СИСТЕМЫ И СРЕДСТВА ИНФОРМАТИКИ

Научный журнал Российской академии наук
(издается под руководством Отделения нанотехнологий и информационных технологий РАН)

Издается с 1989 года
Журнал выходит ежеквартально

Учредитель:
Федеральный исследовательский центр
«Информатика и управление» Российской академии наук

РЕДАКЦИОННЫЙ СОВЕТ

РЕДАКЦИОННАЯ КОЛЛЕГИЯ

Журнал включен в базу данных Russian Science Citation Index (RSCI), интегрированную с Web of Science
Журнал входит в систему Российского индекса научного цитирования (РИНЦ)
Журнал включен в базу данных CrossRef (систему DOI — Digital Object Identifier), в базу данных Ulrich's periodicals directory и в информационную систему «Общероссийский математический портал Math-Net.Ru»

Журнал реферируется в «Реферативном журнале» ВИНИТИ и в системе Google Scholar

Журнал включен в сформированный Минобрнауки России Перечень рецензируемых научных изданий, в которых должны быть опубликованы основные научные результаты диссертаций на соискание ученой степени кандидата наук, на соискание ученой степени доктора наук

http://www.ipiran.ru/journal/collected

СИСТЕМЫ И СРЕДСТВА ИНФОРМАТИКИ

Том 29 № 3 Год 2019

СОДЕРЖАНИЕ

СИСТЕМЫ И СРЕДСТВА ИНФОРМАТИКИ

Том 29 № 3 Год 2019

СОДЕРЖАНИЕ

СИСТЕМЫ И СРЕДСТВА ИНФОРМАТИКИ, 2019. Т. 29. № 3. С. 4–15

ВЫБОР РАЗМЕРНОСТЕЙ ДЛЯ СМЕСИ ВЕРОЯТНОСТНЫХ АНАЛИЗАТОРОВ ГЛАВНЫХ КОМПОНЕНТ

М. П. Кривенко[1]

Аннотация: Рассматриваются вопросы выбора структурных параметров, характеризующих модель смеси вероятностных анализаторов главных компонент, а именно: числа элементов смеси и размерностей этих элементов. Среди набора используемых на практике подходов в случае задачи обучаемой классификации данных фактически доступными остаются лишь методы управления выборкой. Для реализации выбора размерностей предлагается использовать комбинацию известных методов выбора размерностей принятой модели. Смесь вероятностных анализаторов главных компонент позволяет моделировать объемные данные с помощью относительно небольшого числа свободных параметров. Число свободных параметров можно контролировать с помощью выбора латентной размерности данных.

Ключевые слова: вероятностный анализ главных компонент (PPCA); смеси PPCA; критерий выбора модели; бутстреп; перепроверка

DOI: 10.14357/08696527190301

1 Введение

Модель смеси анализаторов главных компонент позволяет детализировать описание реальных данных и тем самым создать предпосылки для повышения качества классификации [1]. Однако по мере роста размерности пространства данных, да еще при естественном желании увеличить число элементов смеси, объем данных, необходимых для надежного определения параметров модели, становится непомерно большим. Поэтому так важны возможности смеси вероятностных анализаторов главных компонент моделировать данные больших размеров с относительно небольшим числом свободных параметров. Число свободных параметров можно контролировать с помощью выбора скрытой пространственной размерности для каждого элемента смеси в отдельности. В этой связи центральной становится проблема выбора структурных параметров (число элементов смеси и для каждого элемента смеси количество главных компонент, которые необходимо оставить).

Базовая вероятностная модель анализа главных компонент (PPCA, probabilistic principal component analysis) для сниженной размерности k основывается на представлении

$$\mathbf{y} = \mathbf{W}\mathbf{x} + \boldsymbol{\mu} + \boldsymbol{\varepsilon}\,,$$

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

где $\mathbf{y}$ — d-мерная наблюдаемая переменная, $\mathbf{y} \sim N(\boldsymbol{\mu}, \mathbf{C}(k))$; $\mathbf{W}$ — $(d \times k)$-матрица преобразования; $\mathbf{x}$ — k-мерная латентная переменная, $\mathbf{x} \sim N(\mathbf{0}, \mathbf{I})$; $\boldsymbol{\varepsilon}$ — d-мерная переменная, $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$; $\mathbf{C}(k) = \mathbf{W}\mathbf{W}^{\mathrm{T}} + \sigma^2\mathbf{I}$. Здесь d — исходная размерность данных, k — сниженная размерность данных, $\boldsymbol{\mu}$, $\mathbf{W}$ и σ^2 суть параметры модели, при $k < d$ число скалярных параметров модели PPCA равно $dk + d + 1$.

После того как найдены оценки параметров $\hat{\boldsymbol{\mu}}$, $\hat{\sigma}^2$ и $\hat{\mathbf{W}}$, можно рассматривать случайную нормально распределенную величину $\mathbf{y} \sim N(\hat{\boldsymbol{\mu}}, \hat{\mathbf{C}}(k))$, где

$$\hat{\mathbf{C}}(k) = \hat{\mathbf{W}}\hat{\mathbf{W}}^T + \hat{\sigma}^2\mathbf{I}\,,$$

и осуществлять построение байесовского классификатора.

Связь вероятностной модели со стандартным PCA (principal component analysis) позволяет моделировать сложные структуры данных с помощью комбинации локальных подмоделей PPCA и реализации смеси вероятностных анализаторов главных компонент (PPCAM). Этот подход позволяет определять все параметры модели путем максимизации правдоподобия, в ходе которого автоматически происходит разбиение данных и определение соответствующих главных осей. Логарифм правдоподобия для такой модели смеси есть

$$L = \sum_{i=1}^{n} \ln\{p(\mathbf{y}_i)\} = \sum_{i=1}^{n} \ln \left\{ \sum_{j=1}^{m} \pi_j p(\mathbf{y}_i|j) \right\},$$

где $p(\mathbf{y}_i|j)$ отвечает элементарной PPCA(j)-модели, а π_j — соответствующий вес элемента смеси с $\pi_j \geq 0$ и $\sum_{j=1}^{m} \pi_j = 1$. С каждым j-м элементом смеси связаны свои параметры $\boldsymbol{\mu}_j$, $\mathbf{W}_j$ и σ_j^2, поэтому общее число параметров модели PPCAM равно

$$(m-1) + \sum_{j=1}^{m} (dk_j + d + 1) = d\sum_{j=1}^{m} k_j + m(d+2) - 1\,.$$

Можно разработать итерационный EM (expectation-maximization) алгоритм для оценивания всех параметров модели π_j, $\boldsymbol{\mu}_j$, $\mathbf{W}_j$ и σ_j^2. Согласно [2], обновления для параметров принимают обычный вид для смеси нормальных распределений.

2 Байесовская селекция моделей

Центральная проблема при использовании смеси PPCA — это задание структурных параметров, т. е. выбор модели смеси с определенными значениями m и $k_1, \ldots, k_m$. Вероятностный характер модели позволяет привлечь для этого байесовские принципы: структурная модель выбирается в соответствии с вероятностью, которая присваивается наблюденным данным, что полностью аналогично

байесовской классификации. Она ранжирует модели пропорционально тому, насколько они прогнозируют имеющиеся данные. При этом сложные модели оказываются менее вероятными, т. е. воплощается принцип бритвы Оккама.

Вероятность набора данных $\mathbf{D} = \{\mathbf{y}_1, \ldots, \mathbf{y}_n\}$, описываемых некоторой моделью M, вычисляется путем интегрирования по неизвестным значениям параметров θ этой модели:

$$I = p(\mathbf{D}|M) = \int\limits_{\Theta} p(\mathbf{D}|\theta, M)p(\theta|M)\, d\theta\,, \tag{1}$$

где $p(\theta|M)$ — априорная плотность параметра θ; $p(\mathbf{D}|\theta)$ — плотность данных при определенном значении θ, или функция правдоподобия θ. Здесь θ — в общем случае некоторый набор скалярных параметров и его размерность — d_M. Априорное распределение $p(\theta|M)$ позволяет учесть имеющуюся информацию о значениях параметров, но становится отягощающим фактором, когда этой информации нет.

Величина $p(\mathbf{D}|M)$ называется обоснованностью (evidence) модели M. Она является маргинальной плотностью распределения, дает представление о вероятности увидеть данные, которые действительно наблюдались, рассчитанной до того, как они стали доступны.

Для сравнения двух моделей M_1 и M_2 используется так называемый байесовский фактор B_{12}:

$$B_{12} = \frac{\mathrm{pr}(M_1|\mathbf{D})}{\mathrm{pr}(M_2|\mathbf{D})} = \frac{\mathrm{pr}(\mathbf{D}|M_1)\mathrm{pr}(M_1)}{\mathrm{pr}(\mathbf{D}|M_2)\mathrm{pr}(M_2)}\,.$$

При применении байесовского подхода ключевым становится вычисление обоснованности модели.

Прямые методы. Аналитически получить для различных моделей значения (1) удается только для тривиальных ситуаций. Точная аналитическая оценка интеграла возможна также для экспоненциального семейства распределений с соответствующими сопряженными априорными распределениями (см., в частности, [3, гл. 10]), но здесь возникают вопросы относительно целесообразности использования таких распределений, в особенности это касается сопряженного распределения. Поэтому приходится обращаться к численным методам, но их классические варианты могут быть крайне неэффективны. Одна из причин заключается в том, что с ростом размера выборки подынтегральная функция приобретает ярко выраженный максимум, а описать область ее больших значений сложно. Другая причина — это проявление в реальных задачах проклятия размерности. Представление о различных стратегиях численного интегрирования для оценки интеграла (1) можно получить из [4].

Асимптотические приближения. Метод Лапласа. Полезное приближение для (1) получается, если предположить, что апостериорная плотность, пропорциональная $p(\mathbf{D}|\theta, M)p(\theta|M)$, является остроконечной функцией, достигающей

своего максимума в апостериорной моде $\tilde{\theta}$. Это так, если функция правдоподобия $p(\mathbf{D}|\theta, M)$ также остроконечна в области своего максимума $\hat{\theta}$, что имеет место для больших выборок. Пусть $l(\theta) = \log p(\mathbf{D}|\theta, M)p(\theta|M)$. Тогда, рассматривая квадратичную аппроксимацию для $l(\theta)$, получим представление для $p(\mathbf{D}|\theta, M)p(\theta|M)$ в виде нормальной плотности со средним $\tilde{\theta}$ и ковариационной матрицей

$$\tilde{\Sigma} = \left(-\mathbf{D}^2 l\left(\tilde{\theta}\right)\right)^{-1},$$

где $\mathbf{D}^2 l(\tilde{\theta})$ — матрица Гессе. После интегрирования получаем приближение по методу Лапласа:

$$\tilde{I} = (2\pi)^{d_M/2} \left|\tilde{\Sigma}\right|^{1/2} p\left(\mathbf{D}|\tilde{\theta}, M\right) p\left(\tilde{\theta}, M\right).$$

Как следует из [5], метод Лапласа обеспечивает адекватные приближения, если правдоподобие не сильно отличается от нормального и речь идет о скромных размерностях. Авторы указанной работы не берутся быть более конкретными, но, по их мнению, выборки менее $5d_M$ вызывают беспокойство, а размером более $20d_M$ оказываются достаточными.

Метод Лапласа имеет множество модификаций как в общем случае (см. ссылки в [5]), так и применительно к PCA (см., например, [6–8]).

Асимптотические приближения. Критерий Шварца. Можно избежать введения плотностей $p(\theta|M)$, используя в выражении для байесовского фактора B_{12} величины S:

$$S = \ln p\left(\mathbf{D}|\hat{\theta}_1, M_1\right) - \ln \mathrm{pr}\left(D|\hat{\theta}_2, M_2\right) - \frac{1}{2}(d_1 - d_2)\ln n,$$

где $\hat{\theta}_k$ — оценка максимального правдоподобия; d_k — размерность параметра θ при M_k. При $n \to \infty$ величина S, часто называемая критерием Шварца, может рассматриваться как грубое приближение к логарифму байесовского фактора. Минус удвоенный критерий Шварца часто называют байесовским информационным критерием (BIC, Bayesian information criterion). В отличие от приближений по методу Лапласа, для которых относительная погрешность порядка $O(n^{-1})$, относительная ошибка $\exp S$ в приближении B_{12} обычно равна $O(1)$. Таким образом, даже для очень больших выборок это не дает истинного значения, но обеспечивает его разумное указание.

Критерий Шварца привлекателен тем, что его можно применять в качестве стандартной процедуры, даже если априорные распределения трудно установить точно. В этом смысле он предоставляет полезную разведочную информацию для практических исследований.

Другие методы. Самостоятельную группу процедур составляют методы интегрирования типа Монте-Карло (см. [5, подразд. 4.2 и 4.3]. Подобные

процедуры не всегда реально точны и требуют больших вычислительных затрат, но в сложных случаях могут оказаться единственными доступными для получения потенциально точных результатов.

Обобщения. Обычно практическое построение модели включает в себя гораздо больше, чем сравнение двух вариантов. Использование в этом случае попарного сравнения гипотез порождает типичные проблемы, но их можно избежать, по крайне мере в принципе, если принять байесовский подход и вычислить апостериорные вероятности всех конкурирующих моделей, которые следуют непосредственно из байесовских факторов. Затем можно сделать итоговый вывод, который учитывает неопределенность модели простым и формально оправданным способом.

Если для каждой гипотезы есть мера полезности Δ, то

$$\mathrm{pr}(\Delta|\mathbf{D}) = \sum_{k=0}^{K} \mathrm{pr}\left(\Delta|\mathbf{D}, M_k\right) \mathrm{pr}\left(M_k|\mathbf{D}\right) \ . \tag{2}$$

Тогда выбор одной модели и условное продолжение вывода на ее основе могут быть разумными, если одно из значений $\mathrm{pr}(M_k|\mathbf{D})$ близко к единице или если в сумме преобладают модели, значения которых аналогичны. Если нет, то анализы, обусловленные одной выбранной моделью, не в состоянии полностью учесть неопределенность в отношении структуры и поэтому могут недооценивать неопределенность, связанную с их выводами.

Несмотря на наличие общей стратегии для устранения неопределенности модели, существуют три основных препятствия для ее широкого применения. Во-первых, сложность расчета байесовских факторов. Второе препятствие состоит в том, что число слагаемых в (2) может быть огромным. Третье препятствие заключается в том, что для каждой модели должны быть указаны предварительные распределения параметров. Существуют различные возможные способы обойти это. Один из подходов состоит в том, чтобы использовать критерий Шварца, что дает точное приближение для некоторых конкретных априорных предположений, оказывающихся разумными (см., например, [5]). Другой способ — указать априорное распределения для одной или нескольких объемлющих моделей, в которые вложены все или большинство рассматриваемых моделей, а затем получить априорные значения для вложенных моделей, обусловив ограничения, которые их определяют (см., например, [9]).

Упростить использование (2) можно с позиций общих алгоритмических идей, которые приводят к методам окна Оккама и MC^3. Следует обратить внимание на то, что эти упрощения не являются приближением (2), принятым для удобства вычислений, а скорее методологическим решением. Поэтому итоговый результат может оказаться несколько иным, чем для (2).

Бутстреп-методы предоставляют возможность обойти упомянутые ранее технические трудности. Действительно, если определиться с моделью данных, то, получая повторные случайные выборки из соответствующего распределения, можно найти требуемые характеристики либо, если это удастся, аналитически,

либо, если это не так, прибегая к методам Монте-Карло. Таким образом может быть получена оценка для качества модели.

Если для различия между двумя моделями использовать расстояние Кульбака–Лейблера и критерий бутстреп-выбора модели, построенный с помощью непосредственной подстановки состоятельных оценок вместо параметров, то согласно [10] оценка этого различия будет иметь отрицательное смещение, примерно эквивалентное числу параметров в подбираемой модели. Интересно, что усложнение процедуры оценивания путем комбинирования непараметрического и параметрического бутстреп-методов приводит к практическому исчезновению смещения.

Метод перепроверки состоит в разделении исходных данных на две подвыборки: калибровочную размера $n - s$ и проверочную размера s; первая используется для подбора модели, а вторая — для оценивания ожидаемого расхождения между смоделированными и реальными данными. Обычно принимают малое значение для s и многократно повторяют формирование различных калибровочных подвыборок с последующим оцениванием эффективности по соответствующим оставшимся проверочным подвыборкам и осреднением полученных результатов. Более подробное описание методов перепроверки дано в [11].

3 Выбор размерностей в задаче обучаемой классификации

Рассмотрим проблему выбора модели в задаче обучаемой классификации данных размерности d для G классов, $G \geq 2$. Пусть для описания данных каждого класса с индексом g применяется PPCAM, характеризуемая числом элементов смеси m_g и сниженными размерностями элементов смеси $k_{g1}, \ldots, k_{gm_g}$, при этом $m_g \leq m_{\max}$, где $m_{\max}$ задано. Тогда для определенного класса число оцениваемых параметров есть $d\sum_{j=1}^{m_g} k_{gj} + m_g(d+2) - 1$, а общее число моделей:

$$\sum_{m_g=1}^{m_{\max}} d^{m_g} = \frac{d(d^{m_{\max}} - 1)}{d-1}\,.$$

Для полной задачи классификации число оцениваемых параметров — $\sum_{g=1}^{G}\left(d\sum_{j=1}^{m_g} k_j + m_g(d+2) - 1\right)$, а общее число моделей — $(d(d^{m_{\max}} - 1)/(d-1))^G$.

Данные формулы для общего количества различных по структуре моделей приводят к большим значениям, что практически исключает возможность их полного сравнительного анализа. Даже в случае структурно несложной задачи дискриминации данных о ферментах, рассмотренной в [12], для $d = 4$, $G = 4$, $m_{\max} = 10$ получаем для полной задачи классификации общее число моделей порядка $3{,}8 \cdot 10^{24}$, а для упрощенной — $2\,560\,000 \approx 2{,}6 \cdot 10^6$.

Рассматриваемая вероятностная модель классифицируемых данных представляет собой смесь распределений (как основа байесовской классификации),

каждый элемент которой есть, в свою очередь, также смесь уже нормальных распределений (как средство описания в рамках отдельного класса неоднородных данных с распределением, фактически отличающимся от нормального). Это кардинально отличается от обычно рассматриваемой в задаче выбора модели для классификации данных смеси нормальных распределений, когда класс описывается просто нормальным распределением (см., например, [13]).

Из-за эффекта проклятия размерности реальным остается лишь последовательный упорядоченный перебор от простых к более сложным моделям с последующим оцениванием их обоснованности и остановка по достижении удовлетворительного результата, что полностью соответствует принципу бритвы Оккама.

В случае задачи классификации в качестве меры полезности Δ естественно использовать вероятность ошибки (или вероятность правильной классификации). Она полностью оправдана с содержательной точки зрения, но создает практически непреодолимые трудности при попытках ее аналитического вычисления. Как следствие, усложнение модели и реальные возможности методов оценивания обоснованности моделей приводят к необходимости обращаться к методам управления обработкой выборки.

В области PCA отправной для выбора размерностей в задаче обучаемой классификации становится проблема выбора числа главных компонент, которые необходимо оставить. Рассматривая PPCA, можно использовать байесовский выбор модели истинной размерности данных. Для его реализации обычно вводятся априорные распределения, включающие гиперпараметр для контроля «жесткости» принимаемых предположений, а для вычисления интеграла выбирается подходящая параметризация и применяется метод Лапласа. В результате можно получить несложную и практичную оценку (см., например, [6]). При этом достоинства полученной оценки демонстрируются с помощью моделирования отдельных случаев. Обращают на себя внимание следующие моменты:

- предлагаемые решения предназначены для оценки плотности, т. е. касаются точности представления данных, и неизвестно, как они проявят себя в задаче классификации;
- методы перепроверки были включены в сравнительный анализ и практически так же правильно выбирают размерность при PCA;
- в итоге вывод об эффективности подхода принимается с помощью моделирования.

Что касается смеси байесовских моделей PCA, то данный вопрос затрагивался в [14], но фактически без раскрытия способов вычисления обоснованности для смеси. Анонсированные результаты свидетельствуют скорее в пользу использования полноразмерных моделей PCA, а не стремления добиваться снижения размерности индивидуально для каждого элемента смеси.

Цель перекрестной проверки состоит в том, чтобы найти подходящее число компонент для модели PCA. В [15] описаны шесть методов перекрестной проверки для задач PCA. Четыре из них обычно используются в стандартном

программном обеспечении, а два дополнительных являются новыми предложениями, направленными на преодоление некоторых потенциальных проблем для используемых в настоящее время методов. Некоторые из существующих методов были описаны в литературе ранее, но ни один из них не был сравнен друг с другом в деталях; немногие из них являются тривиальными расширениями первоначальной идеи перекрестной проверки, поэтому вопрос о том, дают ли они значимые результаты, еще предстоит тщательно показать на практике. Не надо забывать, что данные предпочтения получены методом моделирования только для PPCA и для отдельных конкретных ситуаций, выбор которых сам по себе нетривиальная задача. Кроме этого критерием эффективности выступает точность прогноза, а не преимущества применения модели для решения целевой задачи классификации данных.

Перекрестная проверка — это опробованный и проверенный подход к выбору числа компонент в анализе главных компонент, однако его основной недостаток кроется в стоимости вычислений. Это приводит к тому, что появляются критерии аппроксимации перекрестной проверки, например критерий сглаживающей аппроксимации SACV (smoothing approximation of cross-validation) и обобщенный критерий GCV (generalized cross-validation) [16]. Свойства соответствующих методов оцениваются с помощью моделирования, которые дают многообещающие результаты. Критерий SACV обеспечивает наилучшие результаты при жесткой структуре данных, тогда как критерий GCV лучше в схемах с шумом.

Таким образом, способы оценивания структурных параметров выбранной модели данных формируются в непростых условиях: необходимость привлечения сложных априорных предположений и возникающие из-за этого проблемы полного их задания, отсутствие конкретных наработок в области селекции моделей смеси анализаторов главных компонент, затруднительность решения сопутствующих аналитических и вычислительных задач. Поэтому приходится обращаться к отдельным приемам, реализующим общие идеи и позволяющим получать гарантированные по качеству решения.

Использование опробованных методов селекции моделей для смесей распределений и отдельно для PPCA приводит к упрощенной схеме действий:

- подобрать число элементов смеси для каждого класса самостоятельно (например, с помощью одного из информационных критериев) в предположении, что данные являются полноразмерными;
- при найденных значениях числа элементов смеси для каждого класса провести в отдельности снижение размерности байесовскими или иными методами;
- привлекая методы управления обработкой выборки для полученного решения, оценить вероятности правильной классификации.

Подобный способ не гарантирует наилучшего решения, но позволяет прояснить, возможно ли снижение размерности и приводит ли оно к повышению качества классификации.

Объединение методологически обоснованного байесовского подхода и практичных методов управления обработкой выборки дает следующую итерационную процедуру:

- сформировать набор рассматриваемых моделей смеси анализаторов главных компонент и среди них наиболее подходящую модель с помощью одного из информационных критериев (в первую очередь речь идет о BIC);
- руководствуясь оценкой вероятности правильной классификации попытаться с помощью метода перепроверки улучшить найденную модель методом покоординатного спуска для структурных параметров.

Описанная процедура в рамках реальной задачи может модифицироваться. Так, если ситуация оказывается слишком сложной (конкретные значения величин $d, G, = 4, m_{\max}$ не позволяют реализовать планы), то приходится идти на упрощение при формировании исходного набора моделей и/или в ходе реализации покоординатного спуска.

4 Заключение

Байесовский подход играет важную методологическую роль в современной вычислительно-интенсивной статистической практике. Это относится к ограниченному, но важному классу проблем научного вывода, а также к оценке неопределенности, когда изначально рассматривается множество моделей. Здесь важно подчеркнуть следующее:

- байесовские факторы универсальны, в частности они не требуют, чтобы альтернативные модели были вложенными, допускают их последовательное формирование;
- существуют несколько методов для вычисления байесовских факторов, включая асимптотические приближения, которые легко вычислить, или средства стандартных пакетов обработки данных;
- критерий Шварца (или BIC) дает грубое приближение к логарифму байесовского фактора, который прост в использовании и не требует оценки априорных распределений; он хорош как разведочное средство, подходит для подведения итогов научных исследований;
- имеются алгоритмы, позволяющие учитывать неопределенность модели, когда первоначально рассмотренный класс моделей очень велик;
- байесовские факторы полезны для руководства эволюционным процессом построения модели;
- важно и целесообразно оценить чувствительность выводов к ранее использованным распределениям, обычно обнаруживается, что выводы являются надежными по отношению к предыдущим в качественном смысле, но понятно, что это не гарантируется.

Байесовские факторы имеют много сильных и слабых сторон в байесовском подходе в целом. Основным преимуществом является их прочная логическая основа, которая обеспечивает большую гибкость. Вопрос о том, сколько усилий следует предпринять, прежде чем делать выводы, возникает в любой проблеме анализа данных. Это часть искусства прикладной статистики, оказывающая помощь исследователю при решении, стоит ли продолжать полный байесовский анализ.

Литература

1. *Кривенко М. П.* Обучаемая классификация данных с учетом анализа главных компонент // Информатика и её применения, 2018. Т. 12. Вып. 3. С. 56–61.
2. *Tipping M. E., Bishop C. M.* Mixtures of probabilistic principal component analyzers // Neural Comput., 1999. Vol. 11. Iss. 2. P. 443–482.
3. *Zellner A.* An introduction to Bayesian inference in econometrics. — New York, NY, USA: Wiley, 1971. 431 p.
4. *Evans M., Swartz T.* Approximating integrals via Monte Carlo and deterministic method. — New York, NY, USA: Oxford University Press, 2000. 290 p.
5. *Kass R. E., Raftery A. E.* Bayes factors // J. Am. Stat. Assoc., 1995. Vol. 90. Iss. 430. P. 773–795.
6. *Minka T. P.* Automatic choice of dimensionality for PCA // Advances in neural processing systems 13 / Eds. T. K. Leen, T. G. Dietterich, V. Tresp. — MIT Press, 2000. P. 598–604. http://papers.nips.cc/paper/1853-automatic-choice-of-dimensionality-for-pca.pdf.
7. *Hoyle D. C.* Automatic PCA dimension selection for high dimensional data and small sample sizes // J. Mach. Learn. Res., 2008. Vol. 9. P. 2733–2759.
8. *Nakajima S., Sugiyama M., Babacan D.* On Bayesian PCA: Automatic dimensionality selection and analytic solution // 28th Conference (International) on Machine Learning Proceedings. — Bellevue, WA, USA, 2011. P. 497–504. http://www.icml-2011.org/papers/337_icmlpaper.pdf?CFID=122408014&CFTOKEN=5f7f69b335b8fcd0-38A2E56E-A506-DB5F-F9185D08D5EE991A.
9. *Raftery A. E.* Approximate Bayes factors and accounting for model uncertainty in generalized linear models. — University of Washington, Department of Statistics, 1993. Technical Report 255. 45 p. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.142.9000&rep=rep1&type=pdf.
10. *Chung H.-Y., Lee K.-W., Koo J.-Y.* A note on bootstrap model selection criterion // Stat. Probabil. Lett., 1996. Vol. 26. Iss. 1. P. 35–41.
11. *Arlot S., Celisse A.* A survey of cross-validation procedures for model selection // Statistics Surveys, 2010. Vol. 4. P. 40–79.
12. *Кривенко М. П.* Обучаемая классификация неполных клинических данных // Информатика и её применения, 2017. Т. 11. Вып. 3. С. 27–33.
13. *Jacques J., Bouveyron C., Girard S., Devos O., Duponchel L.* Gaussian mixture models for the classification of high-dimensional vibrational spectroscopy data // J. Chemometr., 2010. Vol. 24. Iss. 11-12. P. 719–727.

14. *Bishop C. M.* Bayesian PCA // Advances in neural information processing systems 11 / Eds. M. J. Kearns, S. A. Solla, D. A. Cohn. — MIT Press, 1998. P. 382–388. http://papers.nips.cc/paper/1549-bayesian-pca.pdf.
15. *Bro R., Kjeldahl K., Smilde A. K., Kiers H. A. L.* Cross-validation of component models: A critical look at current methods // Anal. Bioanal. Chem., 2008. Vol. 390. Iss. 5. P. 1241–1251.
16. *Josse J., Husson F.* Selecting the number of components in principal component analysis using cross-validation approximations // Comput. Stat. Data An., 2012. Vol. 56. Iss. 6g. P. 1869–1879.

Поступила в редакцию 16.07.19

SELECTING THE DIMENSIONALITY FOR MIXTURE OF PROBABILISTIC PRINCIPAL COMPONENT ANALYZERS

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 considers the problems of choosing structural parameters characterizing the model of a mixture of probabilistic principal component analyzers, namely, the number of elements of the mixture and the dimensions of these elements. Among the set of approaches used in practice for the task of classifying data, only sampling management methods are actually available. To implement the choice of dimensions, it is proposed to use a combination of the known methods for model selecting. The mixture of probabilistic principal component analysis allows one to model bulk data using a relatively small number of free parameters. The number of free parameters can be controlled by selecting the latent dimension of the data.

Keywords: probabilistic principal component analysis (PPCA); mixtures of PPCA; model selection criterion; bootstrap; cross-validation

DOI: 10.14357/08696527190301

References

1. Krivenko, M. P. 2018. Obuchaemaya klassifikatsiya dannykh s uchetom analiza glavnykh component [Supervised learning classification of data taking into account principal component analysis]. *Informatika i ee Primeneniya — Inform. Appl.* 12(3):56–61.
2. Tipping, M. E., and C. M. Bishop. 1999. Mixtures of probabilistic principal component analyzers. *Neural Comput.* 11(2):443–482.
3. Zellner, A. 1971. *An introduction to Bayesian inference in econometrics.* New York, NY: Wiley. 431 p.

4. Evans, M., and T. Swartz. 2000. *Approximating integrals via Monte Carlo and dterministic method.* New York, NY: Oxford University Press Inc. 290 p.
5. Kass, R. E., and A. E. Raftery. 1995. Bayes factors. *J. Am. Stat. Assoc.* 90(430):773–795.
6. Minka, T. P. 2000. Automatic choice of dimensionality for PCA. *Advances in neural processing systems 13*. Eds. T. K. Leen, T. G. Dietterich, and V. Tresp. MIT Press. 598–604. Available at: http://papers.nips.cc/paper/1853-automatic-choice-of-dimensionality-for-pca.pdf (accessed May 14, 2019).
7. Hoyle, D. C. 2008. Automatic PCA dimension selection for high dimensional data and small sample sizes. *J. Mach. Learn. Res.* 9:2733–2759.
8. Nakajima, S., M. Sugiyama, and D. Babacan. 2011. On Bayesian PCA: Automatic dimensionality selection and analytic solution. *28th Conference (International) on Machine Learning Proceedings*. Bellevue, WA. 497–504. Available at: http://www.icml-2011.org/papers/337_icmlpaper.pdf?CFID=122408014&CFTOKEN=5f7f69b335b8fcd0-38A2E56E-A506-DB5F-F9185D08D5EE991A (accessed May 14, 2019).
9. Raftery, A. E. 1993. Approximate Bayes factors and accounting for model uncertainty in generalized linear models. Technical Report 255. University of Washington, Department of Statistics. 45 p. Available at: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.142.9000&rep=rep1&type=pdf (accessed May 14, 2019).
10. Chung, H.-Y., K.-W. Lee, and J.-Y. Koo. 1996. A note on bootstrap model selection criterion. *Stat. Probabil. Lett.* 26(1):35–41.
11. Arlot, S., and A. Celisse. 2010. A survey of cross-validation procedures for model selection. *Statistics Surveys* 4:40–79.
12. Krivenko, M. P. 2017. Obuchaemaya klassifikatsiya nepolnykh klinicheskikh dannykh [Supervised learning classification of incomplete clinical data]. *Informatika i ee Primeneniya — Inform. Appl.* 11(3):27–33.
13. Jacques, J., C. Bouveyron, S. Girard, O. Devos, and L. Duponchel. 2010. Gaussian mixture models for the classification of high-dimensional vibrational spectroscopy data. *J. Chemometr.* 24(11-12):719–727.
14. Bishop, C. M. 1998. Bayesian PCA. *Advances in neural information processing systems 11*. Eds. M. J. Kearns, S. A. Solla, and D. A. Cohn. — MIT Press. 382–388. Available at: http://papers.nips.cc/paper/1549-bayesian-pca.pdf (accessed May 14, 2019).
15. Bro, R., K. Kjeldahl, A. K. Smilde, and H. A. L. Kiers. 2008. Cross-validation of component models: A critical look at current methods. *Anal. Bioanal. Chem.* 390(5):1241–1251.
16. Josse, J., and F. Husson. 2012. Selecting the number of components in principal component analysis using cross-validation approximations. *Comput. Stat. Data An.* 56(6g):1869–1879.

Received July 16, 2019

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

УСЛОВНО-ОПТИМАЛЬНОЕ ЛИНЕЙНОЕ ОЦЕНИВАНИЕ НОРМАЛЬНЫХ ПРОЦЕССОВ В ВОЛЬТЕРРОВСКИХ СТОХАСТИЧЕСКИХ СИСТЕМАХ*

***И. Н. Синицын*[1], *В. И. Синицын*[2]**

Аннотация: На основе теории условно-оптимального оценивания (фильтрации и экстраполяции) Пугачёва и предыдущих исследований авторов разработаны два приближенных условно-оптимальных метода для фильтрации и экстраполяции нормальных процессов в вольтерровских стохастических системах (ВСтС), приводимых к СтС с аддитивными и параметрическими шумами. Сформулированы подходы к синтезу фильтров Пугачёва для ВСтС с аддитивными и параметрическими шумами путем эквивалентной замены ВСтС с аддитивными шумами. Подробно рассмотрены одномерные тестовые примеры. Результаты допускают непосредственное обобщение на случаи автокоррелированных шумов и нелинейных функций межвидового взаимодействия, эредитарных ВСтС, а также дискретных и непрерывно-дискретных ВСтС.

Ключевые слова: вольтерровская СтС (ВСтС); метод аналитического моделирования (МАМ); метод канонических разложений (МКР); метод нормальной аппроксимации (МНА); метод статистический линеаризации (МСЛ); стохастическая система (СтС); стохастический процесс (СтП); условно-оптимальные фильтры и экстраполяторы Пугачёва; фильтры и экстраполяторы Калмана

DOI: 10.14357/08696527190302

1 Введение

На основе [1] в [2] разработаны два эффективных метода аналитического моделирования (МАМ) эквивалентных нормальных (гауссовских) стохастических процессов (СтП) в многоуровневых ВСтС. Первый основан на методе нормальной аппроксимации (МНА) для многомерной ВСтС с аддитивными и параметрическими гауссовскими и негауссовскими белыми шумами для функций межвидового взаимодействия произвольного вида, в том числе разрывных. Второй основан на методе статистической линеаризации (МСЛ) функций межвидового взаимодействия и сведении исходной ВСтС к эквивалентной ВСтС

*Работа выполнена при финансовой поддержке РАН (проект 0063-2018-0008).

