения размерности (*8*)

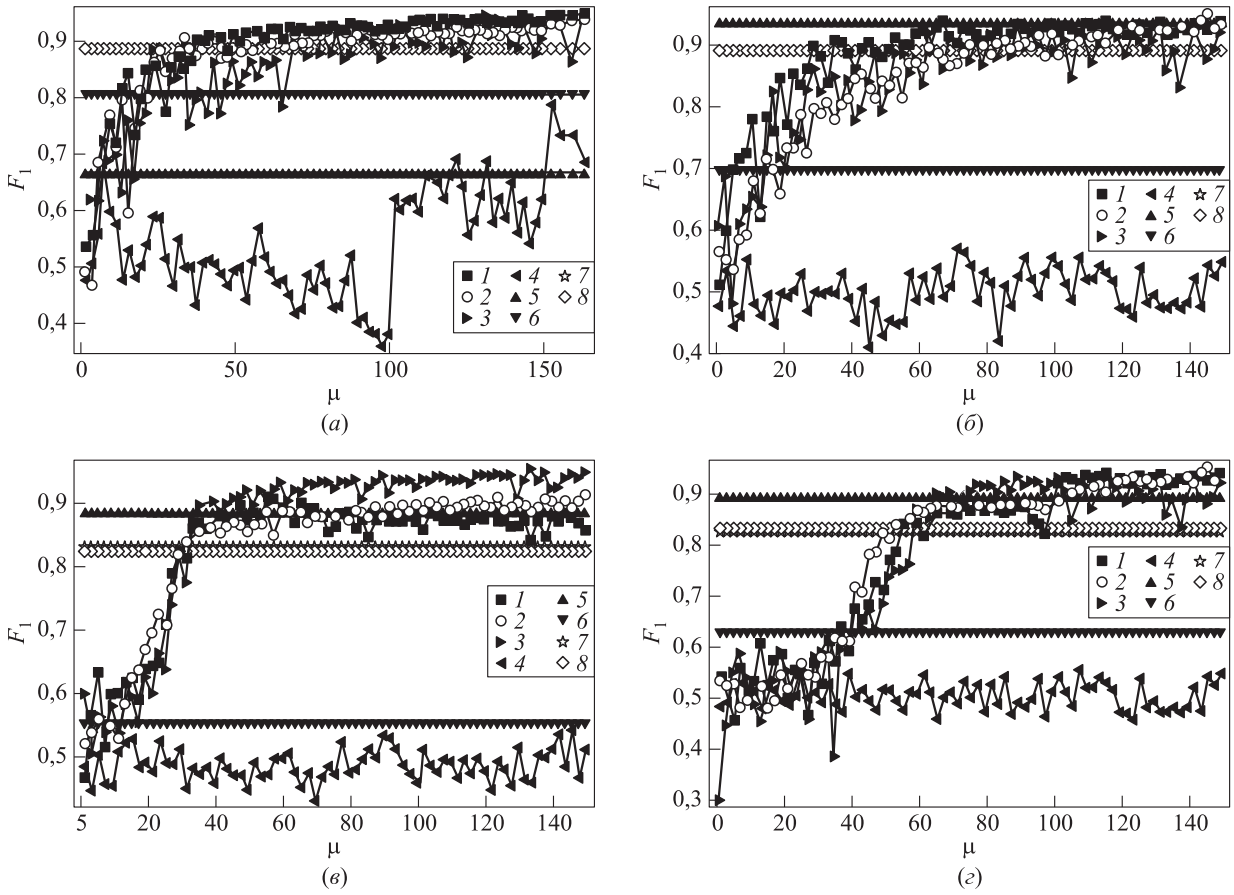


Рис. 2 Зависимость F_1 от μ при различных размерностях выборки $n = 6$ (*а*); 60 (*б*); 300 (*в*) и 600 (*г*) при использовании t-SNE (*1* — Student; *2* — KNN (k nearest neighbor); *3* — Random; *4* — PCA) и других методов снижения размерности (*5* — LLE; *6* — PCA; *7* — ISOMAP), а также без применения снижения размерности (*8*)

улучшит сходимость градиентного метода, используемого для минимизации (4), по сравнению с инициализациями PCA и random.

4.1 Исследование свойств алгоритма на синтетических данных

В данном подразделе для эмпирического исследования свойств предлагаемого алгоритма использовались синтетические выборки $\mathbf{X} = [\mathbf{x}_1 \cdots \mathbf{x}_m]^\mathsf{T}$. Для любого вектора $\mathbf{x}_i$ компоненты были сгенерированы как стандартные нормальные распределения на гранях гиперкуба. При этом эффективная размерность выборки составляла d, а оставшиеся признаки были шумовыми. Далее выборка сворачивалась в спираль по одной из размерностей. Это делалось для того, чтобы реализовать предположение о существовании многообразия меньшей размерности, в котором содержится выборка. Генерировалось одинаковое количество объектов разных классов, а на обучение и контроль выборка разбивалась в соотношении $1 : 4$.

В этом подразделе описывается исследование качества классификации с применением предлагаемого алгоритма в зависимости от основных его параметров и специфики выборки. Для сравнения предлагаемого алгоритма и его исследования рассматривается классификация в комбинации с другими методами снижения размерности: PCA [23], LLE [16], ISOMAP [14], а также без применения снижения размерности. Для построения классификатора использовался метод логистической регрессии на основе Stochastic Gradient Descent [24].

На рис. 1 изображена зависимость меры качества F_1 от размерности вложения k при различных значениях эффективной размерности d. На графиках видно, что качество значительно ухудшается при увеличении k независимо от соотношения k и d. Эксперимент проведен при постоянных $m = 500$, $n = 20$, $S_b = 100$, $S_s = 400$ и $\mu = 150$.

На рис. 2 изображена зависимость F_1 от веса μ меток класса y в стартовой выборке при различных значениях размерности выборки n. Из них можно сделать вывод, что качество классификации повы-

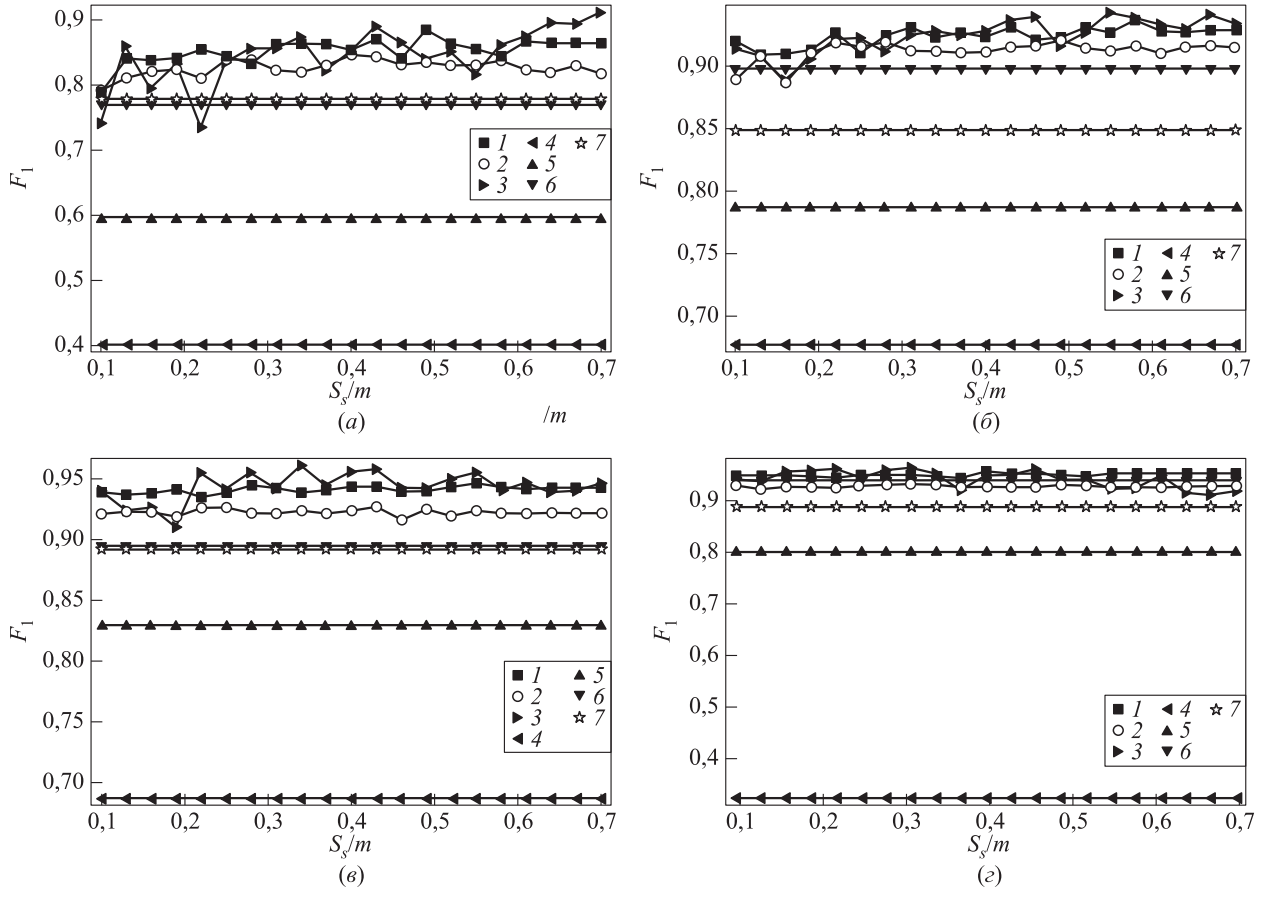


Рис. 3 Зависимость F_1 от S_s/m при различных размерах выборки $m = 1000$ (*а*); 2000 (*б*); 3000 (*в*) и 4000 (*г*) при использовании t-SNE (*1* — Student; *2* — KKN; *3* — Random) и других методов снижения размерности (*4* — LLE; *5* — PCA; *6* — ISOMAP), а также без применения снижения размерности (*7*)

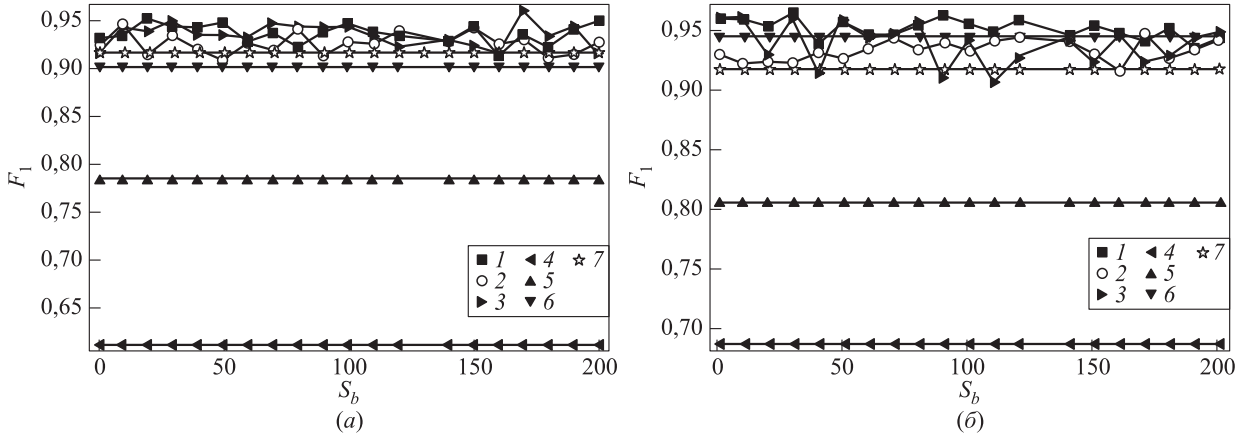


Рис. 4 Зависимость F_1 от S_b при различных размерах выборки $m = 500$ (*а*) и 1000 (*б*) при использовании t-SNE (*1* — Student; *2* — KKN; *3* — Random) и других методов снижения размерности (*4* — LLE; *5* — PCA; *6* — ISOMAP), а также без применения снижения размерности (*7*)

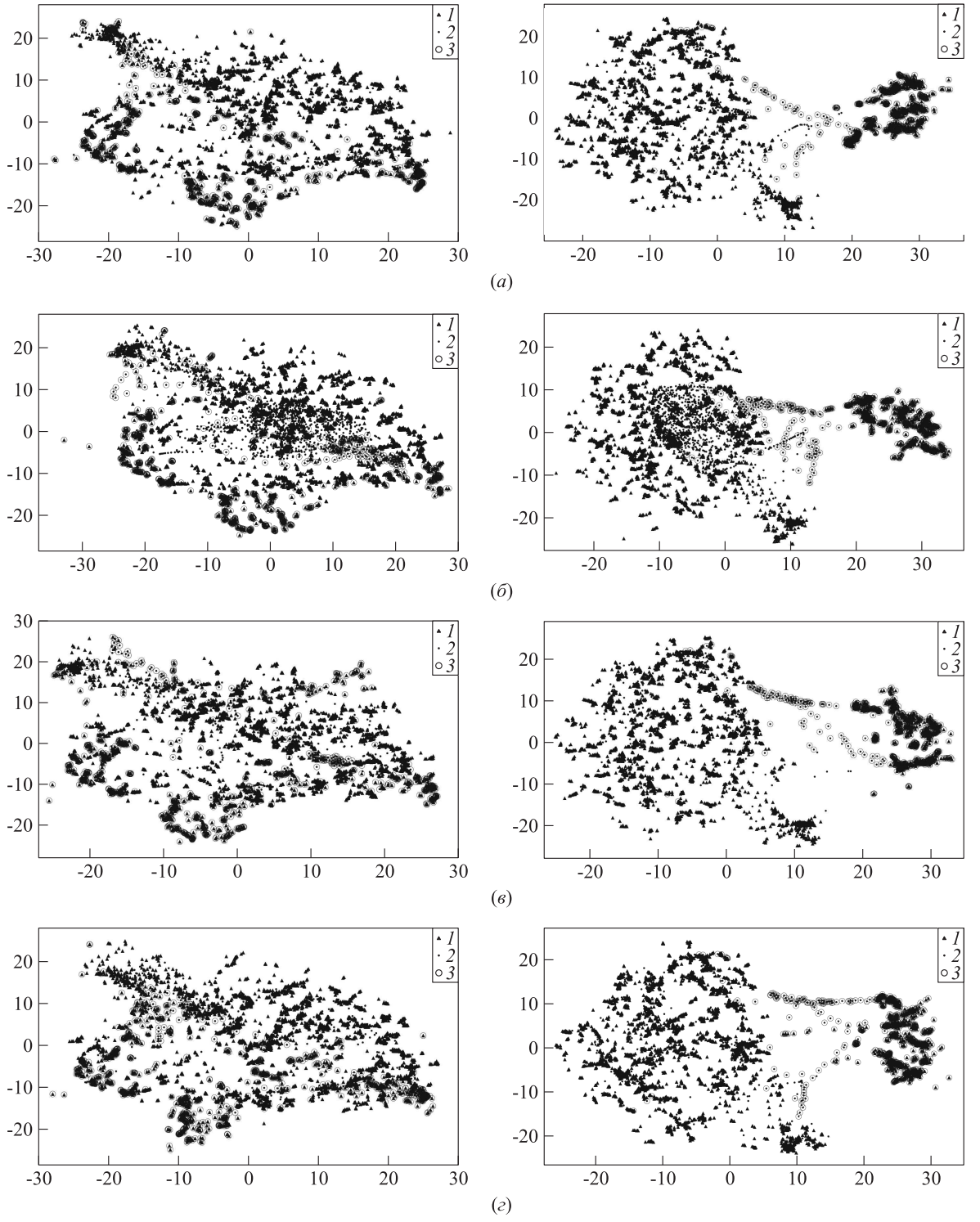


Рис. 5 Демонстрация вложения, выполненного предлагаемым методом ((*а*) Student; (*б*) Softmax; (*в*) PCA; (*г*) Random) при $\mu = 0$ (левый столбец) и 10 (правый столбец): *1* — обучающая выборка; *2* — тестовая выборка; *3* — плагиат

шается с ростом μ, при этом скорость роста падает с ростом n. Также видно, что разработанный метод при достаточно больших значениях μ показывает в среднем лучшие результаты среди всех рассмотренных методов снижения размерности, а также превосходит по качеству классификацию в исходном пространстве. Эксперимент проведен при постоянных $m = 500$, $k = 3$, $S_b = 100$ и $S_s = 400$. В этом эксперименте все исходные признаки были информативными.

Для исследования зависимости качества классификации от величины отношения размера стартовой части к размеру выборки S_s/m был поставлен эксперимент, где при постоянных $n = 6$, $k = d = 3$ и $\mu = 150$ исследовалась зависимость меры качества F_1 от размера выборки m и размера начального вложения S_s. При этом размер дополнительно вкладываемых блоков S_b принимался заведомо бóльшим размера выборки m, так что дополняющая часть не разбивалась на блоки. На рис. 3 выведены результаты. Можно видеть, что зависимость от этих параметров незначительна. При этом скорость работы алгоритма увеличивается при наличии разбиений на стартовую и дополняющую части. Таким образом, показано, что предложенная модификация алгоритма позволяет значительно ускорить его работу без существенного снижения качества.

На графиках рис. 3 также видно, что методы инициализации с помощью (6) и случайной инициализации дают лучшие результаты, в то время как метод инициализации PCA показал результаты порядка 0,5, по причине чего было принято решение не выносить его на рисунок.

Целью эксперимента, результаты которого приведены на рис. 4, было исследование зависимости значения функции качества классификации от S_b. Он был проведен при постоянных $n = 6$, $k = d = 3$, $\mu = 150$ и $S_s = 200$. В результате было обнаружено, что предлагаемый метод устойчив относительно параметра S_b.

4.2 Задача обнаружения внутреннего плагиата

Целью данной части эксперимента был анализ предложенного метода снижения размерности в применении к реальным данным задачи внутреннего плагиата. Рассматривается набор документов. Каждый документ рассматривается как последовательность сегментов s_i, каждый из которых описывается вектором признаков $\mathbf{x}$. В данной работе в качестве сегментов рассматриваются предложения. Каждому s_i поставлена в соответствие метка класса $y_i \in \{0, 1\}$: $y_i = 1$, если s_i — заимствованный сегмент, иначе $y_i = 0$. Задача распознавания внутреннего плагиата ставится как задача восстановления меток y_i по документу.

Иллюстрация вложения реальных данных. Для демонстрации работы алгоритма на реальных данных из предоставленного корпуса [25] `part1` выделен один из документов. Выделенные из него объекты были разделены на обучающую и тестовую выборки. Каждой из них соответствуют непрерывные части текста. Это разделение необходимо для демонстрации работы предложенной модификации и не учитывается при применении оригинального t-SNE. На рис. 6 приведен результат применения оригинального непараметрического метода t-SNE к объектам, выделенным из выбранного документа. На нем видно, что объекты, соответствующие заимствованным частям текста, имеют очаги концентрации в исходном пространстве, что свидетельствует об информативности выбранных признаков. Рисунки 5 и 6 имеют безразмерные оси, полученные в результате нелинейных отображений. Физического смысла эти оси не несут.

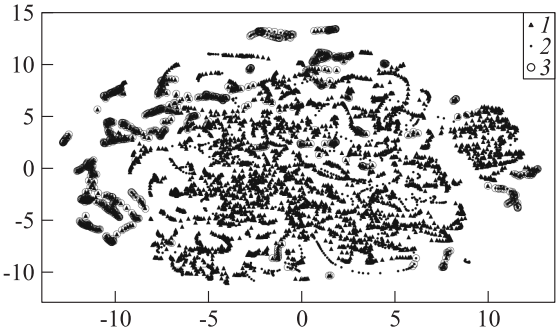


Рис. 6 Визуализация документа с использованием оригинального алгоритма t-SNE: *1* — обучающая выборка; *2* — тестовая выборка; *3* — плагиат

На рис. 5 представлены результаты вложения данных выбранного документа с использованием предложенного алгоритма при различных методах начальной инициализации и при различных значениях веса μ. При выполнении вложений были зафиксированы параметры $S_s = 500$ и $S_b = 200$. В эксперименте данные из выбранного документа были разделены на обучающую и тестовую выборки. В тестовую часть попали образы предложений, которые образовывали в исходном тексте непрерывную цепочку. Обучающая часть вкладывалась с учетом ее разметки, а тестовая — без учета.

Результаты. Из полученных графиков можно сделать вывод, что предложенная модификация при больших значениях веса μ принимает на себя часть ответственности за классификацию. Она склонна

разделять и кластеризовать тестовую выборку по целевому признаку. Таким образом, исходя из описанных выше свойств t-SNE, любой построенный в результирующем пространстве классификатор получает свойство классификатора ближайших соседей с адаптивной константой, подстраиваемой под локальную геометрию выборки. Следует отметить также, что при больших значениях μ минимизация целевой функции (4) требует больше шагов градиентного алгоритма. Таким образом, этот параметр следует выбирать с оглядкой на время работы программы. Авторы рекомендуют значение порядка характерной величины координат векторов обучающей выборки.

5 Заключение

В работе была предложена модификация непараметрического метода снижения размерности t-SNE, состоящая в воплощении возможности выполнения вложения поэтапно, решении проблемы непросмотренных объектов и внедрении возможности учета разметки при выполнении вложения для классификации. Был проведен вычислительный эксперимент на синтетических данных, показывающий эффективность предложенного метода в применении к задаче классификации. Была определена зависимость качества классификации с применением описанного метода от его параметров, экспериментально обосновано использование поэтапного обучающего вложения. Полученные значения качества сравнивались с результатами классификации с применением других методов снижения размерности, а также без их применения.

Была показана устойчивость алгоритма к введенным параметрам размера начальной части S_s и максимального размера блоков S_b, что облегчает его использование на практике. Также явно продемонстрирована зависимость свойств метода от параметра веса разметки выборки μ.

Проанализировано признаковое пространство задачи внутреннего плагиата. Проиллюстрированы свойства предложенного алгоритма относительно данных задачи внутреннего плагиата. Продемонстрирована эффективность предложенных методов инициализации при вложении образов объектов, которые не были использованы при выполнении начального вложения.

Литература

1. *Fefferman C., Mitter S., Narayanan H.* Testing the manifold hypothesis // J. Am. Math. Soc., 2016. Vol. 29. No. 4 P. 983–1049.
2. *Van der Maaten L., Hinton G.* Visualizing data using t-SNE // J. Mach. Learn. Res., 2008. Vol. 9. P. 2579–2605.
3. *Narayanan H., Mitter S.* Sample complexity of testing the manifold hypothesis // Advances in neural information processing systems / Eds. J. D. Lafferty, C. K. I. Williams, J. Shawe-Taylor, *et al.* — Curran Associates, Inc., 2010. Vol. 23. P. 1786–1794.
4. *Zu Eissen S. M., Stein B.* Intrinsic plagiarism detection // European Conference on Information Retrieval. — Springer, 2006. P. 565–569.
5. *Kuznetsov M. P., Motrenko A. P., Kuznetsova M. V., Strijov V. V.* Methods for intrinsic plagiarism detection and author diarization // Working Notes of CLEF / Eds. K. Balog, L. Cappellato, N. Ferro, C. Macdonald. — Évora, Portugal: CEUR-WS, 2016. Vol. 1609. P. 912–919.
6. *Stamatatos E.* Intrinsic plagiarism detection using character n-gram profiles // SEPLN Workshop on Uncovering Plagiarism, Authorship, and Social Software Misuse, 2009. P. 38–46.
7. *Muhr M., Kern R., Zechner M., Granitzer M.* External and intrinsic plagiarism detection using a cross-lingual retrieval and segmentation system // Working Notes for CLEF Conference / Eds. M. Braschler, D. Harman, E. Pianta, N. Ferro. — Padua, Italy: CEUR-WS, 2010. Vol. 1176. http://ceur-ws.org/Vol-1176/CLEF2010wn-PAN-MuhrEt2010.pdf.
8. *Kestemont M., Luyckx K., Daelemans W.* Intrinsic plagiarism detection using character trigram distance scores // Working Notes for CLEF Conference / Eds. V. Petras, P. Forner, P. Clough, N. Ferro. — Amsterdam, The Netherlands: CEUR-WS, 2011. Vol. 1177. http://ceur-ws.org/Vol-1177/CLEF2011wn-PAN-KestemontEt2011.pdf.
9. *Potthast M., Eiselt A., Cedeño L. A., Stein B., Rosso P.* Overview of the 3rd international competition on plagiarism detection // Working Notes for CLEF Conference / Eds. V. Petras, P. Forner, P. Clough, N. Ferro. — Amsterdam, The Netherlands: CEUR-WS, 2011. Vol. 1177. http://ceur-ws.org/Vol-1177/CLEF2011wn-PAN-PotthastEt2011a.pdf.
10. *Fodor I. K.* A survey of dimension reduction techniques. Center for Applied Scientific Computing, Lawrence Livermore National Laboratory, 2002. Technical Report. P. 1–18.
11. *Brooke J., Hirst G.* Paragraph clustering for intrinsic plagiarism detection using a stylistic vector-space model with extrinsic features // Working Notes for CLEF Conference / Eds. P. Forner, J. Karlgren, C. Womser-Hacker, N. Ferro. — Rome, Italy: CEUR-WS, 2012. Vol. 1178. http://ceur-ws.org/Vol-1178/CLEF2012wn-PAN-BrookeEt2012.pdf.
12. *Brooke J., Hammond A., Hirst G.* Unsupervised stylistic segmentation of poetry with change curves and extrinsic features // 1st NAACL-HLT Workshop on Computational Linguistics for Literature Proceedings, 2012. Stroudsburg, PA, USA: Association for Computational Linguistics. P. 26–35.

13. *Gorban A. N., Kégl B., Wunsch D. C., et al.* Principal manifolds for data visualization and dimension reduction. — Springer, 2008. 58 p.

14. *Tenenbaum J. B., De Silva V., Langford J. C.* A global geometric framework for nonlinear dimensionality reduction // Science, 2000. Vol. 290. Iss. 5500. P. 2319–2323.

15. *Belkin M., Niyogi P.* Laplacian eigenmaps and spectral techniques for embedding and clustering // Advances in neural information processing systems / Eds. T. G. Dietterich, S. Becker, Z. Ghahramani. — NIPS Foundation, Inc., 2001. Vol. 14. P. 585–591.

16. *Roweis S. T., Saul L. K.* Nonlinear dimensionality reduction by locally linear embedding // Science, 2000. Vol. 290. Iss. 5500. P. 2323–2326.

17. *Donoho D. L., Grimes C.* Hessian eigenmaps: Locally linear embedding techniques for high-dimensional data // P. Natl. Acad. Sci. USA, 2003. Vol. 100. No. 10. P. 5591–5596.

18. *Zhang Z., Zha H.* Principal manifolds and nonlinear dimensionality reduction via tangent space alignment // J. Shanghai University (English Edition), 2004. Vol. 8. No. 4. P. 406–424.

19. *Weinberger K. Q., Saul L. K.* Unsupervised learning of image manifolds by semidefinite programming // Int. J. Comput. Vision, 2006. Vol. 70. No. 1. P. 77–90.

20. *Chen C., Zhang J., Fleischer R.* Distance approximating dimension reduction of Riemannian manifolds // IEEE T. Syst. Man Cy. B, 2010. Vol. 40. No. 1. P. 208–217.

21. *Van der Maaten L.* Learning a parametric embedding by preserving local structure // RBM, 2009. Vol. 500. P. 26.

22. *Van der Maaten L.* Accelerating t-SNE using tree-based algorithms // J. Mach. Learn. Res., 2014. Vol. 15. No. 1. P. 3221–3245.

23. *Kim H., Park H., Zha H.* Distance preserving dimension reduction for manifold learning // SIAM Conference (International) on Data Mining Proceedings, 2007. P. 527–532.

24. *Bottou L.* Stochastic gradient descent tricks // Neural networks: Tricks of the trade / Eds. G. Montavon, G. B. Orr, K.-R. Muller. — Lecture notes in computer science ser. — 2nd ed. — Berlin–Heidelberg: Springer, 2012. Vol. 7700. P. 421–436.

25. *Potthast M., Stein B., Barrón-Cedeño A., Rosso P.* An evaluation framework for plagiarism detection // 23rd Conference (International) on Computational Linguistics Posters, 2010. P. 997–1005.

Поступила в редакцию 20.02.17

IMPROVING CLASSIFICATION QUALITY FOR THE TASK OF FINDING INTRINSIC PLAGIARISM

I. O. Molybog[1,2], A. P. Motrenko[2], and V. V. Strijov[3]

[1]Center for Energy Systems, Skolkovo Institute of Science and Technology, Skolkovo Innovation Center, 3 Nobel Str., Moscow 143026, Russian Federation

[2]Moscow Institute of Physics and Technology, 9 Institutskiy Per., Dolgoprudny, Moscow Region 141700, Russian Federation

[3]A. A. Dorodnicyn Computing Center, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, 40 Vavilov Str., Moscow 119333, Russian Federation

Abstract: The paper addresses the classification problem in multidimensional spaces. The authors propose a supervised modification of the t-distributed Stochastic Neighbor Embedding Algorithm. Additional features of the proposed modification are that, unlike the original algorithm, it does not require retraining if new data are added to the training set and can be easily parallelized. The novel method was applied to detect intrinsic plagiarism in a collection of documents. The authors also tested the performance of their algorithm using synthetic data and showed that the quality of classification is higher with the algorithm than without or with other algorithms for dimension reduction.

Keywords: data analysis; dimension reduction; nonlinear dimension reduction; manifold learning; intrinsic plagiarism detection

DOI: 10.14357/19922264170307

Acknowledgments

This publication is funded by the Russian Foundation for Basic Research (project No. 16-07-01155).

References

1. Fefferman, C., S. Mitter, and H. Narayanan. 2016. Testing the manifold hypothesis. *J. Am. Math. Soc.* 29(4):983–1049.
2. Van der Maaten, L., and G. Hinton. 2008. Visualizing data using t-SNE. *J. Mach. Learn. Res.* 9(Nov):2579–2605.
3. Narayanan, H., and S. Mitter. 2010. Sample complexity of testing the manifold hypothesis. *Advances in neural information processing systems*. Eds. J. D. Lafferty, C. K. I. Williams, J. Shawe-Taylor, *et al*. Curran Associates, Inc. 23:1786–1794.
4. Zu Eissen, S. M., and B. Stein. 2006. Intrinsic plagiarism detection. *European Conference on Information Retrieval*. Springer. 565–569.
5. Kuznetsov, M. P., A. P. Motrenko, M. V. Kuznetsova, and V. V. Strijov. 2016. Methods for intrinsic plagiarism detection and author diarization. *Working Notes of CLEF*. Eds. K. Balog, L. Cappellato, N. Ferro, and C. Macdonald. Évora, Portugal: CEUR-WS. 1609:912–919.
6. Stamatatos, E. 2009. Intrinsic plagiarism detection using character n-gram profiles. *SEPLN Workshop on Uncovering Plagiarism, Authorship, and Social Software Misuse*. 38–46.
7. Muhr, M., R. Kern, M. Zechner, and M. Granitzer. 2010. External and intrinsic plagiarism detection using a cross-lingual retrieval and segmentation system. *Working Notes for CLEF Conference*. Eds. M. Braschler, D. Harman, E. Pianta, and N. Ferro. Padua, Italy: CEUR-WS. Vol. 1176. Available at: http://ceur-ws.org/Vol-1176/CLEF2010wn-PAN-MuhrEt2010.pdf (accessed September 15, 2017).
8. Kestemont, M., K. Luyckx, and W. Daelemans. 2011. Intrinsic plagiarism detection using character trigram distance scores. *Working Notes for CLEF Conference*. Eds. V. Petras, P. Forner, P. Clough, and N. Ferro. Amsterdam, The Netherlands: CEUR-WS. Vol. 1177. Available at: http://ceur-ws.org/Vol-1177/CLEF2011wn-PAN-KestemontEt2011.pdf (accessed September 15, 2017).
9. Potthast, M., A. Eiselt, L. A. Cedeõo, B. Stein, and P. Rosso. 2011. Overview of the 3rd international competition on plagiarism detection. *Working Notes for CLEF Conference*. Eds. V. Petras, P. Forner, P. Clough, and N. Ferro. Amsterdam, The Netherlands: CEUR-WS. Vol. 1177. Available at: http://ceur-ws.org/Vol-1177/CLEF2011wn-PAN-PotthastEt2011a.pdf (accessed September 15, 2017).
10. Fodor, I. K. 2002. A survey of dimension reduction techniques. Center for Applied Scientific Computing, Lawrence Livermore National Laboratory. Technical Report. 1–18.
11. Brooke, J., and G. Hirst. 2012. Paragraph clustering for intrinsic plagiarism detection using a stylistic vector-space model with extrinsic features. *Working Notes for CLEF Conference*. Eds. P. Forner, J. Karlgren, C. Womser-Hacker, and N. Ferro. Rome, Italy: CEUR-WS. Vol. 1178. Available at: http://ceur-ws.org/Vol-1178/CLEF2012wn-PAN-BrookeEt2012.pdf (accessed September 15, 2017).
12. Brooke, J., A. Hammond, and G. Hirst. 2012. Unsupervised stylistic segmentation of poetry with change curves and extrinsic features. *Ist NAACL-HLT Workshop on Computational Linguistics for Literature Proceedings*. Stroudsburg, PA: Association for Computational Linguistics. 26–35.
13. Gorban, A. N., B. Kégl, D. C. Wunsch, *et al*. 2008. *Principal manifolds for data visualization and dimension reduction*. Springer. 58 p.
14. Tenenbaum, J. B., V. De Silva, and J. C. Langford. 2000. A global geometric framework for nonlinear dimensionality reduction. *Science* 290(5500):2319–2323.
15. Belkin, M., and P. Niyogi. 2001. Laplacian eigenmaps and spectral techniques for embedding and clustering. *Advances in neural information processing systems*. Eds. T. G. Dietterich, S. Becker, and Z. Ghahramani. NIPS Foundation, Inc. 14:585–591.
16. Roweis, S. T., and L. K. Saul. 2000. Nonlinear dimensionality reduction by locally linear embedding. *Science* 290(5500):2323–2326.
17. Donoho, D. L., and C. Grimes. 2003. Hessian eigenmaps: Locally linear embedding techniques for high-dimensional data. *P. Natl. Acad. Sci. USA* 100(10):5591–5596.
18. Zhang, Z., and H. Zha. 2004. Principal manifolds and nonlinear dimensionality reduction via tangent space alignment. *J. Shanghai University (English Edition)* 8(4):406–424.
19. Weinberger, K. Q., and L. K. Saul. 2006. Unsupervised learning of image manifolds by semidefinite programming. *Int. J. Comput. Vision* 70(1):77–90.
20. Chen, C., J. Zhang, and R. Fleischer. 2010. Distance approximating dimension reduction of Riemannian manifolds. *IEEE T. Syst. Man Cy. B* 40(1):208–217.
21. Van der Maaten, L. 2009. Learning a parametric embedding by preserving local structure. *RBM* 500:26.
22. Van der Maaten, L. 2014. Accelerating t-SNE using tree-based algorithms. *J. Mach. Learn. Res.* 15(1):3221–3245.
23. Kim, H., H. Park, and H. Zha. 2007. Distance preserving dimension reduction for manifold learning. *SIAM Conference (International) on Data Mining Proceedings*. 527–532.
24. Bottou, L. 2012. Stochastic gradient descent tricks. *Neural networks: Tricks of the trade*. Eds. G. Montavon, G. B. Orr, and K.-R. Muller. Lecture notes in computer science ser. 2nd ed. Berlin–Heidelberg: Springer. 7700:421–436.
25. Potthast, M., B. Stein, A. Barrón-Cedeño, and P. Rosso. 2010. An evaluation framework for plagiarism detection. *23rd Conference (International) on Computational Linguistics Posters*. 997–1005.

Received February 20, 2017

Contributors

Molybog Igor O. (b. 1995) — apprentice researcher, Skolkovo Institute of Science and Technology, Center for Energy Systems, Skolkovo Innovation Center, 3 Nobel Str., Moscow 143026, Russian Federation; student, Moscow Institute of Physics and Technology, 9 Institutskiy Per., Dolgoprudny, Moscow Region 141700, Russian Federation; i.molybog@skoltech.ru

Motrenko Anastasia P. (b. 1992) — PhD student, Moscow Institute of Physics and Technology, 9 Institutskiy Per., Dolgoprudny, Moscow Region 141700, Russian Federation; anastasiya.motrenko@phystech.edu

Strijov Vadim V. (b. 1967) — Doctor of Science in physics and mathematics, leading scientist, A. A. Dorodnicyn Computing Centre, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, 40 Vavilov Str., Moscow 119333, Russian Federation; strijov@ccas.ru

ОПРЕДЕЛЕНИЕ ЗАИМСТВОВАНИЙ В ТЕКСТЕ БЕЗ УКАЗАНИЯ ИСТОЧНИКА*

К. Ф. Сафин[1], М. П. Кузнецов[2], М. В. Кузнецова[3]

Аннотация: Для задачи поиска заимствований в тексте существуют два подхода: обнаружение «внешних» и «внутренних» заимствований. При поиске внешних заимствований известен корпус, из которого возможны заимствования. При поиске внутренних заимствований исследуемый текст анализируется изолированно, т. е. возможные источники заимствований неизвестны. Данная работа посвящена поиску внутренних заимствований в тексте. Предполагается, что большая часть текста написана одним автором. Необходимо выделить участки текста, написанные другим автором, если таковые имеются. В работе предлагается алгоритм, строящий статистику сегментов текста, по которой определяется факт зависимости. Эксперимент проводится на коллекции конкурса PAN-2011.

Ключевые слова: обработка естественного языка; детектирование внутренних заимствований; поиск выбросов в статистике

DOI: 10.14357/19922264170308

1 Введение

Текстовые заимствования являются большой проблемой в сфере образования и научных исследований [1]. Развитие сети Интернет, в частности, и информационных технологий, в целом, сделало возможным некорректное заимствование информации.

В задаче обнаружения заимствований существуют два глобальных подхода: выявление «внешних» (external plagiarism detection) и «внутренних» (intrinsic plagiarism detection) заимствований. При поиске внешних заимствований предполагается, что в распоряжении исследователя есть некоторый корпус, из которого возможны заимствования. Таким образом, задача состоит в попарном сравнении участков подозрительного текста и текстов из корпуса заимствований.

Задача поиска «внутренних» заимствований состоит в анализе исключительно подозрительного текста. Алгоритмы должны анализировать стиль письма и выделять характерные признаки, свойственные данному автору.

Алгоритмы разбивают исходный текст на сегменты и сравнивают текст сегмента со всем текстом. Разбиение проводится по предложениям [2, 3], или же определяется окно заданной ширины, согласно которому производится сегментирование текста [4–7]. Выбор меры схожести сегмента со всем текстом или, наоборот, меры различия является ядром алгоритма. Работы [2–4] используют стилистические, синтаксические, лексические характеристики: частотность частей речи, порядок следования частей речи в предложении, пунктуацию, среднюю длину предложения и подобные признаки. Возможно использование символьных n-грамм (чаще других применяют 3-граммы) в качестве признака, а точнее, частот их использования [5, 7].

Метод [8] использует кластеризацию абзацев по частоте встречаемости существительных.

В статье [9] описан алгоритм диаризации текстов, т. е. классификации сегментов текста по авторству, что является обобщением задачи поиска внутренних заимствований.

В 2011 г. был проведен конкурс PAN-2011, посвященный поиску заимствований в текстах. Метод Oberreuter [6], ставший победителем в конкурсе PAN-2011, использует функцию, характеризующую письменный стиль автора. Функция строит вектор частот встречаемости слов во всем документе и в выделенном сегменте. Эти векторы используются для определения величины отклонения сегмента от всего текста. Данный алгоритм показал результат 0,32 по F1-мере. Это демонстрирует тот факт, что алгоритма, решающего данную задачу в большинстве случаев, до сих пор нет.

Предлагаемый алгоритм строит статистическое описание текста, которое используется для нахождения заимствованных сегментов текста. Статисти-

*Работа поддержана РФФИ (проект 16-07-01155).

[1]Московский физико-технический институт; ЗАО «Анти-плагиат», kamil.safin@phystech.edu

[2]ООО «Форексис», mikhail.kuznecov@phystech.edu

[3]Московский физико-технический институт; ЗАО «Анти-плагиат», kuznetsova@ap-team.ru

ка должна удовлетворять следующим условиям: на оригинальных сегментах текста иметь небольшой разброс значений по сравнению со значением на всем тексте, а на заимствованных сегментах иметь значительные отличия.

В работе используются данные конкурса PAN-2011 [10]. Для оценки работы алгоритма определяются микро- и макромеры качества precision и recall (точность и полнота) и затем вычисляется F1-мера как среднее гармоническое для precision и recall. Данная мера представляет качество работы алгоритма.

2 Постановка задачи

Пусть D — коллекция текстовых документов, d — текстовый документ, t_i — сегмент текста ($d = \bigcup t_i$, $d \in D$). Среди сегментов текста t_i необходимо выделить те, значение статистики которых $\sigma(\mathbf{t}_i)$ превосходит некоторый заданный порог значений δ_{susp}.

Описание выборки. В работе используется блок текстовых документов конкурса PAN-2011 [10]. В текстах присутствуют сегменты настоящих, имитированных и искусственных заимствований. Каждый сегмент текста соответственно полностью взят из другого источника, либо заимствованный текст переписан человеком другими словами, либо специально обученный алгоритм строит текст, стараясь повторить стиль автора.

Выборка состоит из 4753 текстов, разделенных на 10 частей, к каждому из текстов прилагается файл с экспертной разметкой заимствованных сегментов.

Для анализа корпуса была собрана подвыборка корпуса, состоящая из 30 документов, которые были просмотрены вручную. Анализ показал, что большая часть документов содержит в себе заимствования, сильно отличающиеся от остального текста по тематике и набору используемых слов. К примеру, в текст по экономике вставляется фрагмент, вырезанный из художественного текста.

Также тексты корпуса были исследованы на то, какая доля заимствований содержится в каждом тексте (отношение длины заимствованных фрагментов к длине текста в символах) и сколько различных фрагментов заимствований присутствует в тексте. На рис. 1 приведены гистограммы результатов.

Как видно, тексты в среднем содержат от 1 до 7 фрагментов заимствований. В большинстве текстов доля заимствований не превышает 4%–5%, что усложняет задачу поиска этих заимствований.