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

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

с аддитивными линейными и параметрическими шумами. Получена совместная система уравнений для вероятностных моментов первого и второго порядка. Рассмотрены вопросы аналитического моделирования стационарных регулярных и стохастических режимов. В качестве тестовых примеров получены уравнения нелинейного корреляционного МАМ одной и двух популяций в стохастической среде. Изучены стационарные режимы и их устойчивость. В [3] представлены точные и приближенные МАМ процессов в нелинейных ВСтС в условиях аддитивных и параметрических белых шумов. Изучена устойчивость стационарных регулярных и стохастических режимов по первым двум вероятностным моментам. Рассмотрены вопросы аналитического моделирования одномерных распределений с инвариантной мерой в двумерной дифференциальной ВСтС.

На основе методов канонических разложений (МКР) в [4] разработаны приближенные нелинейные корреляционные МАМ гармонических, квазигармонических и широкополосных режимов в обобщенных нелинейных вольтерровских системах в условиях аддитивных и параметрических гармонических, квазигармонических и широкополосных возмущений. Подробно рассмотрена стохастическая задача В. Вольтерры «R хищников – R жертв» с одним трофическим уровнем. В качестве тестовых примеров рассмотрены задачи нелинейного корреляционного МАМ одномерных и двумерных нелинейных ВСтС в условиях аддитивных и параметрических гармонических, квазигармонических и широкополосных возмущений. Изучена устойчивость регулярных и стохастических режимов. Рассмотрены важные частные случаи.

Рассмотрим развитие [2–4] на случай условно-оптимального линейного оценивания (фильтрации и экстраполяции) нормальных СтП в ВСтС, приводимых к линейным СтС с аддитивными и параметрическими шумами. Статья содержит введение, заключение и 3 раздела. В разд. 2 приведены необходимые сведения из теории условно-оптимального линейного оценивания. В разд. 3 изложены основные результаты в области синтеза условно-оптимальных линейных фильтров и экстраполяторов Пугачёва для обработки информации в ВСтС с аддитивными и параметрическими шумами. Сформулированы принципы эквивалентной замены ВСтС с параметрическими шумами на ВСтС с аддитивными шумами. В разд. 4 рассмотрен ряд тестовых задач для одномерных ВСтС. Заключение содержит основные выводы и возможные обобщения.

2 Условно-оптимальное линейное оценивание в линейных дифференциальных стохастических системах с параметрическими шумами

2.1. Фильтрация. Линейные непрерывные (дифференциальные) системы с параметрическими шумами служат подходящей математической моделью линейных динамических систем со случайными параметрами. На основе таких моделей СтС удается описать специальные эффекты возникновения стохасти-

ческих параметрических колебаний, появления дрейфов и пр. В таких СтС применение линейного фильтра Калмана–Бьюси невозможно. Поэтому целесообразно использовать линейные фильтры Пугачёва [4].

Следуя [5], рассмотрим две взаимосвязанные векторные дифференциальные системы стохастических дифференциальных уравнений Ито с негауссовскими в общем случае белыми шумами:

$$dX_t = (aY_t + a_1X_t + a_0)\,dt + \left(c_{10} + \sum_{r=1}^{n_y} c_{1r}Y_r + \sum_{r=1}^{n_x} c_{1,n_y+r}X_r\right)dW\,; \qquad (1)$$

$$dY_t = (bY_t + b_1X_t + b_0)\,dt + \left(c_{20} + \sum_{r=1}^{n_y} c_{2r}Y_r + \sum_{r=1}^{n_x} c_{2,n_y+r}X_r\right)dW\,, \qquad (2)$$

где a_0, b_0, a, a_1, b, b_1 и c_{ij} $(i = 1, 2,\ j = 1, \ldots, n_x)$ — векторно-матричные функции t, не зависящие от вектора состояния $X_t = [X_1 \cdots X_{n_x}]^{\mathrm{T}}$ и вектора наблюдения $Y_t = \left[Y_1 \cdots Y_{n_y}\right]^{\mathrm{T}}$. Класс допустимых фильтров Пугачёва определим линейным уравнением

$$d\hat{X}_t = (aY_t + a_1\hat{X}_t + a_0)\,dt + \beta_t\left[dY_t - \left(bY_t + b_1\hat{X}_t + b_0\right)dt\right]. \qquad (3)$$

Для определения β_t в (3) необходимо найти математическое ожидание m_t и ковариационную матрицу K_t случайного вектора $Q_t = [X_1 \cdots \hat{X}_{n_x} Y_1 \cdots Y_{n_y}]^{\mathrm{T}}$ и ковариационную матрицу R_t ошибки $\tilde{X}_t = \hat{X}_t - X_t$. Для этого воспользуемся корреляционными уравнениями. Эти уравнения в данном случае имеют вид:

$$\left.\begin{aligned} &\dot{m}_t = \bar{a}m_t + \bar{a}_0; \\ &\dot{K}_t = \bar{a}K_t + K_t\bar{a}^{\mathrm{T}} + c_0\nu c_0^{\mathrm{T}} + \sum_{r=1}^{n_y+n_x}\left(c_0\nu c_r^{\mathrm{T}} + c_r\nu c_0^{\mathrm{T}}\right)m_r + \\ &\qquad + \sum_{r,s=1}^{n_y+n_x} c_r\nu c_s^{\mathrm{T}}\left(m_rm_s + k_{rs}\right), \end{aligned}\right\} \qquad (4)$$

где

$$\bar{a} = \begin{bmatrix} b & b_1 \\ a & a_1 \end{bmatrix};\ \bar{a}_0 = \begin{bmatrix} b_0 \\ a_0 \end{bmatrix};\ c_r = \begin{bmatrix} c_{2r} \\ c_{1r} \end{bmatrix};\ c_0 = \begin{bmatrix} c_{10} \\ c_{20} \end{bmatrix},\ (r = 0, 1, \ldots, n_y + n_x).$$

Уравнения (4) с соответствующими начальными условиями определяют все элементы m_r, k_{rs} матрицы-столбца m_t и матрицы K_t $(r, s = 1, \ldots, n_y + n_x)$.

Уравнение для матрицы ошибки $\tilde{X}_t$ фильтрации R_t определяется матричным уравнением Риккати:

$$\dot{R}_t = a_1 R_t + R_t a_1^{\mathrm{T}} - \left[R_t b_1^{\mathrm{T}} + \left(c_{10} + \sum_{r=1}^{n_y+n_x} c_{1r} m_r \right) \nu \left(c_{20}^{\mathrm{T}} + \sum_{r=1}^{n_y+n_x} c_{2r}^{\mathrm{T}} m_r \right) + \right.$$
$$\left. + \sum_{r,s=1}^{n_y+n_x} c_{1r} \nu c_{2s}^{\mathrm{T}} k_{rs} \right] \kappa_{11}^{-1} \left[b_1 R_t + \left(c_{20} + \sum_{r=1}^{n_y+n_x} c_{2r} m_r \right) \nu \left(c_{10}^{\mathrm{T}} + \sum_{r=1}^{n_y+n_x} c_{1r}^{\mathrm{T}} m_r \right) + \right.$$
$$\left. + \sum_{r,s=1}^{n_y+n_x} c_{2r} \nu c_{1s}^{\mathrm{T}} k_{rs} \right] + \left(c_{10} + \sum_{r=1}^{n_x+n_y} c_{1r} m_r \right) \nu \left(c_{10}^{\mathrm{T}} + \sum_{r=1}^{n_x+n_y} c_{1r}^{\mathrm{T}} m_r \right) +$$
$$+ \sum_{r,s=1}^{n_y+n_x} c_{1r} \nu c_{1s}^{\mathrm{T}} k_{rs} \,. \quad (5)$$

Здесь κ_{11} определяется формулой

$$\kappa_{11} = \left(c_{20} + \sum_{r=1}^{n_y+n_x} c_{2r} m_r \right) \nu \left(c_{20}^{\mathrm{T}} + \sum_{r=1}^{n_y+n_x} c_{2r}^{\mathrm{T}} m_r \right) + \sum_{r=1}^{n_y+n_x} c_{2r} \nu c_{2s}^{\mathrm{T}} k_{rs}, \quad (6)$$

$W = W(t)$ — векторный СтП с независимыми приращениями с нулевым математическим ожиданием и конечной ковариационной функцией

$$k_w(t_1, t_2) = k(\min(t_1, t_2)\,, \;\; k(t) = k(t_0) + \int_{t_0}^{t} \nu(\tau)\, d\tau\,,$$

где $\nu = \nu(t)$ — матрица интенсивностей СтП $W = W(t)$.

Таким образом, имеет место следующий результат.

Теорема 2.1. *Пусть векторный СтП* $\left[X_t^{\mathrm{T}} Y_t^{\mathrm{T}}\right]^{\mathrm{T}}$ *определяется уравнениями линейной дифференциальной негауссовской СтС с параметрическими шумами* (1), (2) *и обладает конечными одномерными моментами. Тогда уравнение линейного фильтра Пугачева имеет вид* (3). *После нахождения моментов первого и второго порядка* m_r, k_{rs} $(r, s = 1, \ldots, n_y + n_x)$ *и ковариационной матрицы ошибки фильтрации* R_t *путем интегрирования уравнений* (4) и (5) *оптимальный коэффициент* β_t *в уравнении фильтра* (3) *определяется по формуле*:

$$\beta_t = \left\{ R_t b_1^{\mathrm{T}} + \left(c_{10} + \sum_{r=1}^{n_y+n_x} c_{1r} m_r \right) \nu \left(c_{20}^{\mathrm{T}} + \sum_{r=1}^{n_y+n_x} c_{2r}^{\mathrm{T}} m_r \right) + \right.$$
$$\left. + \sum_{r,s=1}^{n_y+n_x} c_{1r} \nu c_{2s}^{\mathrm{T}} k_{rs} \right\} \kappa_{11}^{-1}. \quad (7)$$

Пример 2.1. Рассмотрим случай, когда скалярные уравнения (1) и (2) содержат независимые белые шумы V_1 и V_2:

$$\dot{X}_t = aY_t + a_1X_t + a_0 + (c_{10} + c_{11}X_t + c_{12}Y_t)\,V_1;$$
$$\dot{Y}_t = bY_t + b_1X_t + b_0 + (c_{20} + c_{21}X_t + c_{22}Y_t)V_2,\ \ \nu = \begin{bmatrix}\nu_1 & 0\\ 0 & \nu_2\end{bmatrix}.$$

Обратим внимание на то, что c_{1r} и c_{2r} здесь не те, что в уравнениях (1) и (2). Они представляют собой соответственно первые и вторые элементы матриц-строк, на которые умножается вектор $[V_1V_2]^{\mathrm{T}}$ в (1) и (2). Для простоты оставляем для них обозначения c_{1r} и c_{2r}. Тогда в формулах (6) и (7) имеем:

$$\left.\begin{aligned}\kappa_{11} &= \nu_2\left(c_{20} + c_{21}m_1 + c_{22}m_2\right)^2 + \nu_2\left(c_{21}^2k_{11} + 2c_{22}c_{21}k_{21} + c_{22}^2k_{22}\right);\\ \beta_t &= \kappa_{11}^{-1}b_1R_t.\end{aligned}\right\} \tag{8}$$

При этом уравнения (4) и (5), определяющие m_1, m_2, k_{11}, k_{12}, k_{21}, k_{22} и R_t, имеют следующий вид:

$$\left.\begin{aligned}\dot{m}_1 &= a_1m_1 + am_2 + a_0;\ \ \dot{m}_2 = b_1m_1 + bm_2 + b_0;\\ \dot{k}_{11} &= 2\left(a_1k_{11} + ak_{12}\right) + \nu_1\left(c_{11}m_1^2 + c_{12}m_2 + c_{10}\right)^2 +\\ &\qquad + \nu_1\left(c_{11}^2k_{11} + 2c_{12}c_{11}k_{12} + c_{12}^2k_{22}\right);\\ \dot{k}_{12} &= \left(a_1 + b\right)k_{12} + b_1k_{11} + ak_{22};\\ \dot{k}_{22} &= 2\left(b_1k_{12} + bk_{22}\right) + \nu_2\left(c_{21}m_1 + c_{22}m_2 + c_{20}\right)^2 +\\ &\qquad + \nu_2\left(c_{21}^2k_{11} + 2c_{22}c_{21}k_{12} + c_{22}^2k_{22}\right);\\ \dot{R}_t &= 2a_1R_t - \kappa_{11}^{-1}b_1^2R_t^2 + \nu_1\left(c_{11}m_1 + c_{12}m_2 + c_{10}\right)^2 +\\ &\qquad + \nu_1\left(c_{11}^2k_{11} + 2c_{11}c_{12}k_{12} + c_{12}^2k_{22}\right).\end{aligned}\right\} \tag{9}$$

В частном случае, когда параметрические шумы отсутствуют ($c_{12} = 0$, $c_{11} = c_{12} = c_{21} = c_{22} = 0$), а матрицы $a = 0$ и $b = 0$, соотношения (3), (8) и (9) упрощаются и принимают следующий вид:

$$\dot{\hat{X}}_t = a_1\hat{X}_t + a_0 + \beta_t\left[\dot{Y}_t - \left(b_1\hat{X}_t + b_0\right)\right];$$
$$\dot{R}_t = 2a_1R_t - \kappa_{11}^{-1}b_1^2R_t^2 + \nu_1c_{10}^2;$$
$$\dot{m}_1 = a_1m_1 + a_0;\ \ \dot{m}_2 = b_1m_1 + b_0;$$
$$\dot{k}_{11} = 2ak_{11} + \nu_1c_{10}^2;$$
$$\dot{k}_{12} = a_1k_{12} + b_1k_{11};$$

$$\dot{k}_{22} = 2b_1 k_{12} + \nu_2 c_{20}^2;$$
$$\kappa_{11} = \nu_2 c_{20}^2;$$
$$\beta_t = \kappa_{11}^{-1} b_1 R_t.$$

2.2. Экстраполяция. Рассмотрим задачу условно-оптимальной экстраполяции состояния системы с параметрическими шумами для независимых винеровских шумов W_1 и W_2 в уравнениях состояния и наблюдения:

$$\left.\begin{aligned} dX_t &= (a_1 X_t + a_0)\,dt + \left(c_{10} + \sum_{r=1}^{n_x} c_{1,n_y+r} X_r\right) dW_1; \\ dY_t &= (bY_t + b_1 X_t + b_0)\,dt + \left(c_{20} + \sum_{r=1}^{n_y} c_{2r} Y_r + \sum_{r=1}^{n_x} c_{2,n_y+r} X_r\right) dW_2, \end{aligned}\right\} \quad (10)$$

где $W_1 = W_1(t)$ и $W_2 = W_2(t)$ – независимые процессы с независимыми приращениями, приняв уравнение экстраполятора Пугачёва в следующем виде:

$$d\hat{X}_t = \left[a_1(t+\Delta)\hat{X}_t + a_0(t+\Delta)\right] dt + \\ + \beta_t \left[dY_t - (bY_t + b_1 q_t^{-1}\hat{X}_t + b_0 - b_1 q_t^{-1} h_t)\,dt\right]. \quad (11)$$

В [5, п. 6.1.3] получен следующий результат.

Теорема 2.2. *Пусть векторный СтП* $\left[X_t^{\mathrm{T}} Y_t^{\mathrm{T}}\right]^{\mathrm{T}}$ *определяется уравнениями линейной дифференциальной негауссовской СтС с параметрическими шумами* (10) *и обладает конечными одно- и двумерными моментами. Тогда уравнения непрерывного линейного экстраполятора Пугачёва имеют вид* (11). *Необходимые для вычисления* κ_{11} *по формуле* (6) *и* β_t *по формуле* (7) *моменты второго порядка можно найти из уравнений* (4) *для составного вектора* $\left[Y_t^{\mathrm{T}} X_t^{\mathrm{T}} \hat{X}_t^{\mathrm{T}}\right]^{\mathrm{T}}$. *Роль матриц* c_{1r} *и* c_{2r} *играют матрицы* $[0\, c_{1r}]$ *и* $[c_{2r}\, 0]$, *а матрица* ν — *диагональна.*

Найденный экстраполятор Пугачёва можно представить в виде последовательного соединения фильтра Пугачёва, усилителя с коэффициентом усиления $\varepsilon_t = u(t+\Delta, t)$ и сумматора, вводящего неслучайное слагаемое $h_t = h(t)$, т. е. что

$$\dot{\hat{X}}_t = q_t \hat{X}_t + h_t.$$

Здесь $\hat{X}_t$ — выходной сигнал условно-оптимального фильтра, или условно-оптимальная оценка текущего состояния системы X_t; q_t и h_t определяются из уравнений:

$$\dot{h}_t = a_0(t+\Delta) - q_t a_0 + a_1(t+\Delta) h_t;$$
$$\dot{q}_t = a_1(t+\Delta) q_t - q_t a_1.$$

Тогда уравнение условно-оптимального экстраполятора Пугачёва будет иметь вид:

$$d\hat{X}_t = q_t(a_1\hat{X}_t + a_0)\,dt + q_t\beta_t\left[dY_t - \left(bY_t + b_1\hat{X}_t + b_0\right)dt\right] =$$
$$= \left[a_1(t+\Delta)q_t\hat{X}_t - q_t a_1\hat{X}_t + a_0(t+\Delta) - q_t a_0 + a_1(t+\Delta)h_t\right]dt =$$
$$= \left[a_1(t+\Delta)\left(q_t\hat{X}_t + h_t\right) + a_0(t+\Delta)\right]dt + q_t\beta_t\left[dY_t - \left(b\hat{X}_t + b_0\right)dt\right].$$

В условиях теоремы 2.2 точность экстраполяции вычисляется путем интегрирования следующего уравнения:

$$\dot{R}_t = a_1(t+\Delta)R_t + R_t a_1(t+\Delta)^{\mathrm{T}} -$$
$$- \beta_t\left[\left(c_{20} + \sum_{r=1}^{n_y+n_x} c_{2r}m_r\right)\nu_1\left(c_{20}^{\mathrm{T}} + \sum_{r=1}^{n_y+n_x} c_{2r}^{\mathrm{T}}m_r\right) + \right.$$
$$\left. + \sum_{r=1}^{n_y+n_x} c_{2r}\nu_1 c_{2s}^{\mathrm{T}}k_{rs}\right]\beta_t^{\mathrm{T}} + \left[c_{10}(t+\Delta) + \sum_{r=n_y+1}^{n_y+n_x} c_{1r}(t+\Delta)m_r(t+\Delta)\right]\nu_2(t+\Delta)\times$$
$$\times\left[c_{10}(t+\Delta)^{\mathrm{T}} + \sum_{r=n_y+1}^{n_y+1} c_{1r}(t+\Delta)^{\mathrm{T}}m_r(t+\Delta)\right] +$$
$$+ \sum_{s=m+1}^{n_y+n_x} c_{1r}(t+\Delta)\nu_2(t+\Delta)c_{1s}(t+\Delta)^{\mathrm{T}}k_{rs}.$$

Пример 2.2. В задаче примера 2.1 при $a = a_0 = c_{12} = 0$ и постоянном a_1 условно-оптимальный экстраполятор представляет собой последовательное соединение условно-оптимального фильтра и усилителя с коэффициентом усиления $q = e^{a_1\Delta}$.

3 Условно-оптимальная линейная фильтрация и экстраполяция в вольтерровских системах

3.1. Общий случай. Обобщая [4], будем основываться на следующих исходных стохастических уравнениях Ито для наблюдений (2) и состояния:

$$\dot{X} = A(t,X,Y,V) = A_0(t,V) + A_1(t,X,Y,V)X + A_2(t,X,Y,V)Y\,,$$
$$X(t_0) = X_0. \quad (12)$$

Здесь $V = \dot{W}$, W — n_w-мерный СтП с независимыми приращениями. Следуя [2, 3], во-первых, будем считать функции A_j $(j = 0, 1, 2)$ линейными относительно стохастических возмущений V, положив

$$A_j = A_{j0} + A_{j1}V \ \ (j = 0, 1, 2), \tag{13}$$

во-вторых, предположим, что A_{j0} и A_{j1} допускают статистическую линеаризацию по совокупности переменных X и Y:

$$A_{j0} \approx A_{j00} + A^X_{j01}X + A^Y_{j01}Y; \ \ A_{j1} \approx A_{j10} + A^X_{j11}X + A^Y_{j11}Y\,, \tag{14}$$

где функции A_{j00}, A_{j10}, A^X_{j01}, A^X_{j11}, A^Y_{j01} и A^Y_{j11} в общем случае зависят от математических ожиданий m^X и m^Y и ковариационных и взаимных ковариационных матриц K^X, K^Y и R^{XY}. В результате уравнение (12) приближенно можно будет привести к виду (1). Таким образом, в случае широкополосных ВСтС при условиях (13) и (14) приближенные условно-оптимальные фильтры Пугачёва будут определяться теоремой 2.1, а экстраполяторы Пугачёва — теоремой 2.2.

Замечание 3.1. В случае узкополосных ВСтС целесообразно использовать метод канонических разложений совместно с методом нормальных координат [1, 4, 6].

3.2. Стохастическая система Вольтерры–Лотки. В случае системы Вольтерры–Лотки [7, 8] с аддитивными шумами и независимыми от наблюдений правыми частями в уравнении состояния в силу [2] имеем:

$$\dot{X}_i = \varepsilon_{0i}X_i - \sum_{j=1}^{N} p_{0ij}X_iX_j + \alpha_{0i} + \sigma^\alpha_i V, \ \ X_i(t_0) = X_{i0} \ \ \left(i = \overline{1, N}\right). \tag{15}$$

Заменим (15) приближенным по МСЛ, положив

$$X_iX_j \approx \left(K^X_{ij} - m^X_i m^X_j\right) + m^X_j X_i + m^X_i X_j.$$

Тогда (15) можно будет приближенно представить в векторной форме:

$$\dot{X} = a_0 + a_1X + c_{10}, \ \ X\,(t_0) = X_0.$$

Здесь введены обозначения

$$a_0 = \left[-\sum_{j=1}^{N} p_{0ij}(K^X_{ij} - m^X_i m^X_j)\right]_{i=\overline{1,N}};$$

$$[a_1X]_{i=\overline{1,N}} = \left[\varepsilon_{0i} - \left(\sum_{j=1}^{N} p_{0ij}m_j\right) - m_i\sum_{j=1}^{N} p_{0ij}X_j\right]_{i=\overline{1,N}};$$

$$c_{10} = [\sigma^\alpha_i]_{i=\overline{1,N}}\,.$$

Тогда, очевидно, для линейных наблюдений с аддитивным шумом вида

$$\dot{Y} = bY + b_1 X + b_0 + c_{20} V$$

линейный условно-оптимальный фильтр Пугачёва будет совпадать с субоптимальным фильтром на основе МСЛ.

Таким образом, для белого шума V интенсивности ν искомые фильтрационные уравнения (теорема 2.1) будут иметь вид:

$$\dot{\hat{X}} = a_1 \hat{X} + a_0 + \beta \left[\dot{Y} - \left(bY + b_1 \hat{X} + b_0 \right) \right]; \tag{16}$$

$$\left.\begin{aligned} &\beta = \left(R b_1^{\mathrm{T}} + c_{10} \nu c_{20}^{\mathrm{T}} \right) \kappa_{11}^{-1}; \\ &\dot{R} = a_1 R + R a_1^{\mathrm{T}} - \left(R b_1^{\mathrm{T}} + c_{10} \nu c_{20}^{\mathrm{T}} \right) \kappa_{11}^{-1} \left(b_1 R + c_{20} \nu c_{10}^{\mathrm{T}} \right) + c_{10} \nu c_{10}^{\mathrm{T}} \end{aligned}\right\} \tag{17}$$

при условии

$$\kappa_{11} = c_{20} \nu c_{20}^{\mathrm{T}} \neq 0 \,. \tag{18}$$

В случае, если уравнения состояния и наблюдения содержат члены aY и bY, фильтрационные уравнения теоремы 2.1 упрощаются и принимают вид:

$$\dot{\hat{X}} = \left(aY + a_1 \hat{X} + a_0 \right) + \beta \left[\dot{Y} - \left(bY + b_1 \hat{X} + b_0 \right) \right]. \tag{19}$$

При этом уравнения (17) и (18) сохраняются.

3.3. Эквивалентность фильтров. Фильтры пп. 3.2 и 3.3 могут быть использованы для приближенного сведения ВСтС с параметрическими шумами к эквивалентным ВСтС с аддитивными шумами и оценки методической и инструментальной точности алгоритмов.

В нестационарном случае в качестве критериев эквивалентности для СтС теоремы 2.1 могут быть использованы соотношения (16)–(18) или (17)–(19). При этом уравнение наблюдения, как привило, принимают в простейшем линейном виде без параметрических шумов.

В стационарном случае вместо дисперсий и ковариаций используются соответствующие спектральные характеристики [5].

Более того, согласно (18) можно сделать замену негауссовского шума V эквивалентным гауссовским (нормальным) шумом, положив $c_{20} \nu c_{20}^{\mathrm{T}} = c'_{20} \nu_0 c_{20}^{\prime \mathrm{T}}$.

4 Условно-оптимальная линейная фильтрация и экстраполяция в одномерных вольтерровских стохастических системах

Рассмотрим сначала случай поляризованных шумов в уравнениях состояния и наблюдения, положив в основу следующие уравнения:

$$\dot{X} = \alpha_o + \sigma_0^{\alpha} V_1 + \left[\left(\varepsilon_0 + \sigma_0^{\varepsilon} V_2 \right) - \left(\gamma_0 + \sigma_0^{\gamma} V_3 \right) F(X) \right] X, \quad X(t_0) = X_0 \,, \tag{20}$$

где V_1 — нормальный белый шум интенсивности ν_1, $V_2 = V_3 = V_1$; $F(X)$ — функция взаимодействия (в дальнейшем примем $F(Y) = Y$). Положим, что уравнение состояния (20) не зависит от наблюдения, уравнение наблюдений линейно, а нормальный белый шум V_2 интенсивности ν_2 входит аддитивно:

$$\dot{Y} = b_0 + b_1 X + c_{20} V_2 \quad (c_{20} \neq 0)\,. \tag{21}$$

Пользуясь вторым МНА [2, 3], заменим $F(Y)Y = X^2$ в (20) статистическим линеаризованным выражением:

$$X^2 \approx \left(D_X - m_X^2\right) + 2 m_X X\,,$$

тогда эквивалентное уравнение состояние (20) будет иметь следующий вид:

$$\dot{X} = a_0 + a_1 X + (c_{10} + c_{11} X)\, V\,, \quad X(t_0) = X_0\,. \tag{22}$$

Здесь введены обозначения:

$$\left.\begin{array}{c} a_0 = \alpha_0 - \gamma_0 \left(D_X - m_X^2\right);\ \ a_1 = \varepsilon_0 - 2\gamma_0 m_X; \\ c_{10} = \sigma_0^{\alpha};\quad c_{11} = \sigma_0^{\varepsilon};\quad c_{21} = 0\,. \end{array}\right\} \tag{23}$$

При $\nu_2 \neq 0$, применяя к (21) и (22) соотношения примера 2.1 при условиях (23), получим искомые фильтрационные уравнения (8) и (9):

$$\begin{aligned}
\dot{\hat{X}} &= a_1 \hat{X} + a_0 + \beta \left[\dot{Y} - \left(b_1 \hat{X} + b_0\right)\right]; \\
\beta &= \kappa_{11}^{-1} b_1 R\,; \\
\kappa_{11} &= \nu_2 \left[c_{21}^2 D_X + c_{20} + c_{21} m_X^2\right]; \\
\dot{m}_X &= a_1 m_X + a_0;\ \ \dot{m}_Y = b_1 m_X + b_0; \\
\dot{R} &= 2a_1 R - \kappa_{11}^{-1} b_1^2 R^2 + \nu_1 \left[c_{11}^2 D_X + (c_{11} m_X + c_{10}^2)\right]; \\
\dot{D}_X &= 2a_1 D_X + \nu_1 \left[c_{11}^2 D_X + (c_{11} m_X^2 + c_{10})^2\right]; \\
\dot{D}_Y &= 2b_1 K_{XY} + \nu_2 \left[c_{21}^2 + (c_{21} m_X + c_{20})^2\right]; \\
\dot{K}_{XY} &= a_1 K_{XY} + b_1 D_X.
\end{aligned}$$

Частные случаи условий (23):

(1) аддитивные шумы V_1 и V_2:

$$a_0 = \alpha_0 - \gamma_0 \left(D_X - m_X^2\right);\ \ a_1 = \varepsilon_0 - 2\gamma_0 m_X;\ \ c_{10} = \sigma_0^{\alpha};\ \ \sigma_0^{\gamma} = 0;\ \ \sigma = 0;$$

(2) ε-параметрический шум V_1 и аддитивный шум V_2:

$$a_0 = \alpha_0 - \gamma_0 \left(D_X - m_X^2\right), \;\; a_1 = \varepsilon_0 - 2\gamma_0 m_X, \;\; c_{10} = \sigma_0^{\alpha} = 0, \;\; c_{11} = 0;$$

(3) γ-параметрический шум V_1 и аддитивный шум V_2:

$$a_0 = \alpha_0 - \gamma_0 \left(D_X - m_X^2\right); \;\; a_1 = \varepsilon_0 - 2\gamma_0 m_X; \;\; c_{10} = \sigma_0^{\alpha} = 0; \;\; c_{11} = -2m_X\sigma_0^{\gamma}.$$

Согласно примеру 2.2, условно-оптимальный экстраполятор задачи представляет собой последовательное соединение условно-оптимального фильтра и усилителя с коэффициентом $q = e^{a_1\Delta}$.

5 Заключение

Разработаны два приближенных метода для фильтрации и экстраполяции нормальных процессов в ВСтС, приводимых к СтС с аддитивными и параметрическими шумами на основе теории условно-оптимального оценивания Пугачёва и подходах [2–4].

Сформулированы подходы к синтезу фильтров Пугачёва для ВСтС с аддитивными и параметрическими шумами путем эквивалентной замены ВСтС с аддитивными шумами. Подробно исследованы одномерные тестовые задачи.

Результаты допускают непосредственное обобщение на случай автокоррелированных шумов и эредитарных ВСтС, а также дискретных и непрерывно-дискретных ВСтС. Важное значение имеют нелинейные задачи оценивания и идентификации моделей ВСтС.

Литература

1. *Пугачёв В. С., Синицын И. Н.* Теория стохастических систем. — М.: Логос, 2000; 2004. 1000 с.
2. *Синицын И. Н., Синицын В. И.* Аналитическое моделирование нормальных процессов в вольтерровских стохастических системах // Системы и средства информатики, 2018. Т. 28. № 2. С. 4–19.
3. *Синицын И. Н., Синицын В. И.* Аналитическое моделирование распределений с инвариантной мерой в вольтерровских стохастических системах // Системы и средства информатики, 2018. Т. 28. № 3. С. 4–25.
4. *Синицын И. Н., Синицын В. И.* Аналитическое моделирование процессов в вольтерровских стохастических системах методом канонических разложений // Системы и средства информатики, 2019. Т. 29. № 1. С. 109–127.
5. *Синицын И. Н.* Фильтры Калмана и Пугачева. — 2-е изд., перераб. и доп. — М.: Логос, 2007. 776 с.
6. *Синицын И. Н.* Канонические представления случайных функций и их применение в задачах компьютерной поддержки научных исследований. — М.: ТОРУС ПРЕСС, 2009. 768 с.

7. *Вольтерра В.* Математическая теория борьбы за существование. — М.: Наука, 1976. 286 с.
8. *Свирежев Ю. М.* Нелинейные волны, диссипативные структуры и катастрофы в экологии. — М.: Наука, 1987. 368 с.

Поступила в редакцию 11.02.19

CONDITIONALLY OPTIMAL LINEAR ESTIMATION OF NORMAL PROCESSES IN VOLTERRA STOCHASTIC SYSTEMS

I. N. Sinitsyn and V. I. Sinitsyn

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: On the basis of Pugachev's conditionally optimal estimation (filtering and extrapolation) and previous investigations of the present authors, two estimation approximate conditionally optimal methods for normal stochastic processes in Volterra stochastic systems (VStS) reducible to linear StS with additive and parametric noises are developed. Some approaches for synthesis of Pugachev's filters and extrapolators by replacing parametric noises with equivalent corresponding additive noises are given. Test examples for one-dimensional VStS are presented. The given theory and test examples may be simply generalized to VStS with autocorrelated noises and VStS with hereditary and nonlinear interaction functions.

Keywords: Volterra stochastic systems (VStS); method of analytical modeling (MAM); method of canonical expansions (MCE); method of normal approximation (MNA); method of statistical linearization (MSL); stochastic system (StS); stochastic process (StP); Pugachev conditionally optimal filters and extrapolators; Kalman filters and extrapolators

DOI: 10.14357/08696527190302

Acknowledgments

The work was supported by the Russian Academy of Sciences (project 0063-2018-0008).

References

1. Pugachev, V. S., and I. N. Sinitsyn. 2000, 2004. Teoriya stokhasticheskikh sistem [Stochastic systems: Theory and applications]. Moscow: Logos. 1000 p.
2. Sinitsyn, I. N., and V. I. Sinitsyn. 2018. Analiticheskoe modelirovanie normal'nykh protsessov v vol'terrovskikh stokhasticheskikh sistemakh [Analytical modeling of normal processes in Volterra stochastic systems]. *Sistemy i Sredstva Informatiki — Systems and Means of Informatics* 28(2):4–19.

3. Sinitsyn, I. N., and V. I. Sinitsyn. 2018. Analiticheskoe modelirovanie raspredeleniy s invariantnoy meroy v vol'terrovskikh stokhasticheskikh sistemakh [Analytical modeling of distributions with invariant measure in Volterra stochastic systems]. *Sistemy i Sredstva Informatiki — Systems and Means of Informatics* 28(3):4–25.
4. Sinitsyn, I. N., and V. I. Sinitsyn. 2019. Analiticheskoe modelirovanie protsessov v vol'terrovskikh stokhasticheskikh sistemakh metodom kanonicheskikh razlozheniy. *Sistemy i Sredstva Informatiki — Systems and Means of Informatics* 29(1):109–127.
5. Sinitsyn, I. N. 2007. *Fil'try Kalmana i Pugacheva.* 2nd ed. Moscow: Logos. 776 p.
6. Sinitsyn, I. N. 2009. *Kanonicheskie predstavleniya sluchaynykh funktsiy i ikh primenenie v zadachakh komp'yuternoy podderzhki nauchnykh issledovaniy* [Canonical expansions of random functions and their applications in computer aided support]. Moscow: TORUS PRESS. 768 p.
7. Vol'terra, V. 1976. *Matematicheskaya teoriya bor'by za sushchestvovanie* [Mathematical theory of survival]. Moscow: Nauka. 286 p.
8. Svirezhev, Yu. M. 1987. *Nelineynye volny, dissipativnye struktury i katastrofy v ekologii* [Nonlinear waves, dissipative structures and catastrophes]. Moscow: Nauka. 368 p.

Received February 11, 2019

Contributors

Sinitsyn Igor N. (b. 1940) — Doctor of Science in technology, professor, Honored scientist of RF, principal 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; sinitsin@dol.ru

Sinitsyn Vladimir I. (b. 1968) — Doctor of Science in physics and mathematics, associate professor, principal 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; VSinitsin@ipiran.ru

ИНДЕКС ПРЕИМУЩЕСТВА В БАЙЕСОВСКИХ МОДЕЛЯХ НАДЕЖНОСТИ И БАЛАНСА С БЕТА-ПОЛИНОМИАЛЬНЫМИ АПРИОРНЫМИ ПЛОТНОСТЯМИ*

***А. А. Кудрявцев*[1], *С. И. Палионная*[2], *О. В. Шестаков*[3]**

Аннотация: Работа посвящена исследованию вероятностных характеристик индекса преимущества в байесовских моделях баланса в случае, когда негативные и позитивные факторы, влияющие на функционирование системы, имеют априорные бета-распределение и распределение с плотностью полиномиального вида, например равномерное или параболическое. Результаты работы могут применяться для исследования предельной надежности сложных модифицируемых информационно-коммуникационных систем и других индексов преимущества, например коэффициента готовности и вероятности пребывания в работоспособном состоянии в теории надежности, вероятности того, что вызов не будет потерян, в теории массового обслуживания и пр. Приводимый метод может использоваться для аналогичных постановок задач при исследовании распределений с кусочно-полиномиальными априорными плотностями, например распределения Симпсона, Ирвина–Холла, Бэйтса и др.

Ключевые слова: байесовский метод; смешанные распределения; модели баланса; индекс преимущества; повышение надежности; бета-распределение

DOI: 10.14357/08696527190303

1 Введение

Задача прогнозирования надежности сложных модифицируемых информационных систем (СМИС) была сформулирована и подробно рассмотрена в [1, 2]. При этом в рамках многих задач [3] возникает необходимость использования байесовского подхода к анализу роста надежности СМИС, таких как программная система для компьютера, информационно-вычислительная сеть или административно-информационная система, которые, как правило, изначально не обладают требуемой надежностью. Стоит также учитывать, что надежность системы может

*Работа выполнена при частичной финансовой поддержке РФФИ (проект 17-07-00577).

[1]Московский государственный университет им. М. В. Ломоносова, факультет вычислительной математики и кибернетики, nubigena@mail.ru

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

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

меняться и во время функционирования. Такие изменения прежде всего связаны с модификациями системы, которые могут как увеличить, так и уменьшить надежность. Рассмотрим описанную в книге [2] модель роста надежности, позволяющую иметь дело с параметром, непосредственно интерпретируемым как надежность системы.

Пусть на вход произвольной системы подаются некоторые сигналы (например, команды оператора или внешние воздействия). Реакция системы на поданные сигналы может быть либо правильной (корректной), либо неправильной (некорректной). В каждый момент времени t надежность системы можно характеризовать параметром $p(t)$ — вероятностью того, что на сигнал, поданный на вход системы в момент t, система отреагирует правильно. По смыслу такая характеристика надежности ближе всего к традиционно используемому коэффициенту готовности. В случайные моменты времени $0 = Y_0 \le Y_1 \le Y_2 \le \cdots$ система подвергается (мгновенной) модификации, в результате чего изменяется параметр $p(t)$. Предположим, что траектории процесса $p(t)$ непрерывны справа и кусочно-постоянны, так что $p(t) = p(Y_j)$ при $Y_j \le t < Y_{j+1}$.

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

Задав начальную надежность p_0, рассмотрим модель, определяемую рекуррентным соотношением

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

В такой модели случайные величины η_j (параметры «дефективности») описывают возможное уменьшение надежности из-за некачественных модификаций, в ходе которых вместо исправления существующих дефектов в систему могут быть внесены новые, в то время как величины θ_j (параметры «эффективности») описывают повышение надежности за счет исправления дефектов.

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

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

Изучение предельного значения средней величины $\mathsf{E}p_j$ представляет значительный интерес, поскольку эта величина характеризует асимптотическое значение надежности системы в рамках некоторой рекуррентной модели, задаваемой набором $\{(\theta_j, \eta_j)\}$.

В [3] подробно описаны предпосылки для рандомизации средних параметров «эффективности» / «дефективности», связанные с возможностью рассмотрения больших групп СМИС, и обосновано применение к исследованию средней предельной надежности совокупности СМИС байесовского метода. В [4] байесовский подход к задачам теории надежности и массового обслуживания был обобщен на

ряд всевозможных постановок, в основе которых лежит разделение факторов, влияющих на функционирование системы на условно позитивные (p-факторы) и условно негативные (n-факторы). При этом эффективность системы, очевидно, зависит от индекса баланса — отношения негативного фактора к позитивному, а надежность системы определяется индексом преимущества — отношением позитивного фактора к сумме негативного и позитивного. Изменчивость и неопределенность среды, в которой происходит функционирование системы, обусловливает рандомизацию p-фактора μ и n-фактора λ, а следовательно, индекса баланса $\rho = \lambda/\mu$ и индекса преимущества $\pi = \mu/(\mu+\lambda) = 1/(1+\rho)$.

Далее приводятся вероятностные характеристики (плотность и моменты) индекса преимущества в модели, факторы которой имеют априорные бета-распределение и распределение с плотностью полиномиального вида.

2 Основные результаты

Рассмотрим случайную величину ξ с распределением, сосредоточенным на отрезке $[a_\xi, b_\xi]$, и плотностью, которую можно представить в виде полинома:

$$f_\xi(x) = \sum_{i=0}^{n_\xi} c_{\xi,i} x^i\,,\ \ 0 < a_\xi < x < b_\xi\,. \tag{1}$$

К распределениям с полиномиальными плотностями относятся, например, равномерное распределение

$$n_\xi = 0\,,\quad c_{\xi,0} = (b_\xi - a_\xi)^{-1}$$

и параболическое распределение

$$n_\xi = 2,\ \ c_{\xi,0} = -\frac{6a_\xi b_\xi}{(b_\xi - a_\xi)^3},\ \ c_{\xi,1} = \frac{6(a_\xi + b_\xi)}{(b_\xi - a_\xi)^3},\ \ c_{\xi,2} = -\frac{6}{(b_\xi - a_\xi)^3}.$$

Обозначим неполную бета-функцию через

$$B_x(p,q) = \int_0^x t^{p-1}(1-t)^{q-1}\,dt\,,\ \ 0 < x < 1\,,\ \ p > 0\,,\ \ q > 0\,.$$

Для удобства записи будем обозначать бета-функцию через $B_1(p,q)$.

Рассмотрим случайную величину η, имеющую бета-распределение с параметрами $\alpha > 0$, $\beta > 0$ и плотностью

$$f_\eta(x) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B_1(\alpha,\beta)}\,,\ \ x \in (0,1)\,. \tag{2}$$

Через

$$(\alpha)_i = \alpha(\alpha+1)\cdots(\alpha+i-1)\,,\ \ (\alpha)_0 = 1\,,\ \ \alpha \in \mathbb{R}\,,$$

будем обозначать символ Похгаммера.

Рассмотрим классическую гипергеометрическую функцию Гаусса

$${}_2F_1\left[\begin{matrix}\alpha,\beta\\ \gamma\end{matrix}\right](x) = \sum_{i=0}^{\infty}\frac{(\alpha)_i(\beta)_i}{(\gamma)_i\, i!}x^i\,.$$

По аналогии с 7.2.3 и 7.2.4 из [5] введем в рассмотрение обобщенную гипергеометрическую функцию двух переменных

$${}_s^pG_t^q\left[\begin{matrix}\alpha;\,\beta_1,\,\ldots,\,\beta_p;\,\beta_1',\,\ldots,\,\beta_q'\\ \gamma;\,\delta_1,\,\ldots,\,\delta_s;\,\delta_1',\,\ldots,\,\delta_t'\end{matrix}\right](x,y) = \\ = \sum_{i=0}^{\infty}\sum_{j=0}^{\infty}\frac{(\alpha)_{i+j}(\beta_1)_i\cdots(\beta_p)_i(\beta_1')_j\cdots(\beta_q')_j}{(\gamma)_{i+j}(\delta_1)_i\cdots(\delta_s)_i(\delta_1')_j\cdots(\delta_t')_j}\,\frac{x^iy^j}{i!j!}.$$

Обозначим

$${}_s^p\Delta_t^q\left[\begin{matrix}\alpha;\,\beta_1,\,\ldots,\,\beta_p;\,\beta_1',\,\ldots,\,\beta_q'\\ \gamma;\,\delta_1,\,\ldots,\,\delta_s;\,\delta_1',\,\ldots,\,\delta_t'\end{matrix}\right]\left(\begin{matrix}m,n,\\ x,y\end{matrix}\right) = \\ = B_1(m,n)\frac{x^{\alpha-\beta_1}}{\alpha}\cdot{}_s^pG_t^q\left[\begin{matrix}\alpha;\,\beta_1,\,\ldots,\,\beta_p;\,\beta_1',\,\ldots,\,\beta_q'\\ \gamma;\,\delta_1,\,\ldots,\,\delta_s;\,\delta_1',\,\ldots,\,\delta_t'\end{matrix}\right](1,-x) - \\ - B_1(m,n)\frac{y^{\alpha-\beta_1}}{\alpha}\cdot{}_s^pG_t^q\left[\begin{matrix}\alpha;\,\beta_1,\,\ldots,\,\beta_p;\,\beta_1',\,\ldots,\,\beta_q'\\ \gamma;\,\delta_1,\,\ldots,\,\delta_s;\,\delta_1',\,\ldots,\,\delta_t'\end{matrix}\right](1,-y).$$

Ранее были получены следующие результаты [6].

Лемма 1. *Пусть n-фактор λ имеет плотность полиномиального вида на отрезке $[a_\lambda, b_\lambda]$, $0 < a_\lambda < b_\lambda$, а p-фактор μ имеет бета-распределение с параметрами $k > 0$, $l > 0$, причем λ и μ независимы. Тогда индекс баланса $\rho = \lambda/\mu$ имеет плотность*:

$$f_\rho(x) = \sum_{i=0}^{n_\lambda}\frac{c_{\lambda,i}x^i}{B_1(k,l)}\left[B_{\min\{b_\lambda/x,\,1\}}(k+i+1,l) - B_{a_\lambda/x}(k+i+1,l)\right],\ \ x \geq a_\lambda\,.$$

Лемма 2. *Пусть n-фактор λ имеет бета-распределение с параметрами $m > 0$, $n > 0$, а p-фактор μ имеет плотность полиномиального вида на отрезке*

$[a_\mu, b_\mu]$, $0 < a_\mu < b_\mu$, *причем* λ *и* μ *независимы. Тогда индекс баланса* $\rho = \lambda/\mu$ *имеет плотность*:

$$f_\rho(x) = \sum_{i=0}^{n_\mu} \frac{c_{\mu,i} x^{-i-2}}{B_1(m,n)} \left[B_{\min\{b_\mu x,\,1\}}(m+i+1, n) - B_{a_\mu x}(m+i+1, n) \right],$$

$$0 < x \le \frac{1}{a_\mu}.$$

Леммы 1 и 2 позволяют найти плотность индекса преимущества. Приведем основные результаты для вероятностных характеристик индекса преимущества в случае, когда один из факторов имеет распределение с плотностью полиномиального вида (1), а второй — бета-распределение (2).

Теорема 1. *Пусть n-фактор* λ *имеет плотность полиномиального вида на отрезке* $[a_\lambda, b_\lambda]$, $0 < a_\lambda < b_\lambda$, *а p-фактор* μ *имеет бета-распределение с параметрами* $k > 0$, $l > 0$, *причем* λ *и* μ *независимы. Тогда индекс преимущества* $\pi = \mu/(\mu + \lambda)$ *имеет плотность*

$$f_\pi(x) = \\ = \sum_{i=0}^{n_\lambda} \frac{c_{\lambda,i}(1-x)^i}{B_1(k,l) x^{i+2}} \left[B_{\min\{b_\lambda x/(1-x),\,1\}}(k+i+1, l) - B_{a_\lambda x/(1-x)}(k+i+1, l) \right],$$

$$0 < x \le \frac{1}{a_\lambda + 1} \quad (3)$$

и моменты порядка z *вида*

$$\mathsf{E}\pi^z = \sum_{i=0}^{n_\lambda} \frac{c_{\lambda,i}}{B_1(k,l)} \left({}_1^0\Delta_0^1 \left[\begin{matrix} i+1; 0; z \\ i+2 \end{matrix} \right] \left(\begin{matrix} k+i+1, \\ l, b_\lambda, a_\lambda \end{matrix} \right) + \right.$$

$$\left. + {}_1^2\Delta_0^1 \left[\begin{matrix} k+z;\, k+i+1, 1-l;\, z \\ k+z+1;\, k+i+2 \end{matrix} \right] \left(\begin{matrix} k+i+1, \\ 1, b_\lambda^{-1}, a_\lambda^{-1} \end{matrix} \right) \right). \quad (4)$$

Доказательство. Равенство (3) для плотности индекса преимущества следует из леммы 1 и соотношения

$$f_\pi(x) = \frac{1}{x^2} f_\rho\left(\frac{1-x}{x} \right).$$

Найдем моменты индекса преимущества, воспользовавшись формулой (28), из [7, с. 384]:

$$B_z(b,a) = \frac{z^b}{b} {}_2F_1 \left[\begin{matrix} 1-a, b \\ b+1 \end{matrix} \right](z).$$

Имеем:

$$\int\limits_0^{1/(b_\lambda+1)} x^z f_\pi(x)\,dx = \sum_{i=0}^{n_\lambda} \frac{c_{\lambda,i} b_\lambda^{k+i+1}}{(k+i+1)B_1(k,l)} \times$$

$$\times \int\limits_0^{1/(b_\lambda+1)} \frac{x^{z+k-1}}{(1-x)^{k+1}} \cdot {}_2F_1\left[\begin{matrix}1-l, k+i+1\\ k+i+2\end{matrix}\right]\left(\frac{b_\lambda x}{1-x}\right) dx - \sum_{i=0}^{n_\lambda} \frac{c_{\lambda,i} a_\lambda^{k+i+1}}{(k+i+1)B_1(k,l)} \times$$

$$\times \int\limits_0^{1/(b_\lambda+1)} \frac{x^{z+k-1}}{(1-x)^{k+1}} \cdot {}_2F_1\left[\begin{matrix}1-l, k+i+1\\ k+i+2\end{matrix}\right]\left(\frac{a_\lambda x}{1-x}\right) dx \equiv U_1 - U_2\,.$$

Для первого слагаемого по формуле 1.2.4.3 из [7] имеем:

$$U_1 = \sum_{i=0}^{n_\lambda} \frac{c_{\lambda,i} b_\lambda^{k+i+1}}{(k+i+1)B_1(k,l)} \sum_{u=0}^{\infty} \frac{(1-l)_u (k+i+1)_u b_\lambda^u}{(k+i+2)_u u!} \times$$

$$\times \int\limits_0^{1/(b_\lambda+1)} \frac{x^{z+k+u-1}}{(1-x)^{k+u+1}}\,dx = \sum_{i=0}^{n_\lambda} \frac{c_{\lambda,i} b_\lambda^{k+i+1}}{(k+i+1)B_1(k,l)} \sum_{u=0}^{\infty} \frac{(1-l)_u (k+i+1)_u b_\lambda^u}{(k+i+2)_u u!} \times$$

$$\times \int\limits_0^{1/b_\lambda} \frac{y^{k+u+z-1}}{(1+y)^z}\,dy = \sum_{i=0}^{n_\lambda} \frac{c_{\lambda,i} b_\lambda^{-z+i+1}}{(k+i+1)B_1(k,l)} \sum_{u=0}^{\infty} \frac{(1-l)_u (k+i+1)_u}{(k+i+2)_u (k+u+z) u!} \times$$

$$\times {}_2F_1\left[\begin{matrix}z, k+u+z\\ k+u+z+1\end{matrix}\right]\left(-\frac{1}{b_\lambda}\right) = \sum_{i=0}^{n_\lambda} \frac{c_{\lambda,i} b_\lambda^{-z+i+1}}{(k+i+1)B_1(k,l)} \times$$

$$\times \sum_{u=0}^{\infty} \frac{(1-l)_u (k+i+1)_u}{(k+i+2)_u (k+u+z) u!} \sum_{v=0}^{\infty} \frac{(z)_v (k+u+z)_v}{(k+u+z+1)_v v!} \left(-\frac{1}{b_\lambda}\right)^v =$$

$$= \sum_{i=0}^{n_\lambda} \frac{c_{\lambda,i} b_\lambda^{-z+i+1}}{(k+z)(k+i+1)B_1(k,l)} \cdot {}_1^2G_0^1\left[\begin{matrix}k+z;\ 1-l, k+i+1;\ z\\ k+z+1;\ k+i+2\end{matrix}\right]\left(1, -\frac{1}{b_\lambda}\right).$$

Аналогично

$$U_2 =$$

$$= \sum_{i=0}^{n_\lambda} \frac{c_{\lambda,i} a_\lambda^{k+i+1} b_\lambda^{-k-z}}{(k+z)(k+i+1)B_1(k,l)} \cdot {}_1^2G_0^1\left[\begin{matrix}k+z;\ 1-l, k+i+1;\ z\\ k+z+1;\ k+i+2\end{matrix}\right]\left(\frac{a_\lambda}{b_\lambda}, -\frac{1}{b_\lambda}\right).$$

Далее

$$\int\limits_{1/(b_\lambda+1)}^{1/(a_\lambda+1)} x^z f_\pi(x)\,dx = \sum_{i=0}^{n_\lambda} \frac{c_{\lambda,i} B_1(k+i+1,l)}{B_1(k,l)} \int\limits_{1/(b_\lambda+1)}^{1/(a_\lambda+1)} \frac{(1-x)^i}{x^{i+2-z}}\,dx -$$

$$-\sum_{i=0}^{n_\lambda} \frac{c_{\lambda,i} a_\lambda^{k+i+1}}{B_1(k,l)(k+i+1)} \int\limits_{1/(b_\lambda+1)}^{1/(a_\lambda+1)} \frac{x^{k+z-1}}{(1-x)^{k+1}} \cdot {}_2F_1\left[\begin{matrix} 1-l, k+i+1 \\ k+i+2 \end{matrix}\right]\left(\frac{a_\lambda x}{1-x}\right) dx \equiv$$

$$\equiv U_3 - U_4\,.$$

Для первого слагаемого имеем:

$$U_3 = \sum_{i=0}^{n_\lambda} \frac{c_{\lambda,i} B_1(k+i+1,l)}{B_1(k,l)} \int\limits_{a_\lambda}^{b_\lambda} \frac{y^i}{(1+y)^z}\,dy = \sum_{i=0}^{n_\lambda} \frac{c_{\lambda,i} B_1(k+i+1,l)}{B_1(k,l)} \times$$

$$\times \left[\frac{b_\lambda^{i+1}}{i+1} \cdot {}_2F_1\left[\begin{matrix} z, i+1 \\ i+2 \end{matrix}\right](-b_\lambda) - \frac{a_\lambda^{i+1}}{i+1} \cdot {}_2F_1\left[\begin{matrix} z, i+1 \\ i+2 \end{matrix}\right](-a_\lambda)\right].$$

Второе слагаемое вычисляется аналогично U_1 и U_2:

$$U_4 =$$

$$= \sum_{i=0}^{n_\lambda} \frac{c_{\lambda,i} a_\lambda^{-z+i+1}}{(k+z)(k+i+1)B_1(k,l)} \cdot {}^2_1G^1_0\left[\begin{matrix} k+z;\ 1-l, k+i+1;\ z \\ k+z+1;\ k+i+2 \end{matrix}\right]\left(1, -\frac{1}{a_\lambda}\right) -$$

$$-\sum_{i=0}^{n_\lambda} \frac{c_{\lambda,i} a_\lambda^{k+i+1} b_\lambda^{-k-z}}{(k+z)(k+i+1)B_1(k,l)} \cdot {}^2_1G^1_0\left[\begin{matrix} k+z;\ 1-l, k+i+1;\ z \\ k+z+1;\ k+i+2 \end{matrix}\right]\left(\frac{a_\lambda}{b_\lambda}, -\frac{1}{b_\lambda}\right).$$

Поскольку

$$\mathsf{E}\pi^z = \int\limits_0^{1/(b_\lambda+1)} x^z f_\pi(x)\,dx + \int\limits_{1/(b_\lambda+1)}^{1/(a_\lambda+1)} x^z f_\pi(x)\,dx = U_1 - U_2 + U_3 - U_4,$$

учитывая, что

$${}_2F_1\left[\begin{matrix} z, i+z+1 \\ i+z+2 \end{matrix}\right](x) = {}_2F_1\left[\begin{matrix} i+z+1, z \\ i+z+2 \end{matrix}\right](x) = {}^1_0G^1_0\left[\begin{matrix} i+z+1; 0; z \\ i+z+2 \end{matrix}\right](1, x);$$

$${}^2_1G^1_0\left[\begin{matrix} m;\ 1-n, m+i+1;\ z \\ m+1;\ m+i+2 \end{matrix}\right](x,\, y) = {}^2_1G^1_0\left[\begin{matrix} m;\ m+i+1, 1-n;\ z \\ m+1;\ m+i+2 \end{matrix}\right](x,\, y)\,,$$

получаем (4). Теорема доказана.

Полностью аналогично доказывается утверждение для симметричного случая априорных распределений факторов.

Теорема 2. *Пусть n-фактор λ имеет бета-распределение с параметрами $m > 0$ и $n > 0$, а p-фактор μ имеет плотность полиномиального вида на отрезке $[a_\mu, b_\mu]$, $0 < a_\mu < b_\mu$, причем λ и μ независимы. Тогда индекс преимущества $\pi = \mu/(\mu + \lambda)$ имеет плотность*

$$f_\pi(x) = \sum_{i=0}^{n_\mu} \frac{c_{\mu,i} x^i}{B_1(m,n)(1-x)^{i+2}} \left[B_{\min\{b_\mu(1-x)/x,\,1\}}(m+i+1, n) - \right.$$

$$\left. - B_{a_\mu(1-x)/x}(m+i+1, n) \right], \quad \frac{a_\mu}{a_\mu + 1} \le x < 1,$$

и моменты порядка z вида

$$\mathsf{E}\pi^z = \sum_{i=0}^{n_\mu} \frac{c_{\mu,i}}{B_1(m,n)} \left({}_1^0\Delta_0^1 \left[\begin{matrix} i+z+1; 0; z \\ i+z+2 \end{matrix} \right] \left(\begin{matrix} m+i+1, \\ n, b_\mu, a_\mu \end{matrix} \right) + \right.$$

$$\left. + {}_1^2\Delta_0^1 \left[\begin{matrix} m;\, m+i+1, 1-n;\, z \\ m+1;\, m+i+2 \end{matrix} \right] \left(\begin{matrix} m+i+1, \\ 1, b_\mu^{-1}, a_\mu^{-1} \end{matrix} \right) \right).$$

3 Заключение

Приведенные результаты могут применяться для отыскания средней предельной надежности СМИС (момента первого порядка индекса преимущества) и обобщают полученные ранее результаты [8] для моделей, в которых один из факторов имел априорное бета-распределение, а второй — равномерное распределение. Рассматриваемый метод может использоваться для аналогичных постановок задач при исследовании распределений с кусочно-полиномиальными априорными плотностями, например распределения Симпсона, Ирвина–Холла, Бэйтса и др.

Литература

1. *Gnedenko B. V., Korolev V. Yu.* Random summation: Limit theorems and applications. — Boca Raton, FL, USA: CRC Press, 1996. 288 p.
2. *Королев В. Ю., Соколов И. А.* Основы математической теории надежности модифицируемых систем. — М.: ИПИ РАН, 2006. 108 с.
3. *Кудрявцев А. А., Соколов И. А., Шоргин С. Я.* Байесовская рекуррентная модель роста надежности: равномерное распределение параметров // Информатика и её применения, 2013. Т. 7. Вып. 2. С. 55–59.
4. *Кудрявцев А. А.* Байесовские модели баланса // Информатика и её применения, 2018. Т. 12. Вып. 3. С. 18–27.

5. *Прудников А. П., Брычков Ю. А., Маричев О. И.* Интегралы и ряды. — В 3 т. Т. 3. Специальные функции. Дополнительные главы. — 2-е изд., испр. — М.: ФИЗМАТЛИТ, 2003. 688 с.
6. *Кудрявцев А. А., Палионная С. И., Шоргин С. Я.* Бета-полиномиальные априорные плотности в байесовских моделях надежности // Системы и средства информатики, 2018. Т. 28. Вып. 3. С. 54–61.
7. *Прудников А. П., Брычков Ю. А., Маричев О. И.* Интегралы и ряды. — В 3 т. Т. 1. Элементарные функции. — 2-е изд., испр. — М.: ФИЗМАТЛИТ, 2002. 632 с.
8. *Жаворонкова Ю. В., Кудрявцев А. А., Шоргин С. Я.* Байесовская рекуррентная модель роста надежности: бета-равномерное распределение параметров // Информатика и её применения, 2015. Т. 9. Вып. 1. С. 98–105.

Поступила в редакцию 27.07.19

ADVANTAGE INDEX IN BAYESIAN RELIABILITY AND BALANCE MODELS WITH BETA-POLYNOMIAL *A PRIORI* DENSITIES

***A. A. Kudryavtsev*[1], *S. I. Palionnaia*[1], *and O. V. Shestakov*[1,2]**

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

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

Abstract: This work is devoted to the research of the probabilistic characteristics of the advantage index in Bayesian balance models, when negative and positive factors affecting the functioning of the system have an *a priori* beta-distribution and distribution with polynomial density, for example, uniform or parabolic distribution. The results of the work can be used to research marginal reliability of complex modifiable information-communication systems and other advantage indexes, for example, availability ratio and probability of staying in working condition in reliability theory, probability that the call will not be lost, in the theory of mass service, etc. The given method can be used for similar formulations of the problems in the research of distributions with piecewise polynomial *a priori* densities, for example, Simpson distribution, Irwin–Hall distribution, Bates distribution, etc.

Keywords: Bayesian method; mixed distributions; balance models; advantage index; reliability growth; beta-distribution

DOI: 10.14357/08696527190303

Acknowledgments

The work was partly supported by the Russian Foundation for Basic Research (project 17-07-00577).

References

1. Gnedenko, B. V., and V. Yu. Korolev. 1996. *Random summation: Limit theorems and applications*. Boca Raton, FL: CRC Press. 288 p.
2. Korolev, V. Yu., and I. A. Sokolov. 2006. *Osnovy matematicheskoy teorii nadezhnosti modifitsiruemykh sistem* [Fundamentals of mathematical theory of modified systems reliability]. Moscow: IPI RAN. 108 p.
3. Kudryavtsev, A. A., I. A. Sokolov, and S. Ya. Shorgin. 2013. Bayesovskaya rekurrentnaya model' rosta nadezhnosti: ravnomernoe raspredelenie parametrov [Bayesian recurrent model of reliability growth: Uniform distribution of parameters]. *Informatika i ee Primeneniya — Inform. Appl.* 7(2):55–59.
4. Kudryavtsev, A. A. 2018. Bayesovskie modeli balansa [Bayesian balance models]. *Informatika i ee Primeneniya — Inform. Appl.* 12(3):18–27.
5. Prudnikov, A. P., Yu. A. Brychkov, and O. I. Marichev. 2003. *Integraly i ryady*. V 3 t. T. 3. Spetsial'nye funktsii. Dopolnitel'nye glavy [Integrals and series. In 3 vols. Vol. 3. Special functions. Additional chapters]. 2nd ed. Moscow: FIZMATLIT. 688 p.
6. Kudryavtsev, A. A., S. I. Palionnaia, and S. Ya. Shorgin. 2018. Beta-polinomial'nye apriornye plotnosti v bayesovskikh modelyakh nadezhnosti [Beta-polynomial *a priori* densities in Bayesian reliability models]. *Sistemy i Sredstva Informatiki — Systems and Means of Informatics* 28(3):54–61.
7. Prudnikov, A. P., Yu. A. Brychkov, and O. I. Marichev. 2002. *Integraly i ryady*. V 3 t. T. 1. Elementarnye funktsii [Integrals and series. In 3 vols. Vol. 3. Elementary functions]. 2nd ed. Moscow: FIZMATLIT. 632 p.
8. Zhavoronkova, Iu. V., A. A. Kudryavtsev, and S. Ya. Shorgin. 2015. Bayesovskaya rekurrentnaya model' rosta nadezhnosti: beta-ravnomernoe raspredelenie parametrov [Bayesian recurrent model of reliability growth: Beta-uniform distribution of parameters]. *Informatika i ee Primeneniya — Inform. Appl.* 9(1):98–105.

Received July 27, 2019

Contributors

Kudryavtsev Alexey A. (b. 1978) — Candidate of Science (PhD) in physics and mathematics, associate professor, Department of Mathematical Statistics, Faculty of Computational Mathematics and Cybernetics, M. V. Lomonosov Moscow State University, 1-52 Leninskiye Gory, GSP-1, Moscow 119991, Russian Federation; nubigena@mail.ru

Palionnaia Sofia I. (b. 1995) — student, Faculty of Computational Mathematics and Cybernetics, M. V. Lomonosov Moscow State University, 1-52 Leninskiye Gory, GSP-1, Moscow 119991, Russian Federation; sofiapalionnaya@gmail.com

Shestakov Oleg V. (b. 1976) — Doctor of Science 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; oshestakov@cs.msu.su

АППРОКСИМАЦИЯ КОЭФФИЦИЕНТА УСИЛЕНИЯ НАПРАВЛЕННОСТИ АНТЕННЫ ДЛЯ АНАЛИЗА «НАПРАВЛЕННОЙ ГЛУХОТЫ» В ТРЕХМЕРНОМ ПРОСТРАНСТВЕ*

***О. В. Чухно*[1], *Н. В. Чухно*[2], *Ю. В. Гайдамака*[3], *С. Я. Шоргин*[4]**

Аннотация: Рассматривается проблема «направленной глухоты» — ситуации, возникающей, когда устройство не может обнаружить занятый радиоканал из-за высоконаправленной линии связи между другими взаимодействующими в это время устройствами. Ситуация «глухоты» может возникнуть между устройствами с высоконаправленными антеннами, например работающими в миллиметровом диапазоне длин волн, на этапе доступа на основе конкуренции, в частности в соответствии с протоколами IEEE 802.11ad/ay. Получено аналитическое выражение для вероятности возникновения события «глухота» для нескольких вариантов расположения устройств в трехмерном пространстве (3D) и для предложенной линейной аппроксимации коэффициента усиления направленности антенны. Исследуется предложенная формула для нижней границы вероятности события «глухота» для трех реалистичных диаграмм направленности антенны (ДНА) и четырех вариантов фазированных антенных решеток.

Ключевые слова: миллиметровый диапазон длин волн; направленная глухота; 3D; направленный доступ

DOI: 10.14357/08696527190304

1 Введение

Миллиметровый диапазон длин волн, который уже активно эксплуатируется в сетях 5G, согласно последнему утвержденному консорциумом 3GPP стандарту Release 15 [1] расширяет доступную в сетях 4G полосу пропускания за счет использования нелицензированного спектра и применения многолучевых антенн с фазированными антенными решетками. Особенность высоконаправленных

*Публикация подготовлена при поддержке Программы РУДН «5-100» и при финансовой поддержке РФФИ (проекты 17-07-00845 и 18-07-00576).

[1]Российский университет дружбы народов, olgachukhno95@gmail.com

[2]Российский университет дружбы народов, nvchukhno@gmail.com

[3]Российский университет дружбы народов; Федеральный исследовательский центр «Информатика и управление» Российской академии наук, gaydamaka-yuv@rudn.ru

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

передач в этом диапазоне заключается в необходимости точной настройки приемной и передающей антенн, от которой зависит не только скорость, достигаемая в радиоканале, но и время, затраченное на установление соединения. Одной из актуальных задач здесь является исследование так называемой «*направленной глухоты*» (directional deafness) — ситуации, возникающей, когда устройство не может обнаружить занятый радиоканал из-за высоконаправленной линии связи между другими взаимодействующими в это время устройствами. Ситуация «глухоты» приводит к неоднократным безуспешным повторениям процедуры первичного случайного доступа, которые приводят к чрезмерным задержкам и значительным потерям пакетов и негативно сказываются на производительности системы. В ходе активных обсуждений [2–4] предлагались различные решения проблемы направленной глухоты, однако количественная аналитическая оценка была представлена только в [5], где вероятность возникновения события «глухота» оценена для устройств на плоскости. Между тем новые методы трехмерного формирования луча требуют расширения существующего подхода.

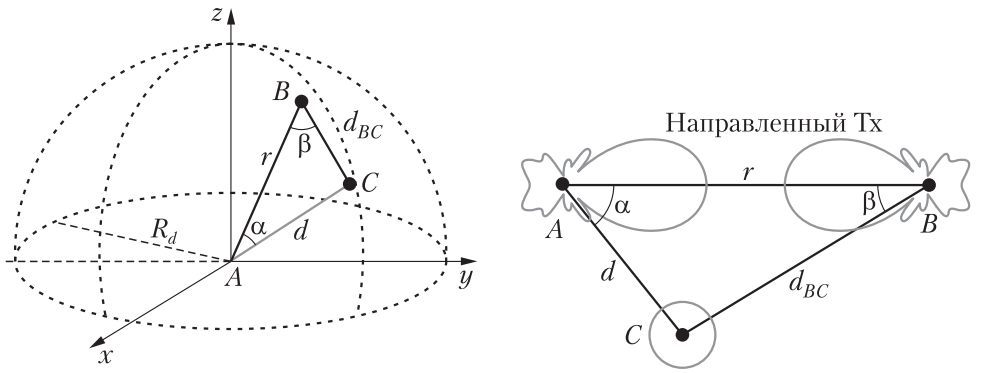
В статье представлено аналитическое решение для оценки вероятности направленной глухоты в трехмерном пространстве с использованием методов стохастической геометрии для моделирования расположения устройств и аналитически управляемой модели ДНА, задаваемой линейной функцией ширины луча по уровню половинной мощности. В результате получено выражение для нижней границы вероятности события «глухота», которое служит аппроксимацией для реалистичных ДНА и для моделирования расположения устройств точечным процессом твердого ядра Матерна.