Критерии качества. В экспериментах используются критерии качества, применявшиеся в PAN-2011 [10]. Обозначим за пару (s, d) последовательность символов, помеченную экспертом как заимствование в документе d. $S = \bigcup s_i$ — совокупность всех заимствованных сегментов. За пару (r, d) обозначим последовательность, помеченную алгоритмом как заимствованную. Аналогично $R = \bigcup r_i$ — совокупность всех сегментов, которые алгоритм классифицировал как заимствованные. Рассмотрим меры качества Precision и Recall:

$$\text{Prec}(S, R) = \frac{1}{|R|} \sum_{r_j \in R} \frac{\left| \bigcup\limits_{s_i \in S} (s_i \cap r_j) \right|}{|r_j|};$$

$$\text{Rec}(S, R) = \frac{1}{|S|} \sum_{s_i \in S} \frac{\left| \bigcup\limits_{r_j \in R} (s_i \cap r_j) \right|}{|s_i|}.$$

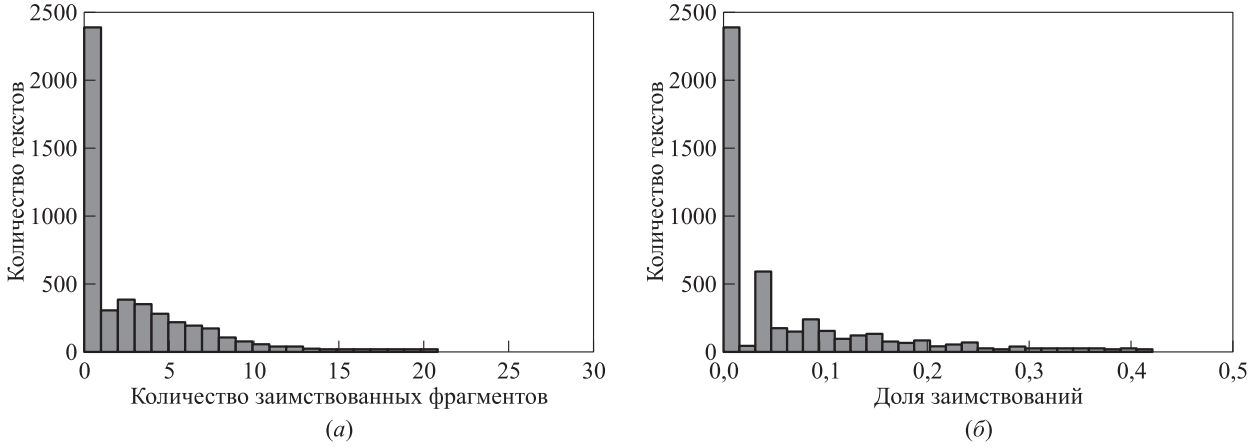


Рис. 1 Распределение текстов по количеству заимствованных фрагментов (*а*) и по доле заимствований (*б*)

Данные величины отражают точность (доля правильного распознавания заимствований по отношению ко всем выделенным сегментам) и полноту (доля правильного распознавания заимствований по отношению ко всем заимствованиям в тексте) работы алгоритма.

Вычисляется F1-мера как среднее гармоническое между Precision и Recall:

$$\mathrm{F1}(S,R) = \frac{\mathrm{Prec}(S,R)\cdot\mathrm{Rec}(S,R)}{\mathrm{Prec}(S,R)+\mathrm{Rec}\,(S,R)}\,.$$

Вычисляется величина гранулярности

$$\mathrm{gran}\,(S,R) = \frac{1}{|S_R|}\sum_{s_i\in S_R}|R_{s_i}|\,,$$

где S_R — множество заимствованных сегментов, обнаруженных алгоритмом; R_s — сегменты, отмеченные алгоритмом, которые детектируют данный сегмент заимствований s:

$$S_R = \{s|s\in S\wedge\exists r\in R : r \text{ detects } s\}\,;$$
$$R_S = \{r|r\in R\wedge r \text{ detects } s\}\,;$$
$$r \text{ detects } s : \text{if } r\cap s\neq\emptyset\,.$$

Таким образом, гранулярность показывает то, насколько мелко алгоритм разбивает заимствованные сегменты текста. Если заимствованные сегменты разделяются алгоритмом на много мелких, то гранулярность будет иметь высокие значения.

По описанным величинам вычисляется итоговая мера качества pladget:

$$\mathrm{pladget}(S,R) = \frac{F1(S,R)}{\log_2(1+\mathrm{gran}(S,R))}\,.$$

Формальная постановка задачи. Для обнаружения заимствований исходный текст d разбивается на сегменты t_i:

$$d = \cup t_i\,.$$

Для каждого сегмента вычисляется вектор признаков $\mathbf{t}_i$ и строится статистика $\sigma(\mathbf{t}_i)$. Затем происходит детектирование выбросов среди значений статистики на основании ее отклонения от среднего значения

$$\sigma_{\mathrm{avr}}(d) = \frac{1}{N}\sum_{i=1}^{N}\sigma(\mathbf{t}_i)\,,$$

где N — число сегментов в тексте. Если отклонение превышает заданный порог δ_{susp}, то сегмент считается заимствованным:

$$|\sigma(\mathbf{t}_i)) - \sigma_{\mathrm{avr}}(d)| > \delta_{\mathrm{susp}}\,.$$

Корпус документов D разбивается на обучающую и тестовую выборки:

$$D = D_{\mathrm{test}}\cup D_{\mathrm{learn}}\,.$$

При обучении параметры алгоритма $\mathbf{w}$ настраиваются таким образом, чтобы улучшить меры качества работы алгоритма. При фиксированном способе разбиения текста мера Granularity не изменяется, так как она зависит от мелкости разбиения. Тогда для увеличения итоговой меры качества Pladget достаточно улучшить F1-меру:

$$\hat{\mathbf{w}} = \arg\max_{\mathbf{w}\in\mathbf{W}}\mathrm{F1}(S,R)\,,$$

т. е. требуется вектор параметров $\mathbf{w} = (l_{\mathrm{segm}}, n, \delta_{\mathrm{susp}})$, максимизирующий F1-меру. Здесь l_{segm} — минимальная длина сегмента; n — ширина окна сглаживания; δ_{susp} — порог выброса. Более подробно параметры описаны в вычислительном эксперименте (см. разд. 5).

3 Базовый эксперимент

Целью базового эксперимента ставилась проверка гипотезы о том, что заимствованные сегменты текста имеют отличные от среднего вектора значения признаков.

В качестве такого признака была выбрана частота встречаемости слов. Каждому слову ставится в соответствие число

$$\mathrm{fr\_class}_w = \log_2\frac{n_{\max}}{n_w}\,, \qquad (1)$$

где $n_{\max}$ — число вхождений наиболее часто употребимого слова в тексте; n_w — частота вхождений слова w в этом предложении.

В качестве основного признака использовались квантили распределения данной величины внутри окна фиксированной ширины.

Обозначим за $m^j = \overline{x^j}$ среднее значение j-го признака для рассматриваемого документа, за r^j — среднеквадратичное отклонение. Тогда нормализованный признак j для сегмента i рассчитывается по формуле:

$$t_i^j = \frac{x^j - m^j}{r^j}\,.$$

За сегменты t_i были выбраны предложения текста. Для каждого предложения t_i строился вектор признаков $\mathbf{t}_i$ и затем подсчитывалось отклонение от усредненного по всему тексту вектора $\mathbf{t}_{\mathrm{avr}}$ в L1-метрике:

$$\sigma(\mathbf{t}_i) = ||\mathbf{t}_i - \mathbf{t}_{\mathrm{avr}}|| = \sum_{j=1}^{l}|t_i^j - t_{\mathrm{avr}}^j|\,. \qquad (2)$$

Эксперимент проводился на одном из текстов конкурсной коллекции PAN-2011.

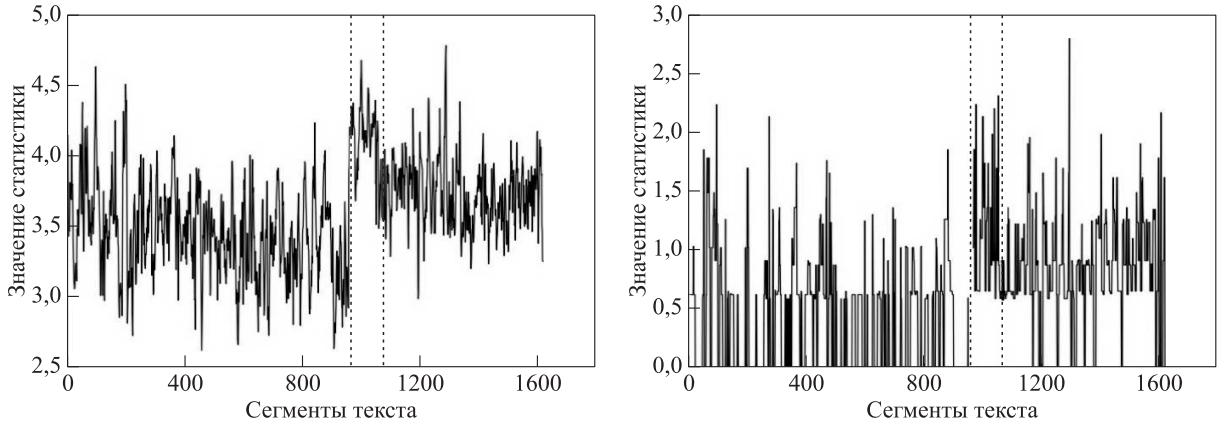


Рис. 2 Отклонение признакового вектора от среднего

На рис. 2 показано отклонение признакового вектора каждого предложения от усредненного вектора. Пунктирными линиями выделены предложения, помеченные экспертом как заимствованные.

Видно, что заимствованные фрагменты имеют характерные выбросы из области средних значений отклонения. Однако некоторые предложения, не являющиеся заимствованными, также сильно отличаются от усредненного признакового вектора. На основании этого можно сделать вывод, что использование только данного признака недостаточно для решения поставленной задачи.

4 Описание алгоритма

Модель. Предлагаемый алгоритм работает с частотными признаками, предоставляющими описание текста. В качестве такого признака выбран признак частоты встречаемости слов, описанный в формуле (1).

Исходный текст подвергается предобработке: удаляются служебные символы, все буквы переводятся в нижний регистр. Также из текста удаляются стоп-слова.

Сегментирование текста. Текст разбивается на предложения. Затем формируется разбиение текста на сегменты t_i: если длина очередного предложения меньше минимальной длины сегмента l_{segm}, к этому предложению добавляется следующее за ним — процесс повторяется, пока длина сегмента t_i не превысит заданную минимальную длину. Минимальная длина сегмента l_{segm} является настраиваемым параметром алгоритма.

Построение статистики и детектирование аномалий. Для каждого сегмента t_i текста строится вектор признаков. Затем строится статистика $\sigma(\mathbf{t}_i)$ на основе отклонения вектора признаков от усредненного по всему тексту вектора (2).

Полученная статистика сглаживается методом скользящего среднего: новые значения статистики $\sigma'(\mathbf{t}_i)$ вычисляются по формуле:

$$\sigma'(\mathbf{t}_i) = \frac{1}{2n+1} \sum_{k=i-n}^{i+n} \sigma(\mathbf{t}_i)\,,$$

где n — ширина сглаживания, которая также является настраиваемым параметром. Значения в крайних точках вычисляются по формулам (N — число сегментов):

$$\sigma'(\mathbf{t}_i) = \frac{1}{i+n+1} \sum_{k=0}^{i+n} \sigma(\mathbf{t}_i)\,;$$

$$\sigma'(\mathbf{t}_i) = \frac{1}{i+n+1} \sum_{k=i-n}^{N} \sigma(\mathbf{t}_i)\,.$$

Полученные значения статистики $\sigma'(\mathbf{t}_i)$ исследуются на выбросы. Если в ряде статистики присутствует аномалия, превышающая заданный порог δ_{susp}, то сегмент t_i, отвечающий этому выбросу, помечается как заимствованный.

Минимальная длина сегмента, ширина окна сглаживания и порог выброса настраиваются на обучающей выборке путем максимизации F1-меры.

5 Вычислительный эксперимент

Алгоритм настраивался на частях 1–5 корпуса PAN-2011 путем максимизации F1-меры. Тестирование проводилось на частях 6–10 корпуса. Оптимальные параметры после настройки: $\hat{l}_{\text{segm}} = 450$; $\hat{n} = 8$; $\hat{\delta}_{\text{susp}} = 0{,}37$.

На рис. 3 и 4 приведены примеры работы алгоритма. Серые участки обозначают заимствованные фрагменты, пунктирными линиями обозначен порог выброса значений стилевой функции.

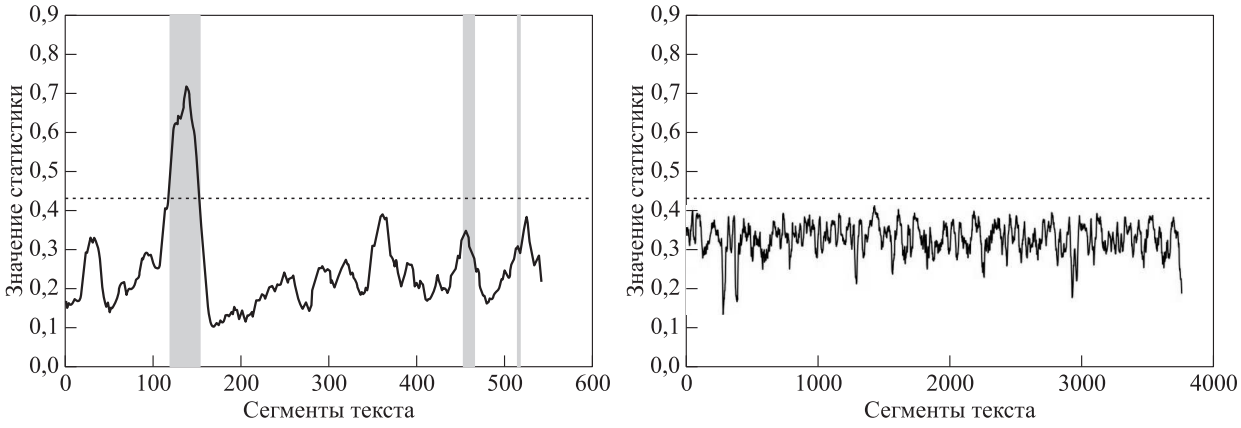


Рис. 3 Результаты на обучающей выборке

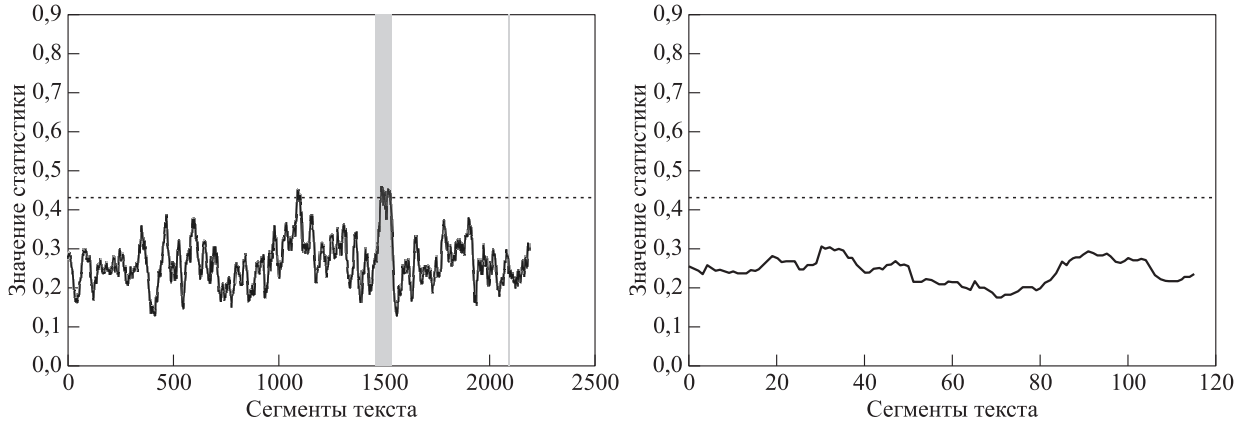


Рис. 4 Результаты на тестовой выборке

Сравнение качества алгоритмов на корпусе PAN-2011

Алгоритм	Precision	Recall	F1	Granularity	Pladget
Предлагаемый авторами	0,27	0,28	0,28	1,04	0,28
Oberreuter	0,34	0,31	0,33	1,00	0,33
Kestemont	0,11	0,43	0,17	1,03	0,17

Результаты работы и сравнение с двумя алгоритмами приведены в таблице.

Описанный алгоритм использует частоты распределения слов. Сегментирование текста происходит по группам предложений. Для определения заимствованных фрагментов исследуются значения статистики каждого сегмента.

На корпусе PAN-2011 алгоритм показал сравнимые результаты с победителем конкурса — алгоритмом Oberreuter. Также было проведено сравнение с алгоритмом Kestemont'a, занявшим второе место на конкурсе. Качество работы предлагаемого алгоритма значительно превышает качество работы алгоритма Kestemont'a.

6 Анализ ошибок

Результаты работы предлагаемого алгоритма зависят от длины документа. При анализе небольших по объему текстов сглаживание приводит к существенной потере информации об аномальных значениях статистики. При малой ширине сглаживания шумовые выбросы вызывают ложное срабатывание алгоритма.

7 Заключение

Предлагаемый алгоритм использует распределение частот слов внутри текста для нахождения

заимствованных сегментов. Сегментирование текста осуществляется по группам предложений. Для каждого сегмента строится статистика. Затем ряд статистики для всего текста сглаживается методом скользящего среднего. Полученные значения исследуются на отклонение от среднего значения для выявления заимствованных сегментов.

Алгоритм был настроен и протестирован на корпусе PAN-2011. Алгоритм Oberreuter [6], модификацией которого является предлагаемый алгоритм, показал на этом же корпусе результаты в 0,32 по F1-мере. Таким образом, описанный алгоритм показал сравнимые результаты при работе с тем же корпусом документов.

Дальнейшие исследования могут быть направлены на более точную настройку параметров алгоритма, подбор параметров в зависимости от длины рассматриваемого текста, поиск новых признаков, которые будут точнее выявлять заимствованные фрагменты, а также поиск новых способов нахождения заимствований.

Авторы выражают свою благодарность доктору физико-математических наук В. В. Стрижову, а также кандидату физико-математических наук Ю. В. Чеховичу за ценные советы при планировании исследования и рекомендации по оформлению статьи.

Литература

1. *Никитов А. В., Орчаков О. А., Чехович Ю. В.* Плагиат в работах студентов и аспирантов: проблема и методы противодействия // Университетское управление: практика и анализ, 2012. № 5. С. 61–68.
2. *Zechner M., Muhr M., Kern R., Granitzer M.* External and intrinsic plagiarism detection using vector space models // CEUR Workshop Proceedings, 2009. Vol. 502. P. 47–55.
3. *Tschuggnall M., Specht G.* Countering plagiarism by exposing irregularities in authors grammars // European Intelligence and Security Informatics Conference. — IEEE, 2013. P. 15–22.
4. *Eissen S. M., Stein B.* Intrinsic plagiarism detection // Advances in information retrieval / Eds. M. Lalmas, A. MacFarlane, S. M. Rüger, *et al.* — Lecture notes in computer science ser. — Springer, 2006. Vol. 3936. P. 565–569.
5. *Stamatatos E.* Intrinsic plagiarism detection using character n-gram profiles // CEUR Workshop Proceedings, 2009. Vol. 502. P. 38–46.
6. *Oberreuter G., L'Huillier G., Ríos S. A., Velásquez J. D.* Outlier-based approaches for intrinsic and external plagiarism detection // Knowlege-based and intelligent information and engineering systems / Eds. A. König, A. Dengel, K. Hinkelmann, *et al.* — Lecture notes in computer science ser. — Springer, 2011. Vol. 6882. P. 11–20.
7. *Bensalem I. Rosso P., Chikhi S.* Intrinsic plagiarism detection using n-gram classes // Conference on Empirical Methods in Natural Language Processing Proceedings. — Stroudsburg, PA, USA: Association for Computational Linguistics, 2014. P. 1459–1464.
8. *Vartapetiance A., Gillam L.* Quite simple approaches for authorship attribution, intrinsic plagiarism detection and sexual predator identification. http://epubs.surrey.ac.uk/id/eprint/766727.
9. *Kuznetsov M., Motrenko A., Kuznetsova R., Strijov V.* Methods for intrinsic plagiarism detection and author diarization. http://ceur-ws.org/Vol-1609/16090912.pdf.
10. *Potthast M., Stein B., Barron-Cedeno A., Rosso P.* An evaluation framework for plagiarism detection // 23rd Conference (International) on Computational Linguistics Proceedings. — Stroudsburg, PA, USA: Association for Computational Linguistics, 2010. P. 997–1005.

Поступила в редакцию 30.01.17

METHODS FOR INTRINSIC PLAGIARISM DETECTION

K. F. Safin[1,2], M. P. Kuznetsov[3], and M. V. Kuznetsova[1,2]

[1]Moscow Institute of Physics and Technology, 9 Institutskiy Per., Dolgoprudny, Moscow Region 141700, Russian Federation
[2]Antiplagiat JSC, 33 Varshavskoe Shosse, Moscow 117105, Russian Federation
[3]"Forecsys" LLC, 42 Vavilov Str., Moscow 119333, Russian Federation

Abstract: There are two ways to find plagiarism in documents: "external" and "intrinsic" plagiarism detection. External plagiarism detection is the task with a known set of possible references. Intrinsic plagiarism detection aims at discovering plagiarism by analyzing only the document by itself. The paper investigates the methods of intrinsic plagiarism detection. The authors developed a plagiarism detection method based on constructing statistics from the features of the document parts and detecting outliers. The proposed algorithm was tested on the PAN-2011 collection for intrinsic plagiarism detection.

Keywords: natural language processing; intrinsic plagiarism detection; outliers detection

DOI: 10.14357/19922264170308

Acknowledgments

The work was supported by the Russian Foundation for Basic Research (project 16-07-01155).

References

1. Nikitov, A. V., O. A. Orchakov, and Ju. V. Chehovich. 2012. Plagiat v rabotakh studentov i aspirantov: Problema i metody protivodeystviya [Plagiarism in works of undergraduate and graduate students: Problem and methods of counteraction]. *Universitetskoe upravlenie: Praktika i analiz* [University Management: Practice and Analysis] 5:61-68.
2. Zechner, M., M. Muhr, R. Kern, and M. Granitzer. 2009. External and intrinsic plagiarism detection using vector space models. *CEUR Workshop Proceedings*. 502:47–55.
3. Tschuggnall, M., and G. Specht. 2013. Countering plagiarism by exposing irregularities in authors grammars. *European Intelligence and Security Informatics Conference Proceedings*. IEEE. 15–22.
4. Eissen, S. M., and B. Stein. 2006. Intrinsic plagiarism detection. *Advances in information retrieval*. Eds. M. Lalmas, A. MacFarlane, S. M. Rüger, *et al.* Lecture notes in computer science ser. Springer. 3936:565–569.
5. Stamatatos, E. 2009. Intrinsic plagiarism detection using character n-gram profiles. *CEUR Workshop Proceedings*. 502:38–46.
6. Oberreuter, G., G. L'Huillier, S. Ríos, and J. Velásquez. 2011. Outlier-based approaches for intrinsic and external plagiarism detection. *Knowlege-based and intelligent information and engineering systems*. Eds. A. König, A. Dengel, K. Hinkelmann, *et al.* Lecture notes in computer science ser. Springer. 6882:11–20.
7. Bensalem, I., P. Rosso, and S. Chikhi. 2014. Intrinsic plagiarism detection using n-gram classes. *Conference on Empirical Methods in Natural Language Processing*. Stroudsburg, PA: Association for Computational Linguistics. 1459–1464.
8. Vartapetiance, A., and L. Gillam. Quite simple approaches for authorship attribution, intrinsic plagiarism detection and sexual predator identification. Available at: http://epubs.surrey.ac.uk/id/eprint/766727 (accessed September 23, 2013).
9. Kuznetsov, M., A. Motrenko, R. Kuznetsova, and V. Strijov. Methods for intrinsic plagiarism detection and author diarization. Available at: http://ceur-ws.org/Vol-1609/16090912.pdf (accessed September 6, 2016).
10. Potthast, M., B. Stein, A. Barrón-Cedeño, and P. Rosso. 2010. An evaluation framework for plagiarism detection. *23rd Conference (International) on Computational Linguistics Proceedings*. Stroudsburg, PA: Association for Computational Linguistics. 997–1005.

Received January 30, 2017

Contributors

Safin Kamil F. (b. 1995) — student, Moscow Institute of Physics and Technology, 9 Institutskiy Per., Dolgoprudny, Moscow Region 141700, Russian Federation; junior researcher, Antiplagiat JSC, 33 Varshavskoe Shosse, Moscow 117105, Russian Federation; kamil.safin@phystech.edu

Kuznetsov Mikhail P. (b. 1989) — Candidate of Science (PhD) in physics and mathematics; analyst, "Forecsys" LLC, 42 Vavilov Str., Moscow 119333, Russian Federation; mikhail.kuznecov@phystech.edu

Kuznetsova Margarita V. (b. 1990) — PhD student, Moscow Institute of Physics and Technology, 9 Institutskiy Per., Dolgoprudny, Moscow Region 141700, Russian Federation; Head of Departament, Antiplagiat JSC, 33 Varshavskoe shosse, Moscow 117105, Russian Federation; kuznetsova@ap-team.ru

ПСИХОЛИНГВИСТИЧЕСКИЙ АНАЛИЗ РУССКОЯЗЫЧНЫХ ТЕКСТОВЫХ СООБЩЕНИЙ НА ОСНОВЕ ИХ ФОНОСЕМАНТИЧЕСКИХ СТАТИСТИЧЕСКИХ ХАРАКТЕРИСТИК*

А. С. Сигов[1], Д. А. Акимов[2], Д. О. Жуков[3], Е. Г. Андрианова[4], В. Е. Сачков[5], В. К. Раев[6]

Аннотация: Рассматривается проблема идентификации типа акцентуации паттерна поведения виртуального субъекта в сети Интернет и социальных сетях на основе статистического анализа текстов, что позволяет сформулировать гипотезу о структурных свойствах его коммуникаций и позволяет построить матрицу вероятностей для отношений между виртуальными масками субъектов. Тексты пользователей рассматриваются как сложные семантико-синтаксические образования, обладающие рядом психолингвистических характеристик. К их числу относятся цельность, а также смысловая направленность сообщения. Кроме того, в тексте, рассматриваемом как продукт речевой деятельности, обладающий большой степенью семантической вариативности, определяемой его темпоральными и сонарными характеристиками, проявляется невербальный характер поведения сетевых субъектов — виртуальных масок и роботизированных агентов. Практическая значимость предлагаемого решения для психолингвистического анализа строится на возрастающем значении развития системы условных знаков, в данном случае условных языков е-коммуникации, для порождения, в свою очередь, управляющих кластеров, регулирующих социальное поведение виртуальных субъектов в Сети. Это предположение строится на гипотезе Кеннета Айверса, в соответствии с которой чем лучше развита система условных знаков, тем больше возможностей она дает для создания новых алгоритмов.

Ключевые слова: психолингвистические характеристики; невербальное поведение; виртуальные маски; процесс мышления; семантический смысл; лингвистический релятивизм

DOI: 10.14357/19922264170309

1 Введение

В настоящее время большой интерес разработчиков информационных сетей вызывает анализ социального аспекта информационного массива (потока) популярных интернет-ресурсов, воздействие которых изменяет семантический смысл и оказывает управляющее воздействие на виртуальных субъектов, равно как и на их кластеры непосредственного взаимодействия. К такой информации относятся данные, отражающие мнения, тенденции, настроения и интересы, преобладающие среди субъектов е-сообщества.

На взгляд авторов, решение таких задач возможно только за счет использования междисциплинарных подходов, в которых методы теоретической информатики должны быть дополнены моделями математической лингвистики естественных языков.

Теоретическим обоснованием рассматриваемой проблемы является гипотеза лингвистической относительности, которая предполагает, что структура языка коммуникации влияет на ментальность пользователей социальных сетей [1] и опосредованно на когнитивные процессы мышления последних.

Воспользуемся нечеткой трактовкой гипотезы Сепира–Уорфа [2], в соответствии с которой процессы мышления, а также используемые в письменной/устной речи лингвистические категории определяются при е-коммуникации как некая форма неязыкового поведения.

*Работа выполнена за счет финансирования Министерством образования и науки Российской Федерации конкурсной части государственных заданий высшим учебным заведениям и научным организациям по выполнению инициативных научных проектов (№ 28.2635.2017/ПЧ).

[1]Московский технологический университет (МИРЭА), assigov@yandex.ru
[2]Московский технологический университет (МИРЭА), akimov_d@mirea.ru
[3]Московский технологический университет (МИРЭА), zhukovdm@yandex.ru
[4]Московский технологический университет (МИРЭА), dtghmflysq@gmail.com
[5]Московский технологический университет (МИРЭА), megawatto@mail.ru
[6]Московский технологический университет (МИРЭА), raev@mirea.ru

Используемый принцип Уорфа, равно как и позднее сформулированная гипотеза Р. Брауна и Э. Леннеберга [3] в отношении цветового восприятия, определяющая разницу в восприятии цветового зрения в различных языках, носит релятивистский характер, равно как и конструктивистский подход, предполагающий, что свойства проявления черт человеческой психики и общие идеи самопроявления в коммуникации в значительной степени подвержены влиянию категорий, сформированных субъектами в процессе социализации, и не зависят от биологических ограничений.

В антологии [4] исследователи лингвистического релятивизма сделали попытку определить связи и границы между мышлением, познанием, языком и культурой, описать степень и виды взаимосвязанности и взаимовлияния. Слобин [5] задавал когнитивный процесс «мышление для речи» как вид процесса, в котором перцептивные данные и другие виды долингвистического мышления переводятся в лингвистические категории для коммуникации с другими субъектами.

Джон Люси выделил основные направления исследований лингвистического релятивизма, и в том числе «областной» подход. При этом подходе выбирается отдельная семантическая область и сравнивается у различных лингвистических и культурных групп (в данном случае групп пользователей и отдельных виртуальных субъектов) с целью обнаружения корреляции между лингвистическими средствами, которые используются в языке для обозначения тех или иных понятий, и характером поведения. С помощью комбинации вышеуказанных подходов и теоретических положений был проведен расчет квалиметрических характеристик процесса коммуникации в сети Интернет на основе анализа отношения (матрицы отношений) виртуальных идентичностей пользователей сети Интернет к тем или иным событиям, явлениям и персонам (субъектам социальной значимости) реального мира. Также учитывалась степень взаимовлияния виртуальных идентичностей или групп идентичностей.

В рамках проведенного исследования проверялась гипотеза о том, что виртуальная идентичность формируется на основе совокупности отношений пользователя к тем или иным сетевым событиям и является формой проявления отношений пользователей между собой, а также доступной информации или источниками информации, представленными в Сети.

Для решения задачи идентификации поведения виртуальной идентичности моделировался некий процесс, в котором пользователь, участвуя во всех информационных взаимодействиях, условно проявляет свои личностные качества посредством из-

Рис. 1 Методологические конструкты и инструментарий анализа сетевых событий

бранного паттерна и маски поведения. Исходя из этого, строилась модель «псевдоличности» (или виртуальный образ) с последующей ее идентификацией в одном или нескольких кластерах сети Интернет. В исследовании была использована идея дополнения лингвистического анализа сетевого поведения субъектов статистическим анализом (рис. 1), выбраны методы анализа, определена область практического использования результатов, обоснована репрезентативность разработанной методики анализа акцентуации виртуальных субъектов.

Классические методы изучения е-коммуникации базируются на семантическом анализе получаемого от пользователей Сети текстового образа взаимодействия или их поведенческого проявления. Ставится задача определения статистически релевантных характеристик языковой среды коммуникации и структурных свойств среды. В дальнейшем при достаточно большом с точки зрения репрезентативности результата числе измерений можно строить матрицу вероятностей для оценки свойств коммуникации между виртуальными масками субъектов.

Областью применения предложенной методики анализа семантического контента на основе формирования словарей «окраски текста» или акцентуации е-коммуникации могут стать аналитические BI (Business Intelligence) запросы в экономических исследованиях, выявление характеристик поведения субъектов Сети в социологии, оценка собы-

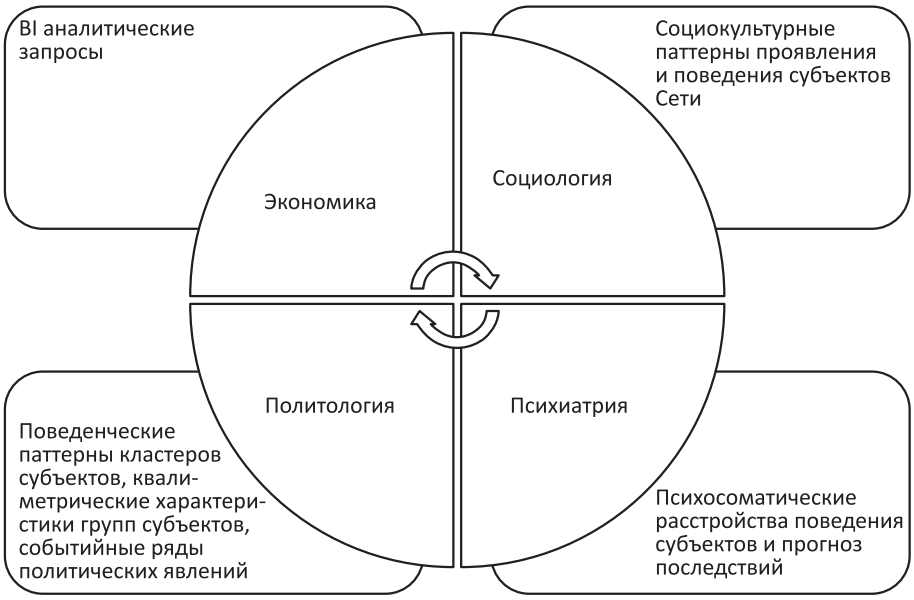


Рис. 2 Проблемный репертуар практических моделей событий Сети

тийных рядов в политологии и анализ расстройств поведения в психиатрии.

Особенности е-коммуникаций социальных сетей и блогов затрудняют получение однозначного результата при изучении только лингвистической составляющей текстов обмена. В исследовании предлагается дополнить существующие подходы, основанные на использовании лингвистического анализа текстов, рядом психолингвистических инструментов (акцентуация) анализа, учитывая при этом их статистическую репрезентативность. Репрезентативность предложенной методики достигается высоким уровнем диверсификации типов виртуальной коммуникации (рис. 2) и ее разнородностью.

По грубым оценкам, контент за единицу времени в 1 с пополняется на 7820 твитов, 1381 графическое изображение, 1558 аудиозвонков, 45 861 запрос Google+; 2 340 000 email, включая 67% спама [6]. В Интернете циркулирует колоссальный объем персональной информации, как правило, текстового содержания в знаковой, образной и звуковой форме (статистические оценки потоков которой показаны на рис. 3), при этом анализу подлежит не просто текст как некий семантический контент, а характер восприятия его физическим объектом либо роботизированным устройством.

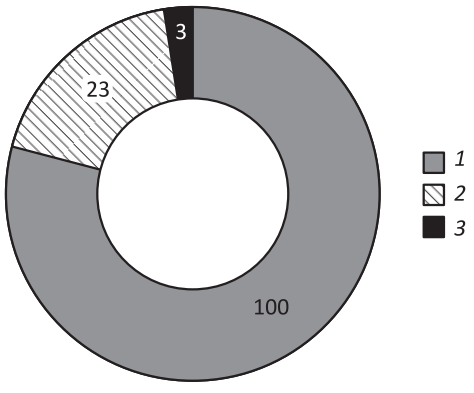


Рис. 3 Приблизительные оценки персональной е-информации в сети Интернет: *1* — общий поток данных; *2* — семантически весомые; *3* — используются часто

2 Методика применения психолингвистического анализа е-коммуникаций на основе словарей окраски текста

Предметом анализа в предлагаемой методике выявления акцентуации виртуальной коммуникации является вербальное и невербальное речевое поведение, сонарные и темпоральные характеристики речевого поведения, лексико-морфологический характер проявления виртуального субъекта.

Для расчета показателей психолингвистических свойств текста используем фоносемантический анализ [7], при этом вычислим процентное совпадение со словарем окраски текста.

Так как сообщение является входной информацией, представленной в виде набора слов, то можно выявить процентное совпадение данного сообщения со словарем [8], используя формулу:

$$S_i = \frac{N_i}{N}, \qquad (1)$$

где S_i — частота появлений некоторой i-й словоформы; N — общее число слов или словосочетаний, встреченных в исследуемом сообщении; i — данная словоформа; N_i — число вхождений данной словоформы во множество всех встреченных слов из словарей.

Примем во внимание, что каждый звук человеческой речи, или фонема, обладает определенным подсознательным значением. Для русского языка эти значения в свое время определил советский ученый, доктор филологических наук А. П. Журавлев [9], который предложил свой вариант фоносемантических значений для каждого звука, или фонемы, русской речи по 25 шкалам. Всем фонемам русского языка по этим шкалам сопоставлены оценки. Для оценки воздействия на человека слова как набора звуков необходимо по соответствующим расчетам определить общее фоносемантическое значение составляющих данное слово звуков по шкалам, разбитым на 3 группы (рис. 4).

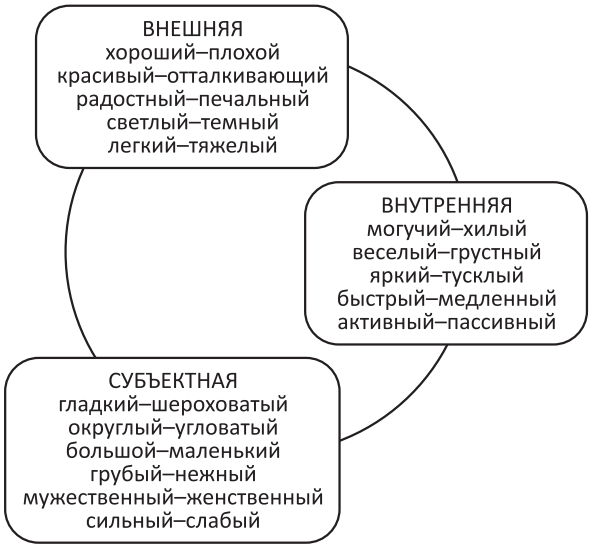


Рис. 4 Фоносемантическое соответствие звуков по шкалам

Данные группы составлены с учетом их частотного использования при анализе взаимодействия виртуальных идентичностей [10] относительно внешних и внутренних критериев оценки.

На основе оценки общего эмоционального состояния виртуальной сущности (субъекта) по отношению к порожденному им тексту выделим следующие группы:

(1) психолингвистические показатели эмоциональной напряженности;

(2) вербальные средства выражения эмоционального напряжения;

(3) вербальные средства выражения мотивационного напряжения.

В общем случае анализ тональности и темпоральности относят к области компьютерной лингвистики, т. е. подразумевается, что можно классифицировать тональность и темпоральность, используя стандартные инструменты обработки естественного языка по типам организации обработки: (1) подходы, основанные на правилах; (2) подходы, основанные на словарях; (3) машинное обучение с учителем; (4) машинное обучение без учителя.

В данной статье приоритет отдан методам, основанным на использовании словарей.

3 Методы анализа текстовых сообщений, основанные на правилах и словарях

Этот метод основан на поиске эмотивной лексики (лексической тональности) в тексте по заранее составленным тональным словарям и правилам с применением лингвистического анализа [11]. По совокупности найденной эмотивной лексики текст может быть оценен по шкале, выражающей объем негативной и позитивной лексики. Данный метод может использовать как списки правил, подставляемые в регулярные выражения, так и специальные правила соединения тональной лексики внутри предложения. Чтобы проанализировать текст, можно воспользоваться следующим алгоритмом: сначала каждому слову в тексте присвоить его значение тональности из словаря (если оно присутствует в словаре), а затем вычислить общую тональность всего текста путем суммирования значения тональностей каждого отдельного предложения.

Основной проблемой методов, основанных на словарях и правилах, считается трудоемкость процесса составления словаря. Для того чтобы получить метод, классифицирующий документ с высокой точностью, термины словаря должны иметь вес, адекватный предметной области документа. Например, слово «огромный» по отношению к объему памяти жесткого диска является положительной характеристикой, но отрицательной по отношению к размеру мобильного телефона. Поэтому данный метод требует значительных трудозатрат, так как для хорошей работы системы необходимо составить большое число правил. Чтобы ускорить процесс составления словарей и правил, данный метод

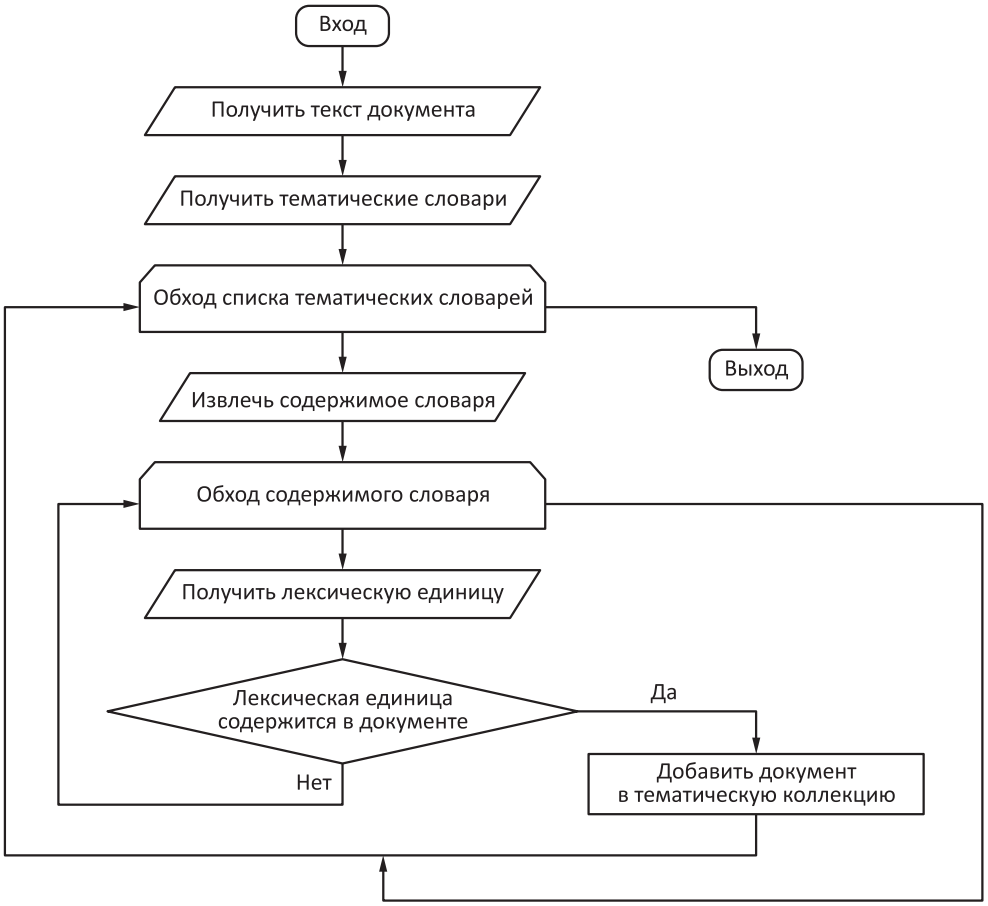


Рис. 5 Алгоритм формирования тематических словарей-тезаурусов

Критерии отбора значимых характеристик [7], или маски акцентуации, виртуального субъекта по типам

Маркер проявления	Характеристика маски виртуального субъекта	Алгоритм расчета
«Светлые», или паранойяльная акцентуация поведения	Тексты мажорной окраски, проявление искусственного оптимизма (все идет «хорошо», «по плану»), «мессианский комплекс»	
«Темные», или эпилептоидная акцентуация поведения	Тексты с большим объемом обсуждения насилия, описания патологической жестокости, выраженным противостоянием «мы и они», «добро и зло» и т. д.	$K_{\text{т}} = \dfrac{\text{число темных словосочетаний}}{\text{число слов}}$
«Печальные», или депрессивная акцентуация поведения	Тексты меланхолического настроения, часто связанные с быстротечностью жизни, с тем, что жизнь — страдания и только смерть способна положить им конец	
«Веселые», или гипертимическая акцентуация поведения	Тексты, представляющие собой описание поведения человека, который сталкивается с препятствиями или опасностями, но успешно преодолевает их и достигает успеха	
«Сложные», или шизотимная акцентуация поведения	Тексты, наполненные философскими понятиями, абстракциями и усложнениями	
«Красивые», или истероидная акцентуация поведения	Тексты с нарочитым описанием эмоциональных аффектов, страстей, страдания и эротизма	

используется с привязкой к конкретной предметной области (например, тематика ресторанов или тематика мобильных телефонов).

Критериями отбора значимых характеристик, или масками акцентуации, виртуального субъекта по типам (см. таблицу): паранойяльной, эпилептоидной, депрессивной, гипертимической и др. — служат агрегации выделенных квалиметрических показателей [12]. Упрощенно тональный словарь представляет собой список слов со значением тональности для каждого слова. Чтобы проанализировать текст, можно воспользоваться следующим алгоритмом (рис. 5): сначала каждому слову в тексте присваивается его значение тональности из словаря (если присутствует), а затем вычисляется общая тональность всего текста путем нахождения усредненных величин либо путем «обучения» классификатора (например, нейронной сети).

Примечание. Под *паранойяльной акцентуацией* подразумевается повышенная подозрительность и болезненная обидчивость, стойкость отрицательных аффектов, стремление к доминированию, непринятие мнения другого и, как следствие, высокая конфликтность, подпадание под власть сверхценных идей и стремление к навязыванию своего мнения. *Эпилептоидный тип акцентуации* связан с такими чертами, как склонность к злобно-тоскливому настроению, раздражительности, агрессивности, внутренней неудовлетворенности, злости, гнева, ярости, жестокости и конфликтности. Личности *с депрессивным типом акцентуации* проявляют лабильность ко всякого рода неприятностям, проявляют неопределенное чувство тяжести, ожидание несчастья. *Гипертимическая акцентуация* характеризуется специфическим поведением, связанным со сменой идей, проявлением словесной ловкости, изворотливости, с направленностью на большое число социальных контактов и отражающим повышенный настрой. Для *демонстративной, или истероидной, акцентуации* характерна поверхностность, наигранность переживаний, «работа на публику», стремление вызвать у аудитории эмоциональный отклик любой ценой, непродуманность речевого поведения. *Шизотимность* поведения виртуального субъекта проявляется в направленности высказываний на себя, замкнутостью на узкий круг вопросов, *акцентуации* на внутреннем мире.

4 Описание методики эксперимента

Шаг 1. По открытым публикациям в социальных сетях определяем архитектуру программных средств и совокупности словарей акцентуации.

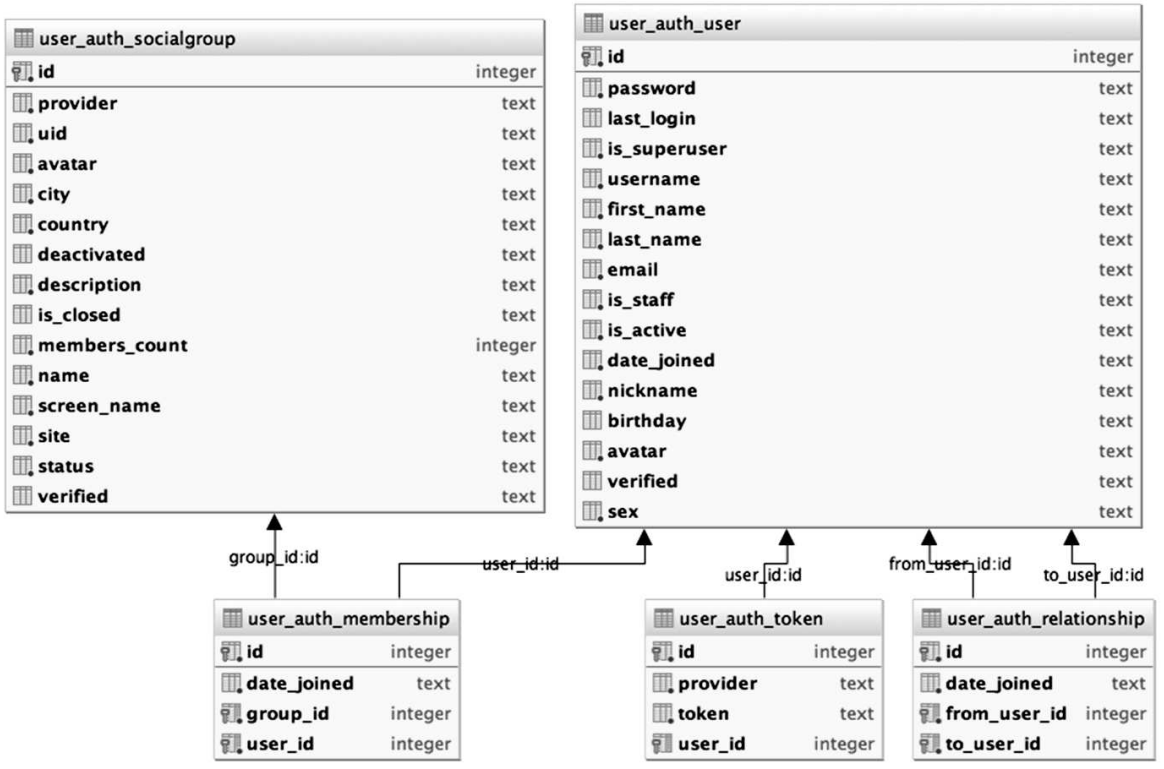


Рис. 6 Архитектура экспериментальной платформы программной системы анализа акцентуации (тональности) новостных групп сообщений пользователей сети «ВКонтакте»

Исследование проводилось на примере социальной сети vk.com с использованием API IBM Watson Tone Analyzer (библиотеки анализа тональности текста IBM), Emotion Analysis (библиотеки анализа эмотивности, или эмоциональной акцентуации, текста IBM), программной библиотеки VK.API (системы для разработчиков сторонних сайтов, которая предоставляет возможность легко авторизовать пользователей «ВКонтакте»), шины сообщений RabbitMQ (платформы, реализующей систему обмена сообщениями между компонентами программной системы, Message Oriented Middleware) и Visual Recognition (библиотеки распознавания изображений).

Архитектура программного решения проведения экспериментов представлена на рис. 6. Для его создания был использован язык программирования Python и язык разработки скриптов ECMAScript.

Эмотивность текста оценивалась с использованием сервиса Text to Speech (TTS).

Шаг 2. Для экспериментальных исследований была выбрана группа Новости RT социальной сети «ВКонтакте». Исследовалась тональность (как доминирующая акцентуация, см. таблицу) комментариев пользователей. Выбор данного сетевого ресурса обусловлен простым алгоритмом его работы, в том числе и для неопытного пользователя, что обусловлено использованием API социальной сети «ВКонтакте». Пользователь сетевого ресурса вводит свой идентификационный номер либо короткое доменное имя, после чего ресурс получает доступ ко всей текстовой информации со стены пользователя, выбирает ключевые слова и позволяет проанализировать данные на основе словарей, маркирующих акцентуацию сообщений, предоставляемых сервисом (http://Indico.io).

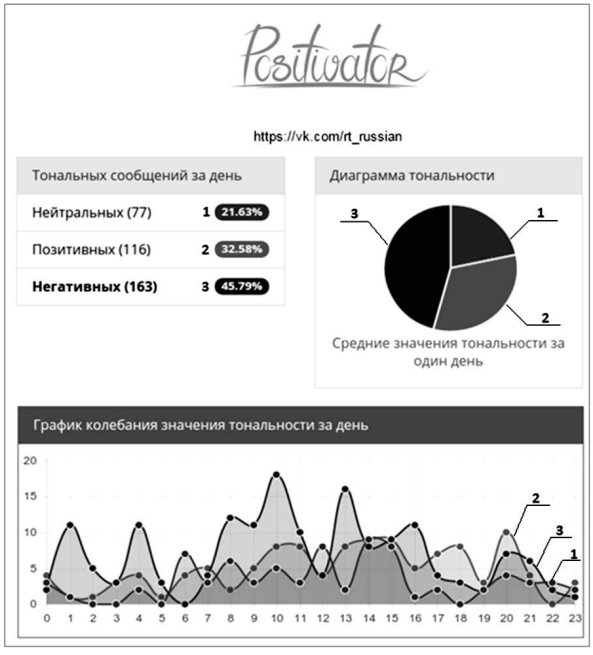


Рис. 7 Оценка тональности (акцентуации) комментариев за день

На следующем шаге рассчитываются суммарные значения позитивных/негативных слов в сообщениях пользователя в процентном соотношении и ставится в соответствие эмоциональный маркер состояния пользователя в текущий момент времени (рис. 7).

Шаг 3. В результате проведенного эксперимента по определению тональности текстов виртуальных субъектов социальной сети «ВКонтакте» на дату обращения были получены следующие результаты: негативных комментариев — 45,79%; позитивных комментариев — 32,58%; нейтральных комментариев — 21,63%.

Негативные тексты объединяют в себе две категории виртуальных субъектов: с депрессивной и эпилептоидной акцентуацией поведения, т. е. тексты с большим объемом обсуждения насилия, описания патологической жестокости, выраженным противостоянием, а также тексты меланхолического настроения.

К *нейтральным* были отнесены тексты виртуальных субъектов с паранойяльной и шизотимной акцентуацией — это по большей части тексты, наполненные философскими понятиями, абстракциями и усложнениями, и тексты мажорной окраски, проявление искусственного оптимизма.

Позитивные тексты получены от виртуальных субъектов с гипертимической акцентуацией речевого поведения, и отчасти сюда были отнесены тексты субъектов с проявлением истероидной акцентуации, иначе это тексты, представляющие собой описание поведения человека, который сталкивается с препятствиями или опасностями, но успешно преодолевает их и достигает успеха, и частично тексты с нарочитым описанием эмоциональных аффектов [13].

Практическое значение разработанной методики заключается в следующем: используя предложенную архитектуру предложенной установки вычислительной сети, можно сформировать карту поведенческих речевых паттернов (см. рис. 7) отдельной социальной сети для проведения мониторинга напряженности социальных настроений на основе тональности текстов произвольных виртуальных идентичностей. Следует заметить, что методом оценки тональности можно варьировать

в зависимости от совокупности выбираемых семантических словарей, при этом диапазон изменений по результатам проведенного эксперимента демонстрировал незначительную выборочную сенситивность, причем без потери значимости для одного и того же текста при использовании иной эталонной коллекции.

5 Заключение

В исследовании в рамках практического эксперимента решена задача идентификации типа акцентуации паттерна поведения виртуального субъекта на основе статистического анализа текста коммуникации и подтверждена гипотеза о структурных свойствах заданной коммуникации. Актуальность предложенной методики определяется еще и тем, что решение поставленной задачи учитывает возрастающее значение развития системы условных знаков, в данном случае условных языков е-коммуникации и управляющих кластеров, позволяющих осуществлять мониторинг и создавать основу для регулирующих воздействий на социальное поведение виртуальных субъектов в Сети.

Результаты экспериментального исследования демонстрируют достаточно высокий уровень релевантности характеристик эмоционального анализа виртуальных субъектов на основе предложенной методики. Разработанная и апробированная методика оценки среднего фона произвольной социальной сети с привязкой по времени позволяет получить карту поведенческих речевых паттернов для проведения мониторинга напряженности социальных настроений на основе тональности текстов произвольных виртуальных идентичностей.

Новизна предложенной методики заключается в идее дополнения лингвистического анализа сетевого поведения субъектов статистическим анализом и в выбранных и адаптированных методах анализа сообщений пользователей. В исследовании определена область практического использования результатов и обоснована репрезентативность разработанной методики анализа акцентуации виртуальных субъектов.

6 Тезаурус

Виртуальная идентичность — коммуникативная репрезентация человека в виртуальной среде, т. е. то, каким образом он позиционирует себя в сети Интернет (паттерн поведения виртуального объекта). Паттерны взаимодействия могут быть представлены в виде групп речевого взаимодействия: подражания, оппонирования, агрессии, конструктивного диалога и т. д.

Конструктивистский подход применительно к виртуальной коммуникации строится на предположении, что свойства человеческой психики и мыслеформы, которыми оперирует виртуальный субъект, в значительной степени подвержены влиянию категорий, сформированных социумом в его непосредственном окружении и усвоенных им в процессе социализации, и, следовательно, не зависят от биологических ограничений.

Лингвистический релятивизм базируется на гипотезе лингвистической относительности и предполагает, что структура языка е-коммуникации влияет на менталитет и способы идентификации его виртуальных агентов, а также на когнитивные процессы реальных субъектов.

Литература

1. *Johansson F., Brynielsson J., Horling P., Malm M., Martenson C., Truve S., Rosell M.* Detecting emergent conflicts through Web Mining and Visualization // 2011 European Intelligence and Security Informatics Conference Proceedings. — IEEE, 2011. P. 346–353.
2. *Kennison S. M.* Introduction to language development. — Los Angeles, CA, USA: SAGE Publications Inc., 2014. 496 p.
3. *Brown R., Lenneber E.* A study in language and cognition // J. Abnorm. Soc. Psych., 1954. Vol. 49. P. 454–462.
4. Rethinking linguistic relativity / Eds. J. J. Gumperz, S. C. Levinson. — Studies in the social and cultural foundations of language ser. — Cambridge: Cambridge University Press, 1999. No. 17. 488 p.
5. *Slobin D. I.* Two ways to travel: Verbs of motion in English and Spanish // Grammatical Constructions: Their form and meaning / Eds. M. Shibatani, S. A. Thompson. — Oxford: Clarendon Press, 1996. P. 195–220.
6. *Barbian G.* Detecting hidden friendship in online Social Networks // 2011 European Intelligence and Security Informatics Conference Proceedings. — IEEE, 2011. P. 269–272.
7. *Горелов И. Н., Седов К. Ф.* Основы психолингвистики. — М.: Лабиринт, 2001. 304 с.
8. *Сидоренко Е. В.* Методы математической обработки в психологии. — СПб.: Речь, 2002. 350 с.
9. *Журавлев А. П.* Звук и смысл. — М.: Просвещение, 1991. 160 с.
10. *Vybornova O., Smirnov I., Sochenkov I., Kiselyov A., Tikhomirov I., Chudova N., Kuznetsova Y., Osipov G.* Social tension detection and intention recognition using Natural Language Semantic Analysis // 2011 European Intelligence and Security Informatics Conference Proceedings. — IEEE, 2011. P. 277–281.

11. *Тихомиров И. А., Смирнов И. В.* Интеграция лингвистических и статистических методов поиска в поисковой машине Exactus // Компьютерная лингвистика и интеллектуальные технологии: Тр. междунар. конф. «Диалог-2008». — М.: РГГУ, 2008. С. 485–491.
12. *Cambria E., Havaci E., Hussain A. A.* Semantic and affective resource for opinion mining and sentiment analysis // 25th Florida Artificial Intelligence Research Society Conference (International) Proceedings. — Palo Alto, CA, USA: AAAI Press, 2012. P. 202–207.
13. *Hoijer H.* The Sapir–Whorf hypothesis // Conference on the Interrelations of Language and Other Aspects of Culture Proceedings: Memoirs of the American Anthropological Association, Comparative Studies of Cultures and Civilizations. — Chicago, IL, USA: University of Chicago Press, 1954. No. 3. P. 92–105.

Поступила в редакцию 25.04.17

PSYCHOLINGUISTIC ANALYSIS OF TEXT MESSAGES IN RUSSIAN BASED ON THEIR PHONOSEMANTIC STATISTICAL CHARACTERISTICS

A. S. Sigov, D. A. Akimov, D. O. Zhukov, E. G. Andrianova, V. E. Sachkov, and V. K. Raev

Moscow Technological University, 78 Vernadsky Av., Moscow 119454, Russian Federation

Abstract: A text as a complex semantic and syntactic formation has a number of psycholinguistic characteristics, which include integrity and semantic orientation. A text can be viewed as a product of speech activity with a high degree of semantic variation determined by its temporal and sonar characteristics. Nonverbal behavior of network entities — virtual masks and robotic agents — reveals itself in texts. The article raises and solves the problem of identifying the type of accentuation of pattern of behavior of a virtual entity based on statistical analysis of text communication, which allows one to formulate a hypothesis about the structural properties of a given communication and build a matrix of probabilities of relationship between virtual masks of subjects. The practical significance of the proposed solution is based on the growing importance of the development of the system of conditional signs, in this case, the conditional languages of e-communication, for the generation of control clusters regulating the social behavior of virtual subjects in the network. This assumption is based on the hypothesis of Kenneth Ivers, according to which, the better the system of conventional signs, the more opportunities to create new algorithms.

Keywords: psycholinguistic characteristics; nonverbal behavior; virtual masks; process of thinking; semantic meaning; linguistic relativism

DOI: 10.14357/19922264170309

Acknowledgments

The work was carried out within the Ministry of Education and Science of the Russian Federation's program of financing the competitive part of public tasks to institutions of higher education and scientific organizations to implement initiative scientific projects (No. 28.2635.2017/PP).

References

1. Johansson, F., J. Brynielsson, P. Horling, M. Malm, C. Martenson, S. Truve, and M. Rosell. 2011. Detecting emergent conflicts through Web Mining and Visualization. *European Intelligence and Security Informatics Conference Proceedings*. IEEE. 346–353.
2. Kennison, S. M. 2014. *Introduction to language development*. Los Angeles, CA: SAGE Publications Inc. 496 p.
3. Brown, R., and E. Lenneber. 1954. A study in language and cognition. *J. Abnorm. Soc. Psych.* 49:454–462.
4. Gumperz, J. J., and S. C. Levinson, eds. 1999. *Rethinking linguistic relativity*. Studies in social and cultural foundations of language ser. Cambridge: Cambridge University Press. No. 17. 488 p.
5. Slobin, D. 1996. Two ways to travel: Verbs of motion in English and Spanish. *Grammatical Constructions: Their form and meaning*. Eds. M. Shibatani and S. A. Thomson. Oxford: Clarendon Press. 195–220.
6. Barbian, G. 2011. Detecting hidden friendship in Online Social Networks. *European Intelligence and Security Informatics Conference Proceedings*. IEEE. 269–272.
7. Gorelov, I. N., and K. F. Sedov. 2001. *Osnovy psikholingvistiki* [Fundamentals of psycholinguistics]. Moscow: Labirint. 304 p.

8. Sidorenko, E. V. 2002. *Metody matematicheskoy obrabotki v psikhologii* [Methods of mathematical processing in psychology]. St. Petersburg: Rech. 350 p.
9. Zhuravlev, A. P. 1991. *Zvuk i smysl* [Sound and meaning]. Moscow: Prosveshchenie, 1991. 160 p.
10. Vybornova, O., I. Smirnov, I. Sochenkov, A. Kiselyov, I. Tikhomirov, N. Chudova, Y. Kuznetsova, and G. Osipov. 2011. Social tension detection and intention recognition using Natural Language Semantic Analysis. *European Intelligence and Security Informatics Conference Proceedings*. IEEE. 277–281.
11. Tihomirov, I. A., and I. V. Smirnov. 2008. Integratsiya lingvisticheskikh i statisticheskikh metodov poiska v poiskovoy mashine Exactus [Integration of linguistic and statistical methods of searching in the search engine Exactus]. *Conference (International) "Dialogue-2008" Proceedings*. Moscow: RGGU. 485–491.
12. Cambria, E., E. Havaci, and A. A. Hussain. 2012. Semantic and affective resource for opinion mining and sentiment analysis. *25th Florida Artificial Intelligence Research Society Conference (International) Proceedings*. Palo Alto, CA: AAAI Press. 202–207.
13. Hoijer, H. 1954. The Sapir–Whorf hypothesis. *Conference on the Interrelations of Language and Other Aspects of Culture Proceedings: Memoirs of the American Anthropological Association, Comparative Studies of Cultures and Civilizations*. Chicago, IL: University of Chicago Press. 3:92–105.

Received April 25, 2017

Contributors

Sigov Alexander S. (b. 1945) — Academician of the Russian Academy of Sciences, President of the Moscow Technological University (MIREA), 78 Vernadsky Av., Moscow 119454, Russian Federation; assigov@yandex.ru

Akimov Dmitry A. (b. 1987) — Candidate of Science (PhD) in technology, associate professor, Moscow Technological University (MIREA), 78 Vernadsky Av., Moscow 119454, Russian Federation; akimov_d@mirea.ru

Zhukov Dmitry O. (b. 1965) — Doctor of Science in technology, professor, Moscow Technological University (MIREA), 78 Vernadsky Av., Moscow 119454, Russian Federation; zhukovdm@yandex.ru

Andrianova Elena G. (b. 1963) — Candidate of Science (PhD) in technology, associate professor, Moscow Technological University (MIREA), 78 Vernadsky Av., Moscow 119454, Russia n Federation; dtghmflysq@gmail.com

Sachkov Valery E. (b. 1989) — PhD student, Moscow Technological University (MIREA), 78 Vernadsky Av., Moscow 119454, Russian Federation; megawatto@mail.ru

Raev Vyacheslav K. (b. 1965) — Doctor of Science in technology, professor, Moscow Technological University (MIREA), 78 Vernadsky Av., Moscow 119454, Russian Federation; raev@mirea.ru

ВЕРОЯТНОСТНАЯ МОДЕЛЬ СОВМЕСТНОГО ИСПОЛЬЗОВАНИЯ РЕСУРСОВ БЕСПРОВОДНОЙ СЕТИ С АДАПТИВНЫМ УПРАВЛЕНИЕМ МОЩНОСТЬЮ*

И. А. Гудкова[1], С. Я. Шоргин[2]

Аннотация: Развивающиеся беспроводные сети последующего поколения (next generation network, NGN) предполагают новые приложения и услуги как для обычных пользователей, так и для устройств межмашинного взаимодействия (machine-to-machine, M2M). Решение проблемы увеличения требований к пропускной способности сети и недостаточности спектра радиочастот, в частности в случае умных городов, может быть достигнуто посредством концепции совместного использования радиочастот (licensed shared access, LSA). Авторы предлагают математическую модель совместного использования ресурсов с адаптивным управлением мощностью. Заложенный в ней алгоритм позволит избежать интерференции M2M-устройств с владельцем спектра, в том числе благодаря тому, что учитывает пространственное расположение устройств и их сессионную активность.

Ключевые слова: беспроводная сеть; умный город; межмашинное взаимодействие; совместное использование радиочастот; адаптивное управление мощностью; случайный процесс; рекуррентный алгоритм; вероятность блокировки; вероятность прерывания обслуживания; среднее число устройств

DOI: 10.14357/19922264170310

1 Введение

Согласно прогнозам развития сетей последующего поколения, уже в 2025 г. беспроводные сети будут перегружены [1], что повлечет за собой необходимость уточнения и разработки новых стратегий использования спектра радиочастот [2]. Широкое распространение получают автономно функционирующие и взаимодействующие друг с другом (M2M) недорогие устройства, являющиеся неотъемлемой частью «умных городов» (smart city). Особенностью M2M-устройств является их дистанционное управление и высокая плотность расположения. Рост числа M2M-устройств существенно сказывается на использовании спектра радиочастот ввиду того, что сети изначально разрабатывались для взаимодействия между людьми (human-to-human, H2H).

Один из вариантов решения проблемы — это динамическое управление спектром в рамках концепции совместного использования радиочастот (LSA) [3–5]. Доступ к спектру получают две стороны — владелец и временный пользователь [6, 7].

В статье исследуется один из сценариев применения системы LSA [8–11], где владелец запрашивает спектр радиочастот изредка на непродолжительное время. В остальное же время спектр доступен M2M-устройствам для передачи данных. Статья имеет следующую структуру.

В разд. 2 описана системная модель совместного использования радиоресурсов с учетом расположения устройств на разном расстоянии от базовой станции [12–15].

В разд. 3 проводится построение математической модели в виде двух случайных процессов (СП), один из которых фиксирует уровень качества канала каждого из активных устройств, а второй, укрупненный, — только суммарное число устройств. Для СП с укрупненными состояниями представлен рекуррентный алгоритм расчета стационарного распределения вероятностей.

В разд. 4 предложены формулы для расчета ключевых показателей эффективности системы — среднего числа устройств и вероятностей блокировки и прерывания обслуживания, приведен пример численного анализа.

*Исследование выполнено при финансовой поддержке Российского научного фонда (проект 16-11-10227).

[1]Российский университет дружбы народов; Институт проблем информатики Федерального исследовательского центра «Информатика и управление» Российской академии наук, gudkova_ia@rudn.university

[2]Институт проблем информатики Федерального исследовательского центра «Информатика и управление» Российской академии наук, sshorgin@ipiran.ru

2 Системная модель совместного использования ресурсов с разноудаленными от базовой станции устройствами

Рассмотрим одну соту беспроводной сети радиуса R с равномерно распределенными по зоне покрытия M2M-устройствами (рис. 1). Устройства с интенсивностью λ переходят в активное состояние и передают данные в восходящем канале. Время передачи данных одним устройством распределено экспоненциально с параметром μ. Каждому устройству в зависимости от дальности расположения от базовой станции (БС) присваивается один из пятнадцати уровней качества канала (channel quality indicator, CQI) — $c = 1, \ldots, 15$, причем чем больше c, тем ближе устройство к БС и выше скорость передачи данных. Объединим устройства с одинаковыми уровнями CQI в логические группы, тогда скорость передачи для всех устройств в группе будет одинаковая. Далее под расстоянием от устройства до БС будем понимать максимально возможное расстояние, на котором может быть расположено устройство с таким же уровнем CQI. Введем дополнительное обозначение: $\eta = 16 - c$; уровень CQI c, величина η и расстояние $\xi_d(\eta) = RL^{-1}\eta$ от устройства до БС являются случайными величинами (СВ). Плотность расстояния от устройства до БС

$$f_{\xi_d(\eta)}(d) = \frac{2d}{R^2},$$

а функция распределения (ФР)

$$F_{\xi_d(\eta)}(d) = \left(\frac{d}{R}\right)^2, \quad 0 \le d \le R.$$

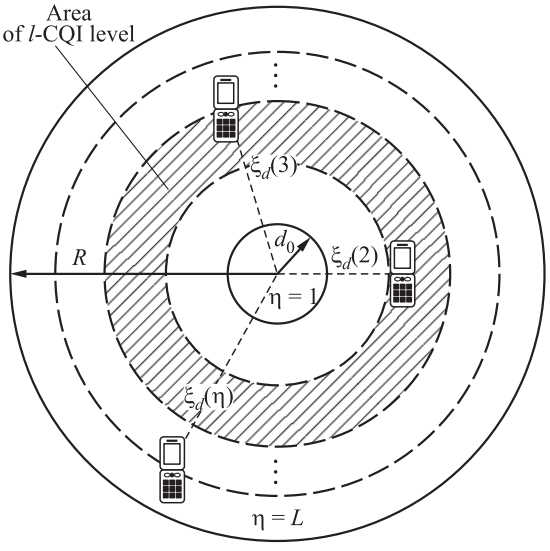


Рис. 1 Пример расположения M2M-устройств в соте

Ряд распределения для параметра η:

$$q_l = \frac{2L - 2l - 1}{L^2}, \quad l = 1, \ldots, L.$$

В качестве примера реализации системы LSA рассмотрим случай использования владельцем спектра радиочастот для воздушной телеметрии. Предположим, что время, в течение которого владелец (аэропорт) не использует спектр, т. е. время, когда полоса доступна для M2M-устройств, и время пролета самолета над сотой, т. е. время, когда полоса недоступна для устройств, распределены по экспоненциальному закону с параметрами α и β соответственно.

Управление радиоресурсами предполагает разделение ресурсов по времени, т. е. деление ширины полосы радиочастот ω не происходит, а передача данных осуществляется на постоянной мощности. Если полоса не требуется аэропорту, то мощность составляет $p_1^{\max}$, в противном случае для регулирования интерференции мощность снижается до значения $p_0^{\max} < p_1^{\max}$. Такое динамическое изменение мощности приводит к изменению достижимой скорости передачи данных $r\left(\xi_{d(\eta)}, p_s^{\max}\right)$, $s = 0, 1$, зависящей также от расстояния между устройством и БС. Согласно формуле Шеннона,

$$r\left(\xi_{d(\eta)}, p_s^{\max}\right) = \omega \ln\left(1 + \frac{G p_s^{\max}}{((R/L)\eta)^{\kappa} N_0}\right), \quad s = 0, 1, \ \eta = 1, \ldots, 15, \quad (1)$$

где N_0 — уровень шума; G — константа затухания сигнала; κ — экспонента затухания сигнала.

Скорость передачи данных каждым активным M2M-устройством не может быть ниже порогового (гарантированного) значения r_0. Если устройству не может быть обеспечена скорость r_0, то запрос на передачу данных будет заблокирован. Если устройство расположено в непосредственной близости от БС, то скорость передачи согласно формуле Шеннона стремится к бесконечности, поэтому определим минимальное расстояние до БС $\xi_d(1) = d_0$, ограничив тем самым максимальную скорость передачи данных $r_s^{\max} = r\left(d_0, p_s^{\max}\right)$. Таким образом, если $\eta = 1$, то достижимая скорость передачи данных $r\left(\xi_{d(\eta)}; p_s^{\max}\right)$, если $\eta = 2, \ldots, L$, то она вычисляется по формуле (1). Максимальное число устройств в соте:

$$K_s = \left\lfloor \frac{r(d_0, p_s^{\max})}{r_0} \right\rfloor, \quad s = 0, 1.$$

Сводный перечень основных обозначений приведен в табл. 1.

Таблица 1 Основные обозначения

Обозначение	Описание
R	Радиус соты, м
ω	Ширина полосы радиочастот, МГц
L	Число уровней качества канала CQI
c	Уровень CQI (СВ)
$\eta = 16 - c$	Величина, обратная уровню CQI c (СВ)
$q_l = \dfrac{2L - 2l - 1}{L^2}$	Вероятность того, что уровень CQI равен l
α^{-1}	Среднее время доступности полосы, с
β^{-1}	Среднее время недоступности полосы, с
k	Число активных устройств
s	Состояние полосы: $s = 1$, если полоса доступна; $s = 0$, если недоступна
$p_0^{\max}$	Максимальное значение мощности сигнала устройства, если полоса недоступна, Вт
$p_1^{\max}$	Максимальное значение мощности сигнала устройства, если полоса доступна, Вт
d_0	Минимальное расстояние от устройства до БС, м
$r\left(\xi_{d(\eta)}, p_s^{\max}\right)$	Достижимая скорость передачи для устройства с уровнем CQI $c = 16 - \eta$, если полоса находится в состоянии s (СВ), бит/с
$r_0^{\max}$	Максимально возможная скорость, если полоса недоступна, бит/с
r_0	Гарантированная скорость передачи данных от устройств, бит/с
$r_1^{\max}$	Максимально возможная скорость, если полоса доступна, бит/с
K_0	Максимальное число устройств, если полоса недоступна
K_1	Максимальное число устройств, если полоса доступна
$\xi_d(\eta)$	Максимальное расстояние от устройства с уровнем CQI $c = 16 - \eta$ до БС (СВ), м
λ	Интенсивность суммарного потока данных от всех устройств в соте, 1/с
μ^{-1}	Среднее время передачи данных от одного устройства, с
$\rho = \dfrac{\lambda}{\mu}$	Суммарная предложенная нагрузка от всех устройств в соте, Эрл

3 Вероятностная модель и стационарное распределение вероятностей состояний беспроводной сети

Перейдем к построению математической модели. Пусть $\xi(t)$ — число активных M2M-устройств; $\eta_i(t)$ — значение параметра η для устройства i; $\zeta(t)$ — состояние полосы в момент времени $t \geq 0$. Тогда функционирование соты опишем СП $\{\xi(t), \eta_1(t), \ldots, \eta_{\xi(t)}, \zeta(t), t \geq 0\}$ над пространством состояний

$$\mathbf{L} = \left\{ (0, s), (k, l_1, \ldots, l_k, s), \right.$$
$$s = 0, 1,\ l_i = 1, \ldots, L,\ i = 1, \ldots, k,\ k = 1, 2, \ldots :$$
$$\left. \sum_{i=1}^{k} \frac{r_0}{\omega \ln\left(1 + G p_s^{\max} / ((RL^{-1} l_i)^{\kappa} N_0)\right)} \leq 1 \right\}.$$

Фрагмент пространства состояний показан на рис. 2.

Перейдем к СП $\{\xi(t), \zeta(t), t \geq 0\}$ с укрупненными состояниями — суммарным числом устройств и состоянием полосы. Пространство состояний такого процесса будет иметь вид:

$$\mathbf{L}_1 = \{(k, s) :\ k = 0, 1, \ldots, K_s,\ s = 0, 1\}\ .$$

Отметим, что при переходе полосы в недоступное состояние происходит снижение мощности передачи данных с $p_1^{\max}$ до $p_0^{\max}$ и прерывание обслуживания $k - K_0$ устройств при условии, что число устройств $k > K_0$. При переходе из недоступного в доступное состояние мощность снова повышается. На рис. 3 представлена диаграмма интенсивностей переходов данного СП.

Обозначим через $P_s(k)$, $s = 0, 1$, условную вероятность того, что $(k+1)$-е M2M-устройство может быть обслужено при условии, что активно k устройств. Можно показать, что вероятности $P_s(k)$ вычисляются по формулам:

$$P_s(0) = $$
$$= F_{\xi_d(\eta)} \left(\min \left\{ R, \left(\frac{G p_s^{\max}}{\left(e^{r_0/\omega} - 1\right) N_0} \right)^{1/\kappa} \right\} \right),$$
$$s = 0, 1\,;$$

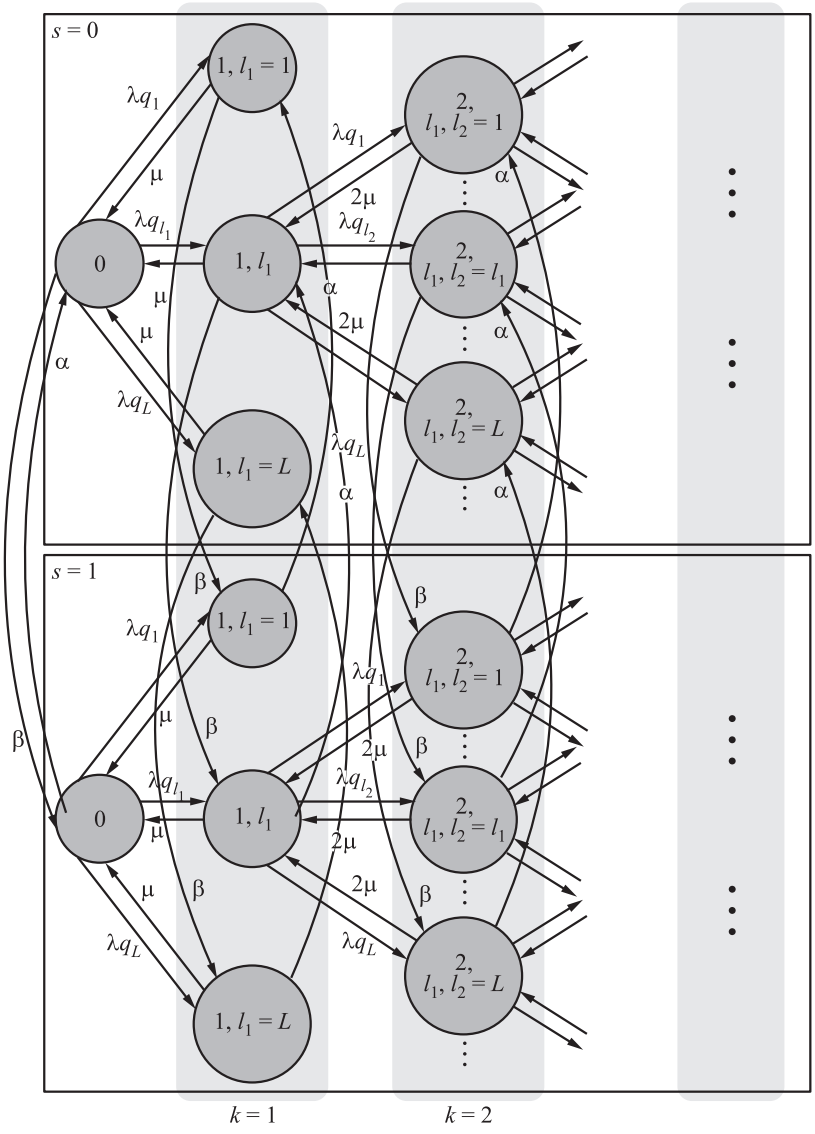


Рис. 2 Фрагмент диаграммы интенсивностей переходов СП с детальными состояниями

$$P_s(k) = \frac{\Phi\left((1-m_{k+1,s})/\tau_{k+1,s}\right)}{\Phi\left((1-m_{ks})/\tau_{ks}\right)}\,,$$
$$k = 1,\ldots,K_s,\ \ s = 0,1\,,$$

где

$$\Phi(x) = \frac{1}{\sqrt{2\pi}}\int\limits_{-\infty}^{x} e^{-t^2/2}\,dt\,;$$

$$m_{ks} = kr_0 E\left[\frac{1}{r\left(d, p_s^{\max}\right)}\right];$$

$$\tau_{ks}^2 = kr_0^2\left(E\left[\left(\frac{1}{r\left(d, p_s^{\max}\right)}\right)^2\right] - \left(E\left[\frac{1}{r\left(d, p_s^{\max}\right)}\right]\right)^2\right);$$

$$E\left[\frac{1}{r\left(d, p_s^{\max}\right)}\right] = \frac{1}{r_s^{\max}}\,F_{\xi_d(\eta)}(d_0) + \int\limits_{d_0}^{R}\frac{1}{\omega\ln(1+Gp_s^{\max}/(x^{\kappa}N_0))}\,f_{\xi_d(\eta)}(x)\,dx\,;$$

$$E\left[\left(\frac{1}{r\left(d, p_s^{\max}\right)}\right)^2\right] = \left(\frac{1}{r_s^{\max}}\right)^2 F_{\xi_d(\eta)}(d_0) + \int\limits_{d_0}^{R}\frac{1}{\omega^2\ln^2(1+Gp_s^{\max}/(x^{\kappa}N_0))}\,f_{\xi_d(\eta)}(x)\,dx\,.$$

Случайный процесс $\{\xi(t), \zeta(t), t \geq 0\}$ является марковским, и для расчета его стационарного рас-

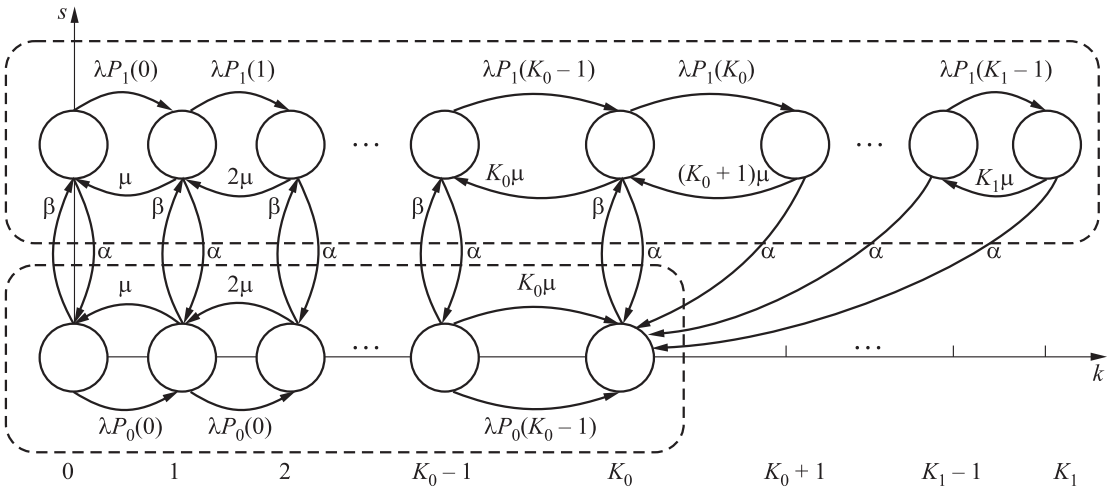


Рис. 3 Диаграмма интенсивностей переходов СП с укрупненными состояниями

пределения вероятностей $p(k,s)$, $(k,s) \in \mathbf{L}_1$ предлагается следующий рекуррентный алгоритм.

1. Значения ненормированных вероятностей $q(k,s)$ вычисляются по формулам:

$$q(0,0) = 1\,;$$
$$q(0,1) = x\,;$$
$$q(k,s) = \delta_{ks} + \gamma_{ks}x\,,\ (k,s) \in \mathbf{L}_1 :\ k > 0\,,$$

где

$$x = \frac{(K_1\mu+\alpha)\delta_{K_1 1} - \lambda P_1(K_1-1)\delta_{K_1-1,1}}{\lambda P_1(K_1-1)\gamma_{K_1-1,1} - (K_1\mu+\alpha)\gamma_{K_1 1}}\,.$$

2. Коэффициенты δ_{ks} и γ_{ks} вычисляются по рекуррентным формулам:

$$\delta_{00} = 1\,,\ \gamma_{00} = 0\,;$$
$$\delta_{01} = 0\,,\ \gamma_{01} = 1\,;$$
$$\delta_{10} = \frac{\lambda P_0(0)+\beta}{\mu}\,,\ \gamma_{10} = -\frac{\alpha}{\mu}\,;$$
$$\delta_{11} = -\frac{\beta}{\mu}\,,\ \gamma_{11} = \frac{\lambda P_1(0)+\alpha}{\mu}\,;$$

$$\delta_{k0} = \frac{\lambda P_0(k-1)+(k-1)\mu+\beta}{k\mu}\,\delta_{k-1,0} -$$
$$-\frac{\lambda P_0(k-2)}{k\mu}\,\delta_{k-2,0} - \frac{\alpha}{k\mu}\,\delta_{k-1,1}\,,\ k = 2,\ldots,K_0,$$

$$\gamma_{k0} = \frac{\lambda P_0(k-1)+(k-1)\mu+\beta}{k\mu}\,\gamma_{k-1,0} -$$
$$-\frac{\lambda P_0(k-2)}{k\mu}\,\gamma_{k-2,0} - \frac{\alpha}{k\mu}\,\gamma_{k-1,1}\,,\ k = 2,\ldots,K_0;$$

$$\delta_{k1} = \frac{\lambda P_1(k-1)+(k-1)\mu+\alpha}{k\mu}\,\delta_{k-1,1} -$$
$$-\frac{\lambda P_1(k-2)}{k\mu}\,\delta_{k-2,1} - \frac{\beta}{k\mu}\,\delta_{k-1,0}\,,$$
$$k = 2,\ldots,K_0+1\,;$$
$$\gamma_{k1} = \frac{\lambda P_1(k-1)+(k-1)\mu+\alpha}{k\mu}\,\gamma_{k-1,1} -$$
$$-\frac{\lambda P_1(k-2)}{k\mu}\,\gamma_{k-2,1} - \frac{\beta}{k\mu}\,\gamma_{k-1,0}\,,$$
$$k = 2,\ldots,K_0+1\,;$$
$$\delta_{k1} = \frac{\lambda P_1(k-1)+(k-1)\mu+\alpha}{k\mu}\,\delta_{k-1,1} -$$
$$-\frac{\lambda P_1(k-2)}{k\mu}\,\delta_{k-2,1}\,,\ k = K_0+2,\ldots,K_1\,,$$
$$\gamma_{k1} = \frac{\lambda P_1(k-1)+(k-1)\mu+\alpha}{k\mu}\,\gamma_{k-1,1} -$$
$$-\frac{\lambda P_1(k-2)}{k\mu}\,\gamma_{k-2,1}\,,\ k = K_0+2,\ldots,K_1\,.$$

3. Значения вероятностей $p(k,s)$ вычисляются по формулам:

$$p(k,s) = \frac{q(k,s)}{\sum_{(i,j)\in\mathbf{L}} q(i,j)}\,,\ (k,s) \in \mathbf{L}_1\,.$$

4 Пример численного анализа и заключение

Зная стационарное распределение вероятностей $p(k,s)$, $(k,s) \in \mathbf{L}_1$, найдем основные показатели эффективности модели — вероятность B блокировки, вероятность Π прерывания обслуживания и среднее число $\overline{K}$ устройств по формулам:

Таблица 2 Исходные данные для численного анализа

Обозначение	Случай 1 (рис. 4)	Случай 2 (рис. 5)	Случай 3 (рис. 6)
R, м	200–400	200; 400	200; 400
ω, МГц	10	10	10
L	15	15	15
α^{-1}, мин	20; 30	20; 30	30
β^{-1}, с	20	20	20
$p_1^{\max}$, дБ·м	23; 42	23–42	23; 42
$p_0^{\max}$, дБ·м	$p_{\max}/2$	$p_{\max}/2$	$p_{\max}/2$
d_0, м	$R/15$	$R/15$	$R/15$
r_0, Мбит/с	1	1	1
λ, 1/с	10	10	2–10
μ^{-1}, с	0,1	0,1	0,1
N_0, дБ·м	-60	-60	-60
G	197,43	197,43	197,43
κ	5	5	5

$$B = \sum_{k=0}^{K_0-1} (1 - P_0(k))\, p(k,0) + \sum_{k=0}^{K_1-1} (1 - P_1(k))\, p(k,1)\,;$$

$$\Pi = \sum_{k=K_0+1}^{K_1-1} \frac{\alpha}{\alpha + k\mu + \lambda P_1(k)} \frac{\binom{k-1}{k-K_0-1}}{\binom{k}{k-K_0}} p(k,1) + \frac{\alpha}{\alpha + K_1\mu} \frac{\binom{K_1-1}{K_1-K_0-1}}{\binom{K_1}{K_1-K_0}} p(K_1,1)\,;$$

$$\overline{K} = \sum_{k=0}^{K_0} k p(k,0) + \sum_{k=0}^{K_1} k p(k,1)\,.$$

Для проведения численного анализа проанализируем передачу данных M2M-устройствами небольшими сессиями, составляющими в среднем 10 с, в высоком качестве на скорости 1 Мбит/с. Рассмотрим небольшой аэропорт, в котором самолеты взлетают раз в 20 (30) мин, среднее время пролета самолета над сотой составляет 20 с. Исходные данные представлены в табл. 2.