Статья организована следующим образом. В разд. 2 изложены основные предположения, построена системная модель, включающая трехмерную ДНА, и поставлена задача исследования. В разд. 3 получено выражение для расчета вероятности события «глухота» для нескольких вариантов расположения устройств в трехмерном пространстве и для предложенной в разд. 2 линейной аппроксимации коэффициента усиления направленности антенны. Численные результаты в разд. 4 иллюстрируют применение формулы для нескольких моделей ДНА и завершаются выводами.

2 Системная модель и основные предположения

A. Пространственное расположение устройств

Система состоит из трех приемо-передающих устройств, оснащенных высоконаправленными антеннами, работающими в диапазоне миллиметровых длин волн. При этом устройство B уже обменивается данными по направленному каналу с устройством A, которое выступает, например, в роли точки доступа, а устройство C пытается установить соединение с A. Предполагается, что устройства расположены в трехмерном пространстве. Распределение распо-



(*а*) (*б*)

Рис. 1 Иллюстрация системной модели в трехмерном пространстве (*а*) и на плоскости (*б*)

ложения B — равномерное в некоторой области интереса вокруг A, которая для простоты полагается сферой радиуса R_d с центром в A, как показано на рис. 1, *а*. Для устройства C задано случайное расстояние d до точки A. Данный подход не ограничен конкретной формой области интереса или распределением местоположений устройств.

Функция $F_r(r)$ и плотность $f_r(r)$ распределения случайного расстояния r известны:

$$F_r(r) = \frac{r^3}{R_d^3}\,; \qquad f_r(r) = \frac{3r^2}{R_d^3}\,, \tag{1}$$

а распределение случайного расстояния d должно быть задано дополнительно.

Проведем плоскость через точки A, B и C (рис. 1, *б*), иллюстрируя задачу для двумерного случая. В $\triangle ABC$ обозначим α — угол между AB и AC, β — угол между AB и BC, d_{BC} — расстояние между B и C. Случайные величины r и α независимы, а β и d_{BC} могут быть выражены через r, α и d.

Б. Модель антенны

В контексте задачи анализа события «глухота» существенным является моделирование трехмерной диаграммы направленности антенны [6]. Предполагается, что ДНА симметрична относительно оси основного луча антенны, т. е. представлена телом вращения. Все устройства передают данные в направленном режиме с узкими идеально выровненными в направлении друг друга лучами. Рассмотрен ненаправленный режим приема, который соответствует, например, операции IEEE 802.11ad/ay после предварительного формирования луча на уровне сектора [7]. На рис. 1, *б* устройства A и B — передатчики Tx (transmitter), устройство C — приемник Rx (receiver), при этом соответствующие ДНА показаны серыми линиями.

Коэффициент усиления в заданном направлении α для произвольно ориентированной антенны задается формулой:

$$D(\alpha) = D_0 \rho(\alpha)\,,$$

где коэффициент D_0 отражает действие антенны в направлении оси максимального усиления, коэффициент $\rho(\alpha) \in [0, 1]$ масштабирует усиление антенны для угла $\alpha \in [0, \pi]$ отклонения от оси максимального усиления, причем $\rho(0) = 1$ соответствует направлению антенны вдоль оси максимального усиления. Коэффициент $\rho(\alpha)$ может быть оценен из результатов измерений для реализованной диаграммы направленности, рассчитан численно для конкретных настроек фазированной решетки либо аппроксимирован аналитически управляемыми моделями. Ниже показан пример линейной аппроксимации коэффициента усиления антенны, согласно которой коэффициент в направлении максимального усиления рассчитывается как отношение площади сферы к площади ДНА в виде конуса:

$$D_0 = \frac{2}{1 - \cos\theta/2}\,,$$

где θ — ширина луча по уровню половинной мощности.

Масштабирующий коэффициент для углового отклонения α от оси максимального усиления может быть аппроксимирован линейной функцией:

$$\rho(\alpha) = \begin{cases} 1 - \dfrac{\alpha}{\theta}\,, & \alpha \le \theta\,; \\ 0 & \text{иначе}\,. \end{cases} \tag{2}$$

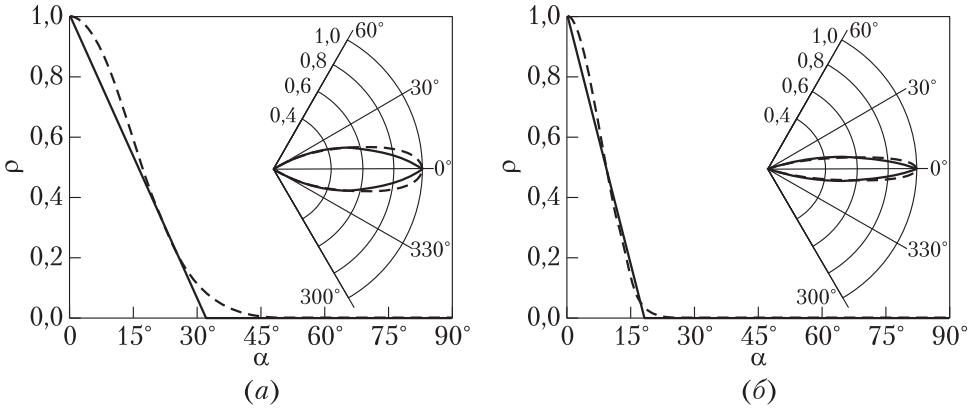


Рис. 2 Иллюстрация предложенной аппроксимации (2) (сплошные линии) для равномерного прямоугольного массива антенных элементов 4×4 (*а*) и 8×8 (*б*) (штриховые линии)

Аппроксимация (2) не учитывает боковые и задние лепестки антенны в отличие от моделей, использующих сектор и круг меньшего радиуса. Однако такие модели не могут применяться для оценки вероятности события «глухота» из-за их тривиального отсечения решений [5]. Предлагаемая аппроксимация проиллюстрирована на рис. 2, где линейная функция (2) сравнивается с реалистичными ДНА для двух фазированных антенных решеток — 4×4 и 8×8 антенных элементов.

В. Модель канала и принимаемая мощность

Принимаемая мощность может быть найдена как

$$P_{\mathrm{rx}} = P_{\mathrm{tx}} G_{\mathrm{tx}} G_{\mathrm{rx}} L^{-1}(d) = \frac{P_{\mathrm{tx}} D_0 \rho(\alpha)}{C d^{\kappa}}, \tag{3}$$

где P_{tx} — мощность передатчика; $G_{\mathrm{tx}} = D_0 \rho(\alpha)$ — коэффициент усиления передающей антенны; $G_{\mathrm{rx}} = 1$ — коэффициент усиления принимающей антенны в режиме ненаправленного приема; d — расстояние между устройствами; C — постоянная распространения волны; κ — экспонента распространения. Параметры C и κ могут быть получены из результатов соответствующих измерений или в простейшем случае прямой видимости взяты из уравнения Фрииса [8] как $\kappa = 2$, $C = (4\pi/\lambda)^2$, где λ — длина волны.

Также стандарты определяют нижний порог P_{thr} принимаемой мощности для протокола физического уровня [7], которая необходима для установления соединения или обнаружения занятого канала между другими взаимодействующими устройствами. Тогда радиус R зоны покрытия вокруг устройства может быть получен из (3) путем замены $d = R$:

$$R = \left[\frac{P_{\mathrm{thr}} D_0 \lambda^2}{(4\pi)^2} \right]^{1/\kappa}.$$

Г. «Направленная глухота»

Предполагается, что устройства A и B установили соединение и активно обмениваются данными во время выделенного периода доступа на основе конкуренции [7], когда устройство C пытается подключиться к A, проверяя наличие свободного канала. В этом случае возможны два исхода.

1. Устройство C обнаруживает сигнал от любого из устройств (т. е. $P_{\mathrm{rx},A} \geq$ $\geq P_{\mathrm{thr}}$ или $P_{\mathrm{rx},B} \geq P_{\mathrm{thr}}$). Тогда устройство C устанавливает таймер и ожидает завершения соединения между A и B, затем возобновляет попытки подключиться к A.

2. Устройство C не обнаруживает сигнала ни от одного из устройств A и B, т. е. $P_{\mathrm{rx},A}$ и $P_{\mathrm{rx},B}$ не превышают порога P_{thr}. В этом случае устройство C продолжает регулярно отправлять запросы через определенные промежутки времени. Ситуация, когда $P_{\mathrm{rx},A} < P_{\mathrm{thr}}$ и $P_{\mathrm{rx},B} < P_{\mathrm{thr}}$, носит название «глухота».

Таким образом, задача исследования состоит в определении вероятности события «глухота», причем наиболее интересен с точки зрения анализа производительности сети случай произвольного расположения устройств.

3 Вероятность события «глухота»

Ниже в виде теорем без доказательств приведены формулы для расчета вероятности события «глухота» для различных распределений углов α и β и расстояний r и d.

Теорема 1. *Для фиксированного расстояния d между A и C и радиуса области интереса R_d вокруг A вероятность $P_D(d)$ события «глухота» может быть получена следующим образом*:

$$P_D(d) = \int_0^{\pi}\int_0^{R_d} I\left(\rho(\alpha) < \frac{d^2}{R^2},\, \rho(\beta < \frac{d_{BC}^2}{R^2}\right) f_r(r) f_\alpha(\alpha)\, dr d\alpha\,,$$

где в $\triangle ABC$ $d_{BC} = \sqrt{r^2 + d^2 - 2rd\cos\alpha}$; $\beta = \arccos\left(r - d\cos\alpha\right)/d_{BC}$.

Доказательство основано на преобразовании неравенств $P_{\mathrm{rx},A} < P_{\mathrm{thr}}$ и $P_{\mathrm{rx}} < P_{\mathrm{thr},B}$ аналогично [5] и опущено здесь для краткости.

Если устройства B и C равномерно распределены в сфере радиуса R_d с центром в A, то распределение α имеет вид [9]:

$$f_\alpha(\alpha) = \frac{1}{\sqrt{\pi}}\,\frac{\Gamma(n/2)}{\Gamma((n-1)/2)}\,\sin^{n-2}\alpha\,,\ \ \alpha \in [0,\pi]\,,$$

где n — размерность пространства. Для трехмерного пространства имеем:

$$f_\alpha(\alpha) = \frac{\sin\alpha}{2}\,,\ \ \alpha \in [0,\pi]\,. \tag{4}$$

Теорема 2. *Если B равномерно распределено в сфере с центром в A, т. е. расстояние r и угол α имеют распределения* (1) *и* (4) *соответственно, тогда вероятность $P_D(d)$ события «глухота» может быть получена следующим образом*:

$$P_D(d) = \frac{3}{2R_d^3}\int_0^{\pi}\int_0^{R_d} I\left(\rho(\alpha) < \frac{d^2}{R^2},\, \rho(\beta) < \frac{d_{BC}^2}{R^2}\right) r^2 \sin\alpha\, dr d\alpha\,. \tag{5}$$

Доказательство следует из теоремы 1.

Теорема 3. *Если B равномерно распределено в сфере с центром в A, тогда для фиксированного расстояния d нижнюю границу вероятности события «глухота» можно найти, используя линейную модель* (2) *диаграммы направленности антенны, в следующем виде*:

$$P_D\left(d|d \le R_d \sin\theta\right) = \frac{d^3 \cot\theta}{32 R_d^3 \sin^2\theta}\left[-\sin\left(6\theta - \frac{4d^2\theta}{R^2}\right) + \right.$$
$$+ 2\sin\left(4\theta - \frac{2d^2\theta}{R^2}\right) + 4\sin\left(2\theta - \frac{d^2 2\theta}{R^2}\right) + \sin\left(4\theta - \frac{d^2 4\theta}{R^2}\right) +$$
$$\left. + \sin\left(-2 + \frac{4d^2\theta}{R^2}\right) - 2\sin\left(\frac{2\theta d^2}{R^2}\right) - 12\theta + 6\pi + \frac{12\theta d^2}{R^2}\right]$$

или

$$P_D\left(d|d > R_d \sin\theta\right) = \frac{\cos\tilde{z}_1 - \cos\tilde{z}_2}{2} +$$
$$+ \frac{d^3 \cot\theta}{64 R_d^3 \sin^3\theta}\left[12\cos\theta\left[\left(\frac{d^2\theta}{R^2} - \theta + \tilde{z}_1\right) + \left(\frac{d^2\theta}{R^2} - \theta - \tilde{z}_2 + \pi\right)\right] + \right.$$
$$+ 6\left[\sin\left(3\theta - \frac{2d^2\theta}{R^2}\right) - \sin\left(\frac{2d^2\theta}{R^2} - \theta\right) - \sin\left(\theta + 2\tilde{z}_1\right) + \sin\left(\theta + 2\tilde{z}_2\right)\right] +$$
$$+ 2\left[\sin\left(5\theta - \frac{2d^2\theta}{R^2}\right) - \sin\left(3\theta + 2\tilde{z}_1\right) - \sin\left(\frac{4d^2\theta}{R^2} + 2\theta^2\right) + \sin\left(3\theta + 2\tilde{z}_2\right)\right] +$$
$$\left. + \sin\left(3\theta + 4\tilde{z}_1\right) - \sin\left(7\theta\frac{4d^2\theta}{R^2}\right) + \sin\left(\frac{4d^2\theta}{R^2} - \theta\right) - \sin\left(3\theta + 4\tilde{z}_2\right)\right], \quad (6)$$

где $\tilde{z}_1 = \max\{\theta(1 - d^2/R^2), z_1\}$, $\tilde{z}_2 = \min\{\pi - \theta(1 - d^2/R^2), z_2\}$, *и* z_1, z_2 *определены следующим образом*:

$$z_{1,2} = \pm 2\arctan\left[\frac{\sqrt{-(R_d^2/d^2)\tan^2\theta + \tan^2\theta + 1} \pm 1}{(R_d/d + 1)\tan\theta}\right]. \quad (7)$$

4 Численные результаты и выводы

В качестве сценария для иллюстрации результатов разд. 3 рассмотрим соединения в рое дронов, которые связываются между собою по протоколу IEEE 802.11ad на частоте 60 ГГц ($\lambda = 0{,}5$ см). По умолчанию предполагаются однородные прямоугольные фазированные антенные решетки, которые содержат изотропные элементы. Мощность передачи фиксирована на уровне $P_{\text{tx}} = 23$ дБм, тогда как $P_{\text{thr}} = -78$ дБм [7].

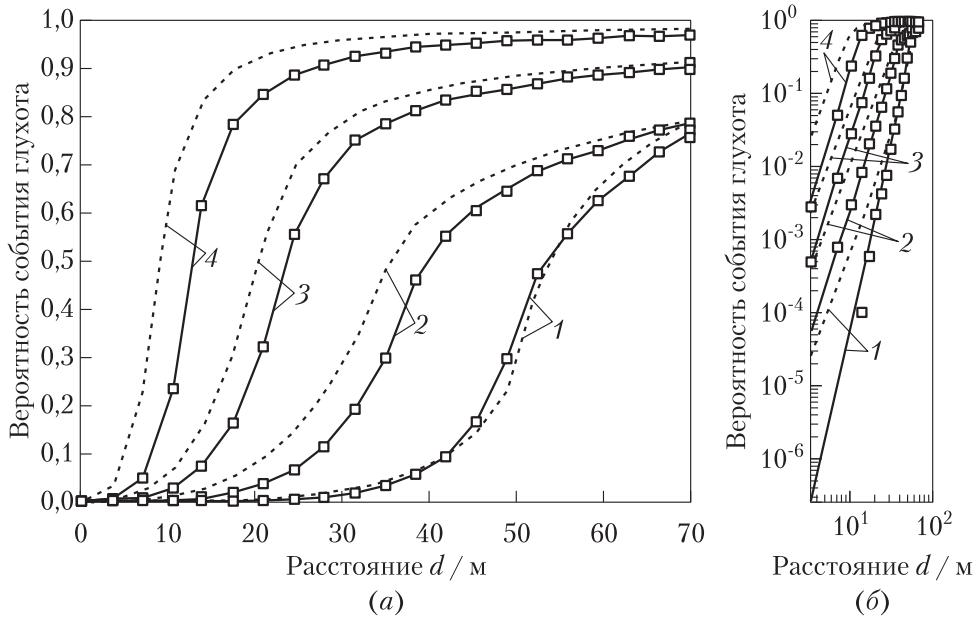


Рис. 3 Вероятность события «глухота» в зависимости от расстояния d для аппроксимации с помощью линейной модели (2) (пунктирные кривые) и реалистичных ДНА (значки — моделирование; сплошные кривые — аналитическое решение) в линейной (*а*) и логарифмической (*б*) шкалах: *1* — 2×2, $\theta = 58°$; *2* — 4×4, $\theta = 32°$; *3* — 8×8, $\theta = 18°$; *4* — 16×16, $\theta = 8°$

Начнем со сравнения результатов для имитационного моделирования расположения устройств методом Монте-Карло (кривая «Моделирование») с точным численным решением (5) для вероятности события «глухота» $P_D(d)$ в случаях реалистичных ДНА (кривая «Аналитическое решение») и линейной модели (2) (кривая «Аппроксимация»). Как показано на рис. 3, *а*, линейная модель представляет адекватную аппроксимацию для широких лучей (2×2), а для более высокой направленности антенн (4×4, 8×8 и 16×16) повторяет поведение конкретной реалистичной ДНА и сходится с ней для больших значений расстояния d.

Для иллюстрации поведения линейной модели для малых значений $P_D(d)$ на рис. 3, *б* графики представлены в логарифмическом масштабе. Когда вероятность события «глухота» падает ниже 10^{-6}, даже если график линейной модели отклоняется от графика реалистичной антенны на величину до двух порядков, первая сохраняет линейность логарифма вероятности события «глухота», что дает закон масштабирования $P_D(d) \sim ad^b$, где b определяется наклоном линейных отрезков на рис. 3, *б*.

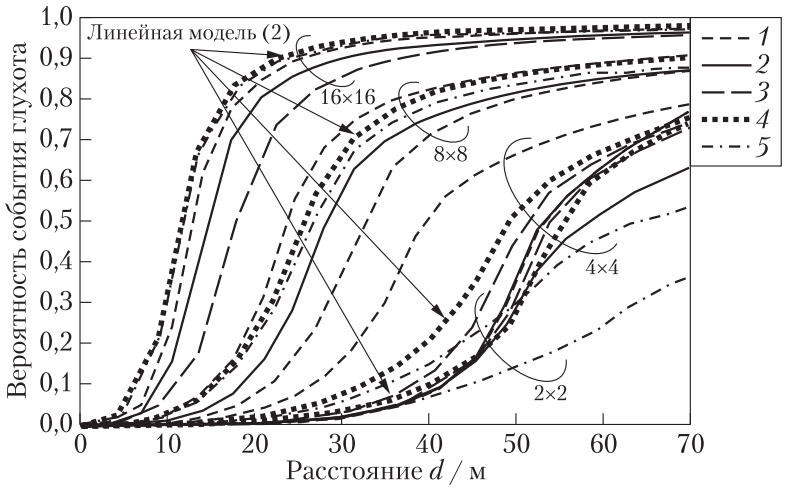


Рис. 4 Вероятность события «глухота» для различных настроек антенны: изотропные (2×2 — $\theta = 58°$; 4×4 — $32°$; 8×8 — $18°$; 16×16 — $\theta = 8°$) и косинусоидальные (2×2 — $\theta = 46°$; 4×4 — $38°$; 8×8 — $18°$; 16×16 — $\theta = 8°$) элементы, распределения Чебышёва (*1* — для изотропной ДНА) и Хэмминга (*2* — для изотропной ДНА; *3–5* — для косинусоидальной ДНА: *3* — распределение; *4* — аппроксимация; *5* — нижняя граница)

Рисунок 4 позволяет сравнить $P_D(d)$ для различных настроек реалистичных антенн и нижнюю границу $P_D(d)$ по формулам (6) и (7), рассчитанную для линейной модели (2).

На рис. 4 также указана ширина θ угла по уровню половинной мощности для рассмотренных антенных решеток. Линейная модель (2) показывает разумную аппроксимацию и может использоваться, когда тип антенного элемента несуществен для системного анализа, в то время как нижняя граница остается близкой к результатам реалистичных антенн для узких лучей.

Рисунок 5 демонстрирует, что масштабирование системы, т. е. изменение радиуса R_d пропорционально радиусу зоны покрытия R, устраняет разрыв между различными значениями угла по уровню половинной мощности θ.

Наконец, для сценария работы роя дронов с заданным ограничением на сближение устройств местоположение устройств моделировалось в соответствии с процессом твердого ядра Матерна типа 1 с параметром r [10, 11]. Как видно на рис. 6, параметр r оказывает незначительное влияние на результаты. Следовательно, предлагаемый аналитический подход обеспечивает достойную аппроксимацию для более реалистичного распределения устройств. Здесь также стоит отметить существенную разницу между трехмерным сценарием направленной глухоты и соответствующими аналитическими результатами, полученными в предположении двумерного сценария [5].

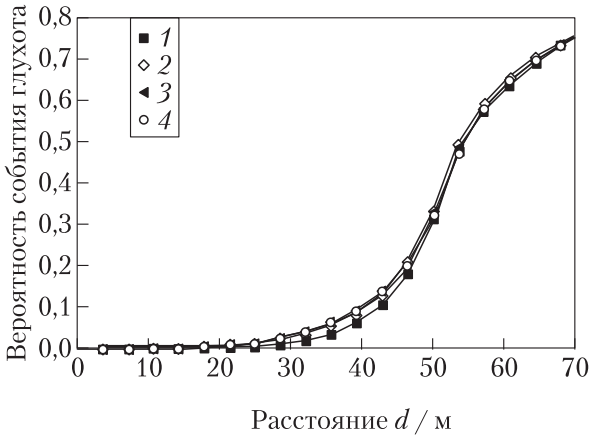


Рис. 5 Влияние масштабирования на вероятность события глухота (3D, распределение Чебышёва): *1* — 2×2, $\theta = 58°$; *2* — 4×4, $\theta = 32°$; *3* — 8×8, $\theta = 18°$; *4* — 16×16, $\theta = 8°$

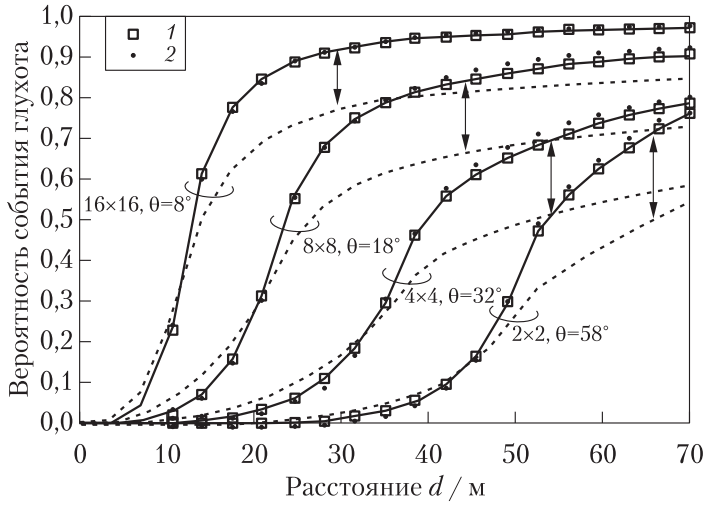


Рис. 6 Сравнение вероятности события глухота для двумерных [5] (пунктирные кривые) и трехмерных (сплошные кривые и значки) случаев и различных значений параметра твердого ядра Матерна: *1* — MHCP = 5; *2* — MHCP = 25

5 Заключение

В заключение стоит подчеркнуть, что «глухота» представляет собой серьезную проблему для любой высоконаправленной системы и может привести к пагубным последствиям для производительности сети связи, как было проде-

монстрировано ранее в [5]. Для более реалистичного распределения устройств в трехмерном пространстве можно получить более точные (до 30% по сравнению с двумерным случаем) оценки для вероятности события «глухота». Предложенная в статье простая аналитически управляемая модель диаграммы направленности антенны, задаваемая линейной функцией ширины луча по уровню половинной мощности, может использоваться в качестве разумной аппроксимации для различных установок антенн, а также при условии распределения устройств в соответствии с процессом твердого ядра.

Литература

1. 3GPP Technical Report 38.211; NR; Physical channels and modulation (Release 15), December 2017.
2. *Gossain H., Cordeiro C., Cavalcanti D., Agrawal D. P.* The deafness problems and solutions in wireless ad hoc networks using directional antennas // 2004 IEEE Global Telecommunications Conference Workshops. — IEEE, 2004. P. 108–113.
3. *Nitsche T., Cordeiro C., Flores A. B., Knightly E. W., Perahia E., Widmer J.* IEEE 802.11ad: Directional 60 GHz communication for multi-gigabit-per-second Wi-Fi // IEEE Commun. Mag., 2014. Vol. 52. No. 12. P. 132–141.
4. *Sim G. H., Nitsche T., Widmer J. C.* Addressing MAC layer ineffciency and deafness of IEEE 802.11ad millimeter wave networks using a multi-band approach // 27th IEEE Annual International Symposium on Personal, Indoor, and Mobile Radio Communications. — IEEE, 2016. P. 1–6.
5. *Galinina O., Pyattaev A., Johnsson K., Andreev S., Koucheryavy Ye.* Analyzing effects of directional deafness on mmWave channel access in unlicensed bands // IEEE Globecom Workshops. — IEEE, 2017. P. 1–7.
6. *Толкачев В. И., Пикалов О. Г., Паньчев С. В., Новиков И. Г.* Расчет коэффициента усиления антенн через трехмерное представление диаграмм направленности и оценка их взаимного влияния // Ракетно-космическое приборостроение и информационные системы, 2018. Т. 5. Вып. 1. С. 75–80.
7. IEEE 802.11 Working Group. Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications. Amendment 3: Enhancements for Very High Throughput in the 60 GHz Band, 2012.
8. *Friis H. T.* A note on a simple transmission formula // P. IRE, 1946. Vol. 34. Iss. 5. P. 254–256. doi: 10.1109/JRPROC.1946.234568.
9. *Cai T., Fan J., Jiang T.* Distribution of angles in random packing on spheres // J. Mach. Learn. Res., 2013. Vol. 14. P. 1837–1864.
10. *Baccelli F., Błaszczyszyn B.* Stochastic geometry and wireless networks: Volume I theory // Found. Trends Netw., 2010. Vol. 3. No. 3-4. P. 249–449. doi: 10.1561/1300000006.
11. *Baccelli F., Błaszczyszyn B.* Stochastic geometry and wireless networks: Volume II applications // Found. Trends Netw., 2011. Vol. 4. No. 1-2. P. 1–312. doi: 10.1561/1300000026.

Поступила в редакцию 02.08.19

APPROXIMATION OF ANTENNA DIRECTIVITY GAIN FOR DIRECTIONAL DEAFNESS ANALYSIS IN THREE-DIMENSIONAL SPACE

***O. V. Chukhno*[1], *N. V. Chukhno*[1], *Yu. V. Gaidamaka*[1,2], *and S. Ya. Shorgin*[2]**

[1]Peoples' Friendship University of Russia (RUDN University), 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: The paper deals with the problem of "directional deafness" that arises when a device cannot detect an occupied radio channel due to the highly directional communication link between other devices interacting at that time. The "deafness" situation can arise between operating in the millimeter band devices, for example, during carrier-sense multiple access stage, in particular, in accordance with the IEEE 802.11ad/ay protocols. An analytical expression has been obtained for the "directional deafness" probability for several variants of the devices location in three-dimensional (3D) space and for the proposed linear approximation of the antenna directivity gain. The proposed formula for the lower bound of the deafness probability for three realistic antenna patterns and four variants of phased antenna arrays is investigated.

Keywords: mmWave; directional deafness; 3D; directional access

DOI: 10.14357/08696527190304

Acknowledgments

The publication was prepared with the support of the "RUDN University Program 5-100" and funded by the Russian Foundation for Basic Research according to the research projects No. 17-07-00845 and No. 18-07-00576.

References

1. 3GPP Technical Report 38.211; NR; Physical channels and modulation (Release 15). December 2017.
2. Gossain, H., C. Cordeiro, D. Cavalcanti, and D. P. Agrawal. 2004. The deafness problems and solutions in wireless ad hoc networks using directional antennas. *IEEE Global Telecommunications Conference Workshops, 2004*. IEEE. 108–113.
3. Nitsche, T., C. Cordeiro, A. B. Flores, E. W. Knightly., E. Perahia, and J. Widmer. 2014. IEEE 802.11ad: Directional 60 GHz communication for multi-gigabit-per-second Wi-Fi. *IEEE Commun. Mag.* 52(12):132–141.
4. Sim, G. H., T. Nitsche, and J. C. Widmer. 2016. Addressing MAC layer inefficiency and deafness of IEEE 802.11ad millimeter wave networks using a multi-band approach.

27th IEEE Annual International Symposium on Personal, Indoor, and Mobile Radio Communications. IEEE. 1–6.

5. Galinina, O., A. Pyattaev, K. Johnsson, S. Andreev, and Y. Koucheryavy. 2017. Analyzing effects of directional deafness on mmWave channel access in unlicensed bands. *IEEE Globecom Workshops*. IEEE. 1–7.
6. Tolkachev, V. I., O. G. Pikalov, S. V. Pan'chev, and I. G. Novikov. 2018. Raschet koeffitsienta usileniya antenn cherez trekhmernoe predstavlenie diagramm napravlennosti i otsenka ikh vzaimnogo vliyaniya [Calculation of antenna directivity gain using a three- dimensional representation of radiation patterns and evaluation of their mutual influence]. *Raketno-kosmicheskoe priborostroenie i informatsionnye sistemy* [Rocket-Space Device Engineering and Information Systems] 5(1):75–80.
7. IEEE 802.11 Working Group. 2012. Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications. Amendment 3: Enhancements for Very High Throughput in the 60 GHz Band.
8. Friis, H. T. 1946. A note on a simple transmission formula. *P. IRE* 34(5):254–256. doi: 10.1109/JRPROC.1946.234568.
9. Cai, T., J. Fan, and T. Jiang. 2013. Distribution of angles in random packing on spheres. *J. Mach. Learn. Res.* 14:1837–1864.
10. Baccelli, F., and B. Błaszczyszyn. 2010. Stochastic geometry and wireless networks: Volume I theory. *Found. Trends Netw.* 3(3-4):249–449. doi: 10.1561/1300000006.
11. Baccelli, F., and B. Błaszczyszyn. 2010. Stochastic geometry and wireless networks: Volume II applications. *Found. Trends Netw.* 4(1-2):1–312. doi: 10.1561/1300000026.

Received August 2, 2019

Contributors

Chukhno Nadezhda V. (b. 1995) — Master student, Applied Probability and Informatics Department, Peoples' Friendship University of Russia (RUDN University), 6 Miklukho-Maklaya Str., Moscow 117198, Russian Federation; nvchukhno@gmail.com

Chukhno Olga V. (b. 1995) — Master student, Applied Probability and Informatics department, Peoples' Friendship University of Russia (RUDN University), 6 Miklukho-Maklaya Str., Moscow 117198, Russian Federation; olgachukhno95@gmail.com

Gaidamaka Yuliya V. (b. 1971) — Doctor of Science in physics and mathematics, professor, Peoples' Friendship University of Russia (RUDN University), 6 Miklukho-Maklaya Str., Moscow 117198, Russian Federation; senior scientist, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, 44-2 Vavilov Str., Moscow 119333, Russian Federation; gaydamaka_yuv@rudn.university

Shorgin Sergey Ya. (b. 1952) — Doctor of Science in physics and mathematics, professor, principal 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; sshorgin@ipiran.ru

СИСТЕМЫ И СРЕДСТВА ИНФОРМАТИКИ, 2019. Т. 29. № 3. С. 52–65

МЕТОД КЛАСТЕРИЗАЦИИ НОВОСТНЫХ СООБЩЕНИЙ СРЕДСТВ МАССОВОЙ ИНФОРМАЦИИ НА ОСНОВЕ ИХ КОНЦЕПТУАЛЬНОГО АНАЛИЗА*

***В. Н. Захаров*[1], *Р. Р. Мусабаев*[2], *А. М. Красовицкий*[3], *Я. Д. Козловская*[4], *Ал-др А. Хорошилов*[5], *Ал-ей А. Хорошилов*[6]**

Аннотация: Изложено решение задачи кластеризации сообщений средств массовой информации (СМИ) на основе разработанной авторами методики автоматического вычисления меры смысловой значимости наименований понятий документов, использующей их статистические, синтаксические и семантические признаки, и технологий автоматического составления декларативных средств для кластеризации документов, базирующихся на методах их семантико-синтаксического и концептуального анализа. На основе предложенной методики вычисления меры смысловой значимости наименований понятий и созданных в процессе проведения настоящего исследования программных и декларативных средств был поставлен эксперимент по обработке представительного массива сообщений СМИ. Анализ полученных результатов показал, что при автоматическом установлении смысловой значимости текстовых наименований понятий использование семантических коррелирующих коэффициентов понятий повышает точность установления смысловой схожести между документами.

Ключевые слова: кластеризация текстов; семантико-синтаксический анализ текстов; концептуальный анализ текстов; декларативные средства; статистическая мера значимых слов документа; семантический корреляционный коэффициент; смысловая близость текстов

DOI: 10.14357/08696527190305

*Статья подготовлена в рамках проекта ПЦФ BR05236839 «Разработка информационных технологий и систем для стимулирования устойчивого развития личности как одна из основ развития цифрового Казахстана».

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

[2]Институт информационных и вычислительных технологий, Алматы, Казахстан, rmusab@gmail.com

[3]Институт информационных и вычислительных технологий, Алматы, Казахстан, akrassovitskiy@gmail.com

[4]Московский авиационный институт (национальный исследовательский университет), yana04029877@mail.ru

[5]Институт проблем информатики Федерального исследовательского центра «Информатика и управление» Российской академии наук, khoroshilov@mail.ru

[6]27 ЦНИИ Министерства обороны России, alex_khor@mail.ru

1 Введение

Современное общество в процессе своего развития порождает огромный объем текстовой информации, относящейся к различным аспектам его деятельности. Целый пласт этой информации относится к так называемой новостной информации — сообщениям о текущих событиях, происходящих в реальном мире. Как правило, новостную информацию формируют СМИ и результатом их основной деятельности является информирование общества обо всех событиях, происходящих в мире. Считается, что для идеального описания конкретного новостного события должна быть справедлива формула, приписываемая еще римскому ритору Квинтилиану:

кто сделал? + что сделал? + какими средствами? + зачем? + когда? + где?

В англоязычной интерпретации эта формула носит название закона пяти W и одного H:

Who? + What? + Where? + Why? + When? + How?,

приписываемого Р. Киплингу. Такое описание «идеального» новостного сообщения должно быть посвящено одному событию, и структура этого новостного сообщения должна включать ряд содержательных компонент (дату, автора, заголовок, формальное описание события, его историю, энциклопедическую справку и др.) [1].

На практике все происходит не так. Идеальных новостных сообщений практически не бывает, а те новостные сообщения, которые появляются на сайтах СМИ, подаются в произвольной форме, и порой пользователю, чтобы получить полную информацию об этих событиях, приходится ознакомиться с большим числом источников.