На рис. 4 показана зависимость вероятности прерывания обслуживания и среднего числа активных устройств от их мощности. Вероятность прерывания обслуживания уменьшается пропорционально увеличению мощности, так как при более высокой мощности для устройств достижима более высокая скорость. При этом вероятность прерывания ниже для более низкой интенсивности отключения полосы. Среднее число устройств увеличивается пропорционально радиусу соты (см.

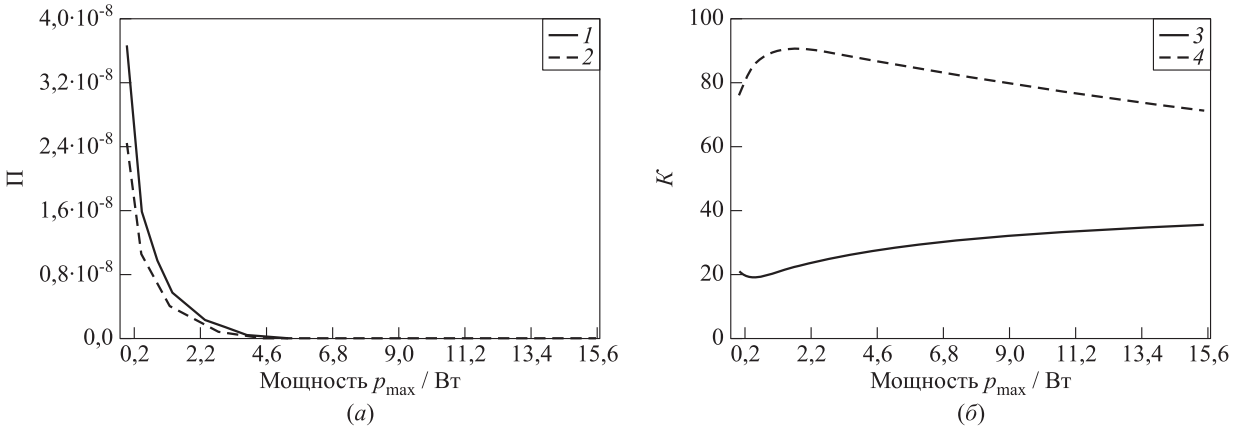


Рис. 4 Показатели эффективности в зависимости от мощности устройств: (*а*) вероятность прерывания обслуживания при $R = 400$ (*1* — $\alpha = 1200$; *2* — $\alpha = 1800$); (*б*) среднее число активных устройств при $\alpha = 1800$ (*3* — $R = 200$; *4* — $R = 400$)

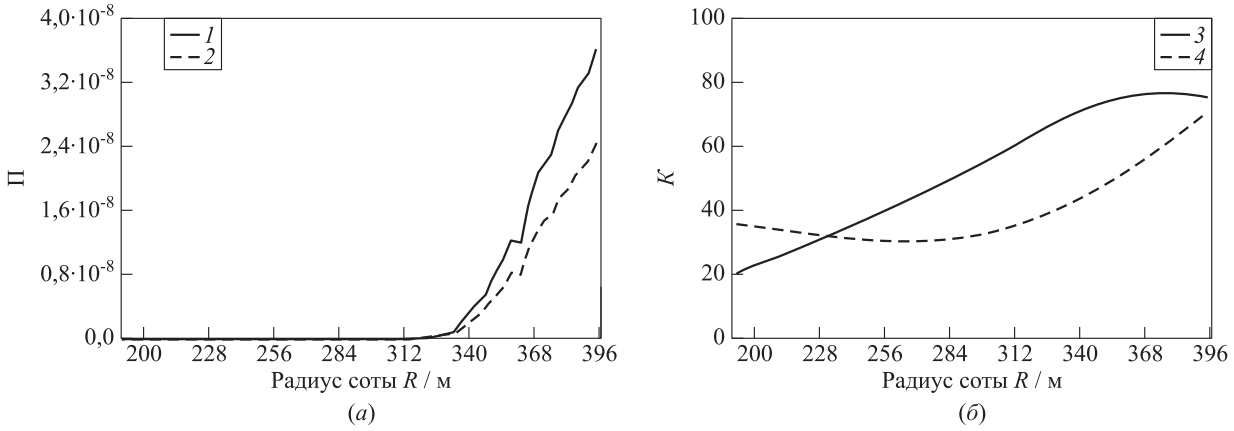


Рис. 5 Показатели эффективности в зависимости от радиуса соты: (*а*) вероятность прерывания обслуживания при $W = 0{,}2$ (*1* — $\alpha = 1200$; *2* — $\alpha = 1800$); (*б*) среднее число активных устройств при $\alpha = 1200$ (*3* — $W = 0{,}2$; *4* — $W = 15{,}85$)

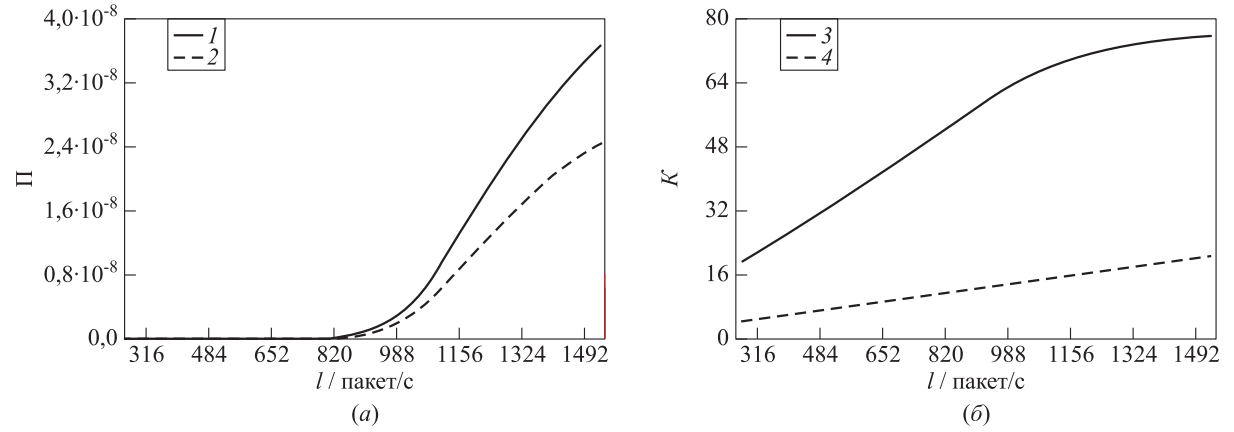


Рис. 6 Показатели эффективности в зависимости от интенсивности потока пакетов данных от устройств: (*а*) вероятность прерывания обслуживания при $R = 400$ (*1* — $\alpha = 1200$; *2* — $\alpha = 1800$); (*б*) среднее число активных устройств (*3* — $R = 400$, $\alpha = 1200$; *4* — $R = 200$, $\alpha = 1800$)

рис. 5). Вероятность прерывания оказывается ниже при меньшей интенсивности изъятия полосы (см. рис. 6).

В заключение отметим, что в статье разработана вероятностная модель совместного использования радиочастот, при помощи которой проведен анализ показателей эффективности применения политики управления мощностью с учетом разноудаленных от БС M2M-устройств.

В дальнейшем предполагается учесть случайную высоту, на которой могут находиться устройства.

Литература

1. Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2016–2021 White Paper. March 28, 2017. http://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/mobile-white-paper-c11-520862.html.
2. *Andrews J., Buzzi S., Choi W., Hanly S. V., Lozano A., Soong C. K., Zhang J. C.* What will 5G be? // IEEE J. Sel. Area. Comm., 2014. Vol. 32. P. 1065–1082.
3. ETSI TR 103 113. Electromagnetic compatibility and Radio spectrum Matters (ERM); System Reference document (SRdoc); Mobile broadband services in the 2 300 MHz – 2 400 MHz frequency band under Licensed Shared Access regime. v1.1.1. July 2013. http://www.etsi.org/deliver/etsi_tr/103100_103199/103113/01.01.01_60/tr_103113v010101p.pdf.
4. ETSI TR 103 154. Reconfigurable Radio Systems (RRS); System requirements for operation of Mobile Broadband Systems in the 2 300 MHz – 2 400 MHz band under Licensed Shared Access (LSA). v1.1.1. October 2014. http://www.etsi.org/deliver/etsi_TS/103100_103199/103154/01.01.01_60/ts_103154v010101p.pdf.
5. ETSI TR 103 235. Reconfigurable Radio Systems (RRS); System architecture and high level procedures for operation of Licensed Shared Access (LSA) in the 2 300 MHz –

2400 MHz band. v1.1.1. October 2015. http://www.etsi.org/deliver/etsi_ts%5C103200_103299%5C103235%5C01.01.01_60%5Cts_103235v010101p.pdf.

6. *Buckwitz K., Engelberg J., Rausch G.* Licensed Shared Access (LSA) — regulatory background and view of Administrations // 9th Conference (International) on Cognitive Radio Oriented Wireless Networks. — IEEE, 2014. P. 413–416.
7. *Ahokangas P., Matinmikko M., Yrjölä S., Mustonen M., Posti H., Luttinen E., Kivimäki A.* Business models for mobile network operators in Licensed Shared Access (LSA) // IEEE Symposium (International) on Dynamic Spectrum Access Networks. — IEEE, 2014. P. 263–270.
8. *Borodakiy V. Y., Samouylov K. E., Gudkova I. A., Ostrikova D. Y., Ponomarenko A. A., Turlikov A. M., Andreev S. D.* Modeling unreliable LSA operation in 3GPP LTE cellular networks // 6th Congress (International) on Ultra Modern Telecommunications and Control Systems and Workshops Proceedings. — Piscataway, NJ, USA: IEEE, 2015. P. 490–496.
9. *Ponomarenko-Timofeev A., Pyattaev A., Andreev S., Koucheryavy Ye., Mueck M., Karls I.* Highly dynamic spectrum management within licensed shared access regulatory framework // IEEE Commun. Mag., 2015. Vol. 54. No. 3. P. 100–109.
10. *Gudkova I. A., Samouylov K. E., Ostrikova D. Y., Mokrov E. V., Ponomarenko-Timofeev A. A., Andreev S. D., Koucheryavy Y. A.* Service failure and interruption probability analysis for Licensed Shared Access regulatory framework // 7th Congress (International) on Ultra Modern Telecommunications and Control Systems and Workshops Proceedings. — Piscataway, NJ, USA: IEEE Computer Society, 2015. P. 123–131.
11. *Samouylov K., Gudkova I., Markova E., Yarkina N.* Queuing model with unreliable servers for limit power policy within Licensed Shared Access framework // Internet of things, smart spaces, and next generation networks and systems / Eds. O. Galinina, S. Balankin, Y. Koucheryavy. — Lecture notes in computer science ser. — Springer, 2016. Vol. 9870. P. 404–413.
12. *Galinina O., Andreev S. D., Gerasimenko M., Koucheryavy Y. A., Himayat N., Yeh S.-P., Talwar S.* Capturing spatial randomness of heterogeneous cellular/WLAN deployments with dynamic traffic // IEEE J. Sel. Area. Comm., 2014. Vol. 32. No. 6. P. 1083–1099.
13. *Ahmadian A., Galinina O., Gudkova I., Andreev S., Shorgin S., Samouylov K.* On capturing spatial diversity of joint M2M/H2H dynamic uplink transmissions in 3GPP LTE cellular system // Internet of things, smart spaces, and next generation networks and systems / Eds. S. Balandin, S. Andreev, Y. Koucheryavy. — Lecture notes in computer science ser. — Springer, 2014. Vol. 9247. P. 407–421.
14. *Samouylov K., Gudkova I., Markova E., Dzantiev I.* On analyzing the blocking probability of M2M transmissions for a CQI-based RRM scheme model in 3GPP LTE // Comm. Com. Inf. Sci., 2016. Vol. 638. P. 327–340.
15. *Gudkova I., Markova E., Masek P., Andreev S., Hosek J., Yarkina N., Samouylov K., Koucheryavy Y.* Modeling the utilization of a multi-tenant band in 3GPP LTE system with Licensed Shared Access // 8th Congress (International) on Ultra Modern Telecommunications and Control Systems and Workshops Proceedings. — Piscataway, NJ, USA: IEEE, 2016. P. 179–183.

Поступила в редакцию 20.04.17

PROBABILITY MODEL FOR ANALYZING LICENSED SHARED ACCESS WITH ADAPTIVE POWER CONTROL IN A WIRELESS NETWORK

I. A. Gudkova[1,2] and S. Ya. Shorgin[2]

[1]Peoples' Friendship University of Russia, 6 Miklukho-Maklaya Str., Moscow 117198, Russian Federation
[2]Institute of Informatics Problems, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, 44-2 Vavilov Str., Moscow 119333, Russian Federation

Abstract: Emerging next generation wireless networks involve new applications and services for human-to-human and machine-to-machine (M2M) devices. The problem of increasing requirements for network capacity and lack of radio spectrum arises. The solution could be found in the licensed shared access framework, e. g., in the case of smart cities. The authors propose a mathematical model of shared access to spectrum with adaptive power control. The algorithm makes it possible to avoid the interference between M2M devices and the spectrum owner due, in part, to the fact that it takes into account the spatial distribution and session activity of devices.

Keywords: wireless network; smart city; machine-to-machine (M2M); licensed shared access (LSA); adaptive power control; stochastic process; recursive algorithm; blocking probability; interruption probability; average number of M2M devices

DOI: 10.14357/19922264170310

Acknowledgments

This work was financially supported by the Russian Science Foundation (grant No. 16-11-10227).

References

1. Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2016–2021 White Paper. March 28, 2017. Available at: http://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/mobile-white-paper-c11-520862.html (accessed June 26, 2017).
2. Andrews, J., S. Buzzi, W. Choi, S. V. Hanly, A. Lozano, C. K. Soong, and J. C. Zhang. 2014. What will 5G be? *IEEE J. Sel. Area. Comm.* 32:1065–1082.
3. ETSI TR 103 113. July 2013. Electromagnetic compatibility and Radio spectrum Matters (ERM); System Reference document (SRdoc); Mobile broadband services in the 2300 MHz – 2400 MHz frequency band under Licensed Shared Access regime. Available at: http://www.etsi.org/deliver/etsi_tr/103100_103199/103113/01.01.01_60/tr_103113v010101p.pdf (accessed June 26, 2017).
4. ETSI TR 103 154. October 2014. Reconfigurable Radio Systems (RRS); System requirements for operation of Mobile Broadband Systems in the 2300 MHz – 2400 MHz band under Licensed Shared Access (LSA). v1.1.1. Available at: http://www.etsi.org/deliver/etsi_TS/103100_103199/103154/01.01.01_60/ts_103154v010101p.pdf (accessed June 26, 2017).
5. ETSI TR 103 235. October 2015. Reconfigurable Radio Systems (RRS); System architecture and high level procedures for operation of Licensed Shared Access (LSA) in the 2300 MHz – 2400 MHz band. v1.1.1. Available at: http://www.etsi.org/deliver/etsi_ts%5C103200_103299%5C103235%5C01.01.01_60%5Cts_103235v010101p.pdf (accessed June 26, 2017).
6. Buckwitz, K., J. Engelberg, and G. Rausch. 2014. Licensed Shared Access (LSA) — regulatory background and view of Administrations. *9th Conference (International) on Cognitive Radio Oriented Wireless Networks*. IEEE. 413–416.
7. Ahokangas, P., M. Matinmikko, S. Yrjölä, M. Mustonen, H. Posti, E. Luttinen, and A. Kivimäki. 2014. Business models for mobile network operators in Licensed Shared Access (LSA). *IEEE Symposium (International) on Dynamic Spectrum Access Networks*. IEEE. 263–270.
8. Borodakiy, V. Y., K. E. Samouylov, I. A. Gudkova, D. Y. Ostrikova, A. A. Ponomarenko, A. M. Turlikov, and S. D. Andreev. 2014. Modeling unreliable LSA operation in 3GPP LTE cellular networks. *6th Congress (International) on Ultra Modern Telecommunications and Control Systems and Workshops Proceedings*. Piscataway, NJ: IEEE. 490–496.
9. Ponomarenko-Timofeev, A., A. Pyattaev, S. Andreev, Ye. Koucheryavy, M. Mueck, and I. Karls. 2015. Highly dynamic spectrum management within licensed shared access regulatory framework. *IEEE Commun. Mag.* 54(3):100–109.
10. Gudkova, I. A., K. E. Samouylov, D. Y. Ostrikova, E. V. Mokrov, A. A. Ponomarenko-Timofeev, S. D. Andreev, and Y. A. Koucheryavy. 2015. Service failure and interruption probability analysis for Licensed Shared Access regulatory framework. *7th Congress (International) on Ultra Modern Telecommunications and Control Systems and Workshops Proceedings*. Piscataway, NJ: IEEE. 123–131.
11. Samouylov, K., I. Gudkova, E. Markova, and N. Yarkina. 2016. Queuing model with unreliable servers for limit power policy within Licensed Shared Access framework. *Internet of things, smart spaces, and next generation networks and systems*. Eds. O. Galinina, S. Balankin, Y. Koucheryavy. Lecture notes in computer science ser. Springer. 9870:404–413.
12. Galinina, O., S. D. Andreev, M. Gerasimenko, Y. A. Koucheryavy, N. Himayat, S.-P. Yeh, and S. Talwar. 2014. Capturing spatial randomness of heterogeneous cellular/WLAN deployments with dynamic traffic. *IEEE J. Sel. Area. Comm.* 32(6):1083–1099.
13. Ahmadian, A., O. Galinina, I. Gudkova, S. Andreev, S. Shorgin, and K. Samouylov. 2014. On capturing spatial diversity of joint M2M/H2H dynamic uplink transmissions in 3GPP LTE cellular system. *Internet of things, smart spaces, and next generation networks and systems*. Eds. S. Balandin, S. Andreev, Y. Koucheryavy. Lecture notes in computer science ser. Springer. 9247:407–421.
14. Samouylov, K., I. Gudkova, E. Markova, and I. Dzantiev. 2016. On analyzing the blocking probability of M2M transmissions for a CQI-based RRM scheme model in 3GPP LTE. *Comm. Com. Inf. Sci.* 638:327–340.
15. Gudkova, I., E. Markova, P. Masek, S. Andreev, J. Hosek, N. Yarkina, K. Samouylov, and Y. Koucheryavy. 2016. Modeling the utilization of a multi-tenant band in 3GPP LTE system with Licensed Shared Access. *8th Congress (International) on Ultra Modern Telecommunications and Control Systems and Workshops Proceedings*. Piscataway, NJ: IEEE. 179–183.

Received April 20, 2017

Contributors

Gudkova Irina A. (b. 1985) — Candidate of Sciences (PhD) in physics and mathematics; associate professor, Peoples' Friendship University of Russia, 6 Miklukho-Maklaya Str., Moscow 117198, Russian Federation; senior scientist, Institute of Informatics Problems, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, 44-2 Vavilov Str., Moscow 119333, Russian Federation; gudkova_ia@rudn.university

Shorgin Sergey Ya. (b. 1952) — Doctor of Science in physics and mathematics, professor; Deputy Director, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences (FRC CSC RAS); principal scientist, Institute of Informatics Problems, FRC CSC RAS; 44-2 Vavilov Str., Moscow 119333, Russian Federation; sshorgin@ipiran.ru

СИСТЕМА МАССОВОГО ОБСЛУЖИВАНИЯ С ОГРАНИЧЕННЫМИ РЕСУРСАМИ И СИГНАЛАМИ ДЛЯ АНАЛИЗА ПОКАЗАТЕЛЕЙ ЭФФЕКТИВНОСТИ БЕСПРОВОДНЫХ СЕТЕЙ*

К. Е. Самуйлов[1], Э. С. Сопин[2], С. Я. Шоргин[3]

Аннотация: Рассматривается многолинейная система массового обслуживания (СМО) с ресурсами ограниченного объема. Поступающая заявка занимает не только прибор, но и некоторый объем ресурсов на все время обслуживания. Помимо потока заявок на систему поступает поток сигналов, при поступлении которых заявки заново разыгрывают объем занимаемых ресурсов. Рассматриваемая система массового обслуживания позволяет описывать функционирование беспроводной сети с учетом перемещения пользователей в течение периода жизни пользовательской сессии. Исследуются две модели перемещения пользователей. В первой пользователи перемещаются независимо друг от друга; следовательно, в соответствующей математической модели поступление сигнала изменяет занимаемый объем ресурсов только одной заявки. Во второй модели пользователи перемещаются совместно, поэтому занимаемый объем ресурсов меняется одновременно у всех заявок.

Ключевые слова: ограниченные ресурсы; сигнал; система массового обслуживания; беспроводная сеть; сети связи 4-го поколения

DOI: 10.14357/19922264170311

1 Введение

Для анализа показателей качества услуг в современных сетях связи 4-го поколения с объектами в движении широко используется имитационное моделирование [1, 2] и простые модели теории массового обслуживания [3] с фиксированным объемом требований.

В работах [4, 5] предлагается анализировать показатели эффективности модели современной беспроводной гетерогенной сети связи в виде СМО ограниченной емкости с требованиями случайного объема. В отличие от моделей, представленных в [6, 7], моделирование беспроводной сети в терминах теории массового обслуживания учитывает процессы установления новых сессий и их завершения, а заданная специальным образом функция распределения случайных требований к радиоресурсам позволяет описать функционирование планировщика в соответствии с выбранной политикой распределения частотных ресурсов и моделью распространения сигнала.

В предложенной в [8] экспоненциальной модели каждая сессия занимает выделенный ей объем частотного ресурса на все время ее длительности. По завершении сессии предполагается освободить некоторый случайный объем ресурсов, отличный от занимаемого, так как местоположение и число пользователей в сети могут с течением времени измениться. Однако данная модель не учитывает изменения в сети, которые могут произойти до завершения обслуживания сессий.

В данной работе исследуются модели, в которых объем занимаемых заявками ресурсов может меняться до завершения обслуживания, при поступлении сигнала. Эта особенность позволяет моделировать функционирование беспроводной сети, в которой пользователи, удаляясь или приближаясь к базовой станции, увеличивают или уменьшают требуемый объем ресурсов в течение периода жизни пользовательской сессии. При этом в связи с тем, что в беспроводных сетях задача поддержания уже принятых сессий имеет более высокий приоритет по сравнению с задачей принятия на обслуживание новых, диспетчеры ресурсов планируют их таким образом, чтобы ни в коем случае не прерывать текущие сессии [9]. Поэтому в исследуемых в данной работе моделях поступление сигнала, изменяющего

*Исследование выполнено при финансовой поддержке Российского научного фонда в рамках научного проекта № 16-11-10227.

[1]Российский университет дружбы народов; Институт проблем информатики Федерального исследовательского центра «Информатика и управление» Российской академии наук, samouylov_ke@rudn.university

[2]Российский университет дружбы народов; Институт проблем информатики Федерального исследовательского центра «Информатика и управление» Российской академии наук, sopin_es@rudn.university

[3]Институт проблем информатики Федерального исследовательского центра «Информатика и управление» Российской академии наук, sshorgin@ipiran.ru

объем занятого заявкой ресурса, не может привести к потере заявки.

Рассматриваются два сценария перемещения пользователей в беспроводной сети. Согласно первому сценарию пользователи перемещаются независимо друг от друга; следовательно, моменты изменения объемов занимаемых заявками ресурсов не зависят друг от друга.

Второй сценарий предполагает одновременное перемещение пользователей в соте, например в общественном транспорте. В этом случае все пользователи перемещаются относительно базовой станции одновременно и достаточно иметь представление о том, каким образом изменяется совокупно занимаемый объем ресурса.

2 Независимое перемещение пользователей

Рассмотрим модель соты, которая может обслуживать одновременно не более N сессий связи с устройствами. Объем доступных частотно-временны́х ресурсов ограничен и не превышает R единиц. Для установления новой сессии каждое устройство требует выделить ему некоторый случайный объем ресурсов $0 \le r \le R$. Если в соте одновременно установлено не более N сессий и требование к ресурсу новой сессии не превышает объема свободного ресурса, то сессия будет установлена, в противном случае она будет отклонена. Пусть в некоторый момент времени один из активных пользователей изменит свое местоположение относительно базовой станции; в этом случае он либо освободит часть занимаемого ресурса, либо базовая станция выделит ему дополнительный ресурс, не превышающий объем доступного в этот момент ресурса соты.

Опишем теперь предложенную модель в терминах теории массового обслуживания. Рассматривается СМО с $N < \infty$ приборами, обладающая некоторым объемом ресурсов $R < \infty$. Введем основные предположения.

1. В систему поступает пуассоновский поток заявок с интенсивностью λ, время обслуживания заявок имеет экспоненциальное распределение с параметром μ.
2. Для обслуживания i-й поступающей заявки требуется r_i ресурса, $r_i \ge 0$, с вероятностью p_{r_i}.
3. Если в момент поступления i-й заявки в системе находится $k < N$ заявок, занимающих $r_\bullet = r_1 + \cdots + r_k$ ресурсов, и $r_i \le R - r_\bullet$, то заявка будет принята к обслуживанию; в противном случае заявка будет потеряна.
4. Каждая заявка, находящаяся на обслуживании, порождает пуассоновский поток сигналов с интенсивностью γ, при поступлении которого она освобождает весь занимаемый ею ресурс, чтобы занять новый объем ресурса.

Пусть в некоторый момент времени $t > 0$ в системе находится $\xi(t)$ заявок, которые занимают $\eta_1(t), \ldots, \eta_{\xi(t)}(t)$ ресурсов. Функционирование системы описывает случайный процесс (СП) $X(t) = \big(\xi(t), \eta_1(t), \ldots, \eta_{\xi(t)}(t)\big)$, однако для дальнейшего анализа модели удобно воспользоваться упрощением, предложенным в [10] для СМО с ресурсами, позволяющими снизить размер пространства состояний системы за счет отслеживания только суммарного объема ресурсов $\delta(t) = \eta_1(t) + \cdots + \eta_{\xi(t)}(t)$. В дальнейшем будем исследовать СП $\tilde{X}(t) = (\xi(t), \delta(t))$.

Рассмотрим подробнее возможные переходы между состояниями системы. Пусть в некоторый момент времени система находится в состоянии (k, r). С вероятностью p_j в систему может поступить заявка, которая займет j единиц ресурса, если $j \le R - r$. Из-за того что неизвестно число занимаемых ресурсов каждой заявкой, невозможно точно определить объем высвобождаемых ресурсов при завершении обслуживания. Поэтому будем считать, что заявка освобождает i единиц ресурса с вероятностью $p_i p_{r-i}^{(k-1)} / p_r^{(k)}$, где $p_r^{(k)}$ является k-кратной сверткой распределения $\{p_i\}$, $i \ge 0$. Данную вероятность можно интерпретировать как вероятность того, что заявка занимает i единиц ресурса при условии, что k заявок суммарно занимают r ресурсов.

В момент поступления сигнала одна из заявок системы сначала освобождает занимаемые ею i единиц ресурса с вероятностью $p_i p_{r-i}^{(k-1)} / p_r^{(k)}$ и занимает j единиц ресурса с нормированной вероятностью $p_j / \sum_{s=0}^{R-r+i} p_s$, поскольку потери при поступлении сигнала по условию не происходят.

Пространство состояний системы описывается множеством

$$\tilde{\mathsf{X}} = \bigcup_{k=0}^{N} \tilde{\mathsf{X}}_k,$$

где $\tilde{\mathsf{X}}_k = \Big\{(k, r) : 0 \le r \le R,\ p_r^{(k)} > 0\Big\}$. Упорядочив состояния в множествах $\tilde{\mathsf{X}}_k$, $0 \le k \le N$, по возрастанию числа ресурсов, введем функции $I(k, r)$, значения которых равны порядковому номеру состояния (k, r) в множестве $\tilde{\mathsf{X}}_k$.

Матрица интенсивностей переходов СП $\tilde{X}(t)$

$$\mathbf{A} = [a((i, j), (k, r))]$$

является блочной трехдиагональной матрицей с диагональными блоками $\mathbf{\Psi}_0, \mathbf{\Psi}_1, \ldots, \mathbf{\Psi}_N$, наддиа-

гональными блоками $\mathbf{\Lambda}_1, \ldots, \mathbf{\Lambda}_N$ и поддиагональными блоками $\mathbf{M}_0, \ldots, \mathbf{M}_{N-1}$, где

$$\Psi_0 = -\lambda \sum_{j=0}^{R} p_j \,;\; \mathbf{\Lambda}_1 = (\lambda p_0, \ldots, \lambda p_r) \,;$$
$$\mathbf{M}_0 = (\mu, \ldots, \mu)^{\mathrm{T}} \,,$$

а остальные матрицы $\{\mathbf{\Psi}_n\}_{1\le n\le N}$, $\{\mathbf{\Lambda}_n\}_{2\le n\le N}$ и $\{\mathbf{M}_n\}_{1\le n\le N-1}$ имеют следующие элементы:

$$\psi_n(I(n,i),I(n,j)) = \begin{cases} -\left[\lambda \sum\limits_{k=0}^{R-i} p_k + n\mu + n\gamma\right], & i=j\,; \\ n\gamma \sum\limits_{s=0}^{i} \dfrac{p_s p_{i-s}^{(n-1)}}{p_i^{(n)}} \dfrac{p_{j-i+s}}{\sum_{k=0}^{R-i+s} p_k}, & i<j\,; \\ n\gamma \sum\limits_{s=i-j}^{i} \dfrac{p_s p_{i-s}^{(n-1)}}{p_i^{(n)}} \dfrac{p_{j-i+s}}{\sum_{k=0}^{R-i+s} p_k}, & i>j\,, \end{cases}$$
$$(n,i),(n,j) \in \mathsf{X}_n^{\sim},\; n = \overline{1,N-1}\,; \quad (1)$$

$$\lambda_n(I(n-1,i),I(n,j)) = \begin{cases} \lambda p_{j-i}\,, & i \le j \le R\,; \\ 0\,, & j<i\,, \end{cases}$$
$$(n-1,i) \in \mathsf{X}_{n-1}^{\sim},\; (n,j) \in \mathsf{X}_n^{\sim},\; n = \overline{2,N}\,; \quad (2)$$

$$\mu_n(I(n+1,i),I(n,j)) = \begin{cases} (n+1)\mu \dfrac{p_{i-j} - p_j^{(n)}}{p_i^{(n+1)}}, & j \le i \le R\,; \\ 0\,, & j>i\,, \end{cases}$$
$$(n+1,i) \in \mathsf{X}_{n+1}^{\sim},\; (n,j) \in \mathsf{X}_n^{\sim},\; n = \overline{1,N-1}\,; \quad (3)$$

$$\psi_N(I(N,i),I(N,j)) = \begin{cases} -[N\mu + N\gamma]\,, & i=j\,; \\ N\gamma \sum\limits_{s=0}^{i} \dfrac{p_s p_{i-s}^{(N-1)}}{p_i^{(N)}} \dfrac{p_{j-i+s}}{\sum_{k=0}^{R-i+s} p_k}, & i<j\,; \\ N\gamma \sum\limits_{s=i-j}^{i} \dfrac{p_s p_{i-s}^{(N-1)}}{p_i^{(N)}} \dfrac{p_{j-i+s}}{\sum_{k=0}^{R-i+s} p_k}, & i>j\,, \end{cases}$$
$$(N,i),(N,j) \in \mathsf{X}_N^{\sim}\,. \quad (4)$$

Стационарные вероятности

$$q_0 = \lim_{t\to\infty} \mathsf{P}\{\xi(t)=0\}\,; \quad (5)$$
$$q_k(r) = \lim_{t\to\infty} \mathsf{P}\{\xi(t)=k,\, \delta(t)=r\}\,,\; (k,r) \in \mathsf{X}_k^{\sim} \quad (6)$$

являются единственным решением системы уравнений равновесия (СУР):

$$\lambda q_0 \sum_{j=0}^{R} p_j = \mu \sum_{j:\,(1,j)\in\mathsf{X}_1^{\sim}} q_1(j)\,;$$

$$\left(\lambda \sum_{j=0}^{R-r} p_j + k\mu + k\gamma\right) q_k(r) = \lambda \sum_{j\ge 0\,,\,(k-1,r-j)\in\mathsf{X}_{k-1}^{\sim}} q_{k-1}(r-j)p_j + (k+1)\mu \sum_{j\ge 0\,,\,(k+1,r+j)\in\mathsf{X}_{k+1}^{\sim}} q_{k+1}(r+j)\frac{p_j p_r^{(k)}}{p_{j+r}^{(k+1)}} + k\gamma \sum_{j:\,(k,j)\in\mathsf{X}_k^{\sim}} q_k(j) \sum_{i=\max(0,j-r)}^{j} \frac{p_i p_{j-i}^{(k-1)}}{p_j^{(k)}} \frac{p_{r-j+i}}{\sum_{s=0}^{R-j+i} p_s},$$
$$1 \le k \le N-1\,,\; (k,r) \in \mathsf{X}_k^{\sim}\,;$$

$$(N\mu + k\gamma) q_N(r) = \lambda \sum_{j\ge 0,\,(N-1,r-j)\in\mathsf{X}_{N-1}^{\sim}} q_{N-1}(r-j)p_j + N\gamma \sum_{j:\,(N,j)\in\mathsf{X}_N^{\sim}} q_N(j) \sum_{i=\max(0,j-r)}^{j} \frac{p_i p_{j-i}^{(N-1)}}{p_j^{(N)}} \frac{p_{r-j+i}}{\sum_{s=0}^{R-j+i} p_s},$$
$$(N,r) \in \mathsf{X}_N^{\sim}\,.$$

Стационарные вероятности (5) и (6) могут быть найдены численно методом UL-разложения СУР в матричном виде

$$\mathbf{q}^{\mathrm{T}}\mathbf{A} = \mathbf{0}^{\mathrm{T}}\,; \quad \mathbf{q}^{\mathrm{T}} \cdot \mathbf{1} = 1\,.$$

Обозначим подвекторы стационарных вероятностей $\mathbf{q}_0 = \{q_0\}$ и $\mathbf{q}_k = \{q_k(r)\}_{(k,r)\in\mathsf{X}_k^{\sim}}$ для всех $1 \le k \le N$, тогда СУР в матричном виде с учетом блочно-трехдиагонального вида матрицы интенсивностей переходов $\mathbf{A}$ примет вид:

$$\mathbf{q}_0\mathbf{\Psi}_0 - \mathbf{q}_1\mathbf{M}_0 = \mathbf{0}\,; \quad (7)$$
$$\mathbf{q}_i\mathbf{\Psi}_i - \mathbf{q}_{i+1}\mathbf{M}_i - \mathbf{q}_{i-1}\mathbf{\Lambda}_i = \mathbf{0}\,,\; 1 \le i \le N-1\,; \quad (8)$$
$$\mathbf{q}_N\mathbf{\Psi}_N - \mathbf{q}_{N-1}\mathbf{\Lambda}_N = \mathbf{0}\,. \quad (9)$$

3 Групповое перемещение пользователей

Теперь рассмотрим сценарий, при котором пользователи перемещаются относительно базовой станции совместно. В этом случае в момент срабатывания сигнала изменяется объем занимаемых ресурсов каждой сессии и, соответственно,

изменяется объем занимаемых в совокупности ресурсов всеми активными сессиями в соте. Важно, что при выделении дополнительных ресурсов прерывания сессий не происходит.

Функционирование СМО описывается пп. 1–3 из предыдущего раздела и п. 4*, который сформулируем следующим образом:

4*. В систему поступает пуассоновский поток сигналов с интенсивностью γ, при поступлении которого заново разыгрывается объем занимаемых всеми заявками ресурсов.

Поведение системы во времени описывает СП $X^*(t) = (\xi)t), \delta(t))$, где $\xi(t)$ — число заявок в системе; $\delta(t)$ — объем совокупно занятых ресурсов. Пространство состояний СП $X^*(t)$ идентично пространству состояний процесса $\tilde{X}(t)$. Обозначим распределение стационарных вероятностей:

$$q_0^* = \lim_{t\to\infty} \mathsf{P}\{\xi(t) = 0\}\,;$$
$$q_k(r) = \lim_{t\to\infty} \mathsf{P}\{\xi(t) = k,\ \delta(t) = r\}\,,\ \ (k,r) \in \mathsf{X}_k^{\sim}\,.$$

Переходы между состояниями системы, соответствующие поступлению новых заявок и завершению облуживания заявок системы, происходят аналогично переходам между состояниями модели с независимым перемещением пользователей; различие возникает в переходах, соответствующих поступлению сигналов. В момент срабатывания сигнала система из состояния (k,j) совершает переход в состояние (k,r) с вероятностью $p_r^{(k)} / \sum_{i=0}^{R} p_i^{(k)}$. Таким образом, СУР СП $X^*(t)$ принимает вид:

$$\lambda q_0^* \sum_{j=0}^{R} p_j = \mu \sum_{j:\ (1,j)\in\mathsf{X}_1^{\sim}} q_1^*(j)\,; \qquad (10)$$

$$\left(\lambda \sum_{j=0}^{R-r} p_j + k\mu + \gamma\right) q_k^*(r) =$$
$$= \lambda \sum_{j\ge 0:\ (k-1;r-j)\in\mathsf{X}_{k-1}^{\sim}} p_j q_{k-1}^*(r-j) +$$
$$+ (k+1)\mu \sum_{j\ge 0:\ (k+1;r+j)\in\mathsf{X}_{k+1}^{\sim}} \frac{p_j p_r^{(k)}}{p_{j+r}^{(k+1)}}\, q_{k+1}^*(r+j) +$$
$$+ \gamma \sum_{i=0}^{R} \frac{p_r^{(k)}}{\sum_{i=0}^{R} p_i^{(k)}}\, q_k^*(j)\,,$$
$$1 \le k \le N-1\,,\ (k,r) \in \mathsf{X}_k^{\sim}\,; \qquad (11)$$

$$(N\mu+\gamma) q_N^*(r) = \lambda \sum_{j\ge 0:\ (N-1;r-j)\in\mathsf{X}_{N-1}^{\sim}} p_j q_{N-1}^*(r-j) +$$
$$+ \gamma \sum_{j:\ (N;r)\in\mathsf{X}_N^{\sim}} \frac{p_r^{(N)}}{\sum_{i=0}^{R} p_i^{(N)}}\, q_N^*(j)\,,\ \ (N,r) \in \mathsf{X}_N^{\sim}. \qquad (12)$$

Теорема 1. *Стационарные вероятности СМО со случайными требованиями и потоком сигналов, изменяющих суммарный объем занимаемых ресурсов, не зависят от интенсивности γ поступления сигналов и имеют вид*:

$$q_k^*(r) = q_0 \frac{\rho^k}{k!}\, p_r^{(k)}\,; \quad q_0^* = \left(\sum_{k=0}^{N}\sum_{r=0}^{R} \frac{\rho^k}{k!}\, p_r^{(k)}\right)^{-1}. \qquad (13)$$

Доказательство теоремы выполняется путем подстановки стационарных вероятностей (13) в СУР (10)–(12).

4 Численный пример

Согласно теореме 1 стационарные вероятности экспоненциальной СМО с групповым перемещением пользователей как частный случай экспоненциальной СМО со случайными требованиями из [8] не зависят от поступающего потока сигналов в систему. Вероятностные характеристики системы, такие как вероятность блокировки B и средний объем занятых ресурсов b, в этом случае могут быть найдены по формулам:

$$B = 1 - G^{-1}(N,R) \sum_{j=0}^{R} p_j G(N-1, R-j)\,;$$
$$b = R - G^{-1}(N,R) \sum_{j=1}^{R} G(N, R-j)\,,$$

полученным по аналогии с [11] с помощью рекуррентного алгоритма вычисления нормировочной константы

$$G(N,R) = \sum_{k=0}^{N}\sum_{r=0}^{R} \frac{\rho_k}{k!}\, p_r^{(k)}\,.$$

Стационарные вероятности (5)–(6) СМО с независимым перемещением пользователей могут быть найдены численно как решения системы матричных уравнений (7)–(9). Вероятностные характеристики системы в этом случае определяются формулами:

$$B = 1 - \sum_{k=0}^{N-1} \sum_{r:\ (k,r)\in\mathsf{X}_k^{\sim}} q_k(r) \sum_{j=0}^{R-r} p_j\,;$$
$$b = \sum_{k=0}^{N} \sum_{r:\ (k,r)\in\mathsf{X}_k^{\sim}} r q_k(r)\,.$$

В качестве примеров распределений требований к ресурсу рассматривались биномиальное распределение $\mathrm{Binom}\,(r,p)$ и геометрическое распределение $\mathrm{Geom}\,(p)$, как и в [12]. Для анализа зависимости вероятностных характеристик СМО с независимым

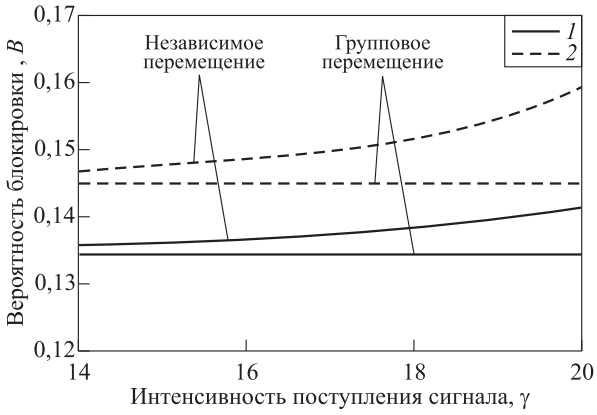


Рис. 1 Зависимость вероятности блокировки от интенсивности поступления сигнала: *1* — Binom; *2* — Geom

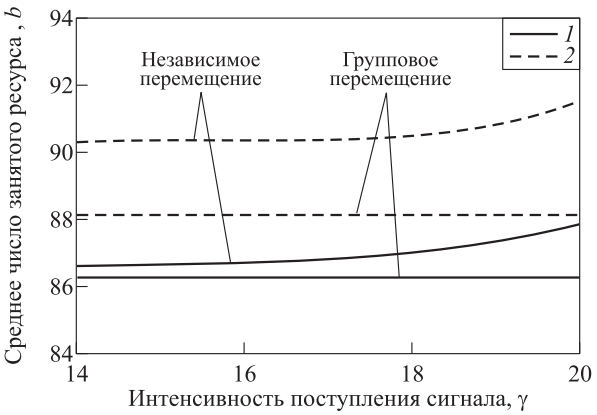


Рис. 2 Зависимость среднего объема занятого ресурса от интенсивности поступления сигнала: *1* — Binom; *2* — Geom

перемещением пользователей от интенсивности γ потока поступающих сигналов в качестве примера рассматриваются:

(1) биномиальное распределение $\mathrm{Binom}\,(r,p)$ требований к ресурсу с параметрами $r \geq 0$ и $0 \leq p \leq 1$, где $p_i = \binom{r}{i} p^i (1-p)^{r-i}$ — вероятность того, что заявка потребует i единиц ресурса, $0 \leq i \leq r$, $p = \overline{m}/r$;

(2) геометрическое распределение $\mathrm{Geom}(p)$ требований к ресурсу с параметром $0 \leq p \leq 1$, где $p_i = p^i(1-p)$ — вероятность того, что заявка потребует i единиц ресурса, $1 \leq i \leq r$, $p = 1/(\overline{m}+1)$.

Для вычисления элементов матриц (1)–(4), которые определяют необходимые компоненты решения СУР (7)–(9), найдем все k-кратные свертки $p_r^{(k)}$ для каждого из предложенных распределений требований к ресурсу. При условии биномиального распределения требований вероятность того, что k заявок системы занимают j единиц ресурса:

$$p_j^{(k)} = \binom{kr}{j} p^j (1-p)^{kr-j};$$

для геометрического закона:

$$p_j^{(k)} = \binom{k+j-1}{k} p^j (1-p)^k.$$

Целочисленный параметр r биномиального распределения требований к ресурсу и параметры p распределений были подобраны таким образом, чтобы математическое ожидание $\overline{m}$ было одинаковым. Максимальное число единиц ресурса, требуемых одной заявке, при биномиальном распределении, таким образом, оказалось $r = 18$, а математическое ожидание для биномиального и геометрического распределений $\overline{m} = 5{,}4$.

Рассматривается пример соты, которая может обслуживать до 100 сессий одновременно, а ресурс выделяется пользователям в процентном соотношении от 100% всего доступного соте ресурса, $N = R = 100$. Средняя продолжительность сессии составляет $\mu = 1$ мин, а среднее число запросов на установление сессии $\lambda = 16$, как оптимальное значение нагрузки.

На рис. 1 и 2 представлены результаты расчета вероятности блокировки системы и среднего объема занимаемых ресурсов в зависимости от поступления сигналов, моделирующих перемещение пользователей в соте.

На рис. 1 можно видеть, что вероятность блокировки в системе с независимым перемещением пользователей растет с ростом γ, несмотря на то что поступление сигнала не может вызвать потери заявки. Наблюдаемый эффект связан с тем, что с повышением интенсивности поступления сигналов заявки интенсивнее используют доступный ресурс системы, как видно на рис. 2, и в результате в системе остается меньше свободного ресурса для принятия новых заявок.

5 Заключение

В работе проведен анализ ресурсной СМО с сигналами, при поступлении которых изменяется объем занимаемых заявками ресурсов. Модель позволяет проводить анализ показателей эффективности беспроводной сети, учитывая перемещение пользователей в радиусе действия. Рассмотрены частные случаи независимого и группового перемещений пользователей. В частности, было доказано, что

при групповом перемещении пользователей показатели качества сети не зависят от интенсивности изменения положения группы относительно базовой станции.

В дальнейшем планируется разработать эффективный вычислительный алгоритм расчета вероятностно-временны́х характеристик модели.

Литература

1. *Boban M., Barros J., Tonguz O. K.* Geometry-based vehicle-to-vehicle channel modeling for large-scale simulation // IEEE T. Veh. Technol., 2014. Vol. 63. No. 9. P. 4146–4164.
2. *Khan M., Han K.* An optimized network selection and handover triggering scheme for heterogeneous self-organized wireless networks // Math. Probl. Eng., 2014. Vol. 2014. No. 2. P. 173068-1–173068-11. https://www.hindawi.com/journals/mpe/2014/173068.
3. *Fowler S., Häll C. H., Yuan D., Baravdish D., Mellouk A.* Analysis of vehicular wireless channel communication via queueing theory model // IEEE Conference (International) on Communications. — Piscataway, NJ, USA: IEEE, 2014. P. 1736–1741.
4. *Наумов В. А., Самуйлов К. Е.* О моделировании систем массового обслуживания с множественными ресурсами // Вестник РУДН. Сер. Математика. Информатика. Физика, 2014. № 3. С. 60–64.
5. *Naumov V., Samouylov K., Sopin E., Andreev S.* Two approaches to analysis of queuing systems with limited resources // Ultra Modern Telecommunications and Control Systems and Workshops Proceedings. — Piscataway, NJ, USA: IEEE, 2014. P. 485–488.
6. *Elshaer H., Boccardi F., Dohler M., Irmer R.* Downlink and uplink decoupling: A disruptive architectural design for 5G networks // IEEE Global Communications Conference Proceedings. — Piscataway, NJ, USA: IEEE, 2014. P. 1798–1803.
7. *Singh S., Zhang X., Andrews J.* Joint rate and SINR coverage analysis for decoupled uplink downlink biased cell associations in HetNets // IEEE T. Wirel. Commun., 2015. Vol. 14. No. 10. P. 5360–5373. doi: 10.1109/TWC.2015.2437378.
8. *Наумов В. А., Самуйлов К. Е., Самуйлов А. К.* О суммарном объеме ресурсов, занимаемых обслуживаемыми заявками // Автоматика и телемеханика, 2016. № 8. С. 125–132.
9. *Bartolini N., Chlamtac I.* Call admission control in wireless multimedia networks // 13th IEEE Symposium (International) on Personal, Indoor and Mobile Radio Communications Proceedings. — Piscataway, NJ, USA: IEEE, 2002. Vol. 1. P. 285–289. doi: 10.1109/PIMRC.2002.1046706.
10. *Naumov V., Samouylov K., Yarkina N., Sopin E., Andreev S., Samuylov A.* LTE performance analysis using queuing systems with finite resources and random requirements // 7th Congress on Ultra Modern Telecommunications and Control Systems Proceedings. — Piscataway, NJ, USA: IEEE, 2015. P. 100–103.
11. *Вихрова О. Г.* К вычислению вероятностных характеристик СМО ограниченной емкости со случайными требованиями к ресурсам // Вестник РУДН. Сер. Математика. Информатика. Физика, 2017. Т. 25. № 3. С. 203–210.
12. *Вихрова О. Г., Самуйлов К. Е., Сопин Э. С., Шоргин С. Я.* К анализу показателей качества обслуживания в современных беспроводных сетях // Информатика и её применения, 2015. Т. 9. Вып. 4. С. 48–55.

Поступила в редакцию 29.06.17

QUEUING SYSTEMS WITH RESOURCES AND SIGNALS AND THEIR APPLICATION FOR PERFORMANCE EVALUATION OF WIRELESS NETWORKS

K. E. Samouylov[1,2], E. S. Sopin[1,2], and S. Ya. Shorgin[2]

[1]Peoples' Friendship University of Russia, 6 Miklukho-Maklaya Str., Moscow 117198, Russian Federation
[2]Institute of Informatics Problems, Federal Research Center "Computer Sciences and Control" of the Russian Academy of Sciences, 44-2 Vavilov Str., Moscow 119333, Russian Federation

Abstract: The paper considers a queuing system with limited resources, random requirements, and signals. Each customer occupies a server and a random amount of resources for the whole service duration. Besides, a Poisson flow of signals arrives to the queue. Signal arrival triggers the resource reallocation process. The model can describe functioning of a wireless network taking into account user movement during a session. Two cases are considered: independent movement of users, when resources are reallocated independently for each session, and joint movement, when all resources are reallocated at once.

Keywords: queuing system; random requirement; signals; limited resources; wireless network; LTE-advanced

DOI: 10.14357/19922264170311

Acknowledgments

This work was financially supported by the Russian Science Foundation (grant No. 16-11-10227).

References

1. Boban, M., J. Barros, and O. Tonguz. 2014. Geometry-based vehicle-to-vehicle channel modeling for large-scale simulation. *IEEE T. Veh. Technol.* 63(9):4146–4164.
2. Khan, M., and K. Han. 2014. An optimized network selection and handover triggering scheme for heterogeneous self-organized wireless networks. *Math. Probl. Eng.* 2014(2):173068-1–173068-11. Available at: https://www.hindawi.com/journals/mpe/2014/173068 (accessed Sptember 11, 2017).
3. Fowler, S., S. Häll, D. Yuan, D. Baravdish, and A. Mellouk. 2014. Analysis of vehicular wireless channel communication via queueing theory model. *IEEE Conference (International) on Communications Proceedings*. Piscataway, NJ: IEEE. 1736–1741.
4. Naumov, V., and K. Samouylov. 2014. O modelirovanii system massovogo obsluzivaniya s mnozhestnennymi resursami [On the modeling of queuing systems with multiple resources]. *Vestnik RUDN. Ser. matematika, fizika, informatika* [RUDN J. Mathematics, information science and physics ser.] 22(3):60–64.
5. Naumov, V., K. Samouylov, E. Sopin, and S. Andreev. 2014. Two approaches to analysis of queuing systems with limited resources. *Ultra Modern Telecommunications and Control Systems and Workshops Proceedings*. Piscataway, NJ: IEEE. 485–488.
6. Elshaer, H., F. Boccardi, M. Dohler, and R. Irmer. 2014. Downlink and uplink decoupling: A disruptive architectural design for 5G networks. *Global Communications Conference Proceedings*. Piscataway, NJ: IEEE. 1798–1803.
7. Singh, S., X. Zhang, and J. Andrews. 2015. Joint rate and SINR coverage analysis for decoupled uplink-downlink biased cell associations in HetNets. *IEEE T. Wirel. Commun.* 14(10):5360–5373. doi: 10.1109/TWC.2015.2437378.
8. Naumov, V., K. Samuilov, and A. Samuilov. 2016. On the total amount of resources occupied by serviced customers. *Automat. Remote Control* 77(8):1419–1427. doi:10.1134/S0005117916080087.
9. Bartolini, N., and I. Chlamtac. 2002. Call admission control in wireless multimedia networks. *13th IEEE Symposium (International) on Personal, Indoor and Mobile Radio Communications Proceedings*. 1:285–289. doi: 10.1109/PIMRC.2002.1046706.
10. Naumov, V., K. Samouylov, N. Yarkina, E. Sopin, S. Andreev, and A. Samuylov. 2015. LTE performance analysis using queuing systems with finite resources and random requirements. *7th Congress (International) on Ultra Modern Telecommunications and Control Systems Proceedings*. Piscataway, NJ: IEEE. 100–103.
11. Vikhrova, O. 2017. K vychisleniyu veroyatnostnykh kharakteristik sistemy massovogo obslyuzivaniya ogranichennoy emkosti so sluchaynymi trebovaniyamy k resursu [About probability characteristics evaluation in queuing system with limited resources and random requirements]. *Vestnik RUDN. Ser. matematika, fizikam informatika* [RUDN J. Mathematics, information science, and physics ser.] 25(3):203–210.
12. Vikhrova, O., K. Samouylov, E. Sopin, and S. Shorgin. 2015. K analizy pokazateley kachestva obsluzhivaniya v sovremennykh besprovodnykh setyakh [On performance analysis of modern wireless networks]. *Informatika i ee Primeneniya — Inform. Appl.* 9(4):48–55.

Received June 29, 2017

Contributors

Samouylov Konstantin E. (b. 1955) — Doctor of Science in technology, professor; Head of Department, Peoples' Friendship University of Russia (RUDN University), 6 Miklukho-Maklaya Str., Moscow 117198, Russian Federation; senior scientist, Institute of Informatics Problems, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, 44-2 Vavilov Str., Moscow 119333, Russian Federation; samouylov_ke@rudn.university

Sopin Eduard S. (b. 1986) — Candidate of Science in physics and mathematics; associated professor, Peoples' Friendship University of Russia (RUDN University), 6 Miklukho-Maklaya Str., Moscow 117198, Russian Federation; senior scientist, Institute of Informatics Problems, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, 44-2 Vavilov Str., Moscow 119333, Russian Federation; sopin_es@rudn.university

Shorgin Sergey Ya. (b. 1952) — Doctor of Science in physics and mathematics, professor; Deputy Director, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences (FRC CSC RAS); principal scientist, Institute of Informatics Problems, FRC CSC RAS, 44-2 Vavilov Str., Moscow 119333, Russian Federation; sshorgin@ipiran.ru

INFORMATICS AND APPLICATIONS 2017. Vol. 11. Iss. 3. P. 106–112

REVISITING JOINT STATIONARY DISTRIBUTION IN TWO FINITE CAPACITY QUEUES OPERATING IN PARALLEL

L. Meykhanadzhyan[1], S. Matyushenko[2], D. Pyatkina[3], and R. Razumchik[4]

Abstract: The paper revisits the problem of the computation of the joint stationary probability distribution p_{ij} in a queueing system consisting of two single-server queues, each of capacity $N \geq 3$, operating in parallel, and a single Poisson flow. Upon each arrival instant, one customer is put simultaneously into each system. When a customer sees a full system, it is lost. The service times are exponentially distributed with different parameters. Using the approach based on generating functions, the authors obtain a new system of equations of a smaller size than the size of the original system of equilibrium equations ($3N-2$ compared to $(N+1)^2$). Given the solution of the new system, the whole joint stationary distribution can be computed recursively. The new system gives some insights into the interdependence of p_{ij} and p_{nm}. If relations between $p_{i-1,N}$ and $p_{i,N}$ for $i = 3, 5, 7, \ldots$ are known, then the blocking probability can be computed recursively. Using the known results for the asymptotic behavior of p_{ij} as $i, j \to \infty$, the authors illustrate this idea by a simple numerical example.

Keywords: two queues; generating function; stationary distribution; paired customers

DOI: 10.14357/19922264170312

1 Introduction

The system with two single-server queues (both limited and unlimited capacity cases) operating in parallel has received significant attention in the literature due its potential application in real-life scenarios (for example, packet switches, packet radio networks, parallel processing systems, inventory control of database systems, etc.). Further, it is assumed that the system consists of two queues (say, queue 1 and queue 2) each with a single server and there is a single Poisson flow of customers arriving at it. Each customer upon arrival is instantly duplicated: one customer goes to queue 1 and the other goes to queue 2. Both queues are working independently, service times follow exponential distribution with different parameters, and the service discipline in a queue is either FCFS (first-come-first-served), LCFS (last-come-first-served), or Random. Despite the simplicity of the structure, even under such markovian assumptions, the system turned out to be notoriously hard to analyze.

A big list of publications on the topic is given in [1], where the authors give an overview of functional equations (and solution approaches), which arise in the analysis of such systems with infinite capacity queues. References to the application related papers are also given. Among the pioneer works in the area, papers [2–5] are worth noticing.

In this paper, the authors revisit the problem of the computation of the joint stationary distribution in the case, when both queues have finite capacity. Under the exponential assumptions (and given additional dedicated Poisson flows to each queue), the matrix algorithm has been proposed already in [3]. Some further considerations, including the study of correlation between the queues' sizes were continued in [14]. In general, the cases, when both of queues are on finite capacity or one of the queues is (see, for example, [6]), have received less attention in the literature. This is presumably due to the fact that in those cases in order to obtain the joint stationary distribution, one can use widely-adopted general techniques: folding algorithm, linear level reduction or block-gaussian elimination algorithms (see, for example, [7, 8]).

Our motivation for revisiting this problem comes from the papers [9–13], where the generating function technique (which utilizes some properties of special functions (Chebyshev and Gegenbauer polynomials)) was applied to the systems with two finite-capacity queues and allowed one to derive new relations for the recursive computation of the joint stationary distribution.

[1]School No. 281 of Moscow, 7 Raduzhnaya Str. Moscow 129344, Russian Federation; lameykhanadzhyan@gmail.com

[2]Peoples' Friendship University of Russia (RUDN University), 6 Miklukho-Maklaya Str., Moscow 117198, Russian Federation; matyushenko_si@rudn.university

[3]Peoples' Friendship University of Russia (RUDN University), 6 Miklukho-Maklaya Str., Moscow 117198, Russian Federation; pyatkina_da@rudn.university

[4]Institute of Informatics Problems, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences; 44-2 Vavilova Str., Moscow 119333, Russian Federation; Peoples' Friendship University of Russia, 6 Miklukho-Maklaya Str., Moscow 117198, Russian Federation; rrazumchik@ipiran.ru

Having applied the same approach for the system considered here, the present authors found that it does not lead to the recursive solution. Yet, it gives an alternative way to compute the joint stationary distribution. Specifically, it requires the solution of the system of linear algebraic equations of the size $(3N-2)$, when the size of both queues is equal to $N \geq 3$ (the exact solutions for $N = 1$ and 2 are obtained in [14]) and is immediately suitable for exact arithmetics implementation. If the whole joint stationary distribution is not of importance, this approach gives the straight way to calculate the blocking probability and new insights into the dependencies between the joint probabilities p_{ij}, which prevent the recursive solution.

The paper is structured as follows. In section 2, the description of the system is given and some known results, which are necessary in what follows, are repeated. Section 3 contains the main contribution of the paper. Here, it is shown how new relations for the joint stationary distribution can be obtained (see Eqs. (5)–(12)). The insights into the interdependence between the joint stationary probabilities is discussed in section 4. Section 5 concludes the paper.

2 System Description

The system under consideration consists of two single-server finite capacity queues (queue 1 and queue 2), operating in parallel independently of each other. By suffering a little a lack of generality, let us assume that the capacities of both queues are equal to $N \geq 3$. There is one incoming Poisson flow of rate λ arriving at the system. Upon arrival, each customer is split into two customers: one enters queue 1 and another enters queue 2. Service time of customers in queue i follows exponential distribution with rate μ_i, $i = 1, 2$. Since we are interested here only in the queue size related characteristics, we allow the service discipline in queues to be either FCFS, or LCFS, or Random. We are interested in the case (as in [3]), when a customer always occupies the place in the queue whenever it is not full. This is much different from the case, when the customer checks the queues' sizes before splitting and leaves the system if at least one queue is full.

Denote by p_{ij} stationary probability of the fact that there are i customers in queue 1 and j customers in queue 2. From [3], it follows that the double generating function for p_{ij},

$$P(u,v) = \sum_{i=0}^{N}\sum_{j=0}^{N} u^i v^j p_{ij}\,, \quad 0 \leq u \leq 1\,,\ 0 \leq v \leq 1\,,$$

has the form:

$$B(u,v)P(u,v) = A(u,v) \tag{1}$$

where

$$B(u,v) = \lambda u^2 v^2 - u(\lambda v + \mu_1 v + \mu_2 v - \mu_2) + \mu_1 v\,;$$

$$A(u,v) = \mu_1 v\,(u-1)\sum_{j=0}^{N} v^j p_{0j}$$

$$+\,\mu_2 u\,(v-1)\sum_{i=0}^{N} u^i p_{i0} + \lambda v^2 u^{N+1}(1-u)\sum_{j=0}^{N} v^j p_{Nj}$$

$$+\,\lambda u^2 v^{N+1}(1-v)\sum_{i=0}^{N} u^i p_{iN}$$

$$+\,\lambda u^{N+1} v^{N+1}(1-u)(1-v)p_{NN}\,.$$

The quadratic polynomial $B(u,v)$ has two roots:

$$u_{1,2} = u_{1,2}(v) = \Big(v(\lambda+\mu_1+\mu_2) - \mu_2$$

$$\mp\sqrt{(v(\lambda+\mu_1+\mu_2)-\mu_2)^2 - 4\lambda\mu_1 v^3}\Big)\Big/(2\lambda v^2)\,.$$

The generating function $P(u,v)$ is the ratio of two polynomial functions. For each value of v, probability generating function $P(u,v)$ is a continuous function of u in the interval $[0,1]$. Then, since the left part in (1) vanishes at points $(u_1(v), v)$, and $(u_2(v), v)$, then the right part must vanish at these points too. In the next section, it will be shown that from this observation, one can obtain the system of linear algebraic equations only for the probabilities $\{p_{0j}, p_{jN}, 0 \leq j \leq N\}$ and $\{p_{j0}, 0 \leq j \leq N-3\}$ which can be solved by any standard numerical method. Once these probabilities are known, the computation of the rest joint stationary probabilities p_{ij} is performed recursively from the system of equilibrium equations.

3 New System of Equations

Both equations $A(u_1(v), v) = 0$ and $A(u_2(v), v) = 0$ share the same unknown quantities. If one expresses term with $\sum_{j=0}^{N} v^j p_{0j}$ from the first equation and put it in the second equation, after collecting common terms, one obtains:

$$\mu_2(v-1)\sum_{i=0}^{N}\left(\frac{u_2^{i+1}-u_1^{i+1}}{u_2-u_1} - u_1u_2\frac{u_2^i-u_1^i}{u_2-u_1}\right)p_{i0}$$

$$+\,\lambda v\,(1-u_1-u_2+u_1u_2)\,\frac{u_2^{N+1}-u_1^{N+1}}{u_2-u_1}\sum_{j=0}^{N} v^{j+1}p_{Nj}$$

$$+\,\lambda v^{N+1}(1-v)\sum_{i=0}^{N}\left(\frac{u_2^{i+2}-u_1^{i+2}}{u_2-u_1}\right.$$

$$\left.-\,u_1u_2\frac{u_2^{i+1}-u_1^{i+1}}{u_2-u_1}\right)p_{iN}$$

$$+\lambda v^{N+1}(1-u_1-u_2+u_1u_2)(1-v)\times\frac{u_2^{N+1}-u_1^{N+1}}{u_2-u_1}p_{NN}=0\,. \quad (2)$$

Instead of cancelling $\sum_{j=0}^{N}v^jp_{0j}$, let us express the term with p_{NN} from the $A(u_1(v),v)=0$ and put it into $A(u_2(v),v)=0$. By doing so, one gets another relation:

$$\begin{aligned}&v\mu_1(1-u_1-u_2+u_1u_2)\frac{{u_2}^{N+1}-{u_1}^{N+1}}{u_2-u_1}\sum_{j=0}^{N}v^jp_{0j}\\&+\mu_2(v-1)\sum_{i=0}^{N}(u_1u_2)^{i+1}\left(\frac{{u_2}^{N-i+1}-u_1^{N-i+1}}{u_2-u_1}\right.\\&\left.-\frac{{u_2}^{N-i}-u_1^{N-i}}{u_2-u_1}\right)p_{i0}\\&+\lambda v^{N+1}(1-v)\sum_{i=0}^{N}(u_1u_2)^{i+2}\left(\frac{u_2^{N-i}-{u_1}^{N-i}}{u_2-u_1}\right.\\&\left.-\frac{u_2^{N-i-1}-{u_1}^{N-i-1}}{u_2-u_1}\right)p_{iN}=0\,. \quad (3)\end{aligned}$$

It is straightforward to see that the roots $u_{1,2}$ admit the following representation:

$$u_1=\frac{\sqrt{\mu_1}\lambda v}{a}(x)\,;\;\; u_2=\frac{\sqrt{\mu_1}\lambda v}{b}(x)$$

where

$$x=\frac{v(\lambda+\mu_1+\mu_2)-\mu_2}{v\sqrt{\mu_1\lambda v}}\,;$$

$$a(x)=\frac{x-\sqrt{x^2-4}}{2}\,;\;\; b(x)=\frac{x+\sqrt{x^2-4}}{2}\,.$$

It can be shown that $|x|>2$ for all $v\in(0,1]$. It is well-known that the fraction $(b(x)^m-a(x)^m)/(b(x)-a(x))$ is in fact a polynomial in v for $m\geq 1$. Thus, $(u_2^m-u_1^m)/(u_2-u_1)$ is a polynomial in v as well. After some tedious algebra (derivation is analogous to the one in [11]), let us find that for $m\geq 1$, the following representation holds:

$$\frac{u_2(v)^m-u_1(v)^m}{u_2(v)-u_1(v)}=\left(\sqrt{\frac{\mu_1}{\lambda}}\right)^{m-1}\sum_{n=\lfloor m/2\rfloor}^{2(m-1)}v^{-n}a_{m,n} \quad (4)$$

where

$$a_{m,n}=\sum_{j=\max\{0,\lceil n-(m-1)\rceil\}}^{\lfloor(2n-(m-1))/3\rfloor}d_{m,2n-(m-1)-2j,j}\,,$$

$$\frac{m-1}{2}\leq n\leq 2(m-1)\,;$$

$$d_{i,m,k}=C_{i-m-1}^{m+1}(0)\binom{m}{k}\left(\frac{\lambda+\mu_1+\mu_2}{\sqrt{\lambda\mu_1}}\right)^{m-k}\left(-\frac{\mu_2}{\sqrt{\lambda\mu_1}}\right)^{k}.$$

Here, $\binom{m}{k}$ is the binomial coefficient and $C_n^m(0)$ denotes the value of Gegenbauer polynomial $C_n^m(x)$ at point $x=0$ (see, for example, [15, p. 175]). Since each fraction $(u_2^m-u_1^m)/(u_2-u_1)$ is a polynomial in v with real coefficients (defined by (4)) and $u_1u_2=\mu_1/\lambda v$, $u_1+u_2=((\lambda+\mu_1+\mu_2)v-\mu_2)/(\lambda v^2)$, both expressions on the left in (2) and (3) are polynomials in v as well with real coefficients depending on λ, μ_1, μ_2 and certain p_{ij}. Due to the lack of space we omit detailed derivations and just state the final result. From the fact that both polynomials (2) and (3) are equal to zero for $v\in(0,1]$, it follows that their coefficients are equal to zero. This leads to two systems of linear algebraic equations (one from (2) and the other from (3)) for the stationary probabilities on the boundaries (p_{0j}, p_{i0}, p_{iN}, and p_{Nj}). Careful inspection shows that from these two systems, one can draw one single system of equations of size $3N-4$ for the probabilities $\{p_{0j},p_{jN},\,0\leq j\leq N\}$ and $\{p_{j0},0\leq j\leq N-3\}$, which can be solved numerically. Specifically, for odd $N\geq 3$, the new system of equations has the form:

$$\begin{aligned}&\sum_{j=0}^{N}p_{0j}a_{N+1,j+(N-1)/2}\\&+\sum_{j=0}^{(N-1)/2-1}p_{jN}\lambda r^{j+2}b_{N-j,N+(N-1)/2-j}\\&-p_{00}\mu_2rb_{N+1,(N-1)/2}=0\,; \quad (5)\\&\sum_{j=i}^{N}p_{0j}a_{N+1,j+(N-1)/2-i}\\&+\sum_{j=0}^{(N-1)/2+(i-1)}p_{jN}\lambda r^{j+2}b_{N-j,N+(N-1)/2-j-i}=0\,,\\&i=\overline{1,\frac{N-1}{2}}\,; \quad (6)\\&\sum_{j=(N+1)/2}^{N}p_{0j}a_{N+1,j-1}+\sum_{i=0}^{N-1}p_{iN}\lambda r^{i+2}b_{N-i,N-i-1}\\&+p_{NN}\lambda r^{N+1}b_{1,0}=0\,; \quad (7)\\&\sum_{j=N-i}^{N}p_{0j}a_{N+1,j+i-(N+1)/2}\\&+\sum_{j=0}^{2i+1}p_{jN}r^{j+2}b_{N-j,(N-1)/2+i-j}=0\,,\\&i=\overline{1,\frac{N-3}{2}}\,; \quad (8)\end{aligned}$$

Algorithm 1 Recursive computation of p_{ij}

For $0 \le i \le N-2$, $p_{i,N-1} \leftarrow p_{i+1,N}\left(1+\rho_1^{-1}+\rho_2^{-1}\right) - p_{iN} - \rho_1^{-1}p_{i+2,N}$
For $1 \le j \le N-2$, $p_{1,j} \leftarrow p_{0,j}\left(\rho_1+\mu_2/\mu_1\right) - p_{0,j+1}\mu_2/\mu_1$
For $2 \le i \le N-1$
  For $1 \le j \le N-1$
  $p_{i,j} \leftarrow p_{i-1,j}\left(\rho_1+1+\mu_2/\mu_1\right) - \rho_1 p_{i-2,j-1} - p_{i-1,j+1}\mu_2/\mu_1$
$p_{N-2,0} \leftarrow p_{N-3,0}\left(\rho_1+1\right) - p_{N-3,1}\mu_2/\mu_1$
For $1 \le j \le N-1$, $p_{N,j} \leftarrow p_{N-1,j}\left(\rho_1+1+\mu_2/\mu_1\right) - \rho_1 p_{N-2,j-1} - p_{N-1,j+1}\mu_2/\mu_1$
For $N-1 \le i \le N$, $p_{i,0} \leftarrow p_{i-1,0}\left(\rho_1+1\right) - p_{i-1,1}\mu_2/\mu_1$

$$\sum_{j=0}^{2i+1} p_{jN} r^j d_{j+1,i} = 0\,, \quad i = \overline{0, \frac{N-3}{2}}\,; \tag{9}$$

$$\sum_{j=1}^{N} p_{jN} \lambda r^{j-1} d_{j+1,(N-1)/2} = 0\,; \tag{10}$$

$$\sum_{j=0}^{N} p_{0j} a_{N+1,j+N-2-i} + \sum_{j=0}^{i} p_{jN} \lambda r^{j+2} b_{n-j,2N-2-i-j} - \sum_{j=0}^{N-3-2i} p_{j0}\mu_2 r^{j+1} b_{N+1-i,N-2-i-j} = 0\,, \quad i = \overline{0, \frac{N-5}{2}}\,; \tag{11}$$

$$\sum_{j=0}^{i} p_{0j} a_{N+1,j+2N-i} - \sum_{j=0}^{i} p_{j0}\mu_2 r^{j+1} b_{N+1-j,2N-i-j} = 0\,, \quad i = \overline{1, N-3}\,, \tag{12}$$

where the following notations are used:

$$r = \sqrt{\frac{\mu_1}{\lambda}}\,;$$

$$a_{ij} = \mu_2 \left(\sqrt{\frac{\mu_1}{\lambda}}\right)^2 C_{i,j}(0) - \mu_1 C_{i,j+1}(0)\,;$$

$$b_{ij} = \sqrt{\frac{\mu_1}{\lambda}}\, C_{i,j}(0) - C_{i-1,j}(0)\,;$$

$$d_{ij} = \sqrt{\frac{\mu_1}{\lambda}}\, C_{i,j}(0) - C_{i+1,j+1}(0)\,.$$

System (5)–(12) consists of $3N-4$ equations in $3N-2$ unknowns. Two additional equations follow from the fact that each queue, when considered independently, operates as the standard $M/M/1/N$ queue and, thus,

$$\sum_{j=0}^{N} p_{0j} = \frac{1-\rho_1}{1-\rho_1^{N+1}}\,; \quad p_{\cdot,N} = \sum_{i=0}^{N} p_{iN} = \rho_2^N \frac{1-\rho_2}{1-\rho_2^{N+1}}.$$

Here and henceforth, $\rho_i = \lambda/\mu_i$, $i = 1, 2$. Once the system (5)–(12), supplemented with these two equations, is solved, all other probabilities p_{ij} can be found recursively (see Algorithm 1).

The relations in Algorithm 1 follow from the system of equilibrium equations for p_{ij} (see, for example, [3, p. 435]). Algorithm 1 is not well suited for the computation of the whole joint stationary distribution p_{ij} because the accuracy of the results heavily depends on the values of initial parameters and sometimes may be low.

4 Relating $p_{i-1,N}$ and $p_{i,N}$

System (5)–(12) gives some insights into the interdependence of p_{ij} and p_{nm}. Specifically, Eqs. (9) and (10) show that it is enough to know the relations between $p_{2,N}$ and $p_{3,N}$, $p_{4,N}$ and $p_{5,N}$, $p_{6,N}$ and $p_{7,N}$, etc. to compute the value of p_{NN}. Indeed, let $p_{i,N} = p_{i-1,N}\alpha_i$. From (9) and (10), for $y_{jN} = p_{jN}/p_{0N}$, one has:

$$y_{1N} = -\frac{d_{1,0}}{r d_{2,0}}\,; \tag{13}$$

$$y_{2i,N} = -\frac{\sum_{j=0}^{2i-1} y_{jN} r^j d_{j+1,i}}{r^{2i} d_{2i+1,i} + \alpha_{2i+1} r^{2i+1} d_{2i+2,i}}\,, \quad i = \overline{1, \frac{N-3}{2}}\,; \tag{14}$$

$$y_{N-1,N} = -\frac{\sum_{j=1}^{N-2} y_{jN} r^{j-1} d_{j+1,(N-1)/2}}{r^{N-2} d_{N,(N-1)/2} + \alpha_N r^{N-1} d_{N+1,(N-1)/2}}\,. \tag{15}$$

From this system and the normalization condition $p_{\cdot,N} = \rho_2^N(1-\rho_2)/(1-\rho_2^{N+1})$, one finds $p_{0N} = p_{\cdot,N}/\sum_{i=0}^{N} y_{iN}$ and $p_{NN} = p_{0N} y_{N,N}$. We are unaware of any general rule for choosing α_i. Yet, some

heuristics can be suggested. Without any loss of generality, further on, we consider $\lambda = 1$. From the results in [5], it follows that if $\mu_2 > \mu_1 > 1$, then for large N, one has:

$$p_{iN} \approx p_{i-1,N} \left(\frac{\Phi(x_{i+1}) - \Phi(x_i)}{\Phi(x_i) - \Phi(x_{i-1})} \right)$$

where

$$x_i = \frac{1}{\sqrt{N}} \left(i - \frac{\mu_2 - \mu_1}{\mu_2 - 1} N \right);$$

$$\Phi(x) = \frac{1}{\sqrt{2\pi}\,\sigma} \int_{-\infty}^{x} e^{-t^2/(2\sigma^2)}\, dt;$$

$$\sigma^2 = \frac{\mu_1^2\mu_2 + \mu_1\mu_2^2 + \mu_1 + \mu_1^2 + \mu_2 + \mu_2^2 - 6\mu_1\mu_2}{(\mu_2 - 1)^3}.$$

Thus, the (approximate) value of p_{NN} can be found from (13)–(15) with

$$\alpha_i = \frac{\Phi(x_{i+1}) - \Phi(x_i)}{\Phi(x_i) - \Phi(x_{i-1})}.$$

It is worth noticing that the value $p_{\cdot,N}(1 - \Phi(x_N))$ gives another approximation for p_{NN} if $\mu_2 > \mu_1 > 1$ and can be quite accurate. We can try one's luck and use the same value of α_i in the overloaded case as well (i. e., when the load of at least one queue is 1) with two minor modifications. Firstly, substitute $-\sigma^2$ instead of σ^2 and, secondly, put $\alpha_i \equiv 1$ whenever $-\sigma^2 < 0$ or $\Phi(x_i) \approx \Phi(x_{i-1})$. With these agreements, by using $p_{i,N} = p_{i-1,N}\alpha_i$ in (13)–(15), the value of p_{NN} can be approximated with $2p_{0N}y_{N,N}$. The data in the table give the idea of the quality of the approximation for the case $N = 35$, $\lambda = 1$, and $\mu_1 = 0.01$ ($\rho_1 = 100$).

Exact values of p_{NN} (solution of (5)–(12)) and approximate values of p_{NN} (solution of (13)–(15)). The case of $N = 35$, $\lambda = 1$, and $\mu_1 = 0.01$

μ_2	p_{NN}	
	Exact value	Approximate value
2.5	$1.1310 \cdot 10^{-15}$	$1.1334 \cdot 10^{-15}$
2	$3.6258 \cdot 10^{-12}$	$3.6380 \cdot 10^{-12}$
1.25	$5.1702 \cdot 10^{-5}$	$5.1934 \cdot 10^{-5}$
1.1	0.0027	0,0027
1.01	0.0217	0.0218
0.9	0.1013	0.1016
0.8	0.1989	0.1972

5 Concluding Remarks

In this paper, it has been shown that the joint stationary probability distribution can be computed using the system of equations of the smaller size (than the original one). The idea follows from the fact that in the finite-capacity case, both roots in the denominator of the generating function are the roots of its numerator (on the contrast to the infinite-capacity case). The drawbacks of the utilized method can be seen when computing the whole joint distribution p_{ij}. Here, the widely-adopted Gaussian elimination and matrix-analytic methods are preferable. Yet, when only the blocking probability is of interest, the utilized method leads to the new computational procedure and some insights into the interdependencies between p_{ij} and p_{nm}. Unlike in some other system with two queues of finite-capacity, here the values of p_{ij} do not allow recursive computation, which is, as clearly seen, due simultaneously happening arrivals. Still the utilized method allows further investigations into the new procedures for the approximate computation of p_{ij} as suggested in [13].

Acknowledgments

This work was supported in part by the Russian Foundation for Basic Research (grants 15-07-03007 and 15-07-03406).

References

1. El-hady, E., J. Brzdek, and H. Nassar. 2017. On the structure and solutions of functional equations arising from queueing models. *Aequationes Math*. 91(3): 445–477.
2. Kingman, J. F. C. 1961. Two similar queues in parallel. *Ann. Math. Stat.* 32(4):1314–1323.
3. Hunter, J. J. 1969. Two queues in parallel. *J. Roy. Stat. Soc. B* 31(3):432–445.
4. Flatto, L., and S. Hahn. 1984. Two parallel queues created by arrivals with two demands I. *SIAM J. Appl. Math*. 44(5):1041–1053.
5. Flatto, L. 1985. Two parallel queues created by arrivals with two demands II. *SIAM J. Appl. Math*. 45(5):861–878.
6. Rao, B. M., and M. J. M. Posner. 1985. Algorithmic and approzimation analyses of the split and match queue. *Commun. Stat. Stochastic Models* 1(3):433–456.
7. Latouche, G., and V. Ramaswami. 1999. *Introduction to matrix analytic methods in stochastic modeling*. Philadelphia, PA: SIAM. 334 p.
8. Bocharov, P. P., C. D'Apice, A. V. Pechinkin, and S. Salerno. 2004. *Queueing theory*. Utrecht: VSP Publishing. 450 p.
9. Bavinck, H., G. Hooghiemstra, and E. De Waard. 1993. An application of Gegenbauer polynomials in queueing theory. *J. Comput. Appl. Math*. 49:1–10.
10. Avrachenkov, K. E., N. O. Vilchevsky, and G. L. Shevlyakov. 2003. Priority queueing with finite buffer size and randomized push-out mechanism. *ACM Conference (International) on Measurement and Modeling of Computer Systems Proceedings*. New York, NY: ACM. 31(1):324–335.

11. Razumchik, R. V. 2014. Analysis of finite capacity queue with negative customers and bunker for ousted customers using Chebyshev and Gegenbauer polynomials. *Asia Pac. J. Oper. Res.* 31(4).

12. Zaryadov, I. S., L. A. Meykhanadzhyan, T. A. Milovanova, and R. V. Razumchik. 2015. Metod nakhozhdeniya statsionarnogo raspredeleniya ocheredi v dvukhkanal'noy sisteme s uporyadochennym vkhodom konechnoy emkosti [On the method of calculating the stationary distribution in the finite two-channel system with ordered input]. *Sistemy i Sredstva Informatiki — Systems and Means of Informatics* 25(3):44–59.

13. Razumchik, R. V. 2015. Algebraic method for approximating joint stationary distribution in finite capacity queue with negative customers and two queues. *Informatika i ee Primeneniya — Inform. Appl.* 9(4):68–77.

14. Hunter, J. J. 1971. Further studies on two queues in parallel. *Aust. NZ J. Stat.* 13(2):83–93.

15. Erdelyi, A., and H. Bateman. 1985. *Higher transcendental functions.* Robert E. Krieger Publishing Co. Vol. II. 414 p.

Received July 15, 2017

Contributors

Meykhanadzhyan Lusine A. (b. 1990) — Candidate of Science (PhD) in physics and mathematics, extended education teaching assistant, School No. 281 of Moscow, 7 Raduzhnaya Str., Moscow 129344, Russian Federation; lameykhanadzhyan@gmail.com

Matyushenko Sergey I. (b. 1963) — Candidate of Science (PhD) in physics and mathematics, associate professor, Peoples' Friendship University of Russia (RUDN University), 6 Miklukho-Maklaya Str., Moscow 117198, Russian Federation; matyushenko_si@rudn.university

Pyatkina Daria A. (b. 1968) — Candidate of Science (PhD) in physics and mathematics, associate professor, Peoples' Friendship University of Russia (RUDN University), 6 Miklukho-Maklaya Str., Moscow 117198, Russian Federation; pyatkina_da@rudn.university

Razumchik Rostislav V. (b. 1984) — Candidate of Science (PhD) in physics and mathematics, leading scientist, Institute of Informatics Problems, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, 44-2 Vavilov Str., Moscow 119333, Russian Federation; associate professor, Peoples' Friendship University of Russia (RUDN University), 6 Miklukho-Maklaya Str., Moscow 117198, Russian Federation; rrazumchik@ipiran.ru

СОВМЕСТНОЕ СТАЦИОНАРНОЕ РАСПРЕДЕЛЕНИЕ ЧИСЛА ЗАЯВОК В СИСТЕМЕ С ДВУМЯ ОЧЕРЕДЯМИ КОНЕЧНОЙ ЕМКОСТИ И ОБЩИМ ВХОДЯЩИМ ПОТОКОМ*

Л. А. Мейханаджян[1], С. И. Матюшенко[2], Д. А. Пяткина[2], Р. В. Разумчик[2,3]

[1]Школа № 281 города Москвы
[2]Российский университет дружбы народов
[3]Институт проблем информатики Федерального исследовательского центра «Информатика и управление» Российской академии наук

Аннотация: Рассматривается система массового обслуживания с входящим пуассоновским потоком и двумя приборами, являющаяся одним из простых вариантов fork-систем. Перед каждым прибором имеется накопитель конечной емкости. При поступлении в систему новой заявки создается ее копия и далее в каждую из очередей поступает по одной заявке. Если в момент поступления заявки накопитель оказывается полностью заполненным, заявка теряется и в систему не возвращается. Времена обслуживания заявок на приборах имеют экспоненциальное распределение с различными параметрами. Хорошо известно, что подобные системы с трудом поддаются аналитическому анализу. В работе предлагается метод нахождения вероятности блокировки, а также совместного стационарного распределения числа заявок в накопителях, основанный на методе производящих функций и использующий некоторые результаты теории специальных функций.