Для решения этой проблемы были созданы новостные агрегаторы — Яндекс. Новости, Рамблер/новости, GoogleNews и др., которые интегрируют новости из тысяч источников в «новостные кластеры», предоставляя пользователям доступ к ранжированным по значению и тематике сюжетам. Сайты таких новостных агрегаторов стали одними из самых популярных ресурсов.

Ключевая проблема новостных агрегаторов заключается в решении задачи кластеризации — формирования групп документов, описывающих близкие новостные сюжеты. На практике результат автоматической кластеризации может быть изменен в зависимости от корпоративной политики (в том числе и вручную редакторами), данных о запросах пользователей или обсуждений в блогах.

2 Обзор методов кластеризации текстов

В настоящее время существует большое число методов кластеризации текстов. Под кластеризацией понимается процесс разделения множества документов на подмножества (кластеры), число и параметры которых заранее не известны. В данном обзоре рассматривается несколько методов, получивших наибольшее распространение. К таким методам можно отнести следующие: метод LSA/LSI

(Latent Semantic Analysis/Indexing), метод STC (Suffix Tree Clustering), метод Scatter/Gather, метод K-means (К-средних) и др. [1–5].

В основе метода K-means лежит итеративный процесс стабилизации центроидов (центра масс) кластеров, которые первоначально выбираются случайным образом для каждого из k кластеров. Каждый документ присваивается тому кластеру, расстояние до центра масс которого от него меньше заданного. Далее на каждой итерации вычисляются центры масс кластеров и документы переприсваиваются другому кластеру до (стабилизации) всех документов.

Метод Suffix Tree Clustering кластеризует тексты в виде суффиксного дерева. Суффиксное дерево — это дерево, содержащее все суффиксы данной строки. Они состоят из вершин, ветвей и дополнительных указателей (Suffix pointer), с помощью которых добиваются линейной скорости построения деревьев. Ветви дерева обозначаются буквами или сочетаниями букв, которые являются частями суффиксов строки. Суффикс вершины получают путем объединения всех букв, находящихся на ребрах дерева (начиная от корневой до данной).

Каждой вершине соответствует фраза. В тех вершинах дерева, которые имеют потомков, есть ссылки на документы, в которых встречается фраза, соответствующая вершине. Базовые кластеры образуются из множеств документов, на которые указывают ссылки. Далее производятся комбинирование базовых кластеров и получение окончательных наборов.

Методы Single Link, Complete Link и Group Average относятся к иерархическим методам, которые делятся на агломеративные и дивизимные методы. Первые объединяют объекты в множества, а вторые наоборот разделяют единые множества объектов на подмножества. Методы Single Link, Complete Link и Group Average относятся к агломеративным иерархическим методам, которые получили широкое распространение.

Метод Self-Organizing Maps (метод самоорганизующихся карт Кохонена) выполняет кластеризацию документов на основе нейросети Кохонена. В результате работы этого метода получается образ документа, представляющий собой карту распределения векторов из обучающей выборки. Эта сеть обучается без учителя на основе принципа самоорганизации.

Метод LSA/LSI давно известен в различных областях науки как метод выявления латентной структуры изучаемых явлений и объектов. В рамках этого метода определяется пространство терминов как пространство элементарных признаков, в котором изначально располагаются документы. Предполагается, что термины должны быть семантически связаны между собой, тогда документы, содержащие семантически близкие термины, сгущаются в определенных местах пространства терминов.

Методы на основе так называемого «мешка слов» (bag-of-words) относятся к наиболее простым алгоритмам кластеризации. В рамках этих методов предварительно по выборке документов строится словарь из всех встречающихся в нем n-грамм (контактно расположенных последовательностей слов), где n меньше или равно заранее заданному значению. Документ представляется в виде векто-

ра, состоящего из набора признаков. Каждому набору из словаря соответствует одна n-грамма. Для каждой n-граммы вектора документа вычисляется его вес по формуле статистической меры TF-IDF (term frequency – inverse document frequency — частота слова – обратная частота документа). Далее производится попарное сопоставление векторов документов путем вычисления косинусной меры близости между их векторами.

Анализ приведенных методов кластеризации документов показывает, что значительная часть этих методов опирается на формальные признаки, которые либо вообще не связаны со смысловой структурой текстов, либо связаны с ней очень отдаленно, и практически не существует методов кластеризации, в явном виде опирающихся на смысловую структуру документов. А между тем описание каждого информационного события включает элементы предикатно-актантной структуры предложений [6–9]. Так, если воспользоваться законом пяти W и одного H, то первое W (who) в предикатно-актантной структуре — это «субъект», второе W (what) в той же структуре — «объект» и т. д.

Исходя из вышеизложенного можно сделать следующий вывод: для адекватного установления смысловой схожести документов в новостном потоке (НП) необходимо разработать методы кластеризации, базирующиеся на теоретических представлениях о смысловой структуре текстов и установлении меры смысловой значимости текстовых наименований понятий. В качестве одного из таких представлений можно использовать фразеологический концептуальный анализ текстов.

3 Концепция смысловой обработки текстовой информации

В основу фразеологического концептуального анализа текстов положено утверждение о том, что смысловое содержание текстов выражается через систему его наименований понятий (слов и фразеологических словосочетаний). С помощью этих понятий формируются смысловые единицы более высоких уровней: предложения и сверхфразовые единства.

Предложения также представляют собой значимые единицы смысла. Основное их свойство — предикативность (наличие у объектов определенных признаков и их отношений). Наличие этих признаков у объектов в текстах выражается через предикатно-актантную структуру предложений. Ее компонентами являются понятия-предикаты (признаки и отношения) и понятия-актанты, выступающие в роли описываемых объектов.

Из предложений формируются сверхфразовые единства. Они представляются в виде последовательностей предложений связного текста и, как правило, формируют более сложные мыслительные образы, которые в новостных сообщениях могут отражать конкретные события реального мира.

Автоматизация процесса смысловой обработки текстовой информации базируется на процедурах семантико-синтаксического и концептуального анализа текстов. Эти процедуры должны опираться на адекватные семантико-синтаксиче-

ские модели текстов, обеспечивающие возможность выявления системы понятий текста и установления смысловых отношений между элементами этой системы.

4 Методы выявления наименований понятий в текстах

Сложность выявления наименований понятий, представленных словосочетаниями, заключается в правильном установлении их границ в текстах и выявлении тех наименований понятий, которые в тексте несут основную смысловую нагрузку. Эти проблемы решаются методами их статистического, синтаксического и концептуального анализа.

Статистические методы позволяют путем назначения весовых коэффициентов установить состав значимых слов и словосочетаний на основе анализа их частот в конкретном тексте и в относительно коротком временно́м текстовом потоке [1].

Синтаксические методы позволяют выявить синтаксическую роль значимых слов и словосочетаний в предложении (принадлежность к словосочетаниям, являющимся в предложении группой подлежащего (субъекта), группой сказуемого (предиката) или группой дополнения (объекта)) [6–10].

Семантические методы позволяют выявить значимые в предметной области слова и словосочетания путем их соотнесения с элементами эталонных словарей или их формализованных семантических признаков с эталонными представлениями этих признаков [6–10].

Общее решение задачи кластеризации текстов может быть сведено к следующим этапам.

1. Определение объектов признакового пространства, под которым будем понимать множество значимых слов и словосочетаний, определяющих содержание текстов в массиве.
2. Вычисление меры смысловой значимости для каждого понятия документа.
3. Вычисление значений меры сходства между текстами документа.
4. Применение методов кластерного анализа для создания групп сходных текстов.
5. Проверка истинности результатов кластерного решения.

В соответствии с этими этапами в первую очередь нужно определить множество слов и словосочетаний, по которым будут оцениваться тексты. Потом необходимо определить, по каким критериям нужно оценивать смысловую значимость каждого элемента признакового пространства, и в соответствии с этими критериями назначить каждому слову или словосочетанию весовой коэффициент их смысловой значимости в признаковом пространстве.

В рамках выполненных исследований по созданию декларативных средств по одному и тому же текстовому массиву были составлены четыре частотных словаря. Статистические данные об объемах частотных словарей, полученных по корпусу текстов СМИ различными методами концептуального анализа, приведены работе [9].

5 Методы определения весовых коэффициентов смысловой значимости наименований понятий в текстовом массиве

В качестве меры смысловой значимости слов и словосочетаний часто используется так называемая статистическая мера TF-IDF [7]:

TF — отношение числа вхождений наименования понятия к общему числу наименований понятий документа. Таким образом оценивается важность наименования понятия t_i в пределах отдельного документа:

$$\mathrm{tf}(t,d) = \frac{n_t}{\sum_k n_k}\,,$$

где n_t — число вхождений наименования понятия t в документ; $\sum_k n_k$ — общее число наименований понятий в данном документе;

IDF — инверсия частоты, с которой некоторое наименование понятия встречается в документах коллекции. Для каждого уникального слова в пределах конкретной коллекции документов существует только одно значение IDF:

$$\mathrm{idf}(t,D) = \log\frac{|D|}{|\{d_i \in D | t \in d_i\}|}\,,$$

где $|D|$ — число документов в коллекции; $|\{d_i \in D | t \in d_i\}|$ — число документов из коллекции D, в которых встречается t (когда $n_t \neq 0$).

Мера TF-IDF является произведением двух сомножителей:

$$\text{tf-idf}(t,d,D) = \mathrm{tf}(t,d)\cdot\mathrm{idf}(t,D)\,.$$

Большой вес в TF-IDF получают наименования понятий с высокой частотой встречаемости в конкретном документе и с относительно низкой частотой в пределах всего корпуса текстов.

Но данная статистическая мера в явном виде не отражает смысловую составляющую наименований понятий. С этой целью была разработана система коррелирующих синтаксических и семантических весовых коэффициентов наименований понятий, восполняющих этот пробел.

Мера смысловой значимости наименований понятий M_3 (с учетом синтаксических коэффициентов) вычисляется по формуле:

$$M_3 = (\mathrm{TF}\cdot\mathbf{IDF})K_dK_zK_s\,.$$

Общая мера смысловой значимости M_6 (с учетом синтаксических и семантических весовых коэффициентов) вычисляется по формуле:

$$M_6 = (\mathrm{TF}\cdot\mathbf{IDF})K_dK_zK_tK_nK_sK_f\,,$$

Таблица 1 Формализованный концептуальный образ документа с различными мерами смысловой значимости понятий

№	Наименование понятия (каждое слово понятия приведено в нормализованной форме)	TF-IDF	M_3	M_6
1	блок петр порошенко	0,00625	0,0075	0,10125
2	верховный рада украина	0,1194	0,365364	0,548046
3	владимир зеленский	0,0022	0,004488	0,040392
4	глава государство	0,0022	0,00396	0,01188
5	глава служба безопасность украина	0,005	0,0075	0,01125
6	законопроект об изменение выборный система	0,05	0,153	0,2295
7	зампредседатель администрация глава государство	0,05	0,09	0,405
8	избирательный законодательство	0,0454	0,08172	0,24516
9	инаугурационный речь	0,0412	0,07416	0,07416
10	кандидат на пост глава партия	0,0523	0,160038	0,240057
11	официальный указ	0,0046	0,0828	0,2484
12	парламентский выбор	0,0301	0,05418	0,05418
13	поддержка население	0,05	0,09	0,09
14	проект изменение в избирательный законодательство	0,13184	0,237312	0,355968
15	служба безопасность украина	0,0523	0,09414	0,42363
16	социологический опрос	0,0602	0,07224	0,21672

где K_d — коэффициент, учитывающий распознаваемость слов при их нормализации; K_z — коэффициент, учитывающий вхождение в заголовки слов или словосочетаний; K_t — коэффициент, учитывающий вхождение в тезаурус слов или словосочетаний; K_n — коэффициент, учитывающий число слов в словосочетании; K_s — коэффициент, учитывающий синтаксическую роль слова или словосочетания в предложении; K_f — коэффициент, учитывающий принадлежность понятия к фамильно-именной группе, бренду и др.

Для каждого наименования понятия вычислялась его статистическая мера значимости TF-IDF, мера смысловой значимости M_3 и общая мера смысловой значимости M_6. В табл. 1 приведены результаты вычислений весовых коэффициентов понятий формализованного концептуального образа документа [7].

6 Установление смысловой близости между документами средств массовой информации

В качестве исходных данных для проведения эксперимента был использован массив документов СМИ, включающий 3004 документа.

Для выполнения основной задачи данного исследования — автоматического установления степени смысловой близости документов СМИ на основе анализа их

Таблица 2 Результаты сопоставления меры смысловой близости между документами

Первый документ			Второй документ			Мера близости		
№	Общая сумма M_6	Сумма совпавших M_6	№	Общая сумма M_6	Сумма совпавших M_6	μ_1	μ_2	μ_3
206	3,56257	2,47787	4450	1,77294	0,95634	0,285	0,352	**0,414**
323	2,06946	1,29836	355	2,69428	2,05129	0,290	0,392	**0,478**
311	2,86595	1,96964	561	3,12935	2,19645	0,263	0,371	**0,482**
319	2,53975	1,56943	964	2,86341	1,89749	0,275	0,348	**0,409**
267	1,86186	1,10946	1387	2,06846	1,39465	0,244	0,330	**0,402**
1595	2,29526	1,65823	217	2,09746	1,86128	0,337	0,481	**0,641**
1595	3,75619	2,67823	6638	2,96421	2,10674	0,321	0,401	**0,507**
169	3,52678	2,84369	8612	3,35965	2,38138	0,305	0,429	**0,572**
569	2,62487	1,75482	384	2,04827	1,82374	0,344	0,452	**0,595**
388	1,62795	0,98243	614	1,82671	1,24897	0,397	0,502	**0,628**
404	1,48267	0,89247	377	2,17955	1,55217	0,245	0,335	**0,429**

смыслового содержания — были вычислены меры близости между документами методом установления косинуса угла между векторами значений наименований понятий.

Меру близости между документами можно установить по следующей формуле:

$$\mu\left(n_i, n_j\right) = \frac{\sum_t n_{it} n_{jt}}{\sqrt{\sum_m n_{im}^2}\sqrt{\sum_k n_{jk}^2}},$$

где n_{it} — значение веса наименования понятия t в документе n_i; n_{jt} — значение веса наименования понятия t в документ n_j; $\sum_m n_{im}$ — сумма всех значений наименований понятий в документе n_i; $\sum_k n_{jk}$ — сумма всех значений наименований понятий в документе n_j.

В табл. 2 приведены статистические данные о результатах сравнения мер близости, вычисленных для разных мер значимости (мера близости μ_1 вычислена для статистической меры значимости TF-IDF, мера близости μ_2 вычислена для меры смысловой значимости M_3, мера близости μ_3 вычислена для общей меры смысловой значимости M_6). Жирным шрифтом помечены меры близости документов, преодолевших пороговое значение $\mu_3 \geq 0{,}4$.

7 Анализ результатов автоматической кластеризации текстов

Автоматическая кластеризация выполнялась путем вычисления меры смысловой близости между документами. Результат формирования кластеров заключался в том, что вокруг одного из документов формировалась группа других

документов, мера семантической близости которых была больше или равна 0,4. Данный коэффициент меры был ранее установлен в работе [1].

Анализ влияния коррелирующих семантических коэффициентов на результаты кластеризации текстового массива СМИ показал, что:

(1) коэффициент K_d, учитывающий распознаваемость слов при их нормализации, увеличивает вес значимых понятий в среднем на 2%;

(2) коэффициент K_z, учитывающий вхождение в заголовки слов или словосочетаний, увеличивает вес значимых понятий в среднем на 2%;

(3) коэффициент K_t, учитывающий вхождение в тезаурус слов или словосочетаний), увеличивает вес значимых понятий в среднем на 2%;

(4) коэффициент K_n, учитывающий число слов в словосочетании, увеличивает вес значимых понятий в среднем на 2%;

(5) коэффициент K_s, учитывающий синтаксическую роль слова или словосочетания в предложении, увеличивает вес значимых понятий в среднем на 5%;

(6) коэффициент K_f, учитывающий принадлежность понятия к фамильно-именной группе, бренду и др., увеличивает вес значимых понятий в среднем на 1%.

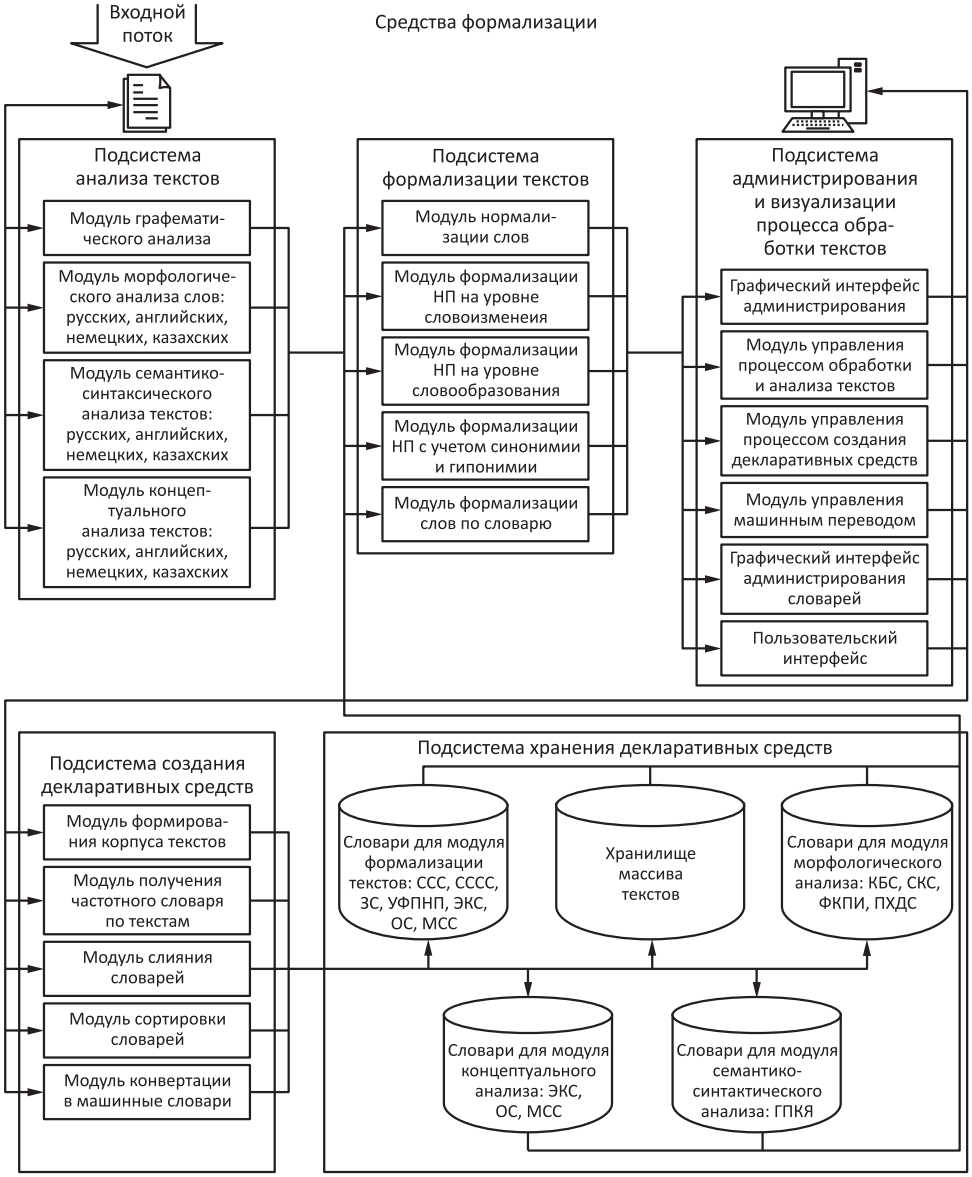
В результате выполнения кластеризации текстового массива СМИ, включающего 3004 документа, было выявлено 189 кластеров стандартным методом установления смысловой значимости наименований понятий, в соответствии с предложенными методами установления смысловой значимости наименований понятий было выявлено 115 кластеров документов, объединяющих такие темы, как «Тарифы ЖКХ», «Энергетика», «Спорт», «Чемпионат по футболу», «Чемпионат мира по боксу», «Легкая атлетика», «Допинг в спорте», «Кандидат на пост главы партии Зеленского», «Предвыборная кампания Владимира Зеленского» и ряд других тем. Выборочная экспертная оценка содержания документов этих кластеров подтвердила правильность предложенных методов.

8 Программная реализация

В рамках выполняемого исследования для обеспечения возможности проверки эффективности описанных методов, моделей и алгоритмов при участии авторов исследования было разработано экспериментальное программное обеспечение, реализующее процесс кластеризации документов СМИ на базе программно-технологической платформы МетаФраз [10]. Общая схема разработанного программного комплекса приведена на рисунке.

9 Заключение

При решении задачи кластеризации сообщений СМИ показана принципиальная возможность ее реализации на основе разработанной авторами методики



Программно-информационная архитектура подсистемы кластеризации текстов

автоматического вычисления меры смысловой значимости наименований понятий документов и технологий автоматического составления декларативных средств для кластеризации документов, базирующихся на методах их семантико-синтаксического и концептуального анализа.

На основе предложенной методики вычисления меры смысловой значимости наименований понятий и созданных в процессе проведения настоящего исследования программных и декларативных средств был проведен эксперимент по обработке представительного массива сообщений СМИ. Анализ полученных результатов показал, что при автоматическом установлении смысловой значимости текстовых наименований понятий использование семантических коррелирующих коэффициентов понятий повышает точность установления смысловой схожести между документами.

Литература

1. *Добров Б. В., Павлов А. М.* Исследование качества базовых методов кластеризации новостного потока в суточном временном окне // Электронные библиотеки: перспективные методы и технологии, электронные коллекции: Труды XII Всеросс. научн. конф. — Казань: КГУ, 2010. С. 287–295.
2. *Киселев М.* Метод кластеризации текстов, основанный на попарной близости термов, характеризующих тексты, и его сравнение с метрическими методами кластеризации // Интернет-математика 2007: Сб. работ участников конкурса научных проектов по информационному поиску / Под ред. П. И. Браславского. — Екатеринбург: Изд-во Урал. ун-та, 2007. С. 74–83.
3. *Васильев В. Г., Кривенко М. П.* Методы автоматизированной обработки текстов. — М.: ИПИ РАН, 2008. 304 с.
4. *Борзых А. И., Брагина Г. А., Хорошилов А. А.* Методы автоматической кластеризации документов в хранилищах научно-технической информации для решения задачи поиска плагиата в текстах документов // Информатизация и связь, 2012. № 8. С. 33–37.
5. *Пархоменко П. А., Григорьев А. А., Астраханцев Н. А.* Обзор и экспериментальное сравнение методов кластеризации текстов // Труды ИСП РАН, 2017. Т. 29. Вып. 2. С. 161–200. doi: 10.15514/ISPRAS-2017-29(2)-6.
6. *Захаров В. Н., Хорошилов А. А.* Автоматическая оценка подобия тематического содержания текстов на основе сравнения их формализованных смысловых описаний // Электронные библиотеки: перспективные методы и технологии, электронные коллекции: Труды XIV Всеросс. научн. конф. — Переславль-Залесский: Ун-т г. Переславля, 2012. С. 189–195. http://ceur-ws.org/Vol-934/paper24.pdf
7. *Захаров В. Н., Хорошилов А. А.* Автоматическое формирование визуального представления смыслового содержания документа // Системы и средства информатики, 2013. Т. 23. № 1. С. 143–158.
8. *Хорошилов А. А.* Методы автоматического установления смысловой близости документов на основе их концептуального анализа // Электронные библиотеки: перспективные методы и технологии, электронные коллекции: Труды XV Всеросс. научн. конф. — Ярославль: ЯрГУ им. П. Г. Демидова, 2013. С. 369–376.
9. *Zakharov V., Krassovitskiy A., Meirambekkyzy Zh., Ualiyeva I., Khoroshilov Al-dr, Khoroshilov Al-ey.* Automatic creation technologies of declarative tools for clustering media documents // Conference (International) on Engineering Technologies and Computer Science Proceedings. — IEEE, 2019. P. 39–42. doi: 10.1109/EnT.2019.00013.

10. *Хорошилов Ал-др А., Никитин Ю. В., Хорошилов Ал-ей А., Будзко В. И.* Автоматическое создание формализованного представления смыслового содержания неструктурированных текстовых сообщений СМИ и социальных сетей // Системы высокой доступности, 2014. Т. 10. № 3. С. 52–70.

Поступила в редакцию 23.07.19

CLUSTERING METHOD OF NEWS MEDIA REPORTS BASED ON CONCEPTUAL ANALYSIS

***V. N. Zakharov*[1], *R. R. Musabaev*[2], *A. M. Krasovitskiy*[2], *Y. D. Kozlovskaya*[3], *Al-dr A. Khoroshilov*[4], *and Al-ey A. Khoroshilov*[5]**

[1]Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, 44-2 Vavilov Str., Moscow 119133, Russian Federation

[2]Institute of Information and Computational Technologies, 125 Pushkin Str., Almaty 050010, Kazakhstan

[3]Moscow Aviation Institute (National Research University), 4 Volokolamskoe Shosse, Moscow 125993, Russian Federation

[4]Institute of Informatics Problems, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, 44-2 Vavilov Str., Moscow 119133, Russian Federation

[5]The 27th Central Research Institute of the Ministry of Defence of the Russian Federation, 5, 1st Khoroshevsky Passage, Moscow 123007, Russian Federation

Abstract: The article describes the solution of a clustering news media reports based on the technique developed by authors of automatic calculation of a measure of semantic meaningfulness of the names of concepts of documents using their statistical, syntactic, and semantic features and technologies of automatic generation of declarative means for clustering documents based on the methods of their semantic-syntactic and conceptual analysis. On the basis of the suggested technique of calculation of a measure of semantic meaningfulness of the names of concepts and the software and declarative means created by the study process, an experiment was conducted to process a representative array of news media reports. The analysis of the results showed that the use of semantic correlating coefficients of concepts improves the accuracy of establishing semantic similarity between documents at automatically establishing the semantic meaningfulness of textual names of concepts.

Keywords: text clustering; semantic-syntactic analysis; conceptual analysis; declarative means; statistical measure of meaningfulness of textual names of documents; semantic correlating coefficient; semantic similarity between documents

DOI: 10.14357/08696527190305

Acknowledgments

The paper has been prepared under the theme "Development of information technologies and systems to stimulate sustainable development of a person as one of the pillars of the development of digital Kazakhstan" (program-targeted financing of the Ministry of Education and Science of the Republic of Kazakhstan, PCF BR05236839).

References

1. Dobrov, B. V., and A. M. Pavlov. 2010. Issledovanie kachestva bazovykh metodov klasterizatsii novostnogo potoka v sutochnom vremennom okne [Basic line for news clusterization methods evaluation]. *Elektronnyye biblioteki: perspektivnyye metody i tekhnologii, elektronnyye kollektsii: Trudy XII Vseross. nauchn. konf.* [Digital Libraries: Advanced Methods and Technologies, Digital Collections: 12th All-Russian Scientific Conference Proceedings]. Kazan. 287–295. Available at: http://rcdl.ru/doc/2010/287-295.pdf (accessed October 15, 2019).
2. Kiselev, M. 2007. Metod klasterizatsii tekstov, osnovannyy na poparnoy blizosti termov, kharakterizuyushchikh teksty, i ego sravnenie s metricheskimi metodami klasterizatsii [Text clustering procedure based on pairwise proximity of key terms and its comparison with metric clustering methods]. *Internet-matematika 2007: sb. rabot uchastnikov konkursa nauchykh proektov po informatsionnomu poisku* [Internet Mathematics 2007: Collection of works of the participants of the contest of scientific projects on information search]. Ed. P. I. Braslavskiy. Ekaterinburg: Ural University Publs. 74–83.
3. Vasil'ev, V. G., and M. P. Krivenko. 2008. *Metody avtomatizirovannoy obrabotki tekstov* [Methods of automated word processing]. Moscow: IPI RAN. 304 p.
4. Borzykh, A. I., G. A. Bragina, and A. A. Khoroshilov. 2012. Metody avtomaticheskoy klasterizatsii dokumentov v khranilishchakh nauchno-tekhnicheskoy informatsii dlya resheniya zadachi poiska plagiata v tekstakh dokumentov [Document automatic clusterization methods in science-technical information storages for plagiarism detecting in documents text problem solving]. *Informatizatsiya i svyaz'* [Informatization and Communication] 8:33–37.
5. Parkhomenko, P. A., A. A. Grigorev, and N. A. Astrakhantsev. 2017. Obzor i eksperimental'noe sravnenie metodov klasterizatsii tekstov [A survey and an experimental comparison of methods for text clustering: Application to scientific articles]. *Proceedings ISP RAN* 29(2):161–200. doi: 10.1551MSPRAS-2017-29(2)-6.
6. Zakharov, V. N., and A. A. Khoroshilov. 2012. Avtomaticheskaya otsenka podobiya tematicheskogo soderzhaniya tekstov na osnove sravneniya ikh formalizovannykh smyslovykh opisaniy [Automatic assessment of similarity of the texts' thematic content on the base of their formalized semantic descriptions comparison]. *Tr. XIV Vseross. nauchn. konf. "Elektronnyye biblioteki: perspektivnyye metody i tekhnologii, elektronnyye kollektsii"* [Digital Libraries: Advanced Methods and Technologies, Digital Collections: 14th All-Russian Scientific Conference Proceedings]. Pereslavl-Zalessky. 189–195. Available at: http://ceur-ws.org/Vol-934/paper24.pdf (accessed July 17, 2019).

7. Zakharov, V.N., and A.A. Khoroshilov. 2013. Avtomaticheskoe formirovanie vizual'nogo predstavleniya smyslovogo soderzhaniya dokumenta [Automatic generation of vizual representation of the document's semantic content]. *Sistemy i Sredstva Informatiki — Systems and Means of Informatics* 23(1):143–158.
8. Khoroshilov, A.A. 2013. Metody avtomaticheskogo ustanovleniya smyslovoy blizosti dokumentov na osnove ikh kontseptual'nogo analiza [Methods for automatically establishing the semantic proximity of documents based on their conceptual analysis]. *Trudy XV Vseross. nauch. konf. "Elektronnyye biblioteki: perspektivnyye metody i tekhnologii, elektronnyye kollektsii"* [Digital Libraries: Advanced Methods and Technologies, Digital Collections: 15th All-Russian Scientific Conference Proceedings]. Yaroslavl': Demidov Yaroslavl' State University. 369–376.
9. Zakharov, V., A. Krassovitskiy, Zh. Meirambekkyzy, I. Ualiyeva, Al-dr Khoroshilov, and Al-ey Khoroshilov. 2019. Automatic creation technologies of declarative tools for clustering media documents. *Conference (International) on Engineering Technologies and Computer Science Proceedings*. IEEE. 39–42. doi: 10.1109/EnT.2019.00013.
10. Khoroshilov, Al-dr A., Yu. V. Nikitin, Al-ey A. Khoroshilov, and V. I. Budzko. 2014. Avtomaticheskoe sozdanie formalizovannogo predstavleniya smyslovogo soderzhaniya nestrukturirovannykh tekstovykh soobshcheniy SMI i sotsial'nykh setey [Automatic construction of a formalized representation of the semantic contents of unstructured texts of mass-media and social networks]. *Sistemy vysokoy dostupnosti* [High Availability Systems] 10(3):52–70.

Received July 23, 2019

Contributors

Zakharov Victor N. (b. 1948) — Doctor of Science in technology, associate professor, scientific secretary, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, 44-2 Vavilov Str., Moscow 119133, Russian Federation; vzakharov@ipiran.ru

Mussabayev Rustam R. (b. 1979) — Head of Laboratory, Institute of Information and Computational Technologies, 125 Pushkin Str., Almaty 050010, Kazakhstan; rmusab@gmail.com

Krassovitsky Alexander M. (b. 1976) — Candidate of Science (PhD) in technology, leading scientist, Institute of Information and Computational Technologies, 125 Pushkin Str., Almaty 050010, Kazakhstan; akrassovitskiy@gmail.com

Kozlovskaya Yana D. (b. 1998) — student, Moscow Aviation Institute (National Research University), 4 Volokolamskoe Shosse, Moscow 125993, Russian Federation; yana04029877@mail.ru

Khoroshilov Aleksandr A. (b. 1952) — Doctor of Science in technology, leading scientist, Institute of Informatics Problems, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, 44-2 Vavilov Str., Moscow 119133, Russian Federation; khoroshilov@mail.ru

Khoroshilov Alexey A. (b. 1958) — Candidate of Science (PhD) in technology, senior scientist, 27th Central Research Institute of the Ministry of Defense of the Russian Federation, 5, 1st Khoroshevsky Passage, Moscow 123007, Russian Federation; alex_khor@mail.ru

ИНДЕКС КОНТЕКСТНОГО НАУЧНОГО ЦИТИРОВАНИЯ*

***И. В. Галина*[1], *М. М. Шарнин*[2]**

Аннотация: Рассматриваются авторский показатель качества научной статьи — индекс контекстного научного цитирования (ИКНЦ) и связь ИКНЦ и индекса научного цитирования (ИНЦ) с другим авторским показателем — мерой подобия (мерой семантического сходства) двух произвольных текстов. Приведены результаты экспериментов, в частности изучена корреляция между ИКНЦ и ИНЦ, зависящая от значения порога семантического подобия. На основе моделирования значений независимых переменных и их коэффициентов регрессии предложена прогностическая математическая вероятностная модель зависимости числа прямых цитирований от числа неявных ссылок и их параметров.

Ключевые слова: автоматизированные системы; индекс контекстного научного цитирования; мера семантического подобия; явные и неявные ссылки

DOI: 10.14357/08696527190306

1 Введение

Настоящая статья представляет часть проекта по разработке семантических методов построения *индекса контекстного научного цитирования*. Конечная цель — автоматизация получения нового показателя качества научной статьи — ИКНЦ, способного прогнозировать будущие значения стандартного *индекса научного цитирования*, который рассчитывается по числу прямых библиографических ссылок на статью. Проект включает создание корпусов естественноязыковых текстов, относящихся к выбранным предметным областям, на основе обработки большого потока сетевых данных для поиска дополнительных неявных связей между документами (неформальных ссылок). На выделенных тематических коллекциях был проведен ряд экспериментов [1–4].

Сбор текстов осуществлялся из открытых источников, в том числе из русской Википедии, из социальных сетей, а также проводился поиск статей по связям цитирования в Google. За счет круглосуточной работы двух серверов постоянно увеличивается объем корпусов, из которых были выделены подклассы экспериментальных (обучающих) выборок по темам «Компьютерная графика,

*Работа выполнена при частичной финансовой поддержке РФФИ (проекты 19-07-00857 и 16-07-00756).