Ключевые слова: система массового обслуживания; fork-система; две очереди; конечная емкость; стационарное распределение

DOI: 10.14357/19922264170312

*Работа частично поддержана грантами РФФИ № 15-07-03007 и № 15-07-03406.

Литература

1. *El-hady E., Brzdek J., Nassar H.* On the structure and solutions of functional equations arising from queueing models // Aequationes Math., 2017. Vol. 91. No. 3. P. 445–477.
2. *Kingman J. F. C.* Two similar queues in parallel // Ann. Math. Stat., 1961. Vol. 32. No. 4. P. 1314–1323.
3. *Hunter J. J.* Two queues in parallel // J. Roy. Stat. Soc. B, 1969. Vol. 31. P. 432–445.
4. *Flatto L., Hahn S.* Two parallel queues created by arrivals with two demands I // SIAM J. Appl. Math., 1984. Vol. 44. No. 5. P. 1041–1053.
5. *Flatto L.* Two parallel queues created by arrivals with two demands II // SIAM J. Appl. Math., 1985. Vol. 45. No. 5. P. 861–878.
6. *Rao B. M., Posner M. J. M.* Algorithmic and approzimation analyses of the split and match queue // Commun. Stat. Stochastic Models, 1985. Vol. 1. No. 3. P. 433–456.
7. *Latouche G., Ramaswami V.* Introduction to matrix analytic methods in stochastic modeling. — Philadelphia, PA, USA: SIAM, 1999. 334 p.
8. *Bocharov P. P., D'Apice C., Pechinkin A. V., Salerno S.* Queueing theory. — Utrecht: VSP Publishing, 2004. 450 p.
9. *Bavinck H., Hooghiemstra G., De Waard E.* An application of Gegenbauer polynomials in queueing theory // J. Comput. Appl. Math., 1993. Vol. 49. P. 1–10.
10. *Avrachenkov K. E., Vilchevsky N. O., Shevlyakov G. L.* Priority queueing with finite buffer size and randomized push-out mechanism // ACM Conference (International) on Measurement and Modeling of Computer Systems Proceedings. — New York, NY, USA: ACM, 2003. Vol. 31. No. 1. P. 324–335.
11. *Razumchik R. V.* Analysis of finite capacity queue with negative customers and bunker for ousted customers using Chebyshev and Gegenbauer polynomials // Asia Pac. J. Oper. Res., 2014. Vol. 31. No. 4. Id. 1450029.
12. *Зарядов И. С., Мейханаджян Л. А., Милованова Т. А., Разумчик Р. В.* Метод нахождения стационарного распределения очереди в двухканальной системе с упорядоченным входом конечной емкости // Системы и средства информатики, 2015. Т. 25. Вып. 3. С. 44–59.
13. *Razumchik R. V.* Algebraic method for approximating joint stationary distribution in finite capacity queue with negative customers and two queues // Информатика и её применения, 2015. Т. 9. Вып. 4. С. 68–77.
14. *Hunter J. J.* Further studies on two queues in parallel // Aust. NZ J. Stat., 1971. Vol. 13. No. 2. P. 83–93.
15. *Erdelyi A., Bateman H.* Higher transcendental functions. Robert E. Krieger Publishing Co., 1985. Vol. II. 414 p.

Поступила в редакцию 15.07.2017

INFORMATICS AND APPLICATIONS 2017. Vol. 11. Iss. 3. P. 113–122

ON PARALLELIZATION OF ASYMPTOTICALLY OPTIMAL DUALIZATION ALGORITHMS

E. V. Djukova[1], A. G. Nikiforov[2], and P. A. Prokofyev[3]

Abstract: The main goal of the paper is to develop and implement an approach to building efficient parallel algorithms for intractable enumeration problems and to apply this approach to one of the central enumeration problems, i. e., dualization. Asymptotically optimal algorithms for dualization are considered to be the fastest among the known ones. They have a theoretical justification of the efficiency on average. The size of enumerated set in the dualization problem grows exponentially with the size of the input; thus, parallel computations are reasonable to be utilized. The authors introduce the static parallelizing scheme for asymptotically optimal algorithms of dualization and present the results of the testing. Statistical processing of the experimental results is conducted in order to determine the kind of distribution of the random variables, representing the size of the subtasks for parallel computation. The conditions, under which the schema demonstrates almost maximum speedup and quite uniform processors load, are discovered.

Keywords: discrete enumeration problem; dualization; asymptotically optimal algorithm; irreducible covering of a Boolean matrix; polynomial-time delay algorithm; parallel dualization algorithm

DOI: 10.14357/19922264170313

1 Introduction

The authors consider dualization, which is the problem of searching for irreducible coverings of a Boolean matrix. Let $L = \|a_{ij}\|_{m\times n}$ be a Boolean matrix and H be a set of columns of L. The set H is called a covering of L if each row of L has at least one unit element in the columns H. A covering H is called irreducible if any proper subset of H is not a covering of L. Let $P(L)$ denote the set of all possible irreducible coverings of L. The problem is to construct $P(L)$.

There are other formulations of dualization, specifically, based on concepts of the theory of Boolean functions and graph and hypergraph theory. Let us present these formulations.

1. Given a conjunctive normal form consisting of m different clauses that implements a monotone Boolean function $F(x_1, \ldots, x_n)$, construct a reduced disjunctive normal form of F.
2. Given a hypergraph $\mathcal{H}$ consisting of n vertices and m edges, find all minimal vertex coverings of $\mathcal{H}$.

The efficiency of enumeration algorithms is characterized by the complexity of a single step [1]. An algorithm has a (quasi-)polynomial-time delay if, for any individual problem, each step of this algorithm (the construction of the current solution) is executed in (quasi-)polynomial time in the input size of the problem. As applied to the search for irreducible coverings, this means that for any $m\times n$ Boolean matrix, the time required for the construction of the next irreducible covering is bounded by a (quasi-)polynomial in m and n. In the general case, no dualization algorithm with a (quasi-)polynomial time delay has yet been constructed and it is not known whether such an algorithm exists.

There are examples of such algorithms for some special cases of dualization [1, 2]. For example, in [1], an algorithm with a time delay $O(n^3)$ was constructed in the case when each row of L has at most two unit elements (in this case, $\mathcal{H}$ is a graph in formulation 2).

Studies concerning the complexity of enumeration problems basically address the possibility of constructing incremental (quasi-)polynomial-time algorithms. In this case, the incremental property means that at every step in the construction of the current solution, an algorithm searches through the set of solutions obtained at the preceding steps and the time taken by this search is (quasi-)polynomial in the input problem size and the number of previously found solutions. An incremental quasi-polynomial-time dualization algorithm was constructed in [3, 4]. For several special cases of dual-

[1]Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, 44-2 Vavilov Str., Moscow 119333, Russian Federation; M. V. Lomonosov Moscow State University, M. V. Lomonosov Moscow State University, 1-52 Leninskiye Gory, GSP-1, Moscow 119991, Russian Federation; edjukova@mail.ru

[2]Technische University of Munich, 21 Arcisstrasse, Munich 80333, Germany; ankifor@gmail.com

[3]Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, 44-2 Vavilov Str., Moscow 119333, Russian Federation; p_prok@mail.ru

ization, incremental polynomial-time algorithms were constructed in [5, 6].

Another approach to the solution of the problem is based on the concept of asymptotically optimal algorithm with a polynomial time delay. This approach was first proposed in [7] and deals with a typical case.

According to this approach, the original enumeration problem Z is replaced by a "simpler" problem Z_1 that has the same input and is solved with a polynomial time delay. The solution set of Z_1 contains the solution set of Z and, second, with increasing input size, the number of solutions of Z_1 is almost always asymptotically equal to the number of solutions of Z. This approach is substantiated by obtaining asymptotics for the typical number of solutions to each of the problems Z and Z_1.

Thus, in contrast to an "exact" algorithm with a polynomial time delay, an asymptotically optimal algorithm can execute redundant polynomial-time steps. A redundant step is a solution of Z_1 that was either found previously or is constructed for the first time but is not a solution to the problem Z. For almost all problems of a given size, the number of redundant steps must have a lower order of growth than the number of all steps of the algorithm as the problem input size increases. Whether or not a step is redundant must be verifiable in a polynomial amount of time in the problem input size.

A number of asymptotically optimal algorithms for constructing irreducible coverings of a Boolean matrix have been proposed in the case when the input matrix satisfies the condition [7–15]:

$$\log m \le (1-\epsilon)\log n\,,\ \ 0<\epsilon<1\,.$$

The following criterion called USM is used to construct $P(L)$ in these algorithms. A set H of r columns of the matrix L is an irreducible covering if and only if the following two conditions hold:

(a) the submatrix L^H of L made up of the columns of H does not contain rows of the form $(0,0,\ldots,0)$; and

(b) L^H contains every row of the form $(1,0,0,\ldots,0,0),(0,1,0,\ldots,0,0),\ldots$ $\ldots,(0,0,0,\ldots,0,1)$; i. e., it contains the identity submatrix of order r.

A set of columns satisfying condition (b) is called consistent. A consistent set of columns is called maximal if it is not contained in any other consistent set of columns.

In the asymptotically optimal dualization algorithm AO1 (see [7]), Z_1 is the problem of constructing a collection of column sets a matrix L satisfying condition (b) in which each set of length r occurs as many times as there are identity submatrices of order r in this set. In fact, all identity submatrices of L are enumerated with a polynomial time delay. Clearly, an irreducible covering can be generated only by a maximal identity submatrix, i. e., by an identity submatrix that is not contained in any other one. A maximal identity submatrix generates a maximal consistent set of columns, i. e., a consistent set of columns that is not contained in any other one.

According to the algorithm AO1, the maximal identity submatrices (the maximal consistent sets of columns) can be enumerated (enumerated with repetition) with a step complexity of $O(qmn)$, where $q=\min\{m,n\}$. As a result of enumerating the identity submatrices, some sets of columns are repeatedly constructed. When obtaining the current maximal identity submatrix Q in time $O(mn)$, the algorithm AO1 checks condition (a) for the set H of columns of L generated by the submatrix Q. If condition (a) holds, then AO1 checks in time $O(mn)$ whether H was constructed at a previous step.

The algorithm AO2 [12], which is a modification of AO1, enumerates (with a polynomial time delay $O(qm^2n)$) only identity submatrices of L that generate coverings. At every step, AO2 constructs an irreducible covering. However, as in AO1, the solutions can repeat. This algorithm takes less redundant steps than AO1. Based on AO2, the algorithms AO2K and AO2M with a reduced execution time were constructed in [15].

The asymptotically optimal algorithm OPT enumerates without repetitions and with a polynomial time delay $O(qm^2n)$ the sets of columns of L satisfying condition (b) and some additional conditions, including the maximality one [13]. Redundant steps in OPT arise due to the construction of maximal consistent sets of columns that are not coverings (do not satisfy condition (a)).

The dualization algorithms RS and MMCS were proposed in [16, 17]. Their description makes use of concepts of hypergraph theory. These algorithms are based on constructing sets of vertices of a hypergraph $\mathcal{H}$ satisfying the "crit" condition, which is equivalent to compatibility condition (b) for the corresponding set of columns of the incidence matrix of $\mathcal{H}$. Thus, the approach proposed in [16, 17] for the construction of dualization algorithms is not new (in fact, RS and MMCS are asymptotically optimal algorithms).

The algorithm RUNC-M [18] is one of the fastest among asymptotically optimal algorithms. As a rule, RUNC-M is less time-consuming than the asymptotically optimal algorithms constructed in [7–13, 15–17]. In this paper, a new implementation of RUNC-M is developed. This implementation works on a number of test tasks significantly faster than the implementation described in [18].

Due to the complexity of dualization, the use of parallel computations is essential. In the development of parallel dualization algorithms, the focus is on deriving theoretical worst case complexity. However, such estimates can be obtained only for some special cases of dualization [19].

In this paper, a new practical parallization scheme for asymptotically optimal dualization algorithms is constructed. The proposed scheme is of static nature and is based on statistical estimations of subtasks size. There exist simple and obvious practical parallelization schemes of asymptotically optimal dualization algorithms. Their main disadvantage is an unbalanced load of processors which produces insufficient speedup.

Let us describe the computational subtasks in question. Let H be an irreducible covering of the Boolean matrix L consisting of columns with indicies $j_1, \ldots, j_r$ where $j_1 < \cdots < j_r$. Then H is called irreducible j_1-covering. The jth computational subtask is to construct all irreducible j-coverings of L. Therefore, let us define the jth subtask size $\nu_j(L)$ as the ratio of the number of irreducible j-coverings to the number of all irreducible coverings. For optimal load balancing, one should know the values of $\nu_j(L)$; however, they become known only after the dualization is completed.

The proposed parallelization scheme is based on processing random r-by-n submatrices of the input matrix where r is a parameter that does not exceed m. The processor load is scheduled only after the calculation of the subtask sizes for a given number of random submatrices.

The validity of estimating $\nu_j(L)$ based on random submatrices is justified statistically. First, let us introduce a special random variable η_r defined on the set of r-by-n submatrices and their irreducible coverings. Its value is defined as the least index of columns in the covering. Next, let us test the statistcal hypothesis that the distribution of η_r is determined by the subtask sizes of the dualization of the matrix L. It is found that according to the Chi-squared test, this hypothesis can be accepted with confidence when $r \geq m/2$.

The scheme is highly scalable (a balanced load and almost maximal speedup). In this paper, the proposed scheme is applied to the algorithm RUNC-M. However, it is also applicable for all dualization algorithms that sequentially construct sets of irreducible 1-coverings, 2-coverings, and so on.

The paper is organized as follows. In section 2, a formal definition of the asymptotically optimal dualization algorithm is given and its basic structure via decision trees is described. In section 3, the algorithm RUNC-M is described, some details about its new implementation are provided, and it is experimentally compared with the previous RUNC-M version from [18]. In section 4, the present approach to parallelizing asymptotically optimal algorithms is described. This approach is applied to the algorithm RUNC-M and tested in section 5. Section 6 contains concluding remarks.

2 Terms and Definitions

Let M_{mn} be a set of $m \times n$ Boolean matrices and $P_{mn}(X) = |X|/|M_{mn}|$ for $X \subseteq M_{mn}$. It is said that $f(L) \approx g(L)$, $m, n \to \infty$, for almost all $L \in M_{mn}$ if

$$\forall \delta > 0,\ \exists \lim_{m,n\to\infty} P_{mn}\left(\left\{L\colon \left|1 - f(L)g(L)^{-1}\right| < \delta\right\}\right) = 1\,.$$

Let us consider the following class of algorithms for enumerating the irreducible coverings of a Boolean matrix $L \in M_{mn}$. Each algorithm A in this class constructs a finite sequence $Q_A(L)$ of column sets of L that contains all elements from $P(L)$. It is assumed that some elements of $Q_A(L)$ can be repeated. At each step, the algorithm A constructs an element of $Q_A(L)$ and checks whether it belongs to $P(L)$. If the constructed element is in $P(L)$, then A additionally verifies in a polynomial time whether it was earlier constructed. Let $N_A(L)$ be a number of steps of the algorithm A (length of $Q_A(L)$).

The algorithm A is asymptotically optimal with a polynomial time delay d if

- d is bounded above by a polynomial in m and n;
- each step in A consists of at most d elementary operations (one matrix element access); and
- $N_A(L) \approx |P(L)|$, $m, n \to \infty$, for almost all $L \in M_{mn}$.

Let $S(L)$ be the set of all identity submatrices of the matrix L. The number of maximal consistent sets of columns is bounded above by $|S(L)|$. The theoretical substantiation of asymptotically optimal dualization algorithms is based on the following statement. If $m \leq n^{1-\varepsilon}$, where $\varepsilon > 0$, then $|S(L)| \approx |P(L)|$, $m, n \to \infty$, for almost all $L \in M_{mn}$ (see [7]).

The column j is said to cover the row i of a matrix L if $a_{ij} = 1$. Let H be a set of columns of L. The set H is said to cover the row i if there exists $j \in H$ covering i. Let the set of columns H be consistent. The column j of L is said to be compatible with the set H if set $H \cup \{j\}$ is consistent; otherwise, this column is called incompatible with the set H.

The work of an asymptotically optimal dualization algorithm can be regarded as an unidirectional traversal of the branches of a decision tree. Each tree vertex is associated with the tuple (H, R, C), where H is the set of columns of L, and R and C are, respectively, the sets of rows and columns describing the submatrix of L, and C and H are disjoint. The vertex $(\emptyset, R_0, C_0)$, where R_0 and C_0 describe the whole matrix L, is the tree root. The leaf vertices are either irreducible coverings or correspond to redundant steps

of the algorithm. Every step of the algorithm represents a transition from one terminal vertex (or root) to another one. A transition from one internal vertex to the next one is performed by adding a column of L to the set H. It is assumed that the number of elementary operations at each step is polynomially bounded in m and n.

Asymptotically optimal algorithms can be classified into two types. Among the first type are the algorithms enumerating the maximal identity submatrices of L. Such algorithms execute redundant steps in which solutions constructed in the preceding steps are constructed once more. The examples of such algorithms are AO1 [7] and AO2 [12]. The algorithms of the second type are based on the enumeration of maximal consistent sets of columns. This class includes the algorithms OPT [13], MMCS, RS [16, 17], PUNC, and RUNC-M [18].

3 Algorithm RUNC-M

The algorithm RUNC-M is described as a recursive procedure RUNCM. The first call RUNCM(L, H_0, R_0, C_0) should be done with the parameters $H_0 = \emptyset$, $R_0 = \{1, \ldots, m\}$, $C_0 = \{1, \ldots, n\}$. Notice that the parameters are passed by value.

PROCEDURE RUNCM(L, H_0, R_0, C_0)
1: $C_0^{\min} = \{j \in C_0 | a_{ij} = 1\}$ where $i \in R_0$ is the index of the row with the least sum $\sum_{j \in C_0} a_{ij}$;
2: **for all** $j \in C_0^{\min}$ **do**
3: $R \leftarrow R_0$
4: $C_0 \leftarrow C_0 \setminus \{j\}$
5: $C \leftarrow C_0$
6: $H \leftarrow H_0 \cup \{j\}$
7: Eliminate from R the rows that are covered by column j
8: **if** $R = \emptyset$ **then**
9: Save the set of columns $H \in P(L)$
10: **else**
11: Eliminate from C the columns that are incompatible with H
12: **call** RUNCM(L, H, R, C)

The following criterion is used for incompatible columns elimination. A row i of the matrix L is called supporting for (H, j), $j \in H$, if $a_{ij} = 1$ and $a_{il} = 0$, $l \in H \setminus \{j\}$. The set of supporting rows for (H, j), $j \in H$, is denoted by $S(H, j)$. A column u is compatible with H if and only if there is no column $j \in H$ such that column u covers all rows from $S(H, j)$.

The present authors developed a new implemetation of the algorithm RUNC-M, which is available at https://github.com/ankifor/dualization-OPT.git.

4 Parallelizing Asymptotically Optimal Algorithms

In this section, a practical parallelization S-scheme that computes the relative subtask sizes by estimating the values $\nu_j(L) = |P_j(L)|/|P(L)|$, $j \in \{1, \ldots, n\}$, is described. The proposed scheme is designed for processing the Boolean matrices in which the number m of rows is significantly greater than the number n of columns.

Let $L \in M_{mn}$ and $r \leq m$. The set of all r-subsets of $\{1, \ldots, m\}$ is denoted by W_m^r. Let $w \in W_m^r$; then L^w denotes the submatrix of L consisting of the rows of L with indicies from w. A function η_r acting from $\Omega_r = \{(L^w, H) : w \in W_m^r, H \in P(L^w)\}$ to $\{1, \ldots, n\}$ is defined such that $\eta_r(L^w, H)$ equals j if $H \in P_j(L^w)$.

Let us choose t random submatrices $L^{w_1}, \ldots, L^{w_t}$, $w_s \in W_m^r$, $s \in \{1, \ldots, t\}$, and build $P(L^{w_s})$ for each of them. Then, let us take u random irreducible coverings $H_1^s, \ldots, H_u^s$ from these sets. Next, let us compose a sample $\vec{x} = (x_1, \ldots, x_N)$, $N = tu$, of values of $\eta_r(L^{w_s}, H_v^s)$, $s \in \{1, \ldots, t\}$, $v \in \{1, \ldots, u\}$, and calculate the frequency $f_r^*(j)$ of occurrence of j in $\vec{x}$. The quantity $f_r^*(j)$ is used as an estimation of $\nu_j(L)$, $j \in \{1, \ldots, n\}$.

A statistical justification of this approach is given below. Also, the values of r, under which the resulting estimates are sufficiently accurate, will be found. On the one hand, the integer r should be as small as possible to reduce the computation time of $f_r^*(j)$. On the other hand, these estimations should be sufficiently reliable.

Let Ω_r be a sample space. The probability of event (L^w, H) is set to $\left(\binom{m}{r} |P(L^w)|\right)^{-1}$. Then, let us denote the probability of event $\eta_r(L^w, H) = j$ by $f_r(j)$.

To test the statistical hypothesis $H_0 : f_r(j) = \nu_j(L)$ about the distribution of the random variable η_r, the Chi-squared test with the statistic

$$Z_r(\vec{x}) = N \sum_{j=1}^{n} \frac{(f_r^*(j) - \nu_j(L))^2}{\nu_j(L)}$$

was used. The p-value is denoted by $\gamma_r^*(\vec{x}) = 1 - \chi_{n-1}^2(Z_r(\vec{x}))$ where χ_{n-1}^2 is the cumulative chi-squared distribution function with $(n-1)$ degrees of freedom. Small values of $\gamma_r^*(\vec{x})$ argue for rejecting H_0.

Now, conduct an experiment. Generate 20 random m-by-n matrices. Then, dualize these matrices and calculate the exact values of $\nu_j(L)$, $j \in \{1, \ldots, n\}$. Let $t = 20$ and $u = 50$. For each matrix L and for each r, $r \in \{10, 13, 15, 18, 20, 25, 30, 35\}$ such that $r < m$, construct a sample $\vec{x}$ from the values η_r and calculate the statistic $Z_r(\vec{x})$ and p-value $\gamma_r^*(\vec{x})$.

The median values of $Z_r(\vec{x})$ and $\gamma_r^*(\vec{x})$ are presented in Table 1 for each configuration 30×150, 40×120,

Table 1 Median values of $(Z_r(\vec{x}), \gamma_r^*(\vec{x}))$ for chi-squared test

r	$m \times n$			
	30×120	40×120	50×100	70×70
10	$(159, < 10^{-4})$	$(167, < 10^{-4})$	$(235, < 10^{-4})$	$(382, < 10^{-4})$
13	$(99, < 10^{-4})$	$(132, < 10^{-4})$	$(157, < 10^{-4})$	$(234, < 10^{-4})$
15	$(77, 0.0134)$	$(112, < 10^{-4})$	$(117, < 10^{-4})$	$(187, < 10^{-4})$
18	$(74, 0.028)$	$(90, 0.0002)$	$(96, < 10^{-4})$	$(147, < 10^{-4})$
20	$(60, 0.0815)$	$(63, 0.0546)$	$(89, < 10^{-4})$	$(131, < 10^{-4})$
25	$(54, 0.315)$	$(60, 0.0876)$	$(50, 0.1382)$	$(85, < 10^{-4})$
30	—	—	—	$(68, 0.0001)$
35	—	—	—	$(54, 0.0478)$

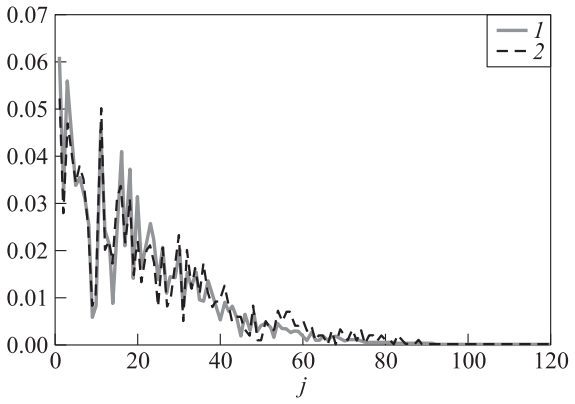


Figure 1 Plots of $\nu_j(L)$ (*1*) and $f_r^*(r)$ (*2*) as functions of j where $m = 30$, $n = 120$, and $r = 15$

50×100, and 70×70 and different values of r. According to Table 1, $\gamma_r^*(\vec{x})$ becomes significant at $r \geq m/2$. That is why, a further increase of r will not make the approximation of $\nu_j(L)$ much more accurate.

Consider the configuration 30×150 as an example. At $r = 15$, the so-called "phase transition" is observed — the function $Z_r(\vec{x})$ stabilizes after this point. The plots of $\nu_j(L)$ and $f_r^*(j)$ for this case are presented in Fig. 1.

To sum up, the experiment shows that $\nu_j(L)$ can be used as an estimate of $f_r(j)$ at $r \geq m/2$. Moreover, the frequency $f_r^*(j)$ is well-known to be a "good" estimate for the probability $f_r(j)$. Thus, $f_r^*(j)$ can be used as an approximation of $\nu_j(L)$ under the conditions stated above.

After computing the estimates $f_r^*(j)$, let us proceed to scheduling the processor load. Let one has $p \leq n$ processors and let the jth subtask be executed by the processor with the index N_j. The vector $\vec{N^p} = (N_1, \ldots, N_n)$ is called schedule. The load balance of the kth processor is defined as

$$\sigma_k(\vec{N^p}) = \sum_{j \in J_n : N_j = k} \nu_j(L)\,.$$

To obtain an efficient schedule, the following optimization problem should be solved:

$$\sigma(\vec{N^p}) = \max_{k \in J_p} \sigma_k(\vec{N^p}) \to \min_{\vec{N^p}}\,. \qquad (1)$$

The following procedure `DistributeTasks` is proposed for finding an approximate solution to problem (1), which is based on a greedy strategy. The parameters of the procedure are the number p of processors, the number n of columns of L, and the vector $\vec{f_r^*} = (f_r^*(1), \ldots, f_r^*(n))$ of estimators for $\nu_j(L)$.

PROCEDURE `DistributeTasks`$(p, n, \vec{f_r^*}) \to (\vec{N^p}, \sigma)$
1: **for all** $k \in \{1, \ldots, p\}$ **do**
2: $\quad \sigma_k \leftarrow 0$
3: **for** $j \in \{1, \ldots, n\}$ **do**
4: $\quad k_0 \leftarrow \operatorname{argmin}_{1 \leq k \leq p} \sigma_k$
5: $\quad N_j \leftarrow k_0$
6: $\quad \sigma_k \leftarrow \sigma_k + f_r^*(j)$

5 Parallel RUNC-M Test Results

Testing was performed on the supercomputer IBM Blue Gene/P of the Lomonosov Moscow State University.

Each computation node contains four PowerPC 450 processor cores running at 850 MHz, 2 GB DRAM, and communication interfaces. The computations were performed in the virtual-node mode (four MPI processes per node, 1 GB limit per process; cannot create additional threads).

Let p be the number of processors and $T_k(p)$ be the algorithm execution time (in seconds) on the kth processor. Let $T(p) = \max_k T_k(p)$ and $T_\Sigma(p) = \sum_k T_k(p)$. The number $s_k(p) = (T_k(p))/(T_\Sigma(p))$ is called the realized load level of the kth processor. The following measures are of interest:

- algorithm speedup $S(p) = T(1)/T(p)$; and
- load balance uniformity $E(p) = S(p)/p$.

Table 2 Parallel RUNC-M execution time in seconds depending on the number of processors

Data	p							
	1	2	4	8	16	32	64	128
$U(30,100)$	3.95	2.03	1.05	0.59	0.37	0.32	0.32	0.32
$U(30,150)$	39.1	20.0	10.4	5.21	3.46	2.32	2.33	2.32
$U(30,200)$	231	116	61.5	32.2	18.8	13.8	13.8	13.8
$U(40,100)$	11.5	5.83	3.05	1.53	0.96	0.95	0.95	0.95
$U(40,150)$	133	67.1	34.8	19.1	10.9	9.44	9.43	9.43
$U(40,200)$	654	328	177	90.5	61.8	40.4	36.8	36.8

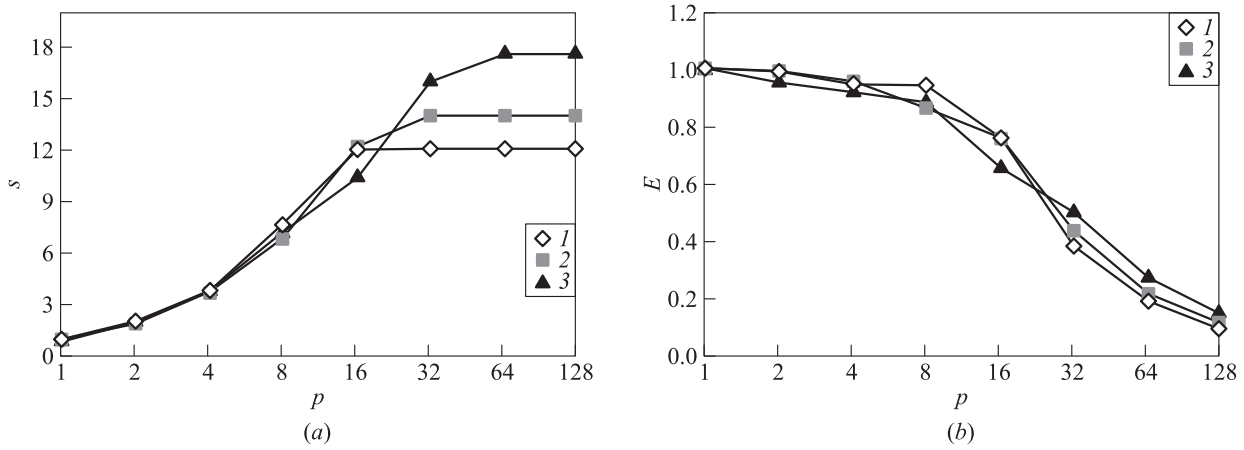


Figure 2 Plots of S (*a*) and E (*b*) as functions of p for $m = 40$ and $n \in \{100, 150, 200\}$: *1* — $m \times n = 40 \times 100$; *2* — 40×150; and *3* — $m \times n = 40 \times 200$

The speedup $S(p) = p$ at $p \geq 1$ is almost maximal. If $E(p)$ is close to unity, then the load balance is considered to be uniform. The measure $s(p)$ is an analog of $\sigma(\vec{N}^p)$ (see formula (1)).

Experiments were carried out on random m-by-n matrices where $m \in \{30, 40\}$ and $n \in \{100, 150, 200\}$. The parameter r of the statistical procedure discribed above is equal to $m/2$. In order to estimate the values of $\nu_j(L)$, the dualization problem is solved for r-by-n submatrices L_w of the matrix L.

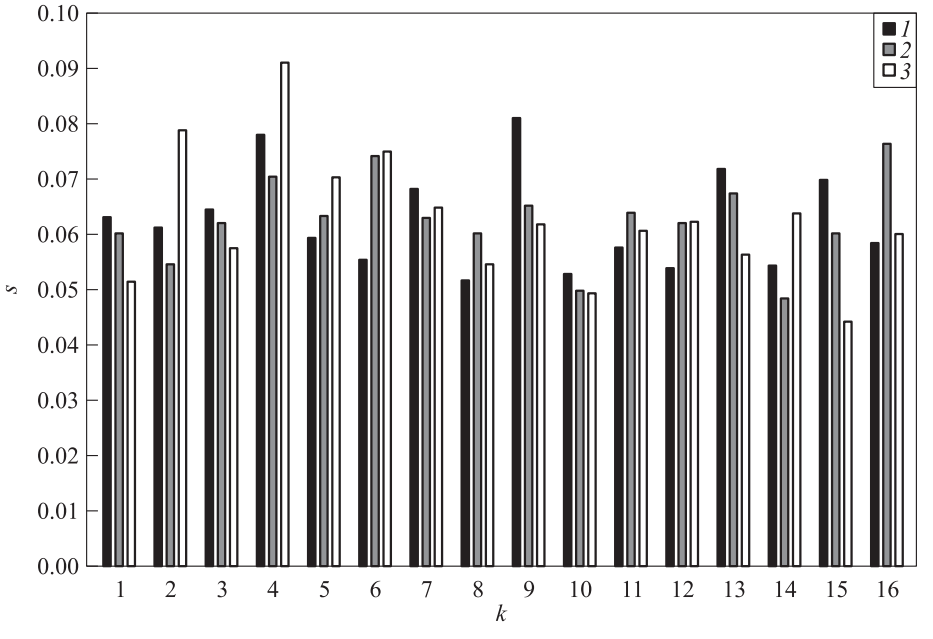
The computation results are presented in Table 2; they include the execution time of the parallel algorithm on different number of processors.

The plots of $S(p)$ and $E(p)$ are given in Fig. 2 for the case when $m = 40$ and $p \in \{1, 2, 4, \ldots, 128\}$. As can be seen, the parallel RUNC-M version has almost linear speedup $S(p)$ and loads processors in a balanced way when the number of processors p is below some threshold p^*. Generally speaking, this threshold depends on the dualized matrix size. For example, p^* equals 32 when $m = 40$ and $n = 200$ and it equals 16 when $m = 40$ and $n = 100$. When the number of processors is greater than the threshold, the execution time $T(p)$ stops to improve. This is because parallelization takes place on the first level of the decision tree built by the algorithm RUNC-M. Under the proposed approach, the size of the computational subtasks are significantly different. Therefore, it is here impossible to distribute tasks in a balanced manner over a large number of processors.

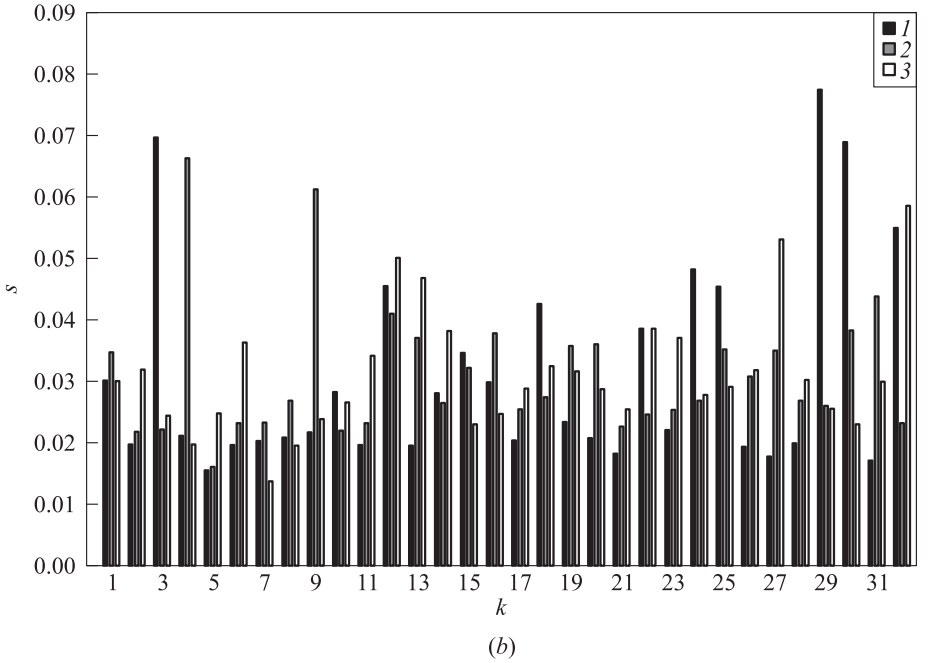
The realized load levels $s_k(p)$ for each processor are presented in Fig. 3. As seen, some realized load levels differ by several fold when $m = 40$, $n = 200$, and $p = 32$. That might be due to insufficient quality of the subtask size estimates $f_r^*(j)$ or nonoptimality of the task distribution schedule. Nevertheless, the variance of the realized load levels is fairly small for $p = 16$ which is in agreement with the high speedup.

6 Concluding Remarks

In this paper, an approach [20] to the parallel algorithm construction for discrete enumeration problems is developed. This approach is based on statistical estimates of computational subtask sizes. Subtasks are assigned to processors in accordance with precalculated schedule. To construct this schedule, the distribution of a special random variable used for estimating the subtask sizes is found. Then, the load balance of processors is optimized. A novel efficient parallelization scheme for



(*a*)



(*b*)

Figure 3 Realized load levels $s_k(16)$ (*a*) and $s_k(32)$ (*b*) as functions of k for $m = 40$ and $n \in \{100, 150, 200\}$: *1* — $m \times n = 40 \times 100$; *2* — 40×150; and *3* — $m \times n = 40 \times 200$

asymptotically optimal dualization algorithms based on the proposed approach is developed. The scheme is applied to the algorithm RUNC-M [18] which is the fastest known dualization algorithm.

The proposed approach to the the construction of parallel dualization algorithms ensures high accuracy of subtask size estimates which under certain conditions leads to highly efficient parallel algorithms.

However, the proposed approach is not that efficient when the number of processors is large because the sizes of computational subtasks can vary significantly (the parallelization is performed at the first level of the deci-

sion tree built by the asymptotically optimal dualization algorithm).

Acknowledgments

The research was supported by the Russian Foundation for Basic Research (projects Nos. 16-01-00445-a and 15-51-05059).

References

1. Johnson, D., and C. Papadimitriou. 1988. On generating all maximal independent sets. *Inform. Process. Lett.* 27(3):119–123.
2. *Eiter, T., G. Gottlob, and K. Makino*. 2003. New results on monotone dualization and generating hypergraph transversals. *SIAM J. Comput.* 32(2):514–537.
3. Fredman, M., and L. Khachiyan. 1996. On the complexity of dualization of monotone disjunctive normal forms. *J. Algorithm.* 21(3):618–628.
4. Khachiyan, L., E. Boros, K. Elbassioni, and V. Gurvich. 2006. An efficient implementation of a quasi-polynomial algorithm for generating hypergraph transversals and its application in joint generation. *Discrete Appl. Math.* 154(16):2350–2372.
5. Boros, E., V. Gurvich, K. Elbassioni, and L. Khachiyan. 2000. An efficient incremental algorithm for generating all maximal independent sets in hypergraphs of bounded dimension. *Parallel Processing Lett.* 10(04):253–266.
6. Boros, E., K. Elbassioni, V. Gurvich, and L. Khachiyan. 2004. Generating maximal independent sets for hypergraphs with bounded edge-intersections. *Latin American Symposium on Theoretical Informatics*. 488–498.
7. Djukova, E. 1977. On an asymptotically optimal algorithm for constructing irredundant tests. *Sov. Math. Dokl.* 18(2):423–426.
8. Djukova, E. 1987. The complexity of the realization of certain recognition procedures. *Comp. Math. Math. Phys.* 27(1):74–83.
9. Djukova, E., and Y. Zhuravlev. 1997. Discrete methods of information analysis in recognition and algorithm synthesis. *Pattern Recognition Image Anal.* 7(2):192–207.
10. Djukova, E., and Y. Zhuravlev. 2000. Discrete analysis of feature descriptions in recognition problems of high dimensionality. *Comp. Math. Math. Phys.* 40(8):1214–1227.
11. Djukova, E. 2003. Discrete recognition procedures: The complexity of realization. *Pattern Recognition Image Anal.* 13(1):8–10.
12. Djukova, E. 2004. On the implementation complexity of the realization of discrete (logical) recognition procedures. *Comp. Math. Math. Phys.* 44(3):532–541.
13. Djukova, E., and A. Inyakin. 2008. Asimptoticheski optimal'noe postroenie tupikovykh pokrytiy tselochislennoy matritsy [Asymptotically optimal irredundent tests enumeration for integer matrix]. *Matematicheskie Voprosy Kibernetiki* [Mathematical Problems of Cybernetics] 17:235–246.
14. Kudryavtsev, V., and A. Andreev. 2010. Test recognition. *J. Math. Sci.* 169(4):457–480.
15. Djukova, E., and P. Prokofyev. 2014. Ob asimptoticheski optimal'nom perechislenii neprivodimykh pokrytiy bulevoy matritsy [On asymptotically optimal enumeration for irreducible coverings of boolean matrix]. *Prikladnaya Diskretnaya Matematika* [Discrete Applied Mathematics] 1:96–105.
16. Murakami, K., and T. Uno. 2011. Efficient algorithms for dualizing large-scale hypergraphs. *CoRR*. abs/1102.3813.
17. Murakami, K., and T. Uno, 2014. Efficient algorithms for dualizing large-scale hypergraphs. *Discrete Appl. Math.* 170:83–94.
18. Djukova, E., and P. Prokofyev. 2015. Asymptotically optimal dualization algorithms. *Comp. Math. Math. Phys.* 55(5):891–905.
19. Khachiyan, L., E. Boros, V. Gurvich, and K. Elbassioni. 2007. Computing many maximal independent sets for hypergraphs in parallel. *Parallel Processing Lett.* 17(2):141–152.
20. Djukova, E., A. Nikiforov , and P. Prokofyev. 2014. Statisticheski effektivnaya skhema rasparallelivaniya algoritmov dualizatsii [Statistically efficient parallel scheme for dualization algorithms]. *Mashinnoe obuchenie i analiz dannykh* [J. Machine Learning Data Anal.] 1(7):843–853.

Received December 7, 2016

Contributors

Djukova Elena V. (b. 1945) — Doctor of Science in physics and mathematics, principal scientist, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, 42 Vavilov Str., Moscow 119333, Russian Federation; associate professor, Faculty of Computational Mathematics and Cybernetics, M. V. Lomonosov Moscow State University, 1-52 Leninskiye Gory, GSP-1, Moscow 119991, Russian Federation; edjukova@mail.ru

Nikiforov Andrey G. (b. 1993) — master student, Faculty of Informatics, Technische University of Munich (TUM), 21 Arcisstrasse, Munich 80333, Germany; ankifor@gmail.com

Prokofyev Petr A. (b. 1982) — Candidate of Science (PhD) in physics and mathematics, scientist, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, 42 Vavilov Str., Moscow 119333, Russian Federation; pprok@mail.ru

О РАСПАРАЛЛЕЛИВАНИИ АСИМПТОТИЧЕСКИ ОПТИМАЛЬНЫХ АЛГОРИТМОВ ДУАЛИЗАЦИИ*

Е. В. Дюкова[1,2], А. Г. Никифоров[3], П. А. Прокофьев[1]

[1]Федеральный исследовательский центр «Информатика и управление» Российской академии наук
[2]Московский государственный университет им. М. В. Ломоносова
[3]Технический университет Мюнхена, г. Мюнхен, Германия

Аннотация: Основной целью работы является разработка и реализация подхода к построению эффективных параллельных алгоритмов для труднорешаемых перечислительных задач и демонстрация этого подхода на примере одной из центральных перечислительных задач — задачи дуализации. Из известных алгоритмов дуализации наиболее быстрыми являются асимптотически оптимальные алгоритмы. Эти алгоритмы имеют теоретическое обоснование эффективности «в среднем». Поскольку число решений дуализации растет экспоненциально с ростом размера входа, актуальным является использование параллельных вычислений. В статье предложена статическая схема распараллеливания асимптотически оптимальных алгоритмов дуализации и осуществлено ее тестирование. Проведена статистическая обработка экспериментов с целью определения вида распределения случайной величины, определяющей объемы подзадач. Выявлены условия, при которых схема демонстрирует ускорение, близкое к максимальному, и достаточно равномерную загрузку процессоров.