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

[2]Институт проблем информатики Федерального исследовательского центра «Информатика и управление» Российской академии наук, mc@keywen.com

визуализация и виртуальное окружение в СССР и России (КГВ)», «Автономные необитаемые подводные аппараты» (АНПА), «Политический экстремизм» и др. Так, по теме АНПА был получен общий словарь словосочетаний объемом 53 000 терминов, из которого выделен словарь ключевых словосочетаний в 870 терминов. Полученный корпус, составлявший в 2016 г. более 100 ГБ (в том числе по тематике КГВ более 10 ГБ текстов, из которых было выделено ядро, содержащее 350 МБ различных неповторяющихся фраз), в 2017 г. был доведен до объема в 240 ГБ. Размер коллекции-2016 по тематике КГВ, составлявший почти 1000 статей, из которой был выделен экспериментальный корпус в 120 связанных между собой статей, в 2017 г. вырос более чем вдвое, а экспериментальный корпус расширился до 250 статей, связанных библиографическими ссылками [2].

2 Индекс контекстного научного цитирования

Методика оценки качества научных статей разрабатывается на основе вероятностной модели влияния (impact) научной статьи на ссылки и идеи в последующих статьях, а также на основе модели представления идеи в виде множества похожих по смыслу фраз [1, 2, 4]. Первая попытка применения ИКНЦ состоялась на корпусе по компьютерной графике (КГВ) [1]. Индекс контекстного научного цитирования рассчитывается автоматически по неявным контекстным ссылкам на статью и связан со статистической вероятностью ожидаемого появления прямых библиографических ссылок. Он позволяет делить новые статьи на группы и ранжировать их по качеству. Неявные ссылки в статье — это упоминания чужих идей и их авторов. Индекс контекстного научного цитирования полезен для обнаружения новых документов (научных статей), так чтобы наиболее значимые документы могли быть представлены в большем количестве. Обычно для обнаружения значимых научных документов используется стандартный ИНЦ, но через ИНЦ невозможно проанализировать новые документы с нулевым значением ИНЦ, а ИКНЦ способен их анализировать, так как учитывает и неявные ссылки на статью.

Вероятностная математическая модель зависимости числа прямых цитирований от числа неявных ссылок и их параметров строится на основе лингвистического процессора, выделяющего неявные ссылки [5]; процессор настраивается с помощью метода машинного обучения так, чтобы корреляция между индексами ИНЦ и ИКНЦ была максимальной. Базой для разработки принципов построения неявных ссылок между документами является подход, позволяющий осуществлять поиск релевантной информации не только по ключевым словам, но и по ассоциациям, опосредованным отношениям. В его основе лежит представление о том, что идеи могут быть адекватно выражены множеством сходных по смыслу фраз или множеством терминов (близким аналогом предлагаемого подхода выражения идей/тем является один из методов тематического моделирования: метод латентного размещения Дирихле — LDA).

При создании текстовых коллекций необходимо решить задачи определения семантического подобия документов и фраз, выявления семантических связей между терминами, а также классификации документов по категориям в автоматизированном режиме. Смысловое сходство фраз определяется с помощью грамматических трансформаций, переводных эквивалентов, синонимии и функциональной синонимии, а также с помощью обнаружения ассоциативных связей, выявленных по авторской методике построения ассоциативного портрета предметной области (АППО). Для структуризации текстов предметной области и построения иерархии категорий используется авторская методика построения ассоциативно-иерархического портрета предметной области (АИППО), где для расчета иерархических связей между значимыми фрагментами текста используются методы тематического моделирования (ТМ), такие как LDA [6]. Для улучшения параметров иерархической кластеризации терминов ТМ применяется совместно с моделями дистрибутивной семантики и семантических векторных пространств. Организация коллекции текстовых документов в виде кластеров производится так, чтобы документы в пределах каждого кластера были похожи. Текст обычно отображается в векторном пространстве (в виде «мешка слов»), и каждый документ представляется как вектор признаков с использованием схемы взвешивания. Затем выполняется кластеризация путем измерения расстояния между векторами признаков. Выделенные по указанным методикам ассоциативные и иерархические связи между значимыми словосочетаниями позволяют разрабатывать более совершенные методы и метрики/меры подобия научных текстов. Создаются тематические коллекции и выборки документов, связанных формальными библиографическим ссылками, а также содержащие сходные тексты, связанные неформально. Неявные связи обнаруживаются с помощью меры семантического подобия (введенный авторами показатель).

Описание методики построения ИКНЦ детально изложено в [4]. При формировании корпусов, содержащих взятые из интернета научные статьи, применяется авторская разработка «KeyCrawler для систем извлечения знаний» [7], осуществляющая направленный семантический поиск и сбор данных из открытых источников по заданным ключевым словам.

Индекс научного цитирования рассчитывается автоматически с помощью лингвистического процессора BREF (основан на Pullenti [5]), выделяющего из коллекции документов библиографические ссылки в виде направленного графа формальных ссылок. Далее подсчитывается число формальных ссылок (входящих дуг графа) на каждую статью. В данном проекте ИНЦ вычисляется не по базам РИНЦ, Scopus, Web of Science, а по текстовым корпусам — тематическим выборкам научных статей.

3 Мера подобия двух произвольных текстов (семантическое сходство)

Мера подобия (мера семантического сходства) — авторский показатель. Вычисление *семантической меры подобия* (*меры неявной связанности*) *между тек-*

стами ведется на созданных тематических корпусах. Этот показатель базируется на соотношении явных (библиографических, прямое цитирование) и неявных ссылок, когда тексты связаны через сходные текстовые фрагменты. Параметры алгоритма расчета данной меры оптимизируются по максимуму корреляции индексов ИНЦ и ИКНЦ. В простейшем случае мера неявного подобия (мера сходства) — measure of implicit connection (MIC) — рассчитывается по следующей формуле:

$$\mathrm{MIC} = \frac{S_{\mathrm{intersection}}}{\min(S_{\mathrm{text1}}, S_{\mathrm{text2}})} \cdot 100\%\,,$$

где $S_{\mathrm{intersection}}$ — сумма весов всех терминов (текстовых фрагментов), включенных в оба текста; S_{text1} — сумма весов ключевых словосочетаний (терминов) первого текста; S_{text2} — сумма весов ключевых словосочетаний (терминов) второго текста. Мера неявного подобия двух идентичных текстов равна 100%. Использование данной меры помогает находить пропущенные связи между статьями.

Выявление имплицитной связанности документов между собой ведется с помощью лингвистического процессора неявных ссылок (ЛПНС), обнаруживающего неявные ссылки с заданными параметрами (пороги минимального и максимального семантического подобия, учет времени появления статей и т. д.). Как и процессор BREF, ЛПНС представляет собой ответвление системы Pullenti [5] и настраивается с помощью машинного обучения так, чтобы корреляция между индексами ИНЦ и ИКНЦ была максимальной. В ЛПНС версии 2018 г. для сравнения семантического подобия текстов используется мера близости терминов на основе Word2Vec (технология нейронных сетей), а также мера WMD (Word Mover's Distance), которая хорошо оценивает расстояние между множествами терминов (для определения семантической близости тем и групп терминов) [8].

Автоматизированный расчет ИКНЦ ведется на основе текущей версии ЛПНС, работающего по методу обнаружения релевантных фраз. *Индекс контекстного научного цитирования вычисляется на материале текстов из тематических коллекций с помощью анализа текста статьи и сравнения его с текстами других статей с использованием меры семантического подобия (semantic similarity measure).* В одной из возможных реализаций алгоритма вычисления ИКНЦ равен числу неявных связей, параметры которых определяются через семантическое сходство, так что корреляция между ИКНЦ и ИНЦ максимизируется. Далее осуществляется преобразование информации об ИКНЦ в различную форму: по статьям (рейтинги статей и трехмерная визуализация статей с учетом значимости/индекса), по авторам, по организациям и т. д.

Рассмотрим один из экспериментов с ИКНЦ и мерой подобия, где использовались два корпуса по компьютерной графике (КГВ), содержащих 120 и 250 статей, связанных официальными библиографическими ссылками. Экспериментальный корпус из 120 связанных статей по теме КГВ, взятых из доступных в интернете российских научных публикаций, был создан в 2016 г. Первоначально нашлось более 1000 статей, откуда затем были автоматически выбраны 120 статей, свя-

занных формальной библиографической ссылкой. Мера подобия выбранных пар менялась от 1,39% до 99%. Большинство рассматриваемых пар имели меру подобия менее 4% и вероятность наличия связи менее 30%. Было принято, что между двумя текстами есть имплицитная связь, если их мера подобия составляет более 8%. Вероятность наличия связи по библиографическим ссылкам между такими текстами более 70%. В 2017 г. этот корпус был расширен до 250 связанных статей. Самая цитируемая статья в обоих корпусах цитируется 7 раз в корпусе из 120 статей и 13 раз в корпусе из 250 статей.

В статьях обеих коллекций представлено два типа связанности: формальная, когда тексты связаны библиографическими ссылками, и имплицитная — через общие (или сходные) текстовые фрагменты. Мера подобия (семантического сходства) основана на соотношении явных и неявных ссылок. Эксперимент показал, что оба типа связанности коррелированы. На первичном корпусе из 120 статей были найдены параметры оптимальной обработки текста, при которых корреляция максимальна и достигает примерно 55%. В простейшем случае вес текстового фрагмента равен длине этого фрагмента, если длина превышает определенный порог. Вес равен нулю, если длина меньше этого порога. Результаты эксперимента показаны на рис. 1. Вертикальная ось — это корреляция между формальной и имплицитной связанностью, а горизонтальная — число букв в пороге длины текстового фрагмента. Корреляция достигает максимума при пороге = 7.

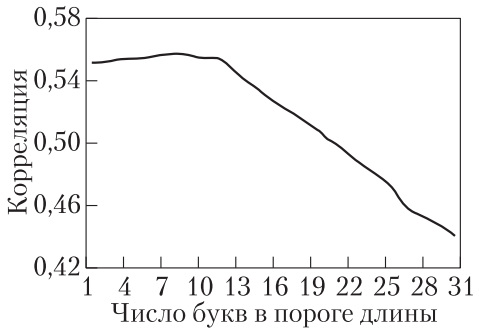


Рис. 1 Корреляция между явной и неявной связанностью в зависимости от порога длины

В эксперименте на коллекции по компьютерной графике ИНЦ равен числу формальных библиографических ссылок, а ИКНЦ рассчитывается как число неявных ссылок на каждый текст коллекции.

Важнейший *параметр неявной связи между двумя текстами* — это минимальное (пороговое) значение меры их подобия. Было принято, что две статьи имеют неявную ссылку между собой, если мера их семантического подобия выше этого порога. Результаты эксперимента на двух коллекциях отражены на рис. 2, где вертикальная ось представляет собой корреляцию между ИКНЦ и ИНЦ, в то время как горизонтальная ось представляет значение порога семантического подобия в сотых долях процента.

Как видно из рис. 2, оптимальный параметр (минимальное пороговое значение семантического сходства) составляет около 0,7% для коллекций из 120 и 250 статей. Другие оптимальные параметры можно найти аналогичным способом. При оптимизации сразу нескольких параметров (например, минимального

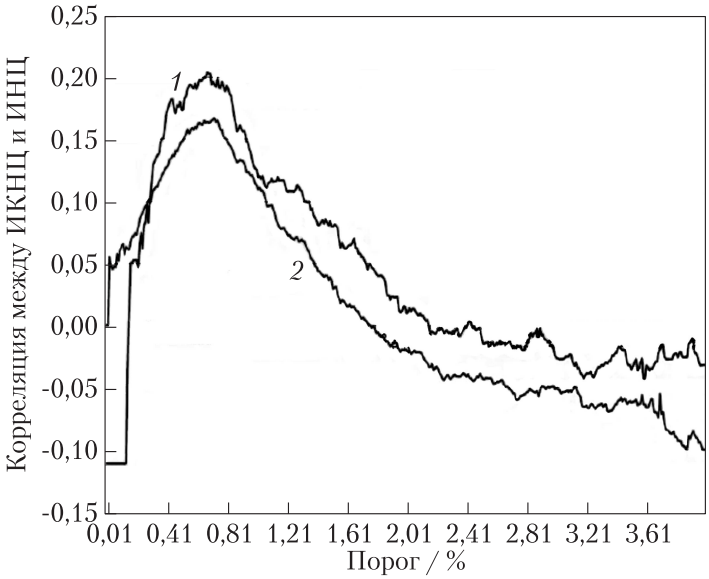


Рис. 2 Корреляция между ИКНЦ и ИНЦ в двух текстовых коллекциях: *1* — коллекция из 120 статей; *2* — коллекция из 250 статей

и максимального порога меры семантического подобия) корреляция между индексами ИКНЦ и ИНЦ достигала 40%. Это обнадеживающий результат, так как оптимальные параметры имеют схожие значения в разных коллекциях, что позволяет расширить объем применения данного метода и использовать его в разных предметных областях. Оптимальные параметры неявных связей, вычисленные при создании коллекций научных статей путем максимизации корреляции между ИКНЦ и ИНЦ, могут применяться для анализа других коллекций и доменов, даже тех, где относительно редко встречаются формальные ссылки (например, в текстах социальных сетей). Таким образом, использование ИКНЦ помогает обнаруживать наиболее значимые статьи и более точно отражать смысловое сходство между ними.

4 Математическая вероятностная модель зависимости числа прямых цитирований от числа неявных ссылок и их параметров

Рассматриваемая ниже вероятностная регрессионная модель зависимости числа формальных цитирований от числа неявных ссылок и их параметров является еще одной из возможных реализаций алгоритма вычисления ИКНЦ, которая позволяет рассчитать ИКНЦ в тех же единицах, что и ИНЦ, а также позволяет оценить качество и точность модели через коэффициент детерминации. Эта

модель рассчитывает предсказанное число формальных цитирований, которое удобно принять за нормализованное значение ИКНЦ и которое более привычно, чем число неявных ссылок.

Данная прогностическая модель строится на основе моделирования значений независимых переменных и их коэффициентов регрессии. Корреляция и регрессия — связанные между собой и широко используемые методы для определения силы связи между переменными. Корреляция представляет собой меру ассоциации, а регрессия обеспечивает средство прогнозирования одной зависимой переменной от других независимых переменных. Специальный лингвистический процессор ищет неявные ссылки с заданными параметрами. В качестве *математической вероятностной модели зависимости числа формальных цитирований от числа неявных ссылок и их параметров используется линейная регрессионная модель*. Регрессия определяется по следующей формуле:

$$\mathrm{Y} = \mathrm{f}(\mathrm{X}, \mathrm{A}) + \mathrm{epsilon};\ \ \mathrm{E}(\mathrm{epsilon}) = 0\,,$$

где A — параметры модели, epsilon — случайная ошибка модели. *Регрессия называется линейной (множественной), если функция регрессии* $\mathrm{f}(\mathrm{X}, \mathrm{A})$ *имеет вид*:

$$\mathrm{f}(\mathrm{X}, \mathrm{A}) = \mathrm{A}_0 + \mathrm{A}_1\mathrm{X}_1 + \mathrm{A}_2\mathrm{X}_2 + \cdots + \mathrm{A}_\mathrm{k}\mathrm{X}_\mathrm{k}\,,$$

где A_i — коэффициенты регрессии; X_i — объясняющие/влияющие/независимые переменные (факторы модели); k — количество факторов модели. Линейная регрессия моделирует связь между зависимой переменной и независимыми переменными путем подгонки линейной функции регрессии к наблюдаемым данным. Число неявных ссылок и их параметры считаются независимыми объясняющими переменными, а предсказанное число формальных цитирований — зависимой переменной.

Линейный коэффициент корреляции (или коэффициент корреляции Пирсона) изменяется от -1 до $+1$ и рассчитывается по формуле:

$$\mathrm{R}(\mathrm{X}, \mathrm{Y}) = \mathrm{sum}((\mathrm{X} - \mathrm{sred} * (\mathrm{X}))(\mathrm{Y} - \mathrm{sred}(\mathrm{Y})))/(\mathrm{sigma}(\mathrm{X}) * \mathrm{sigma}(\mathrm{Y})),$$

где $\mathrm{sigma}(\mathrm{X}) = \mathrm{sqrt}(\mathrm{sum}(\mathrm{X} - \mathrm{sred}(\mathrm{X})^2)$ — стандартное отклонение ряда/выборки X; $\mathrm{sigma}(\mathrm{Y}) = \mathrm{sqrt}(\mathrm{sum}((\mathrm{Y} - \mathrm{sred}(\mathrm{Y})^2))$ — стандартное отклонение ряда/выборки Y; $\mathrm{sred}(\mathrm{X}) = \mathrm{sum}(\mathrm{X})/\mathrm{n}$ — среднее арифметическое.

В формальной математической записи формула коэффициента корреляции имеет вид:

$$r_{xy} = \frac{\mathrm{cov}_{xy}}{\sigma_x \sigma_y} = \frac{\sum (x - \overline{x})(y - \overline{y})}{\sqrt{\sum (x - \overline{x})^2 \sum (y - \overline{y})^2}};$$

$$\overline{X} = \frac{1}{n}\sum_{t=1}^{n} X_t; \quad \overline{Y} = \frac{1}{n}\sum_{t=1}^{n} Y_t\,.$$

В модели линейной регрессии коэффициенты регрессии связаны с соответствующими коэффициентами корреляции. Зная корреляцию, можно рассчитать *коэффициент регрессии* по формуле:

$$A_i = R(X_i, Y) * (\text{sigma}(Y)/\text{sigma}(X_i)) ,$$

где $\text{sigma}(X_i)$ — среднеквадратичное отклонение соответствующего факторного признака, Y — число формальных ссылок. В предлагаемой регрессионной модели значение функции $f(X, A)$ — это предсказанное число формальных цитирований, а объясняющие переменные X_i — факторы, влияющие на число цитирований. Такими факторами являются семантическое подобие текстов (semantic similarity), индекс Хирша автора, рейтинг журнала/конференции/источника и т. д.

Фактор X_1 (семантическое подобие текста рассматриваемой статьи другим текстам) — наиболее прямой показатель качества статьи и ее влияния на другие тексты. При таком подходе результаты исследования можно будет широко применять в различных предметных областях, где автор и рейтинг источника часто неизвестны. Фактор семантического подобия статьи — это некоторая функция, которая рассчитывает число неявных ссылок на статью в зависимости от текста самой статьи, текстов в анализируемой коллекции, параметров алгоритма расчета неявных ссылок. *Формулу фактора подобия* X_1 можно представить в виде некоторой функции следующего вида: $X_1 = F_{sim}$ (текст статьи, тексты коллекции, тексты в интернете, параметры неявных ссылок). Параметры неявных ссылок — это параметры алгоритма расчета семантического подобия текстов, такие как минимальный и максимальный пороги семантического подобия, минимальная длина учитываемых терминов, а также параметры, определяющие необходимость использования других подходов для выявления этих ссылок. В дальнейшем предполагается применение следующих методов: Word2Vec, Word Mover's Distance (WMD), дискриминационные корзины терминов (Discriminative term bucketing) и т. д.

Таким образом, в проекте исследуются различные функции расчета неявных ссылок $F_{sim}()$, при которых достигается максимум корреляции между фактором $X_1 = F_{sim}()$ и числом формальных ссылок Y. Чем выше корреляция, тем качественнее регрессионная модель и тем большую долю вариации Y (т. е. ИНЦ) она объясняет. Для определения качества регрессионной модели используется коэффициент детерминации. *Коэффициент детерминации* R^2 — это доля дисперсии зависимой переменной, объясняемая рассматриваемой моделью зависимости, т. е. объясняющими переменными. Коэффициент детерминации равен квадрату (множественного) коэффициента корреляции.

5 Заключение

Для выявления связей между научными статьями, а также оценки их качества предложены новые показатели — ИКНЦ и мера семантического подобия.

Получены текстовые корпуса, в которых выявлены формальные и неявные ссылки между документами. Предложена методика автоматизированного построения ИКНЦ. Экспериментально выделен один из параметров алгоритма (минимальная длина анализируемых слов), при котором корреляция между ИНЦ и ИКНЦ достигает максимума. Найден один из важных параметров неявной связи между двумя текстами — минимальное (пороговое) значение меры подобия этих текстов. Было принято, что две статьи имеют неявную ссылку между собой, если мера их семантического подобия выше этого порога. Алгоритм расчета ИКНЦ использует меру семантического подобия текстов, настраиваемую по максиму корреляции между библиографическими и неявными ссылками. Описана математическая вероятностная модель зависимости числа прямых цитирований от числа неявных ссылок и их параметров.

В дальнейшем планируется:

(1) улучшить оценку ИКНЦ и меры семантического подобия, используя Word2Vec, Doc2Vec, WMD и другие новые методы;

(2) добавить к оценке параметры даты публикации статей и направления неявных ссылок;

(3) проверить результаты в других предметных областях и на других языках;

(4) построить многоязычный ИКНЦ и рассчитывать неявные ссылки между документами на различных языках, используя базу данных BabelNet.org;

(5) создать улучшенную версию ИКНЦ, которая будет зависеть не только от числа неявных ссылок, но также и от их качества.

Литература

1. *Клименко С. В., Шарнин М. М., Хакимова А. Ф., Золотарев О. В., Мацкевич А. Г.* Методы оценки качества и влияния (impact) научных статей для повышения объективности индекса научного цитирования // Вестник РосНОУ, Сер. 3: Сложные системы: модели, анализ, управление, 2016. Вып. 3. С. 51–59.
2. *Charnine M., Klimenko S.* Semantic cyberspace of scientific papers // Conference (International) on Cyberworlds. — IEEE, 2017. P. 146–149.
3. *Демидов А. О., Шарнин М. М.* Построение развивающейся во времени системы тематических категорий для пополняемого корпуса текстов // Труды Междунар. научн. конф. CPT1617. — Протвино: ИФТИ, 2017. С. 166–171.
4. *Клименко С. В., Шарнин М. М., Галина И. В., Демидов А. О.* Методика построения индекса контекстного научного цитирования (ИКНЦ) // Труды Междунар. научн. конф. по физико-технической информатике CPT2018. — Протвино: ИФТИ, 2018. С. 1–11.
5. *Золотарев О. В., Шарнин М. М., Клименко С. В., Кузнецов К. И.* Система PullEnti — извлечение информации из текстов естественного языка и автоматизированное построение информационных систем // Ситуационные центры и информационно-аналитические системы класса 4i для задач мониторинга и безопасности: Междунар. школа-семинар (Пущино, 21–24 ноября 2015). — Протвино: ИФТИ, 2016. С. 14–21.

6. *Золотарев О. В., Шарнин М. М., Клименко С. В., Мацкевич А. Г.* Исследование методов автоматического формирования ассоциативно-иерархического портрета предметной области // Вестник РосНОУ. Сер. 3: Сложные системы: модели, анализ, управление, 2018. Вып. 1. С. 91–96.
7. *Charnine M.* Keywen: Automated writing tools. — Booktango, 2013. 129 p.
8. *Соколов Е. Г., Казанцев А. С., Шарнин М. М.* Применение технологии нейронных сетей и векторных операций для выявления объектов различных классов и их связей в большом корпусе текстов // Ситуационные центры и информационно-аналитические системы класса 4i для задач мониторинга и безопасности: Труды Междунар. конф. — Протвино: ИФТИ, 2017. С. 131–135.

Поступила в редакцию 08.10.18

THE SCIENCE CONTEXTUAL CITATION INDEX

I. V. Galina and M. M. Charnine

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: A new indicator of the quality of a scientific article — the science contextual citation index (SCCI) and the relationship between the SCCI and the science citation index (SCI) with another author's indicator — a similarity measure (a semantic similarity measure) of two arbitrary texts are considered. The results of experiments with these parameters are given, in particular, the correlation between SCCI and SCI, which depends on the value of the semantic similarity threshold, is studied. Based on modeling the values of independent variables and their regression coefficients, a predictive mathematical probability model is proposed for the dependence of the number of direct citations on the number of implicit references and their parameters.

Keywords: automated systems; science contextual citation index; semantic similarity measure; explicit and implicit reference

DOI: 10.14357/08696527190306

Acknowledgments

The work was partly supported by the Russian Foundation for Basic Research (projects 19-07-00857 and 16-07-00756).

References

1. Klimenko, S. V., M. M. Sharnin, A. F. Khakimova, O. V. Zolotarev, and A. G. Matskevich. 2016. Metody otsenki kachestva i vliyaniya (impact) nauchnykh statey dlya

povysheniya ob"ektivnosti indeksa nauchnogo tsitirovaniya [Methods of assessing the quality and influence (impact) of scientific articles to improve the objectivity of the science citation index]. *Vestnik Ross. nov. univ.* 3:51–59.

2. Charnine, M., and S. Klimenko. 2017. Semantic cyberspace of scientific papers. *Conference (International) on Cyberworlds Proceedings*. IEEE. 146–149.
3. Demidov, A. O., and M. M. Sharnin. 2017. Postroenie razvivayushcheysya vo vremeni sistemy tematicheskikh kategoriy dlya popolnyaemogo korpusa tekstov [Building a developing system of topic categories for an expanding text corpus]. *Conference (International) CPT1617 Proceedings*. Protvino: IFTI. 166–171.
4. Klimenko, S. V., M. M. Sharnin, I. V. Galina, and A. O. Demidov. 2018. Metodika postroeniya indeksa kontekstnogo nauchnogo tsitirovaniya (IKNTs) [The method for construction of science contextual citation index (SCCI)]. *Conference (International) CPT2018 Proceedings*. Protvino: IFTI. 1–11.
5. Zolotarev, O. V., M. M. Sharnin, S. V. Klimenko, and K. I. Kuznetsov. 2016. Sistema PullEnti — izvlechenie informatsii iz tekstov estestvennogo yazyka i avtomatizirovannoe postroenie informatsionnykh sistem [System PullEnti — information extraction from natural language texts and automatic construction of information systems]. *School-Seminar (International) on Situational Centers and Information-Analytical System 4i Class for Monitoring and Security Tasks Proceedings*. Protvino: IFTI. 28–35.
6. Zolotarev, O. V., M. M. Sharnin, S. V. Klimenko, and A. G. Matskevich. 2018. Issledovanie metodov avtomaticheskogo formirovaniya assotsiativno-ierarkhicheskogo portreta predmetnoy oblasti [Research of methods of automatic formation of associative and hierarchical portrait of the subject area]. *Vestnik Ross. nov. univ.* 3(1):91–96.
7. Charnine, M. 2013. *Keywen: Automated writing tools*. Booktango. 129 p.
8. Sokolov, E. G., A. S. Kazantsev, and M. M. Sharnin. 2017. Primenenie tekhnologii neyronnykh setey i vektornykh operatsiy dlya vyyavleniya ob"ektov razlichnykh klassov i ikh svyazey v bol'shom korpuse tekstov [Application of neural networks technology and vector operations for discovering different classes of objects and their relationships in a large collection of texts]. *Conference (International) SCVRT2017 Proceedings*. Protvino: IFTI. 131–135.

Received October 8, 2018

Contributors

Galina Irina V. (b. 1965) — leading engineer, 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;; irn_gl@mail.ru

Charnine Mikhail M. (b. 1959) — Candidate of Science (PhD) in technology, 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; mc@keywen.com

СИСТЕМЫ И СРЕДСТВА ИНФОРМАТИКИ, 2019. Т. 29. № 3. С. 77–91

НАДКОРПУСНЫЕ БАЗЫ ДАННЫХ В ЛИНГВИСТИЧЕСКИХ ПРОЕКТАХ*

***А. Ю. Егорова*[1], *И. М. Зацман*[2], *О. С. Мамонова*[3]**

Аннотация: Рассматривается задача обеспечения лингвистических исследований средствами надкорпусных баз данных (НБД), содержащих выровненные параллельные тексты (каждый из которых включает оригинальный текст и его перевод), а также двуязычные аннотации исследуемых языковых единиц (ЯЕ) и их переводов, сформированные на основе параллельных текстов. Каждая аннотация, формируемая лингвистом, фиксирует некоторую модель перевода ЯЕ. Опыт выполнения в ФИЦ ИУ РАН ряда лингвистических проектов показал, что далеко не все модели перевода, извлекаемые лингвистами из параллельных текстов в процессе лингвистического аннотирования с помощью НБД, описаны в двуязычных словарях и справочниках. Отличительная черта НБД состоит в том, что они позволяют получать новое знание о переводных соответствиях. Оно извлекается лингвистами при сопоставлении и аннотировании предложений оригинального текста и их переводов. Описание функций НБД, которые обеспечивают получение лингвистами нового знания в процессе аннотирования, является основной целью статьи.

Ключевые слова: надкорпусная база данных; лингвистическое аннотирование; языковая единица; корпусная лингвистика; модели перевода

DOI: 10.14357/08696527190307

1 Введение

Надкорпусные базы данных представляют собой информационный ресурс, создаваемый в обеспечение фундаментальных и прикладных лингвистических исследований. Они содержат выровненные параллельные тексты, каждый из которых включает оригинальный текст и его перевод, а также двуязычные аннотации исследуемых ЯЕ и их переводов, сформированные на основе параллельных текстов [1–8].

Каждая аннотация, формируемая лингвистом, описывает некоторую модель перевода ЯЕ. Опыт выполнения в ФИЦ ИУ РАН ряда лингвистических проек-

*Работа выполнена в Институте проблем информатики ФИЦ ИУ РАН при частичной поддержке РФФИ (проект 18-07-00192).

[1]Институт проблем информатики Федерального исследовательского центра «Информатика и управление» Российской академии наук, ann.shurova@gmail.com

[2]Институт проблем информатики Федерального исследовательского центра «Информатика и управление» Российской академии наук, izatsman@yandex.ru

[3]Факультет иностранных языков и регионоведения Московского государственного университета им. М. В. Ломоносова, mamonovaoks@mail.ru

тов показал, что далеко не все модели перевода, извлекаемые из параллельных текстов в процессе аннотирования, описаны в двуязычных словарях и справочниках. Сопоставление извлеченных моделей перевода с уже известными позволяет увидеть и описать новые модели, т. е. получить новое знание (= не описанное в словарях и справочниках) о переводных соответствиях [9–12]. Кроме обнаружения и описания моделей перевода НБД используется для извлечения и описания тех значений ЯЕ, которые не представлены в словарях [13–15].

Новое знание о моделях перевода и значениях ЯЕ извлекается лингвистами в процессе лингвистического аннотирования [16, 17] предложений оригинального текста, содержащих исследуемые ЯЕ, и их переводов. Поэтому одна из основных задач НБД состоит в реализации поиска в тексте тех предложений, которые включают такие единицы. Основная цель статьи состоит в описании функций НБД, обеспечивающих получение лингвистами нового знания в процессе аннотирования и поиска сформированных аннотаций.

2 Сопоставление функций параллельных корпусов и надкорпусных баз данных

Одна из актуальных задач компьютерной лингвистики — формирование информационно-компьютерных инструментов в виде параллельных корпусов и баз данных, содержащих тексты и их переводы. Пользователи, «работающие с параллельными текстами, получают весьма простой и эффективный инструмент сбора материала. Ценность этого инструмента определяется тем, что в лингвистике этап сбора материала является наиболее трудоемким и наименее творческим, а подобные корпусы параллельных текстов позволят сэкономить время и силы для собственно исследовательской работы... Корпус параллельных текстов может быть эффективно использован в различных лингвистических исследованиях (в первую очередь, в области контрастивной лексикологии и двуязычной лексикографии), а также в исследованиях по теории перевода, сравнительного литературоведения, культурологии, автоматической обработки текста и др.» [1].

Функции НБД и корпусов параллельных текстов полностью совпадают в том, что они служат эффективными инструментами накопления, хранения и поиска текстового материала и его перевода. Однако между ними есть и принципиальное отличие: НБД, в отличие от корпусов, дает еще и возможность аннотировать ЯЕ и их переводы, накапливать, хранить и искать аннотации по широкому спектру поисковых параметров [4–6]. Иначе говоря, объектами хранения и поиска в НБД являются одновременно и текстовый материал, и двуязычные аннотации, сформированные лингвистами в процессе его анализа, а объектами хранения и поиска в параллельных корпусах — только исходный текстовый материал.

Например, если лингвист исследует ЯЕ некоторой категории (модальные глаголы, дискурсивные слова, коннекторы или др.), то в параллельном корпусе он может найти предложения, где они встречаются. Потом он может сопоставить

Структура хранения информационных ресурсов

эти предложения с их переводами и провести контрастивный анализ переводных соответствий, а также сравнить употребления исследуемой ЯЕ с ее переводами. Однако параллельные корпусы не дают возможности накапливать результаты анализа, так как они по определению не предназначены для этого. В НБД реализованы возможности лингвистического аннотирования ЯЕ и их переводов, а также функции накопления, хранения и поиска двуязычных аннотаций, описывающих результаты лингвистического анализа. Как будет показано ниже на примере исследования имплицитности в тексте[1], реализация именно этих функций обеспечивает извлечение нового знания лингвистами с помощью НБД. Необходимость поддержки этих функциональных возможностей нашла свое отражение в структуре хранения информационных ресурсов НБД (см. рисунок).

Этот рисунок иллюстрирует две основные особенности структуры хранения информационных ресурсов для лингвистических исследований с помощью НБД. Во-первых, одному хранилищу параллельных текстов могут соответствовать несколько НБД, каждая из которых хранит аннотации ЯЕ одной категории. Это дает возможность адаптировать НБД к исследованию той или иной категории ЯЕ. Во-вторых, связь между хранилищем параллельных текстов и каждой НБД односторонняя, т. е. НБД могут использовать общее для них хранилище параллельных текстов только в режиме чтения. Накопление двуязычных аннотаций, их хранение и поиск реализованы с помощью НБД.

Разработанная структура хранения позволяет использовать единую технологию для пополнения корпусов параллельных текстов Национального корпуса

[1]Приведем пример имплицитности отношения причины в тексте: «не ждите меня, уже поздно» (ср. с «не ждите меня, так как уже поздно», где отношение причины выражено с помощью «так как»).

русского языка (НКРЯ) и формирования хранилища параллельных текстов, используемых всеми существующими и вновь создаваемыми НБД.

3 Формирование параллельных текстов

Подготовка текстов оригинала и перевода каждого произведения включает операции редактирования обоих текстов, которые уже представлены в электронном виде, и их выравнивания. Сначала и оригинал текста, и его перевод проверяются на предмет опечаток. Потом «в текстах удаляется нумерация страниц, знаки табуляции заменяются пробелами, разрывы страниц — знаками абзаца, лишние пробелы и мягкие переносы удаляются, знаки тире унифицируются. Цель этих операций — нормализация знаков форматирования, что в дальнейшем поможет избежать сбоев при использовании программы выравнивания параллельных текстов» [18].

Затем выполняется программа «Евклид» [19], созданная на основе программы выравнивания текстов HunAlign [20]. Процесс выравнивания состоит в соотнесении одного или нескольких предложений текста оригинала и соответствующих им предложений текста перевода. Из предложений оригинала каждого произведения и их переводов формируется множество кортежей. В табл. 1 приведены три примера кортежей из книги И. А. Гончарова «Обломов» в переводе Л. Юргенсон.

Все кортежи в табл. 1 включают по одному предложению в оригинальном тексте и в переводе. Но в общем случае каждый кортеж может состоять из нескольких предложений оригинального текста слева и их переводов справа [18].

Если оригинальный текст имеет два или более переводов на один и тот же язык, то для пополнения корпусов параллельных текстов НКРЯ и формирования хранилища параллельных текстов, используемых всеми НБД, оригинал текста выравнивается с каждым из этих переводов по отдельности. При этом число кортежей, сформированных при выравнивании оригинала с первым переводом, и содержание их левых частей служат основой для выравнивания оригинала с каждым из последующих переводов. Таким образом, после выравнивания

Таблица 1 Кортежи предложений оригинала и перевода

Оригинал	Перевод
Не просите меня петь, я не спою уже больше так...	Ne me demandez pas de chanter, car je ne saurais plus chanter comme ça...
Постойте, еще одно спою... — сказала она, и в ту же минуту лицо ее будто вспыхнуло, глаза загорелись, она опустилась на стул, сильно взяла два–три аккорда и запела.	Si, attendez, sauf une chose encore... dit-elle, et au même instant son visage s'embrasa de couleurs, ses yeux brillèrent, elle prit une chaise, plaqua deux ou trois accords vigoureux et chanta.
Боже мой, что слышалось в этом пении!	Seigneur, qu'entendait-on dans ce chant!

Таблица 2 Оригинал с двумя переводами

Оригинал	Переводы
Не просите меня петь, я не спою уже больше так... [И. А. Гончаров. Обломов (1848–1859)]	Ne me demandez plus de chanter, **car** je ne serais plus capable de chanter ainsi... [Пер. А. Адамов (1959)]
	Ne me demandez pas de chanter, **car** je ne saurais plus chanter comme ça... [Пер. Л. Юргенсон (1988)]

с каждым из переводов по отдельности левые части кортежей одного произведения при соблюдении технологии выравнивания будут совпадать, а правые, как правило, будут отличаться, так как они созданы разными переводчиками (табл. 2).

Таблица 3 Пример разметки словоформы *спою* без снятия омонимии

Лемма	Код	Название
споить	norm	словарная форма
	sg	единственное число
	V	Глагол
	pf	совершенный вид
	act	действительный залог
	indic	изъявительное наклонение
	1p	первое лицо
	tran	переходный
	fut	будущее время
спой	S	Существительное
	inan	неодушевленное
	norm	словарная форма
	sg	единственное число
	dat	дательный падеж
	m	мужской род
спеть	norm	словарная форма
	sg	единственное число
	V	Глагол
	pf	совершенный вид
	act	действительный залог
	indic	изъявительное наклонение
	1p	первое лицо
	tran	переходный
	fut	будущее время

До загрузки выровненных параллельных текстов выполняется метатекстовая и морфологическая разметка согласно Морфологическому стандарту НКРЯ [21]. По этому стандарту «Морфологическая информация, приписываемая произвольному слову в тексте, состоит из четырех “полей” или групп помет: (1) лексема, которой принадлежит словоформа (указывается “словарная запись” данной лексемы и ее принадлежность к той или иной части речи); (2) множество грамматических признаков данной лексемы, или словоклассифицирующие характеристики (например, род для существительного, переходность для глагола); (3) множество грамматических признаков данной словоформы, или словоизменительные характеристики (например, падеж для существительного, число для глагола); (4) информация о нестандартности грамматической формы, орфографических искажениях и т. п. ... В основу метаязыка грамматических помет... положена система сокращенных помет (“тегов”) на основе латинского алфавита» (см. 2-й столбец в табл. 3).

Перечисленные поля включают лингвистическую информацию о словоформах и соответствующих им лексемах. При этом разметка проводится без снятия омонимии. Так, для словоформы *спою* из табл. 1 и 2 при разметке включается информация о трех лексемах *споить*, *спой* (существительное) и *спеть* (см. табл. 3).

Таким образом, часть грамматических признаков (часть речи, залог, переходность, совершенный вид) характеризует лексему, а наклонение, время, число относятся к самой словоформе. На основе этих признаков в НБД реализован грамматический поиск исследуемых ЯЕ. При этом НБД обеспечивает двуязычный поиск, т. е. его можно вести по текстам оригиналов произведений и/или по их переводам.

4 Поиск исследуемых языковых единиц

Поиск в параллельных текстах кортежей с исследуемыми ЯЕ включает два этапа. Сначала выбирается исследуемое направление перевода (табл. 4), т. е. в хранилище параллельных текстов НБД выделяется их подмножество. Например, на 01.06.2019 в хранилище были загружены тексты, содержащие около 14 млн словоупотреблений. Если выбрать в нем только русско-французское направление перевода, то область поиска в хранилище сократится до 4 млн словоупотреблений. Область поиска сократится еще больше, если указать конкретное произведение и его перевод. Отметим, что хранилище НБД содержит тексты разных эпох и жанров — как художественные, так и нехудожественные (в том числе научные, юридические тексты).

После выбора направления перевода открывается окно поискового шаблона, который дает возможность задать лексические и/или грамматические критерии поиска кортежей с исследуемыми ЯЕ. На основе найденных кортежей формируются двуязычные аннотации, содержащие контекст оригинального текста с исследуемой ЯЕ, признаки ЯЕ и ее контекста, а также перевод контекста с эквивалентом ЯЕ и признаки, проставленные для перевода. Отметим, что иногда переводной эквивалент ЯЕ, например коннектора, может отсутствовать. Или наоборот, в переводе коннектор есть, а в оригинале коннектора нет. Далее

Таблица 4 Количественные характеристики хранилища

Направление перевода	Словоупотреблений в оригинальных текстах	Словоупотреблений в переводах	Всего словоупотреблений в направлении
ру-фр	1 659 239	2 284 898	3 944 137
фр-ру	649 245	509 033	1 158 278
ру-нем	910 279	1 168 594	2 078 873
нем-ру	1 390 979	1 220 769	2 611 748
ру-ит	1 515 627	1 889 353	3 404 980
ит-ру	384 111	340 445	724 556

покажем, как поиск кортежей и аннотаций обеспечивает получение лингвистами нового знания об имплицитности в параллельных текстах.

5 Извлечение нового знания с помощью надкорпусных баз данных

В рамках проекта РФФИ «Метод и информационная технология для целенаправленного формирования новых лингвистических типологий» разработаны два метода получения нового лингвистического знания. Первый метод предназначен для целенаправленного обнаружения и заполнения лакун в лингвистических типологиях, которые являются формами представления знаний. Процесс решения этой задачи включает 5 многократно повторяемых стадий, которые в совокупности образуют одну итерацию решения задачи целенаправленного извлечения знаний из параллельных текстов для заполнения лакун [14, 15, 18, 22]:

(1) определение одной или нескольких ЯЕ, исследуемых на текущей итерации, а также поиск кортежей, которые содержат предложения с исследуемыми ЯЕ и их переводными эквивалентами;

(2) лингвистическое аннотирование кортежей, найденных в параллельных текстах на первой стадии;

(3) семантический анализ незавершенных аннотаций, помеченных рубриками, которые указывают на причины незавершенности процесса аннотирования на второй стадии (в том числе на лакуны в существующей типологии исследуемых ЯЕ);

(4) доработка аннотаций, включающих рубрики, которые указывают на лакуны в типологии как на причину незавершенности, и пополнение существующей типологии;

(5) определение значений параметров, характеризующих состояние процесса пополнения типологии на момент завершения каждой итерации.

Второй метод, разработанный А. А. Гончаровым [23, 24], предназначен для получения с помощью НБД нового лингвистического знания о логико-семантических отношениях (ЛСО), которые имплицированы или косвенно выражены в левой части кортежей, но эксплицированы коннектором в правой его части, т. е. в переводе (см. табл. 1 и 2). Реализуемость этого метода проверена экспериментально на примере выровненных русско-французских параллельных текстов объемом около 4 млн словоупотреблений (см. 1-ю строку табл. 4). В рамках данного эксперимента из этих текстов был отобран 381 кортеж, содержащий справа французский коннектор *car*, который используется для выражения ЛСО причины в текстах на французском языке, а слева ЛСО имплицировано или выражено косвенно. При этом была проведена предварительная типологизация отобранных 381 кортежа по четырем видам. Наиболее частотным оказался вид кортежей, в левой части которых ЛСО причины имплицировано (см. 1-ю строку табл. 1 и табл. 2), т. е. не выражено ни коннектором, ни одним из косвенных способов.

К косвенным способам выражения ЛСО относятся: его замена на другое отношение (так, в (1) коннектор *а* выражает прежде всего сопоставительное ЛСО), передача ЛСО неким грамматическим средством (например, приложением, как в (2)) или с помощью пунктуации (в (3) для этого использовано тире).

(1) Она только кляла себя, зачем она вначале не победила стыда и не открыла Штольцу раньше прошедшее, **а** теперь ей надо победить еще ужас [И. А. Гончаров. Обломов (1848–1859)].

Elle se maudit alors de n'avoir pas plus tôt surmonté sa honte et confié son passé à Stolz, **car** maintenant il lui fallait surmonter aussi la terreur [Trad. par Luba Jurgenson (1988)].

(2) Пред отъездом в Москву она, **вообще мастерица одеваться не очень дорого**, отдала модистке для переделки три платья [Л. Н. Толстой. Анна Каренина (1873–1877)].

Avant son départ pour Moscou, elle avait donné trois robes à transformer, **car** elle savait à merveille s'habiller à bon compte [Trad. par Henri Mongault (1952)].

(3) . . .я подумал, что завтра-то уж ему качать головой не придется — заблещу медалью новой, как на строевом смотру [Аркадий Вайнер, Георгий Вайнер. Эра милосердия (1975)].

Je me pris à penser que le lendemain il ne pourrait plus me reprocher quoi que ce soit, **car** je brillerais comme une médaille un jour de revue [Trad. par Jean-Pierre Dussaussois et Evgueni Avrorine, en coll. avec Jean-Georges Synakiewicz. Revu et corrigé par Béatrice Durupt (2005)].

Разработанный метод исследования имплицитных и косвенно выраженных ЛСО включает 5 стадий.

1. Выбор исследуемого ЛСО (в эксперименте это было ЛСО причины) и лингвистическое аннотирование употреблений тех коннекторов русского языка, которые могут считаться прототипическими средствами выражения этого ЛСО (в эксперименте это были коннекторы *потому что*, *поскольку*, *ибо* и т. д.).
2. Анализ полученного массива аннотаций и выявление переводных эквивалентов этих коннекторов во французском языке (для ЛСО причины — это *car*, *parce que*, *puisque*, *comme* и т. д., но в проведенном эксперименте использовался только коннектор *car*).
3. Двуязычный поиск в хранилище кортежей (см. рисунок), удовлетворяющих условию: правая часть кортежа на французском языке должна содержать выявленные на второй стадии переводные эквиваленты (в эксперименте — это коннектор *car*), а левая его часть не должна содержать те коннекторы русского языка, которые выражают исследуемое ЛСО и были аннотированы на первой стадии (*потому что*, *поскольку*, *ибо* и т. д.).
4. Аннотирование кортежей, найденных в хранилище на 3-й стадии (в эксперименте было найдено 444 кортежа, из них 381 кортеж был релевантен цели исследования).

5. Анализ сформированных в НБД аннотаций (в эксперименте сформирована 381 аннотация), включая определение частотности видов кортежей, в которых исследуемое ЛСО в русском языке является имплицитным или выражено одним из косвенных способов (в эксперименте с коннектором *car* было найдено чуть более 40% кортежей с имплицитным ЛСО причины и около 60% кортежей с косвенными способами его выражения).

Этот метод дает возможность определить частотность видов кортежей с имплицитным и выраженным одним из косвенных способов ЛСО любого вида (генерализации, спецификации, противопоставления и т. д.), используя поиск кортежей в хранилище параллельных текстов и двуязычный поиск аннотаций в НБД. Данный метод по сравнению с самым распространенным на сегодняшний день способом выявления имплицитных и косвенно выраженных ЛСО, который использует сплошную ручную разметку от начала до конца каждого текста [25, 26], требует значительно меньших временны́х и человеческих ресурсов для проведения лингвистических исследований.

6 Заключение

Разработанные методы используют поисковые возможности спроектированной М. Г. Кружковым [3–5, 8, 9] НБД, которая обеспечивает двуязычный поиск аннотаций. Эти методы позволяют лингвистам извлекать новое знание, включая уточнение уже существующих типологий [15]. Появляется также возможность формировать новые типологии с нуля, что планируется продемонстрировать при продолжении проекта на примере построения типологии русских конструкций с модальным значением. В существующих методах уточнения лингвистических типологий в процессе аннотирования не допускается никаких изменений в типологии. После завершения работ по аннотированию в типологию могут быть внесены изменения, но в процессе аннотирования она остается неизменной до его окончания [27]. Такие подходы не позволяют формировать типологии с нуля в процессе аннотирования, так как для его обеспечения уже требуется иметь некоторый начальный вариант типологии.

Литература

1. *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.
2. *Loiseau S., Sitchinava D. V., Zalizniak Anna A., Zatsman I. M.* Information technologies for creating the database of equivalent verbal forms in the Russian–French multivariant parallel corpus // Информатика и её применения, 2013. Т. 7. Вып. 2. С. 100–109.
3. *Kruzhkov M. G., Buntman N. V., Loshchilova E. Ju., Sitchinava D. V., Zalizniak Anna A., Zatsman I. M.* A database of Russian verbal forms and their French translation

equivalents // Компьютерная лингвистика и интеллектуальные технологии: По матлам Междунар. конф. «Диалог». — М.: РГГУ, 2014. Вып. 13(20). С. 284–296.

4. *Кружков М. Г.* Информационные ресурсы контрастивных лингвистических исследований: электронные корпуса текстов // Системы и средства информатики, 2015. Т. 25. № 2. С. 140–159.
5. *Зализняк Анна А., Зацман И. М., Инькова О. Ю., Кружков М. Г.* Надкорпусные базы данных как лингвистический ресурс // Корпусная лингвистика-2015: Труды 7-й Междунар. конф. — СПб.: СПбГУ, 2015. С. 211–218.
6. *Зализняк Анна А.* База данных межъязыковых эквиваленций как инструмент лингвистического анализа // Компьютерная лингвистика и интеллектуальные технологии, 2016. Вып. 15(22). С. 854–866.
7. *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. 100–106.
8. *Зацман И. М., Кружков М. Г.* Надкорпусная база данных коннекторов: развитие системы терминов проектирования // Системы и средства информатики, 2018. Т. 28. № 4. С. 156–167.
9. *Zatsman I., Buntman N., Kruzhkov M., Nuriev V., Zalizniak Anna A.* Conceptual framework for development of computer technology supporting cross-linguistic knowledge discovery // 15th European Conference on Knowledge Management Proceedings. — Reading, U.K.: Academic Publishing International Ltd., 2014. Vol. 3. P. 1063–1071.
10. *Zatsman I., Buntman N.* Outlining goals for discovering new knowledge and computerised tracing of emerging meanings // 16th European Conference on Knowledge Management Proceedings. — Reading, U.K.: Academic Publishing International Ltd., 2015. P. 851–860.
11. *Zatsman I., Buntman N., Coldefy-Faucard A., Nuriev V.* WEB knowledge base for asynchronous brainstorming // 17th European Conference on Knowledge Management Proceedings. — Reading, U.K.: Academic Publishing International Ltd., 2016. Vol. 1. P. 976–983.
12. *Zatsman I.* Goal-oriented creation of individual knowledge: Model and information technology // 19th European Conference on Knowledge Management Proceedings. — Reading, U.K.: Academic Publishing International Ltd., 2018. Vol. 2. P. 947–956.
13. *Добровольский Д. О., Зализняк Анна А.* Немецкие конструкции с модальными глаголами и их русские соответствия: проект надкорпусной базы данных // Компьютерная лингвистика и интеллектуальные технологии: По мат-лам Междунар. конф. «Диалог». — М.: РГГУ, 2018. С. 172–184.
14. *Зацман И. М.* Стадии целенаправленного извлечения знаний, имплицированных в параллельных текстах // Системы и средства информатики, 2018. Т. 28. № 3. С. 175–188.
15. *Зацман И. М.* Целенаправленное развитие систем лингвистических знаний: выявление и заполнение лакун // Информатика и её применения, 2019. Т. 13. Вып. 1. С. 91–98.
16. Handbook of linguistic annotation / Eds. N. Ide, J. Pustejovsky. — Dordrecht, The Netherlands: Springer Science + Business Media, 2017. 1468 p.

17. *Гончаров А. А., Инькова О. Ю., Кружков М. Г.* Методология аннотирования в надкорпусных базах данных // Системы и средства информатики, 2019. Т. 29. № 2. С. 148–160.
18. *Гончаров А. А., Зацман И. М.* Информационные трансформации параллельных текстов в задачах извлечения знаний // Системы и средства информатики, 2019. Т. 29. № 1. С. 180–193.
19. *Сичинава Д. В.* Параллельные тексты в составе Национального корпуса русского языка: новые направления развития и результаты // Труды Института русского языка им. В. В. Виноградова, 2015. № 6. С. 194–235.
20. *Varga D., Németh L., Halácsy P., Kornai A., Trón V., Nagy V.* Parallel corpora for medium density languages // Conference (International) on Recent Advances in Natural Language Processing Proceedings. — Shoumen, Bulgaria: INCOMA Ltd., 2005. P. 590–596.
21. Морфологический стандарт Национального корпуса русского языка: http://www.ruscorpora.ru/corpora-morph.html.
22. *Зацман И. М.* Имплицированные знания: основания и технологии извлечения // Информатика и её применения, 2018. Т. 12. Вып. 3. С. 74–82.
23. *Гончаров А. А., Инькова О. Ю.* Методика поиска имплицитных логико-семантических отношений в тексте // Информатика и её применения, 2019. Т. 13. Вып. 3. С. 97–104.
24. *Гончаров А. А., Инькова О. Ю.* Способы выражения причинных отношений в русском языке: опыт анализа с использованием кросслингвистической надкорпусной базы данных // Русская грамматика: активные процессы в языке и речи: Сб. научных трудов Междунар. научн. симпозиума. — Ярославль: ЯГПУ им. К. Д. Ушинского, 2019. С. 385–395.
25. PDTB Research Group. The Penn Discourse Treebank 2.0 annotation manual. Technical Report IRCS-08-01. — Philadelphia, PA, USA: Institute for Research in Cognitive Science, University of Pennsylvania, 2008. https://www.seas.upenn.edu/~pdtb/PDTBAPI/pdtb-annotation-manual.pdf.
26. *Webber B., Prasad R., Lee A., Joshi A.* The Penn Discourse Treebank 3.0 annotation manual, 2019. https://catalog.ldc.upenn.edu/docs/LDC2019T05/PDTB3-Annotation-Manual.pdf.
27. *Zufferey S., Degand L.* Annotating the meaning of discourse connectives in multilingual corpora // Corpus Linguist. Ling., 2013. Vol. 13. Iss. 2. P. 399–423. doi: 10.1515/cllt-2013-0022.

Поступила в редакцию 23.07.19

SUPRACORPORA DATABASES IN LINGUISTIC PROJECTS

***A. Yu. Egorova*[1], *I. M. Zatsman*[1], *and O. S. Mamonova*[2]**

[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

[2]Faculty of Foreign Languages and Area Studies, M. V. Lomonosov Moscow State University, 1 Leninskie Gory, Bldg. 13-14, Moscow 119991, Russian Federation

Abstract: The paper considers the task of providing linguistic studies with means of supracorpora databases containing aligned parallel texts (each includes the original text and its translation) as well as bilingual annotations of the researched linguistic items and their translation equivalents formed on the basis of parallel texts. Each annotation, formed by a linguist, fixes a translation model of a linguistic item. The experience of implementing several linguistic projects at Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences showed that not all translation models that linguists extract from parallel texts during linguistic annotation are described in bilingual dictionaries and handbooks. Thus, supracorpora databases allow researchers to create new knowledge about the translation equivalents of the researched linguistic items. It is extracted by linguists when comparing and annotating the sentences of the original text and their translations. The main aim of the paper is to describe the functions of supracorpora databases that provide linguists with new knowledge in the process of annotation.

Keywords: supracorpora database; linguistic annotation; linguistic unit; corpus linguistics; translation models

DOI: 10.14357/08696527190307

Acknowledgments

The work was fullfilled at the Institute of Informatics Problems, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences with partial support of the Russian Foundation for Basic Research (project 18-07-00192).

References

1. 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.
2. Loiseau, S., D. V. Sitchinava, Anna A. Zalizniak, and I. M. Zatsman. 2013. Information technologies for creating the database of equivalent verbal forms in the Russian–French multivariant parallel corpus. *Informatika i ee Primeneniya — Inform. Appl.* 7(2):100–109.

3. Kruzhkov, M. G., N. V. Buntman, E. Yu. Loshchilova, D. V. Sitchinava, Anna A. Zalizniak, and I. M. Zatsman. 2014. A database of Russian verbal forms and their French translation equivalents. *Computer Linguistics and Intellectual Technologies: Conference (International) "Dialog" Proceedings*. Moscow: RGGU. 13(20):284–296.
4. Kruzhkov, M. G. 2015. Informatsionnye resursy kontrastivnykh lingvisticheskikh issledovaniy: elektronnye korpusa tekstov [Information resources for contrastive studies: Electronic text corpora]. *Sistemy i Sredstva Informatiki — Systems and Means of Informatics* 25(2):140–159.
5. Zaliznyak, Anna A., I. M. Zatsman, O. Yu. In'kova, and M. G. Kruzhkov. 2015. Nadkorpusnye bazy dannykh kak lingvisticheskiy resurs [Supracorpora databases as linguistic resource]. *7th Conference (International) on Corpus Linguistics Proceedings*. St. Petersburg: St. Petersburg State University Publs. 211–218.
6. Zaliznyak, Anna A. 2016. Baza dannykh mezh"yazykovykh ekvivalentsiy kak instrument lingvisticheskogo analiza [A database of cross-linguistic equivalences as an instrument of linguistic analysis]. *Komp'yuternaya lingvistika i intellektual'nye tekhnologii* [Computer Linguistics and Intellectual Technologies] 15(22):854–866.
7. Zaliznyak, Anna A., I. M. Zatsman, and O. Yu. In'kova. 2017. Nadkorpusnaya baza dannykh konnektorov: postroenie sistemy terminov [Supracorpora database on connectives: Term system development]. *Informatika i ee Primeneniya — Inform. Appl.* 11(1):100–106.
8. Zatsman, I. M., and M. G. Kruzhkov. 2018. Nadkorpusnaya baza dannykh konnektorov: razvitie sistemy terminov proektirovaniya [Supracorpora database of connectives: Design-oriented evolution of the term system]. *Sistemy i Sredstva Informatiki — Systems and Means of Informatics* 28(4):156–167.
9. Zatsman, I., N. Buntman, M. Kruzhkov, V. Nuriev, and Anna A. Zalizniak. 2014. Conceptual framework for development of computer technology supporting crosslinguistic knowledge discovery. *15th European Conference on Knowledge Management Proceedings*. Reading, U.K.: Academic Publishing International Ltd. 3:1063–1071.
10. Zatsman, I., and N. Buntman. 2015. Outlining goals for discovering new knowledge and computerised tracing of emerging meanings discovery. *16th European Conference on Knowledge Management Proceedings*. Reading, U.K.: Academic Publishing International Ltd. 851–860.
11. Zatsman, I., N. Buntman, A. Coldefy-Faucard, and V. Nuriev. 2016. WEB knowledge base for asynchronous brainstorming. *17th European Conference on Knowledge Management Proceedings*. Reading, U.K.: Academic Publishing International Ltd. 1:976–983.
12. Zatsman, I. 2018. Goal-oriented creation of individual knowledge: Model and information technology. *19th European Conference on Knowledge Management Proceedings*. Reading, U.K.: Academic Publishing International Ltd. 2:947–956.
13. Dobrovol'skiy, D. O., and Anna A. Zalizniak. 2018. Nemetskie konstruktsii s modal'nymi glagolami i ikh russkie sootvetstviya: proekt nadkorpusnoy bazy dannykh [German constructions with modal verbs and their Russian correlates: A supracorpora database project]. *Computer Linguistics and Intellectual Technologies: Conference (International) "Dialog" Proceedings*. Moscow: RGGU. 17(24):172–184.
14. Zatsman, I. 2018. Stadii tselenapravlennogo izvlecheniya znaniy, implitsirovannykh v parallel'nykh tekstakh [Stages of goal-oriented discovery of knowledge implied in

parallel texts]. *Sistemy i Sredstva Informatiki — Systems and Means of Informatics* 28(3):175–188.

15. Zatsman, I. M. 2019. Tselenapravlennoe razvitie sistem lingvisticheskikh znaniy: vyyavlenie i zapolnenie lakun [Goal-oriented development of linguistic knowledge systems: Identifying and filling lacunae]. *Informatika i ee Primeneniya — Inform. Appl.* 13(1):91–98.
16. Ide, N., and J. Pustejovsky, eds. 2017. *Handbook of linguistic annotation*. Dordrecht, The Netherlands: Springer Science + Business Media. 1468 p.
17. Goncharov, A. A., O. Yu. Inkova, and M. G. Kruzhkov. 2019. Metodologiya annotirovaniya v nadkorpusnykh bazakh dannykh [Annotation methodology of supracorpora databases]. *Sistemy i Sredstva Informatiki — Systems and Means of Informatics* 29(2):148–160.
18. Goncharov, A. A., and I. M. Zatsman. 2019. Informatsionnye transformatsii parallel'nykh tekstov v zadachakh izvlecheniya znaniy [Information transformations of parallel texts in knowledge extraction]. *Sistemy i Sredstva Informatiki — Systems and Means of Informatics* 29(1):180–193.
19. Sitchinava, D. V. 2015. Parallel'nye teksty v sostave Natsional'nogo korpusa russkogo yazyka: novye napravleniya razvitiya i rezul'taty [Parallel texts within the Russian National Corpus: New directions and results]. *Trudy Instituta russkogo yazyka im. V. V. Vinogradova* [Proceedings of the V. V. Vinogradov Russian Language Institute] 6:194–235.
20. Varga, D., L. Németh, P. Halácsy, A. Kornai, V. Trón, and V. Nagy. 2005. Parallel corpora for medium density languages. *Conference (International) on Recent Advances in Natural Language Processing Proceedings*. Shoumen, Bulgaria: INCOMA Ltd. 590–596.
21. Morfologicheskiy standart Natsional'nogo korpusa russkogo yazyka [The morphological standard of the Russian national corpus]. Available at: http://www.ruscorpora.ru/corpora-morph.html (accessed July 16, 2019)
22. Zatsman, I. 2018. Implitsirovannye znaniya: osnovaniya i tekhnologii izvlecheniya [Implied knowledge: Foundations and technologies of explication]. *Informatika i ee Primeneniya — Inform. Appl.* 12(3):74-82.
23. Goncharov, A. A., and O. Yu. Inkova. 2019. Metodika poiska implitsitnykh logiko-semanticheskikh otnosheniy v tekste [Methods for identification of implicit logical-semantic relations in texts]. *Informatika i ee Primeneniya — Inform. Appl.* 13(3):97–104.
24. Goncharov, A. A., and O. Yu. Inkova. 2019. Sposoby vyrazheniya prichinnykh otnosheniy v russkom yazyke: opyt analiza s ispol'zovaniem krosslingvisticheskoy nadkorpusnoy bazy dannykh [Means of expressing causal relations in Russian: Analysis using a cross-linguistic supracorpora database]. *Russian Grammar: Active Processes in Language and Discourse: International Scientific Symposium*. 385–395.
25. PDTB Research Group. 2008. The Penn Discourse Treebank 2.0 annotation manual. Technical Report IRCS-08-01. Philadelphia, PA: Institute for Research in Cognitive Science, University of Pennsylvania. Available at: https://www.seas.upenn.edu/~pdtb/PDTBAPI/pdtb-annotation-manual.pdf (accessed July 16, 2019)
26. Webber, B., R. Prasad, A. Lee, and A. Joshi. 2019. The Penn Discourse Treebank 3.0 annotation manual. Available at: https://catalog.ldc.upenn.edu/docs/LDC2019T05/PDTB3-Annotation-Manual.pdf (accessed July 16, 2019).

27. Zufferey, S., and L. Degand. 2013. Annotating the meaning of discourse connectives in multilingual corpora. *Corpus Linguist. Ling.* 13(2):399–423. doi: 10.1515/cllt-2013-0022.

Received July 23, 2019

Contributors

Egorova Anna Yu. (b. 1991) – 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; ann.shurova@gmail.com

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

Mamonova Oksana S. (b. 1998) — student, Faculty of Foreign Languages and Area Studies, M. V. Lomonosov Moscow State University, 1 Leninskie Gory, Bldg. 13-14, Moscow 119991, Russian Federation; mamonovaoks@mail.ru

СИСТЕМЫ И СРЕДСТВА ИНФОРМАТИКИ, 2019. Т. 29. № 3. С. 92–103

ОШИБКИ В МАШИННОМ ПЕРЕВОДЕ: ПРОБЛЕМЫ КЛАССИФИКАЦИИ

А. А. Гончаров[1], *Н. В. Бунтман*[2], *В. А. Нуриев*[3]

Аннотация: Рассматриваются проблемы классификации ошибок в машинном переводе. В первой части предлагается обзор разных подходов к оценке качества машинного перевода и классификации ошибок, наблюдаемых при работе автоматизированных систем перевода. Вторая часть посвящена описанию оригинальной классификации ошибок машинного перевода, которая была создана таргетированно — для перевода коннекторов в языковой паре русский–французский. На этой языковой паре подобных исследований еще не проводилось. В предлагаемой классификации выделяются две основные группы ошибок — грамматические/лексические ошибки во фрагменте текста с коннектором и ошибки непосредственно в переводе коннектора. В работе использовался параллельный корпус, состоящий из русскоязычных текстов и их референтных переводов (около 4 млн словоупотреблений). Из этих русскоязычных текстов отбирались фрагменты, содержащие коннекторы, которые затем переводились на французский в двух разных системах машинного перевода.

Ключевые слова: классификация; машинный перевод; качество машинного перевода; ошибки в машинном переводе

DOI: 10.14357/08696527190308

1 Введение

Цель настоящей статьи — показать, какие проблемы связаны с оценкой качества машинного перевода. Эти проблемы поднимаются в контексте существующих подходов к классификации ошибок, которые возникают при работе автоматических переводчиков. Недостатки проанализированных подходов учитывались при создании оригинальной таксономии ошибок, разработанной применительно к переводу коннекторов (с русского языка на французский). В данной таксономии выделяются две основные группы ошибок — грамматические/лексические ошибки во фрагменте текста с коннектором и ошибки непосредственно в переводе коннектора. Для генерирования образцов перевода использовались

[1]Институт проблем информатики Федерального исследовательского центра «Информатика и управление» Российской академии наук, a.gonch48@gmail.com

[2]Московский государственный университет им. М. В. Ломоносова, nabunt@hotmail.com

[3]Институт проблем информатики Федерального исследовательского центра «Информатика и управление» Российской академии наук, nurieff.v@gmail.com

две системы: статистическая translate.yandex.ru (Яндекс.Переводчик)[1] и система нейронного машинного перевода translate.google.com (Google's Neural Machine Translation system — далее GNMT) (подробнее о принципах работы систем Яндекс.Переводчик и GNMT см. [1–3]). Корпус примеров машинного перевода составил 3000 параллельных контекстов, где на каждый из переводчиков приходится по 1500.

2 Подходы к оценке качества машинного перевода

Исследования по оценке качества машинного перевода стали одним из перспективных направлений в области компьютерной лингвистики. Об этом свидетельствуют посвященные им монографии «Translation quality assessment: From principles to practice» [4] и «Quality estimation for machine translation» [5], которые вышли в 2018 г. Бóльшая часть этих исследований посвящена разработке автоматизированных метрик для оценки качества машинного перевода, которые основываются на принципе максимально возможной генерализации при построении классификации ошибок (см., например, классификацию http://www.qt21.eu/mqm-definition/issues-list-2015-12-30.html и [6]). Такой принцип работы направлен на стандартизацию классификации ошибок, снижение затрат на оценку качества машинного перевода и повышение коммерческой привлекательности конкретных автоматических переводчиков. Однако применение этого подхода ведет к тому, что некоторые ошибки не учитываются автоматизированными метриками и, следовательно, не могут быть учтены при оценке качества машинного перевода.

Такие недостатки автоматизированных методов оценки качества машинного перевода призван компенсировать подход, предполагающий применение экспертного анализа с привлечением специалистов (переводчиков и лингвистов, см. [7, 8]). Лингвистический анализ ошибок машинного перевода может носить таргетированный характер, когда его объектом становятся ошибки при переводе определенных языковых единиц (как в случае коннекторов). Таксономии ошибок машинного перевода, созданные посредством лингвистического анализа переводов, отличаются более универсальной рубрикацией и в большей степени ориентированы на выявление проблем перевода, возникающих на каждом языковом уровне и в конкретной паре языков.

Одним из примеров такой таксономии служит классификация, предложенная в [7], где авторы отмечают, что, несмотря на явную полезность автоматизированных метрик, они вряд ли могут быть использованы в качестве единственного средства для выявления ошибок машинного перевода; наряду с ними необходимо использовать экспертный лингвистический анализ. Эта типология имеет иерархическую структуру и включает пять укрупненных классов ошибок: «Пропущенные слова» (Missing Words), «Словопорядок» (Word Order), «Неправильные слова»

[1] Данные для системы Яндекс.Переводчик собирались до 2018 г.

(Incorrect Words), «Неизвестные слова» (Unknown Words) и «Пунктуационные ошибки» (Punctuation errors). Ошибки, образующие первый класс, возникают, если в переводе отсутствует подлежащее переводу слово. Они делятся на две группы:

(1) ошибки, возникающие, когда в исходном предложении отсутствующее слово непосредственно участвует в смыслообразовании;

(2) ошибки, возникающие, когда общий смысл предложения понятен вне зависимости от слова, которое отсутствует в переводе, но при этом его отсутствие ставит под вопрос приемлемость переводного варианта с точки зрения языковой естественности.

Ошибки «словопорядка» выделяются в соответствии с тем, насколько необходимо изменить порядок слов в результирующем переводе, чтобы привести его к естественной форме. При этом необходимость изменения словопорядка может возникать:

– на локальном уровне, т. е. необходимо изменить порядок следования отдельных слов и целых словосочетаний в рамках одного синтаксического отрезка (части предложения);

– на более протяженных текстовых отрезках, т. е. слово или словосочетание необходимо переместить в другой синтаксический отрезок в рамках одного предложения.