Ключевые слова: дискретная перечислительная задача; дуализация; асимптотически оптимальный алгоритм; неприводимое покрытие булевой матрицы; алгоритмы с полиномиальной задержкой; параллельный алгоритм дуализации

DOI: 10.14357/19922264170313

Литература

1. *Johnson D., Papadimitriou C.* On generating all maximal independent sets // Inform. Process. Lett., 1988. Vol. 27. No. 3. P. 119–123.
2. *Eiter T., Gottlob G., Makino K.* New results on monotone dualization and generating hypergraph transversals // SIAM J. Comput., 2003. Vol. 32. No. 2. P. 514–537.
3. *Fredman M., Khachiyan L.* On the complexity of dualization of monotone disjunctive normal forms // J. Algorithm., 1996. Vol. 21. No. 3. P. 618–628.
4. *Khachiyan L., Boros E., Elbassioni K., Gurvich V.* An efficient implementation of a quasi-polynomial algorithm for generating hypergraph transversals and its application in joint generation // Discrete Appl. Math., 2006. Vol. 154. No. 16. P. 2350–2372.
5. *Boros E., Gurvich V., Elbassioni K., Khachiyan L.* An efficient incremental algorithm for generating all maximal independent sets in hypergraphs of bounded dimension // Parallel Processing Lett., 2000. Vol. 10. No. 4. P. 253–266.
6. *Boros E., Elbassioni K., Gurvich V., Khachiyan L.* Generating maximal independent sets for hypergraphs with bounded edge-intersections // Latin American Symposium on Theoretical Informatics, 2004. P. 488–498.
7. *Дюкова Е. В.* Об асимптотически оптимальном алгоритме построения тупиковых тестов // Докл. АН СССР, 1977. Т. 233. № 4. С. 527–530.
8. *Дюкова Е. В.* О сложности реализации некорых процедур распознавания // Ж. вычисл. матем. матем. физ., 1987. Т. 27. № 1. Р. 114–127.
9. *Djukova E., Zhuravlev Y.* Discrete methods of information analysis in recognition and algorithm synthesis // Pattern Recognition Image Anal., 1977. Vol. 7. No. 2. P. 192–207.
10. *Дюкова Е., Журавлёв Ю.* Дискретный анализ признаковых описаний в задачах распознавания большой размерности // Ж. вычисл. матем. матем. физ., 2000. Т. 40. № 8. С. 1264–1278.
11. *Djukova E.* Discrete recognition procedures: The complexity of realization // Pattern Recognition Image Anal., 2003. Vol. 13. No. 1. P. 8–10.
12. *Дюкова Е.* О сложности реализации дискретных (логических) процедур распознавания // Ж. вычисл. матем. матем. физ., 2004. Т. 44. № 3. С. 562–572.
13. *Дюкова Е. В., Инякин А. С.* Асимптотически оптимальное построение тупиковых покрытий целочисленной матрицы // Математические вопросы кибернетики, 2008. № 17. С. 235–246.
14. *Kudryavtsev V., Andreev A.* Test recognition // J. Math. Sci., 2010. Vol. 169. No. 4. P. 457–480.
15. *Дюкова Е. В., Прокофьев П. А.* Об асимптотически оптимальном перечислении неприводимых покрытий булевой матрицы // Прикладная дискретная математика, 2014. № 1. С. 96–105.

*Работа поддержана грантами РФФИ № 16-01-00445-а и № 15-51-05059.

16. *Murakami K., Uno T.* Efficient algorithms for dualizing large-scale hypergraphs // CoRR, 2011. abs/1102.3813.
17. *Murakami K., Uno T.* Efficient algorithms for dualizing large-scale hypergraphs // Discrete Appl. Math., 2014. Vol. 170. P. 83–94.
18. *Дюкова Е., Прокофьев П.* Об асимптотически оптимальных алгоритмах дуализации // Ж. вычисл. матем. матем. физ., 2015. Т. 55. № 5. С. 895–910.
19. *Khachiyan L., Boros E., Gurvich V., Elbassioni K.* Computing many maximal independent sets for hypergraphs in parallel // Parallel Processing Lett., 2007. Vol. 17. No. 2. P. 141–152.
20. *Дюкова Е. В., Никифоров А. Г., Прокофьев П. А.* Статистически эффективная схема распараллеливания алгоритмов дуализации // Машинное обучение и анализ данных, 2014. Т. 1. № 7. С. 843–853.

Поступила в редакцию 07.12.2016

INFORMATICS AND APPLICATIONS 2017. Vol. 11. Iss. 3. P. 123–131

STATISTICAL DATA AS INFORMATION SOURCE FOR LINGUISTIC ANALYSIS OF RUSSIAN CONNECTORS

O. Inkova[1] and N. Popkova[2]

Abstract: The aim of this paper is to describe statistical data gathered from the supracorpora database (SCDB) of connectors for further analysis of their formal and functional properties. Until now, these properties have usually been described applying semantic analysis, while corpus data, if used at all, have not been subject to statistical processing. It is automatically generated and verifiable information, collected from texts corpora that can be one of the most reliable tools in the analysis of linguistic units, including connectors. The paper shows what statistics one may obtain from the SCDB and how to use it in the linguistic analysis in case of *tol'ko*, a polyfunctional linguistic unit that can be a part of multicomponent and two-place connectors.

Keywords: annotation of connectors; corpus linguistics; supracorpora databases; parallel texts; statistical data

DOI: 10.14357/19922264170314

1 Introduction

The paper aims to show what statistics the SCDB can provide in order to analyze formal and functional properties of connectors[3]. Until now, semantic analysis has prevailed in the field, and corpus data, if used at all, have not been subject to statistical processing [1–3]. The opportunity to apply corpus and statistical methods of analysis has radically changed the way linguistic research is perceived, as such methods allow one to enhance the reliability and validity of the results achieved. Since quantitative corpus data are generated automatically and, hence, can be easily verified, they may serve as one of the most reliable tools in the analysis of linguistic units.

Therefore, electronic corpora evolve to have statistical tools. The most substantial corpus project in Russia, the "Russian National Corpus" (RNC, www.ruscorpora.ru), offers a wide range of data analysis tools that get updated on a regular basis, with some of them having only recently become available. Being designed both for regular users and researchers, the RNC is a representative corpus containing texts of different types and genres that date back from the 18th to the 21st century. The RNC uses a metamarkup language that affords a possibility of "marking up a text with metatags so as to specify how it was created, its author, topic, genre particulars, etc." [4]. These tags determine which statistical tools could be available for the RNC users. Studying a particular linguistic unit, users not only can get the information about the number of tokens and texts in which the word has been used, but also they can find out what is the topic and the type of every single text, the date when the text was created and its author's name. The data, both in numbers and percentage, are displayed in tables. When the RNC user studies if and how a word or a word combination occurs in texts within a certain timeframe, (s)he may get an automatically generated frequency graph created with the "Graphs" function. Together with graphs, the tables are generated where some numbers are hyperlinks. Frequency data are also available for downloading in ZIP archive format. Nonetheless, such statistical data provide information about texts where the linguistic unit in question occurs, and functional properties of the linguistic unit itself are only represented indirectly. It goes without saying that information about how frequently the unit occurs in texts of a specific genre, or texts written by an author of a specific sex does not suffice to describe a connector, especially its formal properties.

Another corpus project, referred to as "Russkaya korpusnaya grammatika" [Russian corpus grammar], aims "to give a synchronic description of a representative fragment of Russian grammar, which would be grounded in the data collected from the RNC and use quantitative methods of corpus analysis" [5]. In "Conjunction" (http://rusgram.ru/%D0%A1%D0%BE%D1%8E%D0%B7#4), that is thought to be the most relevant chapter for the present study [6], one can find data, both in numbers and percentage, on how frequently coordinate and subordinate conjunctions occur. The percentage indicates the portion of coordinate and subordinate

[1]Institute of Informatics Problems, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, 44-2 Vavilov Str., Moscow 119333, Russian Federation; olyainkova@yandex.ru

[2]Institute of Informatics Problems, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences; 44-2 Vavilov Str., Moscow 119333, Russian Federation; natasha_popkova@mail.ru

[3]Connectors are the functional words of different grammatical classes (coordinate and subordinate conjunctions, some adverbs, prepositions, particles, and "discourse markers" — linguistic units with complex grammatical structure) that serve to connect different parts of the text.

conjunctions found in the RNC and the share of every conjunctive class in the overall number of conjunctions. Data on individual conjunctions within semantic classes are also available; though for some of them, e. g., substitutive conjunctions, no statistical data can be provided.

This paper describes what statistical functions the SCDB has got and what data they help to collect, which is exemplified by 834 annotations of *tol'ko*[1]. This particular word *tol'ko* has been chosen for its polyfunctionality, i. e., it can act as an adverb, a particle, or a conjunction:

– Вы, видно, **только** проснулись [You seem to have **just** woken up] (adverb);

– Каких **только** подарков он ей ни делал [**Just** think how many presents he gave her]! (particle);

– **Только** он вошел, все встали [**Just** after he entered, everybody stood up] (conjunction).

What is more, *tol'ko* can also act as a part of multicomponent and two-place connectors, for example, **Как** только кончается отношение служебное, **так** кончается всякое другое [**As soon as** service relationship is over, any other relationship is over too] (L. Tolstoy), where *tol'ko* acts as a part of the two-place connector *kak tol'ko... tak* [as soon as]. It is worth noting that existing electronic linguistic resources do not allow to run a query for two-place linguistic units as one unit. Besides, such multicomponent units are characterized by a high degree of formal variability. For instance, *kak tol'ko... tak* may also occur as *kak tol'ko... totchas zhe*: Он, **как только** проснулся, **тотчас же** вознамерился встать, умыться [**As soon as** he woke up, he wanted to get up and wash himself] (I. Goncharov). All the corpora we know are not designed to gather data on how multicomponent linguistic units function, in particular, those whose elements, as in the example above, are located separately.

2 Statistical Analysis and Electronic Linguistic Resources

The SCDB, a new electronic linguistic resource, is a superstructure built upon the RNC or, more precisely, upon its parallel Russian-French subcorpus. The database allows to retrieve correspondences between linguistic units from both languages and to annotate them, specifying their characteristics that are relevant for linguistic analysis. For more details on the SCDB of connectors, see [7–10].

In the SCDB, the initial information object to start the annotation process with is a *monoequivalence* (ME), defined as a two-place tuple. Its structure is as follows: A source text fragment containing the linguistic unit under study (Column 1, Fig. 1) and a corresponding translation fragment (Column 2), containing its functionally equivalent fragment (FEF, the term coined in [11]).

The object annotated in the SCDB of connectors is referred to as *discursive realization* (DR) or, in other words, the actual form of the linguistic unit in which it occurred in the text. Along with the source DR, its FEF also gets annotated, which makes the SCDB different from other electronic linguistic resources.

Figure 1 shows the annotated DR *kak tol'ko* and its FEF *dès que*.

context of the DR	DR and its characteristics	context of the FEF	FEF and its characteristics	characteristics of the ME
Как только дело дошло до "честного слова", я махаю руками и сажусь за стол. [**As soon as** it comes to "my word of honour", I wave my hands and sit down at the table.]	**как только** [**as soon as**] <temporary> <with predication> <initial> <CNT q p> <CNT> <Contact>	**Dès qu'on** en arrive à la parole d'honneur, je lève les bras au ciel et m'assieds à mon bureau.	**dès que** <temporary> <with predication> <initial> <CNT q p> <CNT> <Contact>	☐ NB ☑ Cngrn ☐ Dvrg ☑ Type 1 ☐ Type 2 ☐ Type 3

Figure 1 An example of annotation

[1]As of May 1, 2017, the SCDB contains parallel texts (mostly fiction) in Russian and French of 3.5 M tokens, 10,562 Russian-French annotations, and 909 French-Russian annotations.

Table 1 The *tol'ko* cluster: DRs and the frequency of their occurrence

	DRs with *tol'ko*	Frequency of occurrence of the DR
1	Только [just; once]	134
2	Не только (расстояние) но и [not only (distance) but also]	113
3	Лишь только [hardly; scarcely; only when]	47
4	Как только [as soon as; once]	43
5	Не только (расстояние) но даже [not only (distance) but even]	31
6	Не только (расстояние) но [not only (distance) but]	21
7	Если только [if only]	17
8	Вот только [though; only; just]	16
9	Не только [not only]	16
10	Вот только (расстояние) ∅* [only/just (distance) ∅]	16
11	Не только (расстояние) и [not only (distance) and]	10
12	Не только (расстояние) даже [not only (distance) even]	9
13	Не (расстояние) только [only (distance) not; not only]	8
14	Только что [just]	8
15	Не только (расстояние) но даже и [not only (distance) but also]	8
16	Не только (расстояние) но (расстояние) и [not merely (distance) but (distance) even]	8
17	Не только (расстояние) но и вообще [not only (distance) but generally]	6
18	Не только (расстояние) а [not only [distance] but]	6
19	Как только (расстояние) тотчас же [just (distance) at once]	5
20	А не только [and not only; and not merely]	5

*The ∅ symbol means that the second component of a two-place DR is not explicit.

Professional linguists use the SCDB to annotate connectors in three steps. First, they identify a connector's DR, its FEF, and then build an ME. At the same stage, the DR is structurally analyzed and attributed to clusters of its components. For instance, the DR *kak tol'ko* is assigned both to the *tol'ko* cluster and to the *kak* cluster (how clusters are populated is not shown in Fig. 1; however, the SCDB has got this option). Accordingly, these two clusters are supposed to include all the DRs that are present in the SCDB and contain *tol'ko* (Table 1) and *kak*. This annotation procedure allows (a) to retrieve from the SCDB all the combinations of elements of DRs, (b) to generate a list of linguistic units in which a certain element can act as a component, and (c) to find a prototypical form for a connector with a high degree of formal variability based on how frequent its DRs are. For example, in the SCDB, there is a vast variety of DRs with *tol'ko* that have temporal meaning, such as: *как только* [as soon as; once]; *как только... тут же* [as soon as]; *как только... тут и* [as soon as]; *как только... как раз и* [just... at once]; *как только... так* [as soon as; at once; as soon... so]; *едва только... как* [as soon as... then; hardly... before]; *лишь только* [hardly; scarcely; only when]; *лишь только... как* [no sooner... when; as soon as]; *лишь только... лишь только* [as soon as... as soon as][1]. For all these occurrences, the form *kak tol'ko* has been established as prototypical, since more than 60% of DRs are those with *kak tol'ko*. Coupled with semantic analysis, this approach would help to tackle a complicated theoretical question of whether two or more DRs with common elements are formal variations of one connector and if so, which is the prototypical form of this connector or connectors. The question has already been raised in [12–14].

During the second stage of the annotation process, linguists characterize the functioning of the DR in a given context. These characteristics are also grouped into six clusters: **Relations**, **Structure**, **Position**, **Order**, **Status**, and **Disposition**.

The following characteristics have been assigned to the DR with *tol'ko*:

– ⟨**temporary**⟩ means that the connector expresses temporary logical-semantic relation between situations *q* (*дело дошло до «честного слова»* [*it comes to "my word of honour"*]) and *p* (*я махаю руками и сажусь за стол* [*I wave my hands and sit down at the table*]);

– ⟨**with predication**⟩ means that the text fragment (*дело дошло до «честного слова»* [*it comes to "my word of honour"*]), marked by the connector *kak tol'ko*, has a predicative structure, i. e., subject (*дело* [*it*]) and predicate (*дошло* [*comes*]);

[1]The data from the RNC (namely, the Russian-English subcorpus) show that the two-place DRs with *tol'ko* are often not translated in their entirety, only the first part of those DRs tends to get translated.

– ⟨**initial**⟩ means that the connector *kak tol'ko* occupies the initial position in the text fragment *q* (*дело дошло до «честного слова»* [*it comes to "my word of honour"*]) it marks;

– ⟨**CNT q p**⟩ indicates the order in which the text fragments marked by the connector come;

– ⟨**CNT**⟩ means that the DR *kak tol'ko* fulfils a connective function in the text fragment, i. e., it acts as a connector. In case of polyfunctional units like *tol'ko*, the Status cluster and its components allow to register different functions of such units; and

– ⟨**Contact**⟩ indicates that both components of the connector *kak tol'ko* [as soon as] go one after another and are not separated by other words.

The FEF is annotated according to the same scheme. Column 4 in Fig. 1 contains six characteristics of the FEF *dès que* that coincide with those chosen for *kak tol'ko*, although this is not always the case (see Fig. 3 below where characteristics do differ).

At the third and final stage of the annotation process, characteristics of the ME itself are marked in Column 5. The *NB* mark signals that MEs are of special interest for experts. *Cngrn* and *Dvrg* marks are used to describe translation type. When a connector is translated by a connector, congruent (*Cngrn*) translation is marked, while divergent (*Dvrg*) translation is marked when a connector is translated by another linguistic unit or syntactic construction. Marks *Type1*, *Type2*, and *Type3* are used to describe annotation type (see more in [8, 9, 15]). Eight hundred and thirty four Russian-French annotations of *tol'ko* mentioned above were marked as *Type1* since they have been made for the connector as a whole and not for its separate components. Overall, 1,219 Russian-French annotations of *tol'ko* of all three types have been registered. Moreover, 59 French-Russian annotations of *tol'ko* have been made in the SCDB, with 50 of them being of Type1. Therefore, as of May 1, 2017, 1278 annotations for DRs with *tol'ko* have been created in the SCDB.

3 Supracorpora Database Statistical Functions

The SCDB has been designed to have three functions in order to generate statistical data. First, it gives information about how frequent are the DRs, annotations of which are stored in the database. Second, it provides statistical data on every single occurrence of the DR as the queries of it are executed in the SCDB. Third, it allows one to see which patterns are used to translate linguistic units, annotated in the SCDB, and how frequently they occur.

The first function extracts the data displayed in Table 1 that shows which DRs make up the *tol'ko* cluster, it also shows the frequency of their occurrence. Table 1 demonstrates first 20 most frequent DRs, each of which has been annotated in the SCDB five times or more. The most frequent among them is DR *tol'ko* with 134 MEs registered. Overall, the *tol'ko* cluster includes 155 DRs.

The information about the frequency of occurrence of DRs is automatically generated for every cluster of components. During the annotation process, the distribution within clusters is recalculated.

The second function returns individual statistical data, received as a result of queries that are executed in the SCDB. A query template (Fig. 2) is used to set query parameters. Running the query, the user gets not only the number of annotations that comply with the parameters set, but also the list of these annotations that can be used subsequently for further analysis.

In Fig. 2, the field **Кластер РР в оригинале** [Cluster of DRs in the source text] contains *tol'ko*, i. e., only annotations of *tol'ko* and those of the other 166 Russian DRs that contain it (such as *kak tol'ko*, *lish' tol'ko*, etc.) will be selected. Figure 2 shows that in the query, some filters are specified so as to restrict the set of texts and translations, whose fragments appear in the annotations. Moreover, there are three other filters:

(1) DRs with *tol'ko* must only express temporary relation;

(2) DRs with *tol'ko* mark a part of a sentence with a predication, i. e., a text fragment marked by a DR must have a predicative structure; and

(3) annotations must be of Type 1.

There are 34 annotations satisfying such criteria. When the query is completed, annotations are stored in a list. One of the annotations is depicted in Fig. 3.

The third statistical function returns data on patterns used to translate the linguistic units under study in the text fragments that get annotated in the SCDB. With this function, one can evaluate the frequency of different translation patterns.

For example, let us consider French translations of the DR *tol'ko*. Overall, 35 translation patterns have been identified, including the zero pattern, i. e., the absence of translation equivalent. These patterns appear in 137 annotations. Data on seven most frequent translation patterns are displayed in Table 2. Data on the frequency of occurrence of the remaining 28 patterns are reported in the penultimate row of Table 2.

What is of utmost importance is that numbers in Column 3 are hyperlinks leading to the lists of corresponding annotations. Hyperlinks help to visualize

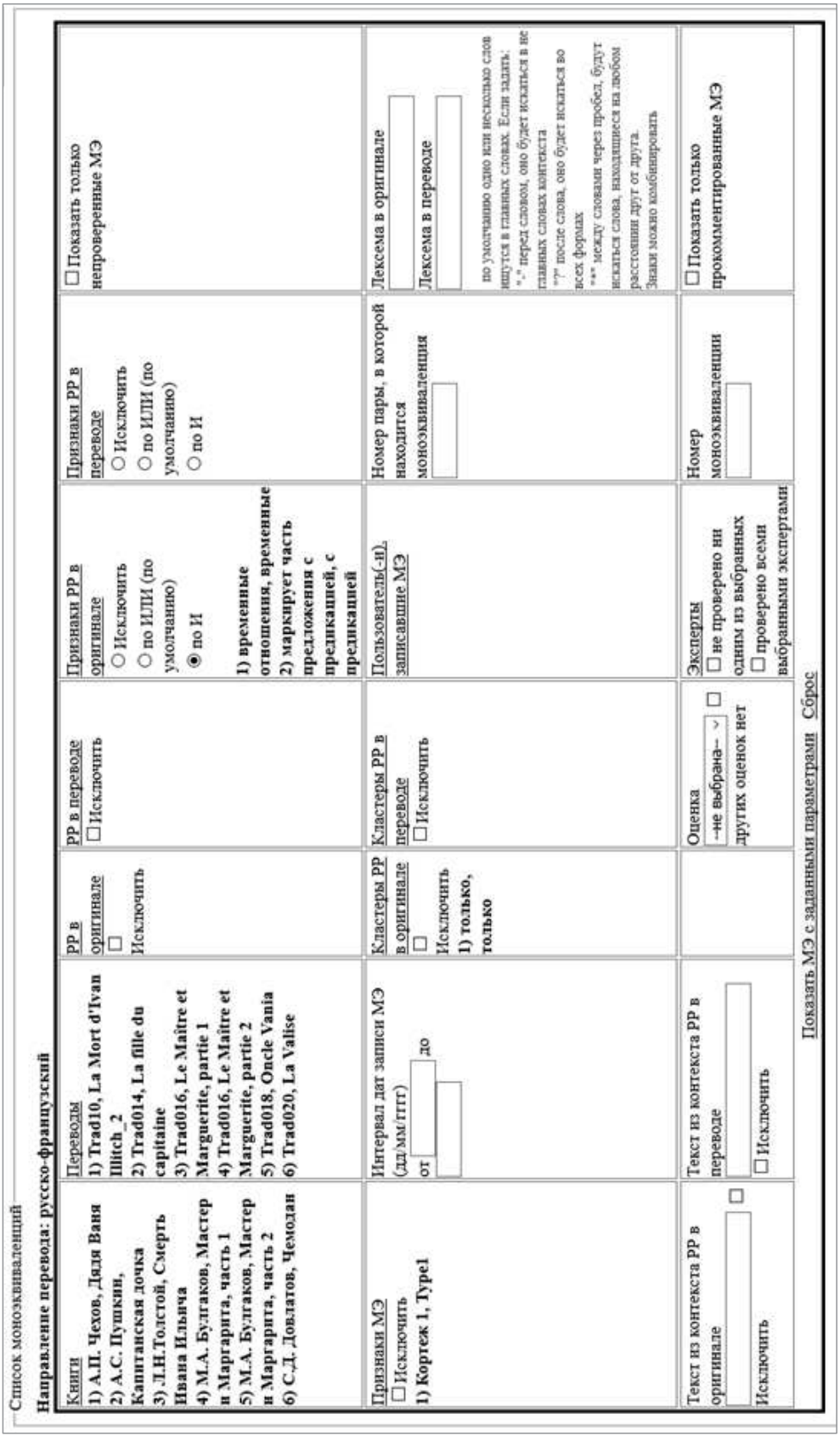


Figure 2 A query template with parameters

№ ME	context of the DR	DR and its characteristics	context of the FEF	FEF and its characteristics	characteristics of the ME
1362	Тогда, **лишь только** процессия вышла на самый верх за цепь, он и появился впервые и притом как человек явно опоздавший. [It was **only when** the procession came to the very top, beyond the file, **that** he had first appeared, and as an obvious latecomer at that. (Tr. L. Volokhonsky, R. Pevear, 1997)]	**лишь только** [only when] <temporary> <with predication> <initial> <CNT q p> <CNT> <Contact>	**Au moment précis,** *en effet*, **où** le cortège franchissait le deuxième cordon de légionnaires et atteignait le sommet, il fut le premier à sortir de la foule et à se précipiter en avant, comme s'il redoutait d'arriver trop tard.	**au moment précis|où** <temporary> <initial> <CNT q p> <CNT> <Contact> <composite sentence> <SubCNT>	☐ Exp ☑ NB ☑ Cngrn ☐ Dvrg ☑ Type 1 ☐ Type 2 ☐ Type 3

Figure 3 An annotation found in the query

Table 2 *Tol'ko*: Frequency distribution of translation patterns

Russian linguistic unit	French translation pattern	Frequency of occurrence of the translation pattern in the SCDB	Frequency of occurrence (percentage)
tol'ko	ne . . . que	30	21.90
	zero	16	11.68
	seulement	15	10.95
	seul	14	10.22
	mais	12	8.76
	sauf que	6	4.38
	juste	6	4.38
	Remaining 28 translation patterns	38	27.73
Number of annotations		137	100

statistics, make it verifiable, and using them gives the possibility to analyze the annotations after they have been selected. This is a big advantage of the SCDB of connectors over other electronic linguistic resources. Our statistical data show that most often *tol'ko* is translated by the restrictive negation *ne . . . que*, i. e., it is its most frequent FEF and appears in 21.9% of annotations.

4 Concluding Remarks

The SCDB of connectors opens up vast opportunities to get statistical data on functional and formal properties of connectors depending on the research objectives and purposes. The main advantage of such data is that they are representative and verifiable because of the hyperlinks going to the annotations one can find after the query is completed.

Furthermore, unlike most data from other electronic linguistic resources, statistical data in the SCDB describe first and foremost the linguistic units themselves and not the texts where they occur. This results from a fine elaboration of annotation parameters: in the SCDB, every parameter characterizes connectors either structurally or functionally. For instance, statistics for the Relations parameter, denoting the relation expressed by the connector, are not only important for the analysis, but also in case of polysemic connectors (vast majority

of connectors) help to define which relation they tend to convey. Statistics for the Structure parameter allow to define the common syntactical structure of the connector. In future, these data will help to answer the question about how frequently the connector occurs at the interclausal level (i. e., in a complex sentence) and at the intersentential level (i. e., between separate utterances). So far, the question has only been raised [16], although it seems to be theoretically important for the description of linguistic means with a connective function, including connectors.

In this regard, the Order parameter also gives important information about how text fragments, linked by a connector, are positioned. And such parameters as Position and Status are used, for example, to identify polyfunctional linguistic units, for which a connective function is only one of the possible functions. The question about how to define functions of a polyfunctional linguistic unit in an utterance remains open. In its turn, the Disposition parameter indicates if there could be any distance between elements of two-place and multicomponent connectors, which is a specific functional property of these connectors. Such kind of statistics allows, for instance, to answer the so far unanswered question: "Is a DR an independent linguistic unit or is it a more or less free combination of components?"

Thus, the linguistic analysis of connectors strongly relies upon statistical data — a pivotal tool in helping researchers to get new insights and to make results more reliable. Statistical data processing opens up new research perspectives and suggests solutions of many yet unsolved questions.

Acknowledgments

The study has been conducted at the Institute of Informatics Problems, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences with financial aid from the Russian Foundation for Basic Research (grant No. 16-06-00070).

References

1. In'kova-Manzotti, O. Yu. 2001. *Konnektory protivopostavleniya vo frantsuzskom i russkom yazykakh. Sopostavitel'noe issledovanie* [Connectors of opposition in French and Russian. A comparative study]. Moscow: Informelektro. 434 p.
2. Chzhon, Kh. Kh. 2003. *Prisoedinitel'nye skrepy v sovremennom russkom yazyke: sintaksis i semantika* [Conjunctive ties in modern Russian: Syntax and semantics]. Moscow: Lomonosov Moscow State University. PhD Thesis. 190 p.
3. Zav'yalov, V. N. 2009. Morfologicheskie i sintaksicheskie aspekty opisaniya struktury soyuzov v sovremennom russkom yazyke [Morphological and syntactical aspects of the conjunctions' structure description in modern Russian]. Vladivostok: DGU. D.Sc. Thesis. 393 p.
4. Natsional'nyy korpus russkogo yazyka [Russian National Corpus]. Available at: http://www.ruscorpora.ru (accessed April 23, 2017).
5. Russkaya korpusnaya grammatika [Russian Corpus Grammar]. Available at: http://rusgram.ru/ (accessed April 28, 2017).
6. Apresyan, V. Yu., and O. E. Pekelis. 2011. *Soyuz* [Conjunction]. Available at: http://rusgram.ru/ (accessed April 28, 2017).
7. Zaliznyak Anna A., I. M. Zatsman, O. Yu. In'kova, and M. G. Kruzhkov. 2015. Nadkorpusnye bazy dannykh kak lingvisticheskiy resurs [Subcorpora databases as linguistic resource]. *Corpus Linguistics: 7th Conference (International) Proceedings*. St. Petersburg: St. Petersburg State University. 211–218.
8. Zatsman, I. M., O. Yu. In'kova, M. G. Kruzhkov, and N. A. Popkova. 2016. Predstavlenie krossyazykovykh znaniy o konnektorakh v nadkorpusnykh bazakh dannykh [Representation of cross-lingual knowledge about connectors in supracorpora databases] *Informatika i ee Primeneniya — Inform. Appl.* 10(1):106–118.
9. In'kova, O. Yu., and M. G. Kruzhkov. 2016. Nadkorpusnye russko-frantsuzskie basy dannykh glagol'nykh form i konnektorov [Supracorpora databases of Russian and French verbal forms and connectors]. *Lingue slave a confronto* [Slavic languages in comparison]. Eds. O. In'kova and A. Trovesi. Bergamo: Bergamo University Press. 365–392.
10. Zaliznyak, Anna A., I. M. Zatsman, and O. Yu. In'kova. 2017. Nadkorpusnaya baza dannykh konnektorov: postroenie sistemy terminov [Supracorpora database of connectors: Developing a terminology]. *Informatika i ee Primeneniya — Inform. Appl.* 11(1):100–108.
11. Dobrovol'skiy, D. O., A. A. Kretov, and S. A. Sharov. 2005. Korpus parallel'nykh tekstov: Arkhitektura i vozmozhnosti ispol'zovaniya [Corpus of parallel texts: Architecture and applications]. *Natsional'nyy korpus russkogo yazyka: 2003–2005* [Russian National Corpus: 2003–2005]. Moscow: Indrik. 263–296.
12. In'kova, O. Yu., and N. A. Popkova. 2016. Struktura dvukhmestnykh konnektorov russkogo yazyka v svete korpusnykh dannykh [The structure of two-part correlative connectors as an object of corpus analysis]. *Computational Linguistics and Intellectual Technologies: Conference (International) "Dialogue" Proceedings*. Moscow: RGGU. 15(22):200–213.
13. In'kova, O. Yu. 2016. K probleme opisaniya mnogokomponentnykh konnektorov russkogo yazyka: ne tol'ko... no i [Towards the problem of the description of multiword connectives of Russian language: Ne tol'ko... no i (not only... but also)]. *Voprosy yazykoznaniya* [Topics in the Study of Language] 2:37–60.
14. Kobozeva, I. M. 2016. Kognitivno-semanticheskiy podkhod k opisaniyu sredstv svyazi predlozheniy (na primere konnektorov so znacheniem neposredstvennogo sledovaniya) [Cognitive-semantical approach to the description

of ways to connect sentences (case of connectors of immediate consecution)]. *Tr. Instituta russkogo yazyka im. V. V. Vinogradova* [V. V. Vinogradov Russian Language Institute of the Russian Academy of Sciences Collections] 11:118–131.

15. Popkova, N. A., O. Yu. In'kova, I. M. Zatsman, and M. G. Kruzhkov. 2015. Metodika postroeniya monoekvivalentsiy v nadkorpusnoy baze dannykh konnektorov [Methodology of constructing monoequivalences in the supracorpora database of connectors]. *Tr. 2-y nauchn. konf. "Zadachi sovremennoy informatiki"* [2nd Scientific Conference "Modern Informatics' Problems" Proceedings]. Moscow: FRC CSC RAS. 143–153.
16. Uryson, E. V. 2012. Soyuzy, konnektory i teoriya valentnostey [Conjunctions, connectors, and the valence theory]. *Computational Linguistics and Intellectual Technologies: Conference (International) "Dialogue" Proceedings*. Moscow: RGGU. 11(1):627–637.

Received July 11, 2017

Contributors

Inkova Olga Yu. (b. 1965) — Doctor of Science in philology, senior scientist, Institute of Informatics Problems, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences; 44-2 Vavilov Str., Moscow 119333, Russian Federation; olyainkova@yandex.ru

Popkova Natalia A. (b. 1992) — junior scientist, Institute of Informatics Problems, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences; 44-2 Vavilov Str., Moscow 119333, Russian Federation; natasha__popkova@mail.ru

СТАТИСТИЧЕСКИЕ ДАННЫЕ КАК ИНФОРМАЦИОННАЯ ОСНОВА ЛИНГВИСТИЧЕСКОГО АНАЛИЗА КОННЕКТОРОВ РУССКОГО ЯЗЫКА*

О. Ю. Инькова, Н. А. Попкова

Институт проблем информатики Федерального исследовательского центра «Информатика и управление» Российской академии наук

Аннотация: Целью статьи является описание статистических данных, которые получены с помощью надкорпусной базы данных (НБД) коннекторов для лингвистического анализа их формальных и функциональных свойств. До настоящего времени эти свойства описывались, как правило, с применением семантического анализа; корпусные данные, если и привлекались, то без их статистической обработки. Именно программно генерируемая и верифицируемая статистическая информация, полученная на основе корпусов текстов, может служить одним из надежных параметров лингвистического анализа языковых единиц, в том числе коннекторов. Описаны виды получаемой с помощью НБД статистики и возможности ее использования для собственно лингвистического анализа на примере языковой единицы *только*, отличающейся своей полифункциональностью, а также способностью входить в качестве составляющей в значительное число многокомпонентных и двухместных коннекторов.

Ключевые слова: аннотирование коннекторов; корпусная лингвистика; надкорпусные базы данных; параллельные тексты; статистические данные

DOI: 10.14357/19922264170314

Литература

1. *Инькова-Манзотти О. Ю.* Коннекторы противопоставления во французском и русском языках. Сопоставительное исследование. — М.: Информэлектро, 2001. 434 с.
2. *Чжон Х. Х.* Присоединительные скрепы в современном русском языке: синтаксис и семантика: Дис. . . . канд. филол. наук. — М.: МГУ им. М. В. Ломоносова, 2003. 190 с.
3. *Завьялов В. Н.* Морфологические и синтаксические аспекты описания структуры союзов в современном русском языке: Дис. . . . д-ра филол. наук. — Владивосток: ДГУ, 2009. 393 с.
4. Национальный корпус русского языка. http://www.ruscorpora.ru.

*Исследование выполнено при финансовой поддержке РФФИ (проект 16-06-00070).

5. Русская корпусная грамматика. http://rusgram.ru.
6. *Апресян В. Ю., Пекелис О. Е.* Союз. — М., 2011. http://rusgram.ru.
7. *Зализняк Анна А., Зацман И. М., Инькова О. Ю., Кружков М. Г.* Надкорпусные базы данных как лингвистический ресурс // Корпусная лингвистика-2015: Тр. междунар. конф. — СПб.: СПбГУ, 2015. С. 211–218.
8. *Зацман И. М., Инькова О. Ю., Кружков М. Г., Попкова Н. А.* Представление кроссязыковых знаний о коннекторах в надкорпусных базах данных // Информатика и её применения, 2016. Т. 10. Вып. 1. С. 106–118.
9. *Инькова О. Ю., Кружков М. Г.* Надкорпусные русско-французские базы данных глагольных форм и коннекторов // Lingue slave in confronto / Eds. O. Inkova, A. Trovesi. — Bergamo: Bergamo University Press, 2016. С. 365–392.
10. *Зализняк Анна А., Зацман И. М., Инькова О. Ю.* Надкорпусная база данных коннекторов: построение системы терминов // Информатика и её применения, 2017. Т. 11. Вып. 1. С. 100–108.
11. *Добровольский Д. О., Кретов А. А., Шаров С. А.* Корпус параллельных текстов: архитектура и возможности использования // Национальный корпус русского языка: 2003–2005. — М.: Индрик, 2005. С. 263–296.
12. *Инькова О. Ю., Попкова Н. А.* Структура двухместных коннекторов русского языка в свете корпусных данных // Компьютерная лингвистика и интеллектуальные технологии: По мат-лам ежегодной Междунар. конф. «Диалог». — М.: РГГУ, 2016. Вып. 15(22). С. 200–213.
13. *Инькова О. Ю.* К проблеме описания многокомпонентных коннекторов русского языка: не только... но и // Вопросы языкознания, 2016. № 2. С. 37–60.
14. *Кобозева И. М.* Когнитивно-семантический подход к описанию средств связи предложений (на примере коннекторов со значением непосредственного следования) // Тр. Института русского языка им. В. В. Виноградова, 2016. Т. 11. С. 118–131.
15. *Попкова Н. А., Инькова О. Ю., Зацман И. М., Кружков М. Г.* Методика построения моноэквиваленций в надкорпусной базе данных коннекторов // Задачи современной информатики: Тр. 2-й научной конф. — М.: ФИЦ ИУ РАН, 2015. С. 143–153.
16. *Урысон Е. В.* Союзы, коннекторы и теория валентностей // Компьютерная лингвистика и интеллектуальные технологии: По мат-лам ежегодной Междунар. конф. «Диалог». — М.: РГГУ, 2012. Вып. 11(18). Т. 1. С. 627–637.

Поступила в редакцию 11.07.2017

INFORMATICS AND APPLICATIONS 2017. Vol. 11. Iss. 3. P. 132–141

INDICATOR EVALUATION OF PROCESSES OF KNOWLEDGE TRANSFER FROM SCIENCE TO TECHNOLOGY

I. M. Zatsman[1], G. V. Lukyanov[2], V. A. Minin[3], V. A. Havanskov[4], and S. K. Shubnikov[5]

Abstract: The article is dedicated to indicator evaluation of information interactions between science and technology. Some of these indicators are defined as single numeric values and some as matrices of numeric values that characterize the intensity of the knowledge flow from different research areas into specific technological branches. The article provides a description of primary information resources, mainly full-text descriptions of patents, which are used to define numerical values of these indicators. It also gives a description of secondary information resources generated as the result of patent documentation processing, including information on references to scientific publications cited in patents. Primary and secondary resources were used to create and test the information model and the corresponding indicators of assessment of interaction between science and technology. This model was applied as a foundation for calculation of numerical values of integral and thematic indicators of the intensity of scientific knowledge flow into the branch of information technologies.

Keywords: information interaction between science and technology; citation of scientific works; intensity of the knowledge flow; indicator assessment; information technology

DOI: 10.14357/19922264170315

Introduction

Science has always been a major driver for technological progress and sustainable socioeconomic development. As a matter of fact, just scientific results mainly determine the nature, character, pace, and scale of technological progress shaping the landscape of a modern society. Thus, it is of utmost importance for strategic planning to reveal functional dependencies between specific scientific (from one side) and specific technological (from the other side) areas, branches, and disciplines. It is essential to underline that the objective of such an "investigation" does not lie in the generally acceptable and absolutely clear but totally imprecise understanding of the social role of science which actually provides next to nothing help as far as management of state scientific and technological resources is concerned. Any rational research of this kind is aimed at finding such models of interactions between science and technology which could support strategic planning with measurable facts and figures more or less strictly reflecting these interactions.

Currently, there is only one intrinsic information resource which comprises implicit data on how science is translated into technology; and this resource is patent documentation. Historically, it turned out that patents contain perfectly documented initial information which makes it possible to extract explicit data on the intangible interplay between science and technology with a subsequent opportunity of finding quantitative assessments for scientific knowledge flow into the branch of technologies. This initial information in patent documentation is presented in the form of references to scientific publications which could be employed for calculation of a comprehensive set of quantitative indicators as essential benchmarks for strategic planning.

It is relevant to mention some favorable conditions to perform such a sophisticated job. Patents are perfectly documented and patent information is generally available online and well suited for computerized processing which makes it possible to solve the concomitant tasks and get together this tremendous science-technology puzzle at a reasonable time-span and limited labor resources.

Thus, patent documentation containing references to scientific publications is the major data resource for the study of information linkages between science and technology, reflecting the process of knowledge transfer

[1]Institute of Informatics Problems, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, 44-2 Vavilov Str., Moscow 119333, Russian Federation; izatsman@yandex.ru

[2]Institute of Informatics Problems, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, 44-2 Vavilov Str., Moscow 119333, Russian Federation; gena-mslu@mail.ru

[3]Institute of Informatics Problems, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, 44-2 Vavilov Str., Moscow 119333, Russian Federation; aleksisss@ya.ru

[4]Institute of Informatics Problems, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, 44-2 Vavilov Str., Moscow 119333, Russian Federation; chavanskov@yandex.ru

[5]Institute of Informatics Problems, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, 44-2 Vavilov Str., Moscow 119333, Russian Federation; sergeysh50@yandex.ru

from different fields of scientific research into corresponding technological areas. The conceived approach to the study of processes of knowledge transfer, described in the article, is based on the analysis of arrays of invention descriptions that contain references to scientific publications cited by authors of inventions.

This approach has been actively developed abroad for more than 30 years now [1–9]. Similar research has been conducted for several years by a dedicated team of experts at the Institute of Informatics of the Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences. The experience acquired by the team to-date has allowed to calculate numerical values of a wide range of indicators by means of computer processing of full-text descriptions of inventions and their annotations (abstracts) [10–18]. The study of knowledge transfer processes has been carried out as a part of official research activities approved by the Russian Government for the Institute of Informatics in the field of "Monitoring and Indicator Assessment of Scientific Activity" [19–23].

This paper summarizes the previously obtained results in the form of a holistic model of interplay between science and technology and presents their further expansion in the following directions:

– consideration of the diversity of scientific document types cited in patented inventions, which gives the opportunity to evaluate and compare the intensity of scientific knowledge transfer through different channels (journal articles, conference proceedings, books, etc.); and

– modeling the process of calculating the values of indicators of intensity of scientific knowledge transfer, which provides the opportunity to clarify the meaning of indicators of different types and to consider different aspects of the citation in the descriptions of inventions (e. g., the citation of scientific publications by the inventors and/or experts).

Assessment of the Intensity of Scientific Knowledge Transfer

The term "indicator of transfer intensity — ITI" is defined as a numerical value or a matrix of numerical values which characterizes a certain aspect of the intensity of scientific knowledge transfer. The article covers a wide range of indicators which are divided into two main categories: integral (ITI-I) and thematic (ITI-T).

Integral ITIs yield a general view on the intensity of knowledge transfer. Thematic ITIs are divided into several types and provide perception on thematically oriented directions of knowledge transfer. The proposed range of ITI-I types makes it possible to characterize the process of knowledge transfer from different points of view.

The scenario of information resource processing aimed at calculating ITIs (presented in Fig. 1) embraces the following components, which are determined according to the objective of the case study of scientific knowledge transfer process:

(1) description of the technologies under scrutiny using the International Patent Classification (IPC);

(2) time-span (year start – year end) for which patent applications were submitted or patents on inventions were granted, whose description texts are processed under the scenario;

(3) classifier of scientific research areas (CSRA) that is used to properly assign corresponding items from CSRA to certain references which are referred to in the descriptions of inventions;

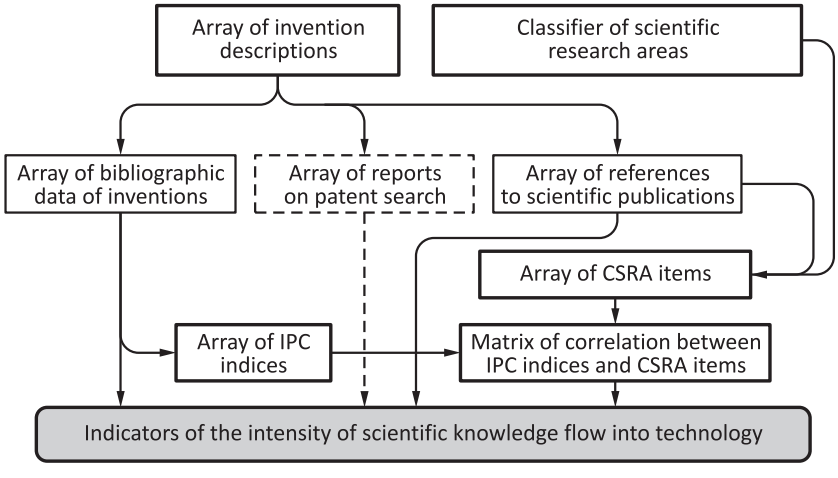


Figure 1 Scenario of information resource processing

Table 1 Information resources for ITI calculation

Information resource	Category
Array of invention descriptions	Primary
Array of bibliographic data of inventions	Primary
International patent classification	Primary
Classifier of scientific research areas	Primary
Array of reports on patent search	Primary
Array of references on scientific publications with assigned CSRA items	Secondary
Array of interconnections between items of CSRA and IPC indices	Secondary

(4) types of cited scientific publications (books, journal articles, conference proceedings, etc.); and

(5) types of ITI-T and method parameters used in determining the indicator values of each type.

Information resources which are used to determine the values of ITIs are divided into two main types:

(1) primary information resources that can be obtained from the Russian Patent Office (RPO): descriptions of inventions and bibliographic data of inventions; and

(2) secondary information resources generated by processing the primary information resources under the scenario presented in Fig. 1.

Conception of the scenario for processing information resources assumes primarily a description of the very technology sector, where the intensity of scientific knowledge transfer is the subject of investigation. It follows from the definition of "scenario" that the technological area under investigation is described using a set of IPC indices. Such a description of a specific technology allows to shape an array of primary information resources, retrieving them from the information system (database) of RPO in the form of full-text descriptions of inventions in patents.

According to the scenario at the first stage, references made by the inventor and/or expert to the cited scientific publications are retrieved, and after that, these references are distributed among the items of CSRA. According to the retrieval results, secondary information resources are created in the form of arrays of references to scientific publications. Each reference is assigned an item (or several items) from CSRA [15].

A list of primary and secondary information resources which are used to define values of ITIs is submitted in Table 1. These information resources supplemented by a system of internal relations between their components make up the basis for modeling the process of determining numerical values of ITIs.

Principles of Classification

The article is dealing with the following three principles (or bases) of ITI classification:

(1) as the first principle of ITI classification, we consider the distribution of transferred knowledge among scientific areas, fields, and disciplines in the dichotomy "basic (fundamental) – applied."

These two categories of indicators reflect how results of basic and applied research are translated into inventions. If a basic science classifier in the scenario is replaced by some applied science classifier, it will mean switching to another distribution scheme of scientific knowledge according to CSRA and, therefore, to a different category of ITI-Ts[1].

This transition could create an effect when two thematically similar publications from the point of view of basic science might be found in absolutely different branches of applied science;

(2) the second basis for classification of ITIs is based on data published by RPO in the register of reports on patent search prepared by experts of RPO during the examination of invention applications. These data include symbolic marking with letters X, Y, A, D, and T under standard ST.14 of the World Intellectual Property Organization to indicate the relationship of a specific reference to the essence of the invention in question [24]; and

(3) to explain the third principle of classification, it is necessary to mention that the number of publications cited in the descriptions of inventions and which belong to the same thematic category according to CSRA may differ significantly from patent to patent.

However, if in the description of an invention there is at least only one such reference, it means a possible knowledge transfer relating to this category of CSRA. If there are several publications that have the same CSRA item within the description of an invention, the second

[1] Values of ITI-Is by definition do not depend on such a replacement.

Table 2 Principles of ITI classification

Principle of classification	Aspect (base-code)	ITI type
Basic (fundamental) or applied science classifier of transferred knowledge (related only to ITI-T)	K — knowledge type	Basic (fundamental) research (FR) Applied research (AR)
Subject of citation: author of the invention and/or patent expert	R — relation type	Expert (E) Author (A) Expert–author (EA)
Method of taking into consideration of CSRA items of publications cited in the invention for each IPC index of invention	C — computation type	Actual (P) Frequency (F)

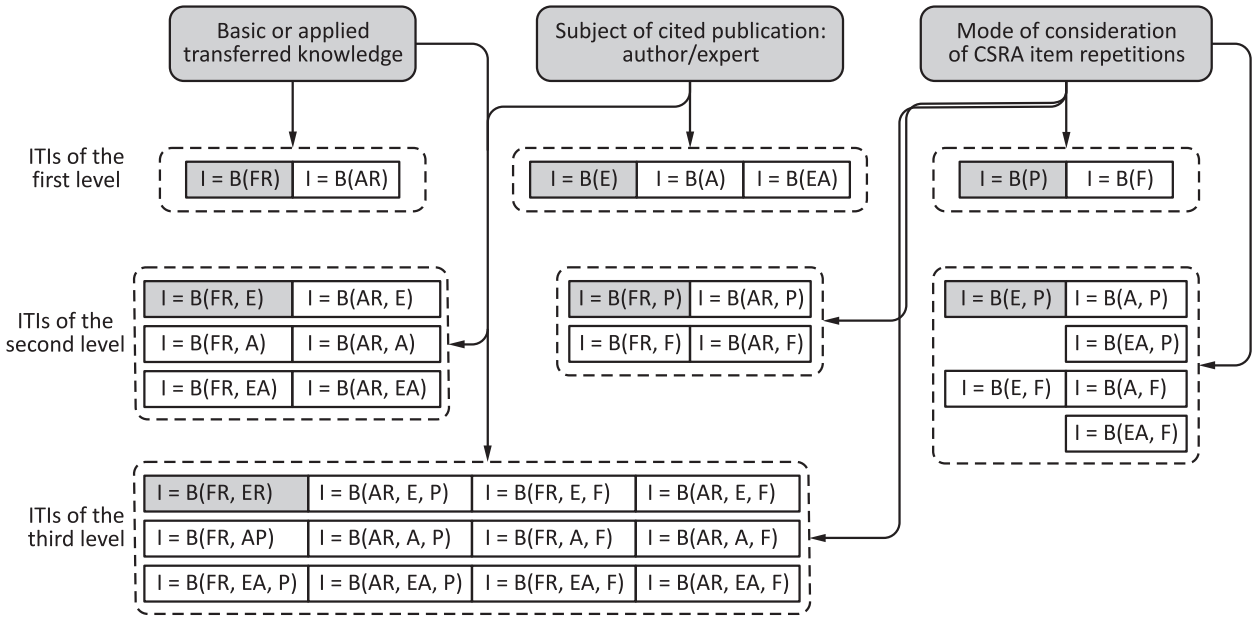


Figure 2 The ITI types

and subsequent references can be considered as recurrent cases of knowledge transfer related to this category of CSRA. Thus, the third principle of classification assumes consideration or not consideration of such repetitions of CSRA items assigned to cited publications in the descriptions of inventions. Then, it is possible to define two types of ITIs: "actual indicators" when you take into account the presence of a particular CSRA item without its repetitions and "frequency indicators" based on the number of repetitions of each CSRA item cited within one specific invention.

Thus, the article is dealing with three principles of ITI classification (presented in Table 2) which in combination provide an opportunity to offer a typology of ITIs which includes 35 types as presented in Fig. 2.

Introduction of the above mentioned three principles (or bases) of ITI classification entails three levels of typology. Simultaneous consideration of all three bases of classification of ITIs can significantly extend their range, namely, from 7 to 35 types. Seven types of ITIs which have been used previously are filled with grey color in Fig. 2 [8]. Strictly speaking, values are assigned only to 12 ITI types of the third level for each of the three bases of classification. Values for the other 23 IPI types are assigned only for one or two bases. This means that each of the indicators is formed as a combination of IPI types of the third level. For example, ITIs of type $I = B(\mathrm{FR})$ of the first level are generated by combination of six types of ITIs of types $B(\mathrm{FR, E, P})$, $B(\mathrm{FR, A, P})$, $B(\mathrm{FR, EA, P})$, $B(\mathrm{FR, E, F})$, $B(\mathrm{FR, A, F})$, and $B(\mathrm{FR, EA, F})$ which were defined by a basic science classifier.

Method of Determining the Values of Indicators

The basic concept of the method, applied to determine the values of ITIs, consists in a matrix of frequency ratios

which describes the intensity of knowledge transfer. According to the scenario, each technological area under scrutiny is described by IPC indices and each scientific area (branch, discipline) is defined by CSRA items. The quantitative characteristics of the intensity of knowledge transfer are the frequency ratios of pairs ⟨IPC index; CSRA item⟩. According to the proposed conception, the higher the number of a certain CSRA item relating to a specific IPC index (frequency ratio of the pair ⟨IPC index; CSRA item⟩), the more is significant impact of that CSRA item on the technological area with the corresponding IPC index.

Definition of the values in each cell of this matrix begins with analysis of a tuple of type ⟨invention; quoted publication⟩. During execution of the scenario, such tuples are generated for each of the invention from the list which is denoted as $P_s = \{p_1, \ldots, p_N\}$ where N is the number of descriptions of inventions that were selected as a result of a search query to the database of RPO. The query specifies values for parameters {IPC} and {Period}. Parameter {IPC} describes every technological area as a set of IPC indexes. Each invention p_i from the list of P_s is defined by a nonempty set of IPC indices $Q_i = \{q_1^i, \ldots, q_{Z(i)}^i\}$ where $Z(i)$ is the number of IPC indices specified in invention p_i.

To create the required set of indicators within each type of ITIs, the following attributes are retrieved from the field of bibliographic data of inventions (in addition to IPC indices): publication date of the initial application; date of issue of the patent; and country of the applicant. Emphasis on the IPC indices is made due to the fact that they are crucial to form the matrix of frequency ratios.

Each description of invention p_i from the list of P_s is determined by a number of references cited in scientific publications $\Pi_i = \{\pi_1^i, \ldots, \pi_{Y(i)}^i\}$ where $Y(i)$ is the number of publications cited in invention p_i. It should be noted that some sets of descriptions of inventions might be empty if there are no references in the description of an invention.

Each ordered pair of IPC index q_m^i and reference π_j^i, that is, a tuple of type $\langle q_m^i, \pi_j^i \rangle$, is denoted by $\lambda_{m,j}^i$ where $m = 1, \ldots, Z(i)$; $j = 1, \ldots, Y(i)$; and $i = 1, \ldots, N$. For invention p_i, a set of tuples $\langle q_m^i, \pi_j^i \rangle$ composes a set of tuples $\Lambda_i = \{\lambda_{m,j}^i,\ m = 1, \ldots, Z(i),\ j = 1, \ldots, Y(i)\}$, generated as a result of linguistic analysis of the description of invention p_i. In the course of distribution of references among CSRA items in cited publications, each reference is assigned as a set of items according to the specified in the CSRA scenario. These items are determined by bibliographic data of publication sources where the cited articles are published. At the same time, with the set of CSRA items, year, country, and type of a cited scientific publication are determined.

Table 3 Example of tuple ⟨IPC index, CSRA item⟩

Patent No.	Issue year	IPC index	CSRA item	CSRA item repetition
2337396	2008	G06F 1/00	02-202	2
			02-205	2
			02-206	1
			02-410	2
			02-440	2
			07-820	2
			07-450	2
2311674	2007	G06F 1/00	02-205	1
			02-410	1
			02-202	1
			07-820	1
			02-440	1

Each reference can be assigned in several CSRA items, which make up a set $R_j^i = \{r_{j,1}^i \ldots, r_{j,K(j)}^i\}$ where $K(j)$ is the number of CSRA items assigned to a certain reference π_j^i retrieved from the description of invention p_i.

This approach could be explained by data presented in Table 3. There are only two patents in Table 3 (just for example) with the same IPC index G06F 1/00. Each patent presented in the table is assigned to its specific set of science classifier items. Since the two patents have the same IPC index (G06F 1/00), both sets of CSRA items with the corresponding numbers of item repetitions are assigned to this index.

Each ordered pair of an IPC index q_m^i and CSRA item $r_{j,k}^i$, that is, a tuple of type $\langle q_m^i, r_{j,k}^i \rangle$ is denoted as $\mu_{m,k}^{i,j}$ where $m = 1, \ldots, Z(i)$; $k = 1, \ldots, K(j)$; $j = 1, \ldots, Y(i)$; and $i = 1, \ldots, N$. Then, for the description of invention p_i, one more set $M_i = \{\mu_{m,k}^{i,j}\}$ can be defined. It includes tuples of type ⟨IPC index, CSRA item⟩ as an ordered combination of IPC indices of invention p_i and CSRA items assigned to all scientific publications cited in invention p_i. The result of constructing any set of tuples M_i is stipulated by the defined in the scenario mode of consideration of reference repetitions retrieved form each invention ("actual indicator" and "frequency indicator"). Distribution of references in cited publications among CSRA items is the final stage of the linguistic analysis of the descriptions of inventions in the framework of the developed model.

Unification of all sets of M_i for all descriptions of inventions P_s yields a set of M_s tuples of type ⟨IPC index, CSRA item⟩. After unification of all M_i, there is a possibility to calculate the frequency ratio of each tuple, i. e., to find out how many times it is found in set M_s. If one groups together the tuples for each pair of IPC–CSRA, then the frequencies of these pairs will reflect the intensity of ties of CSRA items with a specific technological area defined in the scenario with the specified CSRA and the time period of the study of these relations. If we arrange

Table 4 Integral and thematic indicators

IPC index	Applied science classifier items / thematic indicator values								Integral indicator
	13.00.00	20.00.00	27.00.00	28.00.00	45.00.00	47.00.00	50.00.00	76.00.00	
1	2	3	4	5	6	7	8	9	10
G06E	0	0	0	0	0	16	0	0	16
G06F	2	423	59	758	469	517	915	0	3143
G06G	0	0	0	2	15	27	14	0	58
G06K	0	539	4	587	550	602	655	2	2939
G06N	0	19	1	14	30	26	19	1	110
G06Q	0	0	0	8	0	0	1	5	14
G06T	0	8	51	51	8	17	80	0	215

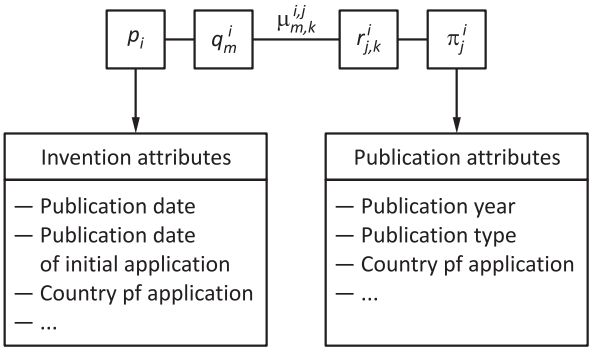


Figure 3 Data fields of invention and publication: p_i — description of invention p_i; q_m^i — IPC index of invention p_i; π_j^i — reference to publication cited in p_i; $r_{j,k}^i$ — science classifier item of publication cited in p_i; and $\mu_{m,k}^{i,j}$ — tuple of IPC index and science classifier item

IPC indices in rows and CSRA items in columns and frequencies of tuples $\langle q_m^i, r_{j,k}^i \rangle$ within total set M_s are placed in cells with addresses $(q_m^i, r_{j,k}^i)$ as presented in Table 4, then we obtain the desired matrix of frequency ratios (columns 2–9 in Table 4).

The values of frequency ratios presented in Table 4 are calculated for one of the indicators of type $I = B(\mathrm{AR}, \mathrm{A}, \mathrm{F})$ as it is designated in Fig. 3. These values are obtained under the following conditions of information resource processing:

- applied science classifier is used [25];
- only references made by the authors of inventions are taken into consideration; and
- repeated items of science classifier is applied.

The sum of all frequencies in each row (in cells from columns 2–9 in Table 4) makes up the value of the integral indicator ITI-I (column 10), reflecting the total impact of the identified research areas (branches, disciplines) on the technological area appropriate for the IPC index of this row under the three above mentioned conditions.

This linguistic-mathematical approach is distinguished by the fact that each tuple $\mu_{m,k}^{i,j} = \langle q_m^i, r_{j,k}^i \rangle$ is accompanied by data from several fields of the descriptions of inventions (publication date of the application or the patent, country of the applicant, etc.) as well as by related publications (year, type, country of publication, etc.) as presented in Fig. 3.

Concluding Remarks

Investigation of the process of knowledge transfer from different areas of scientific research into technology is important from a scientific as well as from practical points of view, because it allows to identify those scientific areas, branches, and disciplines which have exercised and thus might with substantial probability exercise in the future a significant influence on technological advances of utmost interest.

The proposed method of calculating matrices of frequency ratios can be used for strategic planning to study the extent of expected influence of scientific research on the development of technologies of interest if RPO information resources with retrospective data for a sufficiently long period of time and adequate methods of statistical forecasting are employed. The indicators devised for studying the influence of exploratory research on technologies should be regarded as information objects representing various aspects of the process of knowledge transfer and, first of all, the intensity to which results of basic and applied science are translated into high-technologies. The feasibility of the method was tested on the example of calculating indicator values of type $I = B(\mathrm{AR}, \mathrm{A}, \mathrm{F})$.

This research entails some concomitant results with significant capabilities of contribution to planning, organization, and support of long-term research at federal level with definitely specified objectives, structure, outreach, and allocated financial and human resources. Such an approach to the management of scientific activities actually expedites viable efforts of consolidating

financial, organizational, and other resources in specific scientific areas which most probably will provide subsequent technological advances crucial for economic and social development in the current environment.

Acknowledgments

The work was done under financial support of the Russian Foundation for Basic Research (project No. 16-07-00075).

References

1. Narin, F., and E. Noma. 1985. Is technology becoming science? *Scientometrics* 7(3-6):369–381.
2. Mansfield, E. 1991. Academic research and innovation. *Res. Policy* 20(1):1–12.
3. Schmoch, U. 1993. Tracing the knowledge transfer from science to technology as reflected in patent indicators. *Scientometrics* 26(1):193–211.
4. Mansfield, E. 1995. Academic research underlying industrial innovations: Sources, characteristics and financing. *Rev. Econ. Statistics* 77(1):55–62.
5. Narin, F., and D. Olivastro. 1998. Linkage between patents and papers: An interim EPO/US comparison. *Scientometrics* 41(1-2):51–59.
6. Mansfield, E. 1998. Academic research and industrial innovation: An update of empirical findings. *Res. Policy* 26(7-8):773–776.
7. Tijssen, R. J. W., R. K. Buter, and Th. N. Van Leeuwen. 2000. Technological relevance of science: An assessment of citation linkages between patents and research papers. *Scientometrics* 47(2):389–412.
8. Van Looy, B., E. Zimmermann, R. Veugelers, A. Verbeek, J. Mello, and K. Debackere. 2003. Do science–technology interactions pay on when developing technology? An exploratory investigation of 10 science-intensive technology domains. *Scientometrics* 57(3):355–367.
9. European Commission. 2003. 3rd European Report on Science & Technology Indicators. Luxembourg: Office for Official Publications of the European Communities. 451 p.
10. Zatsman, I. M., and S. K. Shubnikov. 2007. Printsipy obrabotki informatsionnykh resursov dlya otsenki innovatsionnogo potentsiala napravleniy nauchnykh issledovaniy [The principles of processing of information resources for estimation of innovative potential of the directions of scientific research]. *Tr. IX Vseross. nauchn. konf. "Elektronnye biblioteki: perspektivnye metody i tekhnologii, elektronnye kollektsii* [9th All-Russian Scientific Conference "Digital Libraries: Advanced Methods and Technologies, Digital Collections" Proceedings]. Pereslavl-Zalessky: Pereslavl University. 35–44.
11. Arkhipova, M. Yu., I. M. Zatsman, and S. Yu. Shul'ga. 2010. Indikatory patentnoy aktivnosti v sfere informatsionno-kommunikatsionnykh tekhnologiy i metodika ikh vychisleniya [Indicators of patent activities in the sphere of information and communication technologies and a technique of their computation]. *Ekonomika, statistika i informatika. Vestnik UMO* [Economics, statistics, and informatics: UMO Bull]. 4:93–104.
12. Minin, V. A., I. M. Zatsman, M. G. Kruzhkov, and T. P. Norekyan. 2013. Metodologicheskie osnovy sozdaniya informatsionnykh sistem dlya vychisleniya indikatorov tematicheskikh vzaimosvyazey nauki i tekhnologiy [Methodological bases for creating information systems calculating indicators of thematic linkages between science and technologies]. *Informatika i ee Primeneniya — Inform. Appl.* 7(1):70–81.
13. Minin, V. A., I. M. Zatsman, V. A. Havanskov, and S. K. Shubnikov. 2013. Arkhitekturnye resheniya dlya sistem vychisleniya indikatorov tematicheskikh vzaimosvyazey nauki i tekhnologiy [Information system conceptual decisions for assessment of linkages between science and technologies]. *Sistemy i Sredstva Informatiki — Systems and Means of Informatics* 23(2):260–283.
14. Zatsman, I. M., V. A. Havanskov, and S. K. Shubnikov. 2013. Metod izvlecheniya bibliograficheskoy informatsii iz polnotekstovykh opisaniy izobreteniy [Method of bibliographic information extraction from full-text descriptions of inventions]. *Informatika i ee Primeneniya — Inform. Appl.* 7(4):52–65.
15. Havanskov, V. A., and S. K. Shubnikov. 2014. Poisk i rubritsirovanie ssylok na tsitiruemye publikatsii v elektronnykh bibliotekakh polnotekstovykh opisaniy izobreteniy [Search and classifying references to cited publications in electronic libraries of full-text descriptions of inventions]. *Tr. XVI Vseross. nauchn. konf. "Elektronnye biblioteki: perspektivnye metody i tekhnologii, elektronnye kollektsii* [16th All-Russian Scientific Conference "Electronic Libraries: Perspective Methods, and Technologies, Electronic Collections" Proceedings]. Dubna: JINR. 165–173.
16. Minin, V. A., I. M. Zatsman, V. A. Havanskov, and S. K. Shubnikov. 2014. Indikatory tematicheskikh vzaimosvyazey nauki i tekhnologiy: ot teksta k chislam [Indicators of thematic science–technology linkages: From text to numbers]. *Informatika i ee Primeneniya — Inform. Appl.* 8(3):114–125.
17. Minin, V. A., I. M. Zatsman, V. A. Havanskov, and S. K. Shubnikov. 2015. Indikatory tematicheskikh vzaimosvyazey nauki i informatsionno-komp'yuternykh tekhnologiy v nachale XXI veka [Indicators for thematic linkages between science and information and computer technologies at the beginning of the XXI century]. *Informatika i ee Primeneniya — Inform. Appl.* 9(2):111–120.
18. Minin, V. A., I. M. Zatsman, V. A. Havanskov, and S. K. Shubnikov. 2016. Intensivnost' tsitirovaniya nauchnykh publikatsiy v izobreteniyakh po informatsionno-komp'yuternym tekhnologiyam, patentuemykh v Rossii otechestvennymi i zarubezhnymi zayavitelyami [Intensity of citation of scientific publications in inventions on information and computer technologies patented in Russia by domestic and foreign applicants]. *Informatika i ee Primeneniya — Inform. Appl.* 10(2):107–122.
19. Zatsman, I. M., and O. S. Kozhunova. 2007. Semanticheskiy slovar' sistemy informatsionnogo monitoringa v sfere nauki: zadachi i funktsii [Semantic dictionary of information monitoring system in science: Tasks and functions].

Sistemy i Sredstva Informatiki — Systems and Means of Informatics 17(1):124–141.

20. Zatsman, I., and O. Kozhunova. 2008. Evaluating for institutional academic activities: Classification scheme for R&D indicators. *10th Conference (International) on Science and Technology Indicators: Book of abstracts*. Vienna: ARC GmbH. 428–431.
21. Zatsman, I., and O. Kozhunova. 2009. Evaluation system for the Russian Academy of Sciences: Objectives-resources-results approach and R&D indicators. *2009 Atlanta Conference on Science and Innovation Policy Proceedings*. Eds. S. E. Cozzens and P. Catalan. Available at: http://smartech.gatech.edu/bitstream/1853/32300/1/104-674-1-PB.pdf (accessed July 14, 2017).
22. Zatsman, I., and A. Durnovo. 2010. Incompleteness problem of indicators system of research programme. *11th Conference (International) on Science and Technology Indicators: Book of abstracts*. Leiden: CWTS. 309–311.
23. Zatsman, I. M., and A. A. Durnovo. 2011. Modelirovanie protsessov formirovaniya ekspertnykh znaniy dlya monitoringa programmno-tselevoy deyatel'nosti [Modeling of processes for creation of expert knowledge for monitoring of goal-oriented programme activities]. *Informatika i ee Primeneniya — Inform. Appl.* 5(4):84–98.
24. Standart ST.14. 2016. Rekomendatsii po vklyucheniyu ssylok, tsitiruemykh v patentnykh dokumentakh [Standard ST.14. Recommendations for the inclusion of references cited in patent documents]. Available at: http://www.rupto.ru/docs/interdocs/stand_wipo/03_14_01.pdf (accessed July 14, 2017).
25. Gosudarstvennyy rubrikator nauchno-tekhnicheskoy informatsii (GRNTI) [The State Classifier of Scientific and Technical Information]. Available at: http://grnti.ru (accessed July 14, 2017).

Received July 14, 2017

Contributors

Zatsman Igor M. (b. 1952) — Doctor of Science in technology, Head of Department, Institute of Informatics Problems, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, 44-2 Vavilov Str., Moscow 119333, Russian Federation; izatsman@yandex.ru

Lukyanov Gennady V. (b. 1952) — Candidate of Military Science (PhD), associate professor, leading scientist, Institute of Informatics Problems, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, 44-2 Vavilov Str., Moscow 119333, Russian Federation; gena-mslu@mail.ru

Minin Vladimir A. (b. 1941) — Doctor of Science in physics and mathematics, consultant, Institute of Informatics Problems, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, 44-2 Vavilov Str., Moscow 119333, Russian Federation; aleksisss@ya.ru

Havanskov Valerij A. (b. 1950) — scientist, Institute of Informatics Problems, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, 44-2 Vavilov Str., Moscow 119333, Russian Federation; chavanskov@yandex.ru

Shubnikov Sergej K. (b. 1955) — senior scientist, Institute of Informatics Problems, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, 44-2 Vavilov Str., Moscow 119333, Russian Federation; sergeysh50@yandex.ru

ИНДИКАТОРНОЕ ОЦЕНИВАНИЕ ПРОЦЕССОВ ПЕРЕНОСА ЗНАНИЙ ИЗ ОБЛАСТИ НАУЧНЫХ ИССЛЕДОВАНИЙ В СФЕРУ ТЕХНОЛОГИЧЕСКОГО РАЗВИТИЯ*

И. М. Зацман, Г. В. Лукьянов, В. А. Минин, В. А. Хаванских, С. К. Шубников

Институт проблем информатики Федерального исследовательского центра «Информатика и управление» Российской академии наук

Аннотация: Данная работа посвящена индикаторному оцениванию информационных связей науки и технологий. Индикаторы связей определяются как число или матрица числовых значений, которые

* Работа выполнена при поддержке РФФИ, проект № 16-07-00075.

характеризуют интенсивность и различные аспекты процесса переноса знаний из разных областей исследований в сферу технологий. Дано описание первичных информационных ресурсов, используемых для определения значений этих индикаторов, включая полнотекстовые описания изобретений. Приводится описание вторичных информационных ресурсов, генерируемых в процессе обработки полнотекстовых описаний, включая информацию о ссылках на научные публикации, цитируемые в описаниях. Исходные и вторичные ресурсы использовались при создании и апробации информационной модели индикаторного оценивания связей науки и технологий. На ее основе были определены значения интегральных и тематических индикаторов интенсивности переноса научных знаний в сферу разработки информационных технологий.

Ключевые слова: информационные взаимосвязи науки и технологий; цитирование научных работ; интенсивность процесса переноса знаний; индикаторное оценивание; информационные технологии

DOI: 10.14357/19922264170315

Литература

1. *Narin F., Noma E.* Is technology becoming science? // Scientometrics, 1985. Vol. 7. No. 3-6. P. 369–381.
2. Mansfield E. Academic research and innovation // Res. Policy, 1991. Vol. 20. No. 1. P. 1–12.
3. *Schmoch U.* Tracing the knowledge transfer from science to technology as reflected in patent indicators // Scientometrics, 1993. Vol. 26. No. 1. P. 193–211.
4. *Mansfield E.* Academic research underlying industrial innovations: Sources, characteristics and financing // Rev. Econ. Statistics, 1995. Vol. 77. No. 1. P. 55–62.
5. *Narin F., Olivastro D.* Linkage between patents and papers: An interim EPO/US comparison // Scientometrics, 1998. Vol. 41. No. 1-2. P. 51–59.
6. *Mansfield E.* Academic research and industrial innovation: An update of empirical findings // Res. Policy, 1998. Vol. 26. No. 7-8. P. 773–776.
7. *Tijssen R. J. W., Buter R. K., Van Leeuwen Th. N.* Technological relevance of science: An assessment of citation linkages between patents and research papers // Scientometrics, 2000. Vol. 47. No. 2. P. 389–412.
8. *Van Looy B., Zimmermann E., Veugelers R., Verbeek A., Mello J., Debackere K.* Do science–technology interactions pay on when developing technology? An exploratory investigation of 10 science-intensive technology domains // Scientometrics, 2003. Vol. 57. No. 3. P. 355–367.
9. European Commission. 3rd European Report on Science & Technology Indicators. Luxembourg: Office for Official Publications of the European Communities, 2003. 451 p.
10. *Зацман И. М., Шубников С. К.* Принципы обработки информационных ресурсов для оценки инновационного потенциала направлений научных исследований // Электронные библиотеки: перспективные методы и технологии, электронные коллекции: Тр. IX Всеросс. научн. конф. — Переславль: Университет города Переславля, 2007. С. 35–44.
11. *Архипова М. Ю., Зацман И. М., Шульга С. Ю.* Индикаторы патентной активности в сфере информационно-коммуникационных технологий и методика их вычисления // Экономика, статистика и информатика: Вестник УМО, 2010. № 4. С. 93–104.
12. *Минин В. А., Зацман И. М., Кружков М. Г., Норекян Т. П.* Методологические основы создания информационных систем для вычисления индикаторов тематических взаимосвязей науки и технологий // Информатика и её применения, 2013. Т. 7. Вып. 1. С. 70–81.
13. *Минин В. А., Зацман И. М., Хаванских В. А., Шубников С. К.* Архитектурные решения для систем вычисления индикаторов тематических взаимосвязей науки и технологий // Системы и средства информатики, 2013. Т. 23. № 2. С. 260–283.
14. *Зацман И. М., Хаванских В. А., Шубников С. К.* Метод извлечения библиографической информации из полнотекстовых описаний изобретений // Информатика и её применения, 2013. Т. 7. Вып. 4. С. 52–65.
15. *Хаванских В. А., Шубников С. К.* Поиск и рубрицирование ссылок на цитируемые публикации в электронных библиотеках полнотекстовых описаний изобретений // Электронные библиотеки: перспективные методы и технологии, электронные коллекции: Тр. XVI Всеросс. научн. конф. — Дубна: ОИЯИ, 2014. С. 165–173.
16. *Минин В. А., Зацман И. М., Хаванских В. А., Шубников С. К.* Индикаторы тематических взаимосвязей науки и технологий: от текста к числам // Информатика и ёе применения, 2014. Т. 8. Вып. 3. С. 114–125.
17. *Минин В. А., Зацман И. М., Хаванских В. А., Шубников С. К.* Индикаторы тематических взаимосвязей науки и информационно-компьютерных технологий в начале XXI века // Информатика и её применения, 2015. Т. 9. Вып. 2. С. 111–120.
18. *Минин В. А., Зацман И. М., Хаванских В. А., Шубников С. К.* Интенсивность цитирования научных публикаций в изобретениях по информационно-компьютерным технологиям, патентуемых в России отечественными и зарубежными заявителями // Информатика и её применения, 2016. Т. 10. Вып. 2. С. 107–122.
19. *Зацман И. М., Кожунова О. С.* Семантический словарь системы информационного мониторинга в сфере науки: задачи и функции // Системы и средства информатики, 2007. Т. 17. № 1. С. 124–141.
20. *Zatsman I., Kozhunova O.* Evaluating for institutional academic activities: Classification scheme for R&D indicators // 10th Conference (International) on Science and

Technology Indicators: Book of abstracts. — Vienna: ARC GmbH, 2008. P. 428–431.

21. *Zatsman I., Kozhunova O.* Evaluation system for the Russian Academy of Sciences: Objectives-resources-results approach and R&D indicators // 2009 Atlanta Conference on Science and Innovation Policy Proceedings / Eds. S. E. Cozzens, P. Catalan. http://smartech.gatech.edu/bitstream/1853/32300/1/104-674-1-PB.pdf.
22. *Zatsman I., Durnovo A.* Incompleteness problem of indicators system of research programme // 11th Conference (International) on Science and Technology Indicators: Book of abstracts. — Leiden: CWTS, 2010. P. 309–311.
23. *Зацман И. М., Дурново А. А.* Моделирование процессов формирования экспертных знаний для мониторинга программно-целевой деятельности // Информатика и её применения, 2011. Т. 5. Вып. 4. С. 84–98.
24. Стандарт ST.14. Рекомендации по включению ссылок, цитируемых в патентных документах // Справочник по информации и документации в области промышленной собственности. — WIPO, 2016. С. 3-14-1–3-14-12. http://www.rupto.ru/docs/interdocs/stand_wipo/03_14_01.pdf.
25. Государственный рубрикатор научно-технической информации (ГРНТИ). http://grnti.ru.

Поступила в редакцию 14.07.2017

ОБ АВТОРАХ

Акимов Дмитрий Александрович (р. 1987) — кандидат технических наук, доцент Московского технологического университета (МИРЭА)

Андрианова Елена Гельевна (р. 1963) — кандидат технических наук, доцент Московского технологического университета (МИРЭА)

Ганебных Сергей Николаевич (р. 1968) — научный сотрудник Федерального исследовательского центра «Информатика и управление» Российской академии наук

Гудкова Ирина Андреевна (р. 1985) — кандидат физико-математических наук, доцент Российского университета дружбы народов; старший научный сотрудник Института проблем информатики Федерального исследовательского центра «Информатика и управление» Российской академии наук

Драницына Маргарита Александровна (р. 1983) — аспирант кафедры математической статистики факультета вычислительной математики и кибернетики Московского государственного университета им. М. В. Ломоносова

Дюкова Елена Всеволодовна (р. 1945) — доктор физико-математических наук, главный научный сотрудник Федерального исследовательского центра «Информатика и управление» Российской академии наук; доцент факультета вычислительной математики и кибернетики Московского государственного университета им. М. В. Ломоносова

Жуков Дмитрий Олегович (р. 1965) — доктор технических наук, профессор Московского технологического университета (МИРЭА)

Захарова Татьяна Валерьевна (р. 1962) — кандидат физико-математических наук, доцент кафедры математической статистики факультета вычислительной математики и кибернетики Московского государственного университета им. М. В. Ломоносова; старший научный сотрудник Института проблем информатики Федерального исследовательского центра «Информатика и управление» Российской академии наук

Зацман Игорь Моисеевич (р. 1952) — доктор технических наук, заведующий отделом Института проблем информатики Федерального исследовательского центра «Информатика и управление» Российской академии наук

Инькова Ольга Юрьевна (р. 1965) — доктор филологических наук, старший научный сотрудник Института проблем информатики Федерального исследовательского центра «Информатика и управление» Российской академии наук

Кириков Игорь Александрович (р. 1955) — кандидат технических наук, директор Калининградского филиала Федерального исследовательского центра «Информатика и управление» Российской академии наук

Ковалёв Сергей Протасович (р. 1972) — доктор физико-математических наук, ведущий научный сотрудник Института проблем управления им. В. А. Трапезникова Российской академии наук

Колесников Александр Васильевич (р. 1948) — доктор технических наук, профессор кафедры телекоммуникаций Балтийского федерального университета имени Иммануила Канта; старший научный сотрудник Калининградского филиала Федерального исследовательского центра «Информатика и управление» Российской академии наук

Королев Виктор Юрьевич (р. 1954) — доктор физико-математических наук, профессор, заведующий кафедрой математической статистики факультета вычислительной математики и кибернетики Московского государственного университета им. М. В. Ломоносова; ведущий научный сотрудник Института проблем информатики Федерального исследовательского центра «Информатика и управление» Российской академии наук; профессор Университета Дианьзи города Ханчжоу (Китай)

Кривенко Михаил Петрович (р. 1946) — доктор технических наук, профессор, ведущий научный сотрудник Института проблем информатики Федерального исследовательского центра «Информатика и управление» Российской академии наук

Кузнецов Михаил Павлович (р. 1989) — кандидат физико-математических наук; аналитик ООО «Форексис»

Кузнецова Маргарита Валерьевна (р. 1990) — аспирант Московского физико-технического института; руководитель отдела ЗАО «Анти-плагиат»

Ланге Андрей Михайлович (р. 1979) — кандидат физико-математических наук, научный сотрудник Федерального исследовательского центра «Информатика и управление» Российской академии наук

Ланге Михаил Михайлович (р. 1945) — кандидат технических наук, ведущий научный сотрудник Федерального исследовательского центра «Информатика и управление» Российской академии наук

Листопад Сергей Викторович (р. 1984) — кандидат технических наук, старший научный сотрудник Калининградского филиала Федерального исследовательского центра «Информатика и управление» Российской академии наук

Лукьянов Геннадий Викторович (р. 1952) — кандидат военных наук, доцент, ведущий научный сотрудник Института проблем информатики Федерального исследовательского центра «Информатика и управление» Российской академии наук

Матюшенко Сергей Иванович (р. 1963) — кандидат физико-математических наук, доцент Российского университета дружбы народов

Мейханаджян Лусине Акобовна (р. 1990) — кандидат физико-математических наук, педагог дополнительного образования школы № 281 города Москвы

Минин Владимир Александрович (р. 1941) — доктор физико-математических наук, консультант Института проблем информатики Федерального исследовательского центра «Информатика и управление» Российской академии наук

Молибог Игорь Олегович (р. 1995) — стажер-исследователь Сколковского института науки и технологий, Центр энергетических систем; студент Московского физико-технического института

Мотренко Анастасия Петровна (р. 1992) — аспирант Московского физико-технического института

Никифоров Андрей Геннадьевич (р. 1993) — студент магистратуры факультета информатики Технического университета Мюнхена (Германия)

Попкова Наталия Александровна (р. 1992) — младший научный сотрудник Института проблем информатики Федерального исследовательского центра «Информатика и управление» Российской академии наук

Прокофьев Петр Александрович (р. 1982) — кандидат физико-математических наук, научный сотрудник Федерального исследовательского центра «Информатика и управление» Российской академии наук

Пяткина Дарья Анатольевна (р. 1968) — кандидат физико-математических наук, доцент Российского университета дружбы народов

Раев Вячеслав Константинович (р. 1937) — доктор технических наук, профессор Московского технологического университета (МИРЭА)

Разумчик Ростислав Валерьевич (р. 1984) — кандидат физико-математических наук, ведущий научный сотрудник Института проблем информатики Федерального исследовательского центра «Информатика и управление» Российской академии наук; доцент Российского университета дружбы народов

Самуйлов Константин Евгеньевич (р. 1955) — доктор технических наук, профессор, заведующий кафедрой Российского университета дружбы народов; старший научный сотрудник Института проблем информатики Федерального исследовательского центра «Информатика и управление» Российской академии наук

Сафин Камиль Фанисович (р. 1995) — студент Московского физико-технического института; младший исследователь ЗАО «Анти-плагиат»

Сачков Валерий Евгеньевич (р. 1989) — аспирант Московского технологического университета (МИРЭА)

Сигов Александр Сергеевич (р. 1945) — академик Российской академии наук, Президент Московского технологического университета (МИРЭА)

Сопин Эдуард Сергеевич (р. 1986) — кандидат физико-математических наук, доцент Российского университета дружбы народов; старший научный сотрудник Института проблем информатики Федерального исследовательского центра «Информатика и управление» Российской академии наук

Стрижов Вадим Викторович (р. 1967) — доктор физико-математических наук, ведущий научный сотрудник Вычислительного центра им. А. А. Дородницына Федерального исследовательского центра «Информатика и управление» Российской академии наук

Хаванcков Валерий Александрович (р. 1950) — научный сотрудник Института проблем информатики Федерального исследовательского центра «Информатика и управление» Российской академии наук

Шоргин Сергей Яковлевич (р. 1952) — доктор физико-математических наук, профессор; заместитель директора Федерального исследовательского центра «Информатика и управление» Российской академии наук (ФИЦ ИУ РАН); главный научный сотрудник Института проблем информатики ФИЦ ИУ РАН

Шубников Сергей Константинович (р. 1955) — старший научный сотрудник Института проблем информатики Федерального исследовательского центра «Информатика и управление» Российской академии наук

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 предоставленной автором(ами) рукописи не должна нарушать положений глав 69, 70 раздела VII части IV Гражданского кодекса, которые определяют права на результаты интеллектуальной деятельности и средства индивидуализации, в том числе авторские права, в РФ.

   Ответственность за нарушение авторских прав, в случае предъявления претензий к редакции журнала, несут авторы статей.

   Направляя рукопись в редакцию, авторы сохраняют свои права на данную рукописи и при этом передают учредителям и редколлегии журнала неисключительные права на издание статьи на русском языке (или на языке статьи, если он отличен от русского) и на перевод ее на английский язык, а также на ее распространение в России и за рубежом. Каждый автор должен представить в редакцию подписанный с его стороны «Лицензионный договор о передаче неисключительных прав на использование произведения», текст которого размещен по адресу http://www.ipiran.ru/publications/licence.doc. Этот договор может быть представлен в бумажном (в 2-х экз.) или в электронном виде (отсканированная копия заполненного и подписанного документа).

   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 15 страниц указанного формата. При превышении указанного объема редколлегия вправе потребовать от автора сокращения объема рукописи.

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://www.translit.net/ru/bgn/ бесплатной программой транслитерации русского текста в латиницу.

   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, ФИЦ ИУ 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 simultaneous 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 15 pages of the specified format. If the specified size exceeded, the editorial board is entitled to require the author to reduce the manuscript.

   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.net/ru/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:
FRC CSC RAS, 44, block 2, Vavilov Str., 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