В данной классификации самый обширный класс ошибок машинного перевода образуют «Неправильные слова». Эти ошибки возникают, когда система не находит правильного переводного соответствия какому-либо исходному слову или словосочетанию. Выделяются 5 типов таких слов:

(1) переводные эквиваленты, нарушающие смысл предложения;

(2) грамматически неправильные переводные эквиваленты (для флективных языков);

(3) избыточные слова (при переводе разговорной речи);

(4) эквиваленты, выбор которых не соответствует стилистическому регистру текста;

(5) неизвестные системе машинного перевода идиоматические выражения, которые в результате переводятся буквально.

Ошибка «неизвестного слова» возникает, если системе не удается идентифицировать исходное слово, что происходит, когда обрабатываются неизвестные слова (и основы) или неизвестные формы известных основ. Упоминая пунктуационные ошибки, авторы не заостряют на них внимания. В целом, отмечается, что присутствие в переводе ошибок одного типа не исключает наличия в нем ошибок других типов; кроме того, одна ошибка может повлечь за собой другую или ряд других.

Другая таксономия ошибок машинного перевода разработана в [9]. Здесь выделяются орфографические, лексические, грамматические, семантические и дискурсивные ошибки. Орфографические ошибки представлены тремя типами: пунктуационными, ошибками употребления заглавных букв и ошибками в написании слов. На лексическом уровне ошибки связаны с опущениями, добавлениями и отсутствием перевода. Случаи опущения и добавления рассматриваются применительно к типу слов, которые они непосредственно затрагивают: (а) «полнозначные» слова (content words), которые передают смысл предложения; (б) служебные слова, отвечающие за грамматические связи в предложении. Связанные с опущениями ошибки возникают, если в переводном тексте на выходе отсутствует слово, подлежащее переводу. Ошибками добавления в переводе, наоборот, считаются «лишние» — не обязательные для перевода — слова. Если автоматической системе не удается найти ни одного перевода-кандидата для определенного исходного слова, то в переводном тексте на выходе она оставляет его как есть, т. е. система допускает ошибку отсутствия перевода. Грамматические ошибки затрагивают морфологический и синтаксический уровни языка, они сгруппированы по двум категориям: ошибки неверного выбора (misselection) и неверного порядка (misordering). Ошибки неверного выбора имеют морфологический характер, они могут возникать из-за неверного распознавания частей речи; неправильного выбора глагольного времени; отсутствия согласования в лице, роде, числе; выбора системой полной формы артикля или предлога вместо усеченной и т. д. Ошибки неверного порядка связаны с проблемами синтаксической организации текста в переводе. Семантические ошибки — ошибки в значении и выборе слова. Авторы подразделяют их на смысловые (confusion of senses), ошибки выбора (wrong choice) и ошибки на сочетание и идиоматику (collocational and idiomatic). Смысловые ошибки возникают, когда система переводит исходное слово эквивалентом в том значении, которое не соответствует данному контексту. Ошибка выбора имеет место, если выбор перевода-кандидата совершенно некорректен (как в случае омонимов). Ошибки на сочетание и идиоматику являются результатом нарушения правил композиционной семантики. Дискурсивные ошибки или ошибки речевого уровня обусловлены неспособностью системы выбрать наиболее естественный в данном речевом контексте вариант. Также выделяются ошибки стилистические, ошибки вариативности и ошибки избыточного перевода. Стилистические ошибки связаны с нарушениями стилистического регистра. Ошибки вариативности возникают из-за того, что система не способна распознать диалектные варианты на лексическом или грамматическом уровне. Ошибками избыточного перевода считаются случаи, когда система переводит некоторые последовательности слов, которые, согласно принятой в переводящем языке норме, не должны переводиться (например, названия книг или фильмов для отдельных языков).

Строго говоря, классификации ошибок машинного перевода сейчас разрабатываются и используются многими специалистами [10–15]. В рамках данного исследования предлагается оригинальная классификация ошибок машинного пе-

ревода, разработанная специально для перевода коннекторов в языковой паре русский–французский. На этой языковой паре подобных исследований еще не проводилось. Кроме того, одним из основных преимуществ данного исследования является объем используемого параллельного корпуса, состоящего из русскоязычных текстов и их референтных переводов, выполненных профессиональными переводчиками (около 300 тыс. предложений и около 4 млн словоупотреблений), что дает возможность провести репрезентативное сравнение референтного и машинного переводов. Обычно в подобных исследованиях используются тестовые корпусы гораздо меньшего объема (ср., например, [15], где англо-французский корпус составляет всего 100 предложений).

В следующем разделе остановимся подробнее на принципах и особенностях формирования предлагаемой классификации ошибок машинного перевода.

3 Предлагаемая классификация ошибок машинного перевода

Классификация разрабатывалась для оценки качества машинного перевода текстовых фрагментов, содержащих коннекторы, с русского на французский. Отбор этих фрагментов осуществлялся с учетом частотности употребления и структуры коннекторов. Было решено проанализировать по 500 контекстов для однословных, многоэлементных и двухкомпонентных коннекторов[1] (подробнее о критериях отбора и структуре материала см. [16]).

В ходе первого этапа работы (2016–2017 гг.) анализировались результаты перевода при помощи системы Яндекс.Переводчик. На тот момент классификация включала 8 рубрик для обозначения разных типов ошибок, а также рубрику «Ошибок нет» (NoError), применявшуюся к допустимым вариантам перевода фрагмента текста с русского языка на французский. Были выявлены различные типы ошибок как во фрагментах текста, содержащих коннектор, так и в самих коннекторах (подробное описание списка см. в [3]).

На втором этапе работы (2018 г.), когда анализировались результаты машинного перевода, полученные при помощи системы GNMT, первоначальная классификация ошибок итерационно уточнялась. После завершения этого этапа обновленная классификация ошибок включала уже 15 рубрик, не считая рубрики «Ошибок нет».

В таблице сопоставляются оба варианта классификации, причем те признаки (и их коды), которые были добавлены на втором этапе работы, для наглядности выделены полужирным шрифтом. Видно, что помимо 6 случаев, когда признаки, обозначающие ту или иную ошибку, не менялись, и 5 случаев, когда на втором этапе работы признаки были добавлены, наблюдаются еще две возможности уточнения классификации: один признак может специфицироваться (разделение

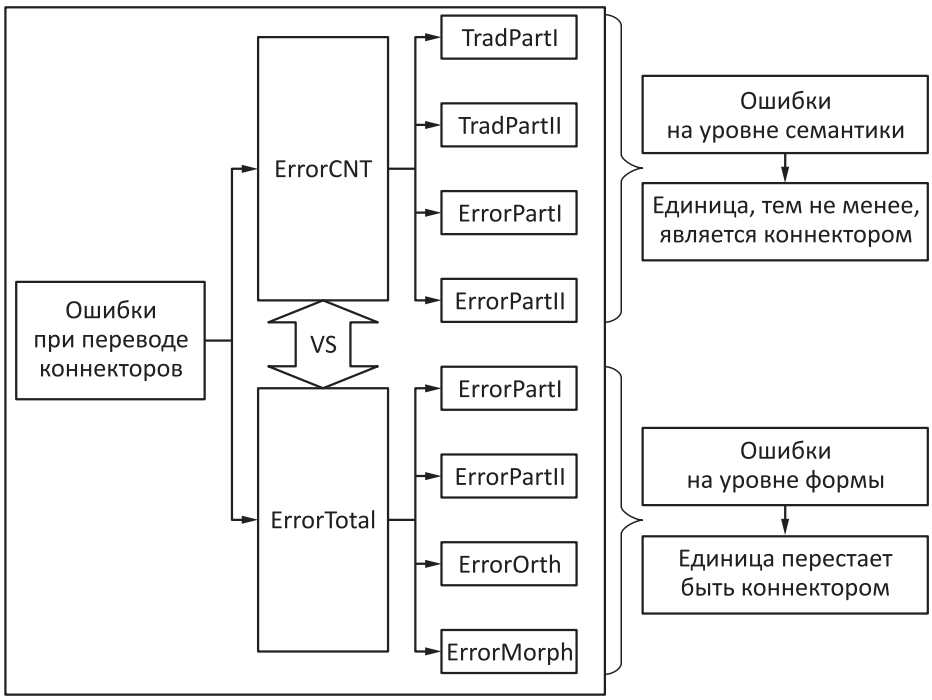
[1] Элементом называется минимальная составляющая коннектора, компонентом — часть коннектора, маркирующая каждый из двух или более соединяемых им фрагментов текста. Например, *не только. . . но и* состоит из двух компонентов, каждый из которых содержит два элемента.

Сопоставление двух версий классификации ошибок в машинном переводе

<table>
<tr><td rowspan="2"></td><td colspan="2">Название рубрики</td><td colspan="2">Код рубрики</td></tr>
<tr><td>На 1-м этапе работы</td><td>На 2-м этапе работы</td><td>На 1-м этапе работы</td><td>На 2-м этапе работы</td></tr>
<tr><td rowspan="8">Рубрики, относящиеся к переводу фрагмента текста с коннектором</td><td colspan="2">Все предложение аграмматично</td><td colspan="2">AgramTotal</td></tr>
<tr><td colspan="2">Локальная аграмматичность во фрагменте текста, не вводимом коннектором</td><td colspan="2">AgramLocal</td></tr>
<tr><td colspan="2">Локальная аграмматичность во фрагменте текста, вводимом коннектором</td><td colspan="2">AgramPostCNT</td></tr>
<tr><td>—</td><td>Лексическая ошибка во фрагменте текста</td><td>—</td><td>ErrorLex</td></tr>
<tr><td colspan="2">Русское слово кириллицей</td><td colspan="2">Cyrillic</td></tr>
<tr><td>—</td><td>Русское слово латинским шрифтом</td><td>—</td><td>Latin</td></tr>
<tr><td>—</td><td>Пропуск фрагмента текста</td><td>—</td><td>Lacuna</td></tr>
<tr><td rowspan="6">Рубрики, относящиеся к переводу коннектора</td><td>—</td><td>Переведена первая часть неоднословного коннектора</td><td>—</td><td>TradPartI</td></tr>
<tr><td>—</td><td>Переведена вторая часть неоднословного коннектора</td><td>—</td><td>TradPartII</td></tr>
<tr><td rowspan="2">Часть неоднословного коннектора переведена ошибочно</td><td>Первая часть неоднословного коннектора переведена ошибочно</td><td rowspan="2">ErrorPart</td><td>ErrorPartI</td></tr>
<tr><td>Вторая часть неоднословного коннектора переведена ошибочно</td><td>ErrorPartII</td></tr>
<tr><td rowspan="2">Семантическая ошибка в выборе коннектора</td><td>Семантическая ошибка в выборе коннектора</td><td rowspan="2">ErrorCNT</td><td>ErrorCNT</td></tr>
<tr><td>Коннектор ошибочно заменен языковой единицей, не являющейся коннектором</td><td>ErrorMorph</td></tr>
<tr><td colspan="2">Орфографическая ошибка в форме коннектора</td><td colspan="2">ErrorOrth</td></tr>
<tr><td colspan="2">Коннектор переведен несуществующей единицей</td><td colspan="2">ErrorTotal</td></tr>
</table>

ErrorPart на ErrorPartI и ErrorPartII) либо из массива случаев, обозначаемых неким признаком, выделяется группа, для которой вводится новый признак (добавление признака ErrorMorph для случаев, когда при переводе не просто выбирается единица с другой семантикой, но такая единица при этом не является коннектором).

Ошибки, касающиеся фрагмента текста, анализируются в той мере, в какой они способствуют пониманию возможных причин ошибок в переводе коннектора. В силу этого грамматические ошибки рассматриваются с учетом их локализации по отношению к коннектору: во вводимом коннектором фрагменте (AgramPostCNT) или за его пределами (AgramLocal). Если же они есть и там и там, то такому контексту присваивается рубрика AgramTotal. Это же касается и лексических ошибок, которые можно было бы рассматривать в рамках более мелких групп. Однако для анализа перевода коннекторов важно понимать лишь то, идет ли речь о выборе неправильной лексической единицы (ErrorLex) или об отсутствии перевода (Cyrillic и Latin), выражающегося в том числе в пропуске языковой единицы или целого фрагмента текста (Lacuna). Все эти рубрики спо-



Основные сочетания ошибок, при которых ошибка в части коннектора способна вызвать ошибку в коннекторе в целом

собны функционировать автономно, и никакие из них не обусловливают наличие каких-либо других.

Несколько иначе обстоит дело с рубриками, описывающими ошибки в переводе непосредственно коннекторов, которые, хотя и могут функционировать независимо друг от друга, часто используются в сочетаниях. Это вызвано во многом тем, что некоторые из них описывают ошибки в части коннектора (ErrorPartI, ErrorPartII, TradPartI, TradPartII), некоторые — и в части, и во всем коннекторе (ErrorMorph, ErrorOrth), а некоторые — только в коннекторе целиком (ErrorCNT, ErrorTotal).

Следствием ошибок в части коннектора может стать ошибочный выбор коннектора в целом (ErrorCNT) или вовсе генерация полностью ошибочной единицы (ErrorTotal). Это подтверждается тем, что ErrorCNT сочетается с какими-либо из рубрик, обозначающих наличие ошибок в переводе коннекторов в 27,8% случаев, а ErrorTotal — в 79,8%. На рисунке отражены основные механизмы сочетания ошибок машинного перевода коннекторов. Принимаются во внимание только те ситуации, когда наличие ошибки в части коннектора обусловливает ошибочность коннектора в целом. Именно поэтому сочетания рубрик ErrorCNT и ErrorOrth, ErrorTotal и TradPartI, ErrorTotal и TradPartII — возможные, но не обладающие этим свойством — на рисунке не отражаются.

4 Заключение

В статье предложена оригинальная таргетированная, т. е. созданная специально с целью изучения конкретного языкового явления, классификация. Такой подход позволяет специфицировать и/или вводить обобщающие признаки. Подобные классификации могут разрабатываться для исследования широкого круга языковых явлений (не только коннекторов). Таким образом, появляется возможность конкретизировать описание ошибок машинного перевода в зависимости от целей исследования, в целом сохраняя достаточный уровень генерализации, чего, как показано в разд. 2, достичь удается нечасто. Эти характеристики позволяют классификации оставаться простой и удобной в использовании, будучи при этом приложимой к решению конкретных прикладных и теоретических задач.

Литература

1. *Wu Y., Schuster M., Chen Z., et al.* Google's neural machine translation system: Bridging the gap between human and machine translation. 2016. https://arxiv.org/pdf/1609.08144.pdf.
2. *Johnson M., Schuster M., Le Q. V., Krikun M., Wu Y., Chen Zh., Thorat N., Viégas F., Wattenberg M., Corrado G., Hughes M., Dean J.* Google's multilingual neural machine translation system: Enabling zero-shot translation // T. Association Computational Linguistics, 2017. Vol. 5. P. 339–351.

3. *Nuriev V., Buntman N., Inkova O.* Machine translation of Russian connectives into French: Errors and quality failures // Информатика и её применения, 2018. Т. 12. Вып. 2. С. 105–113.
4. Translation quality assessment: From principles to practice / Eds. J. Moorkens, Sh. Castilho, F. Gaspari, S. Doherty. — Cham, Switzerland: Springer, 2018. 292 p.
5. *Specia L., Scarton C., Paetzold G. H.* Quality estimation for machine translation. — San Rafael, CA, USA: Morgan & Claypool, 2018. 148 p.
6. *Lommel A.* Metrics for translation quality assessment: A case for standardising error typologies // Translation quality assessment: From principles to practice / Eds. J. Moorkens, Sh. Castilho, F. Gaspari, S. Doherty. — Cham, Switzerland: Springer, 2018. P. 109–127.
7. *Vilar D., Xu J., D'Haro L., Ney H.* Error analysis of statistical machine translation output // 5th Conference (International) on Language Resources and Evaluation Proceedings. — Genoa, Italy: European Language Resources Association, 2006. http://www.lrec-conf.org/proceedings/lrec2006/pdf/413_pdf.pdf.
8. *Zhou Wang B., Liu S., Li M., Zhang D., Zhao T.* Diagnostic evaluation of machine translation systems using automatically constructed linguistic check-points // 22nd Conference (International) on Computational Linguistics Proceedings. — Stroudsburg, PA, USA: Association for Computational Linguistics. — Manchester, 2008. Vol. 1. P. 1121–1128.
9. *Costa Â., Ling W., Luís T., Correia R., Coheur L.* A linguistically motivated taxonomy for machine translation error analysis // Machine Translation, 2015. Vol. 29. P. 127–161.
10. *Guillou L., Hardmeier C.* PROTEST: A test suite for evaluating pronouns in machine translation // 10th Conference (International) on Language Resources and Evaluation Proceedings. — Portoroz, 2016. P. 636–643.
11. *Burchardt A., Macketanz V., Dehdari J., Heigold G., Peter J. T., Williams P.* A linguistic evaluation of rule-based, phrase-based, and neural MT engines // Prague Bull. Math. Linguistics, 2017. Vol. 108. Iss. 1. P. 159–170.
12. *Popović M.* Comparing language related issues for NMT and PBMT between German and English // Prague Bull. Math. Linguistics, 2017. Vol. 108. Iss. 1. P. 209–220.
13. *Burlot F., Yvon F.* Evaluating the morphological competence of machine translation systems // 2nd Conference on Statistical Machine Translation Proceedings. — Copenhagen, 2017. P. 43–55.
14. *Comelles E., Arranz V., Castellón I.* Guiding automatic MT evaluation by means of linguistic features // Digit. Scholarsh. Hum., 2017. Vol. 32. Iss. 4. P. 761–778.
15. *Isabelle P., Cherry C., Foster G.* A challenge set approach to evaluating machine translation // Conference on Empirical Methods in Natural Language Processing Proceedings. — Copenhagen, 2017. P. 2476–2486.
16. *Бунтман Н. В., Гончаров А. А., Зацман И. М., Нуриев В. А.* Количественный анализ результатов машинного перевода с использованием надкорпусных баз данных // Информатика и её применения, 2018. Т. 12. Вып. 4. С. 100–109.

Поступила в редакцию 25.07.19

MACHINE TRANSLATION ERRORS: PROBLEMS OF CLASSIFICATION

A. A. Goncharov[1], *N. V. Buntman*[2], *and V. A. Nuriev*[1]

[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
[2]M. V. Lomonosov Moscow State University, GSP-1, Leninskie Gory, Moscow 119991, Russian Federation

Abstract: The paper considers the problems of classifying machine translation errors. Its first part reviews some approaches to evaluation of machine translation quality and to classification of errors that machine translation systems tend to make. The other part of the paper describes an original taxonomy of machine translation errors — the targeted one. It has been devised specifically to classify the errors central to translation of connectives (from Russian into French). To date, there have been no such studies for this pair of languages. The proposed classification includes two groups of errors: (*i*) grammatical/lexical errors in the translation of the text chunk where a given connective occurs; and (*ii*) errors in the translation of a connective itself. This study uses a parallel Russian–French corpus that stores Russian source texts and their reference — made by professional humans — translations into French. The corpus totals 300 thousand sentences (about 4 million words). The source texts where connectives occur have been used to generate machine translations by two automated systems.

Keywords: classification; machine translation; quality of machine translation; machine translation errors

DOI: 10.14357/08696527190308

References

1. Wu, Y., M. Schuster, Z. Chen, *et al.* 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. Available at: https://arxiv.org/pdf/1609.08144.pdf (accessed July 18, 2019).
2. Johnson, M., M. Schuster, Q. V. Le, M. Krikun, Y. Wu, Zh. Chen, N. Thorat, F. Viégas, M. Wattenberg, G. Corrado, M. Hughes, and J. Dean. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *T. Association Computational Linguistics* 5:339–351.
3. Nuriev, V., N. Buntman, and O. Inkova. 2018. Machine translation of Russian connectives into French: Errors and quality failures. *Informatika i ee Primeneniya — Inform. Appl.* 12(2):105–113.
4. Moorkens, J., Sh. Castilho, F. Gaspari, and S. Doherty, eds. 2018. *Translation quality assessment. From principles to practice.* Cham, Switzerland: Springer. 292 p.

5. Specia, L., C. Scarton, and G. H. Paetzold. 2018. *Quality estimation for machine translation*. San Rafael, CA: Morgan & Claypool. 148 p.
6. Lommel, A. 2018. Metrics for translation quality assessment: A case for standardising error typologies. *Translation quality assessment: From principles to practice*. Eds. J. Moorkens, Sh. Castilho, F. Gaspari, and S. Doherty. Cham, Switzerland: Springer. 109–127.
7. Vilar, D., J. Xu, L. D'Haro, and H. Ney. 2006. Error analysis of statistical machine translation output. *5th Conference (International) on Language Resources and Evaluation Proceedings*. Italy, Genoa: European Language Resources Association. Available at: http://www.lrec-conf.org/proceedings/lrec2006/pdf/413_pdf.pdf (accessed July 18, 2019).
8. Zhou Wang, B., S. Liu, M. Li, D. Zhang, and T. Zhao. 2008. Diagnostic evaluation of machine translation systems using automatically constructed linguistic check-points. *22nd Conference (International) on Computational Linguistics Proceedings*. Manchester. 1:1121–1128.
9. Costa, Â., W. Ling, T. Luís, R. Correia, and L. Coheur. 2015. A linguistically motivated taxonomy for machine translation error analysis. *Machine Translation* 29:127–161.
10. Guillou, L., and C. Hardmeier. 2016. PROTEST: A test suite for evaluating pronouns in machine translation. *10th Conference (International) on Language Resources and Evaluation Proceedings*. Portoroz. 636–643.
11. Burchardt, A., V. Macketanz, J. Dehdari, G. Heigold, J. T. Peter, and P. Williams. 2017. A linguistic evaluation of rule-based, phrase-based, and neural MT engines. *Prague Bull. Math. Linguistics* 108(1):159–170.
12. Popović, M. 2017. Comparing language related issues for NMT and PBMT between German and English. *Prague Bull. Math. Linguistics* 108(1):209–220.
13. Burlot, F., and F. Yvon. 2017. Evaluating the morphological competence of machine translation systems. *2nd Conference on Statistical Machine Translation Proceedings*. Copenhagen. 43–55.
14. Comelles, E., V. Arranz, and I. Castellón. 2017. Guiding automatic MT evaluation by means of linguistic features. *Digit. Scholarsh. Hum.* 32(4):761–778.
15. Isabelle, P., C. Cherry, and G. Foster. 2017. A challenge set approach to evaluating machine translation. *Conference on Empirical Methods in Natural Language Processing Proceedings*. Copenhagen. 2476–2486.
16. Buntman, N. V., A. A. Goncharov, I. M. Zatsman, and V. A. Nuriev. 2018. Kolichestvennyy analiz rezul'tatov mashinnogo perevoda s ispol'zovaniem nadkorpusnykh baz dannykh [Using supracorpora databases for quantitative analysis of machine translations]. *Informatika i ee Primeneniya — Inform. Appl.* 12(4):96–105.

Received July 25, 2019

Contributors

Goncharov Alexander A. (b. 1994) — 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; a.gonch48@gmail.com

Buntman Nadezhda V. (b. 1957) — Candidate of Science (PhD) in philology, associate professor, M. V. Lomonosov Moscow State University, GSP-1, Leninskie Gory, Moscow 119991, Russian Federation; nabunt@hotmail.com

Nuriev Vitaly A. (b. 1980) — Candidate of Science (PhD) in philology, 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; nurieff.v@gmail.com

СИСТЕМЫ И СРЕДСТВА ИНФОРМАТИКИ, 2019. Т. 29. № 3. С. 104–113

ХАРАКТЕРИЗАЦИЯ ПОСЛЕДОВАТЕЛЬНОСТНЫХ САМОСИНХРОННЫХ ЭЛЕМЕНТОВ*

***Ю. А. Степченков*[1], *Ю. Г. Дьяченко*[2], *Н. В. Морозов*[3], *Д. Ю. Степченков*[4], *Д. Ю. Дьяченко*[5]**

Аннотация: Специфика функционирования самосинхронных (СС) схем предъявляет особые требования к процедуре их характеризации. Процедура должна учитывать дисциплину формирования информационных и фазовых сигналов на основе задаваемых пользователем атрибутов входов и выходов характеризуемого элемента. Предложена методика уточнения процесса характеризации для последовательностных СС-элементов, основанная на использовании векторов определения статических значений или направлений переключения входов и выходов. Алгоритмизация и реализация предложенного подхода в новой версии системы автоматизированной характеризации интегральных библиотек (САХИБ) повысили ее эффективность и обеспечили достоверную характеризацию всех типов последовательностных элементов из библиотеки СС-элементов для 65-нанометровой КМОП (комплеменарный металл-оксид-полупроводник) технологии. Автоматическое дополнение в процессе характеризации моделей последовательностных элементов конструкциями анализа порядка изменения сигналов на их входах и предупреждения о некорректной последовательности входов облегчает и ускоряет проектирование СС цифровых схем.

Ключевые слова: самосинхронная схема; временные параметры; характеризация; моделирование; триггер; начальное состояние

DOI: 10.14357/08696527190309

1 Введение

Современный тренд микроэлектроники — разработка надежных энергоэффективных цифровых устройств — открывает широкие перспективы для внед-

*Исследование выполнено в рамках проекта № КП19-260 (Механизмы обеспечения отказоустойчивости современных высокопроизводительных и высоконадежных применений), финансируемого Минобрнауки России.

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

[2]Институт проблем информатики Федерального исследовательского центра «Информатика и управление» Российской академии наук, diaura@mail.ru

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

[4]Институт проблем информатики Федерального исследовательского центра «Информатика и управление» Российской академии наук, stepchenkov@mail.ru

[5]Институт проблем информатики Федерального исследовательского центра «Информатика и управление» Российской академии наук, diaden87@gmail.com

рения СС-схем, являющихся естественно надежными и обладающими рядом преимуществ по сравнению с синхронными аналогами [1, 2]. Их эффективное проектирование невозможно без встраивания дополнительных библиотек СС-элементов в промышленные системы проектирования сверхбольших интегральных схем (СБИС) и, соответственно, без подготовки стандартизованных файлов описания моделей элементов, содержащих электрические и энергетические параметры элементов встраиваемых библиотек. Традиционно для этого используются программные средства характеризации.

Известные в настоящее время программные средства характеризации элементов стандартных библиотек, например [3, 4], разрабатывались для синхронной схемотехники и не позволяют получить адекватные модели для СС-элементов. Поэтому разработка эффективного программного средства, обеспечивающего получение адекватных параметров элементов базиса проектирования СС-схем и достоверность характеристик проектируемых СС КМОП СБИС и систем на кристалле, является актуальной задачей.

В 2012–2016 гг. в Институте проблем информатики Российской академии наук (ИПИ ФИЦ ИУ РАН, отдел 52) была разработана система характеризации СС-элементов на КМОП-транзисторах САХИБ [5]. Однако ее опытная эксплуатация выявила некоторые недостатки, снижающие эффективность характеризации последовательностных СС-схем.

2 Проблемы характеризации самосинхронных элементов

Характеризация элемента основана на автоматизированном электрическом моделировании его принципиальной схемы. Точность и адекватность полученных результатов моделирования обеспечиваются двумя факторами:

(1) корректностью определения пары «активный вход – активный выход», такой что изменение уровня сигнала именно на активном входе и только на нем приводит к требуемому переключению выбранного выхода при фиксированных значениях остальных входов элемента;

(2) адекватностью и реализуемостью начального состояния элемента, которое должно быть статическим в отсутствие переключений активного входа.

При характеризации последовательностных элементов (триггеров и схем на их основе) возникает проблема задания начального состояния, поскольку в каждый момент времени состояние ячейки памяти может определяться не только текущим значением сигналов в цепях схемы элемента, но и предшествующим их значением.

2.1 Особенности работы триггерных самосинхронных элементов

Опытная эксплуатация системы характеризации СС-элементов САХИБ [5] показала, что начальное состояние триггерных элементов не всегда вычисляется

корректно. Но этот недостаток может быть устранен с учетом особенностей их функционирования. Особенность работы триггерных СС-элементов заключается в соблюдении правильной последовательности поступления сигналов на входы элемента. При характеризации СС-триггеров с входом управления и информационным входом в виде унарного, бифазного или парафазного без спейсера сигнала необходимо учитывать, что при корректной последовательности изменений состояний входов переключение информационного входа не приводит к изменению состояния выходов триггера.

В триггерах с парафазным со спейсером (ПФС) информационным входом такая зависимость существует и ее надо учитывать. Если при этом триггер с ПФС-входом имеет и вход управления, то зависимость выхода от информационного входа должна определяться при рабочем состоянии входа управления.

Полученные реальные зависимости выходов от входов будут описывать поведение триггера в «легальных» условиях, в которых дисциплина формирования сигналов соответствует требованиям СС-схем. Ниже перечислены зависимости, для которых должны быть определены задержки для разных типов триггеров.

Однотактный и двухтактный D-триггер и RS-триггер с бифазным или парафазным без спейсера входом

1. Зависимость обоих компонентов бифазного информационного выхода от входа управления.
2. Зависимость индикаторного выхода от входа управления, переключающегося либо из спейсера в рабочее значение, либо из рабочего значения в спейсер.
3. Зависимость обоих компонентов бифазного информационного выхода от входа установки.
4. Зависимость индикаторного выхода от входа СС-установки для случаев, когда установка инициируется или завершается.
5. Зависимость выхода инверсии входа управления ЕВ (если он есть) от входа управления, переключающегося либо из спейсера в рабочее значение, либо из рабочего значения в спейсер.

Однотактный и двухтактный RS-триггер с информационным ПФС-входом

1. Зависимость обоих компонентов бифазного информационного выхода от ПФС-входа.
2. Зависимость индикаторного выхода от ПФС-входа, переключающегося либо из спейсера в рабочее значение, либо из рабочего значения в спейсер.
3. Зависимость обоих компонентов бифазного информационного выхода от входа установки.
4. Зависимость индикаторного выхода от входа СС-установки, когда установка инициируется или завершается.

2.2 Задание начального состояния триггера

Начальное состояние СС-триггера определяется состояниями бистабильных ячеек (БЯ), из которых он состоит. В однотактном триггере состояние единственной БЯ однозначно определяется состоянием входов и выходов триггера. В двухтактных триггерах состояние первой БЯ должно быть:

(1) идентично состоянию второй (выходной) БЯ для случаев зависимости информационного и индикаторного выходов от входов установки и зависимости индикаторного выхода от информационного входа и входов управления при переключении в рабочую фазу;

(2) противоположно состоянию второй БЯ для случаев зависимости информационного выхода от информационного входа и входов управления и зависимости индикаторного выхода от информационного входа и входов управления при переключении в спейсер.

Для задания пользователем начального состояния входов и выходов используется вектор определения статических значений или направлений переключения входов и выходов характеризуемого элемента:

$$\langle\text{имя\_активного\_входа}\rangle : \langle\text{имя\_активного\_выхода}\rangle[\langle\text{имя\_входа\_1}\rangle = \\ = I_1, \ldots, \langle\text{имя\_входа\_}N\rangle = I_N, \\ \langle\text{имя\_выхода\_1}\rangle = O_1, \ldots, \langle\text{имя\_выхода\_}M\rangle = O_M],$$

где $\langle\text{имя\_входа\_1}\rangle, \ldots, \langle\text{имя\_входа\_}N\rangle$ — имена всех входов элемента; $\langle\text{имя\_выхода\_1}\rangle, \ldots, \langle\text{имя\_выхода\_}M\rangle$ — имена всех выходов элемента; $I_1, \ldots, I_N$ — статические значения входов или направление переключения активного входа элемента; $O_1, \ldots, O_M$ — начальные состояния выходов или направление переключения активного выхода элемента. При задании значений входов и выходов элемента используется следующий алфавит:

0, 1 — статическое значение входа или начальное значение выхода;

R — направление переключения активного входа или выхода из 0 в 1;

F — направление переключения активного входа или выхода из 1 в 0;

* — значение входа или выхода не важно.

Порядок перечисления входов и выходов в векторе может быть любым.

Например, для триггера R0CE10 с бифазным информационным входом (R, S), входом управления E и входом СС-установки C, изображенного на рис. 1, могут быть заданы следующие векторы:

```
//--- Переключение триггера в спейсер
E : I [E=F,R=*,S=*,C=1,I=R,Q=*,QB=*]
//--- Переключение триггера в рабочую фазу
E : I [E=R,R=*,S=*,C=1,I=F,Q=*,QB=*]
E : Q [E=R,R=1,S=0,C=1,I=F,Q=F,QB=R]
```

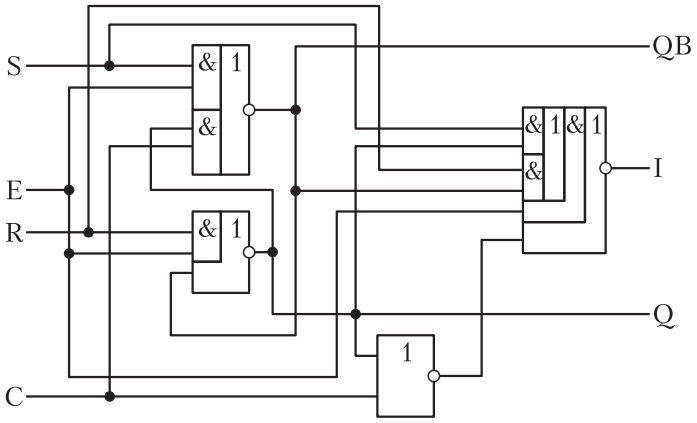


Рис. 1 Функционально-логическая схема триггера R0CE10

```
E : QB [E=R,R=1,S=0,C=1,I=F,Q=F,QB=R]
E : Q [E=R,R=0,S=1,C=1,I=F,Q=R,QB=F]
E : QB [E=R,R=0,S=1,C=1,I=F,Q=R,QB=F]
//--- Установка триггера
C : I [E=0,R=*,S=*,C=F,I=F,Q=*,QB=*]
C : I [E=0,R=*,S=*,C=R,I=R,Q=*,QB=*]
C : Q [E=0,R=*,S=*,C=F,I=F,Q=F,QB=R]
C : QB [E=0,R=*,S=*,C=F,I=F,Q=F,QB=R]
```

Информационный вход данного триггера не имеет спейсера, поэтому фазовыми переключениями триггера «дирижирует» вход управления.

2.3 Фиксирование нарушений в работе самосинхронного триггера

В реальных схемах СС-дисциплина формирования сигналов может нарушаться из-за того, что либо схема не является СС-схемой, либо ее топологическая реализация привнесла слишком большие и разные по величине задержки в критические цепи. Поэтому модель СС-триггера должна включать проверки соотношения времен поступления сигналов на входы триггера и изменения входов при определенных значениях выходов и инициировать предупреждения при нарушениях СС-дисциплины.

К нарушениям СС-дисциплины относятся следующие:

– информационный ПФС-вход переключился в рабочую фазу при рабочем значении индикаторного выхода;

– информационный ПФС-вход переключился в спейсер при спейсере на индикаторном выходе;

- вход управления (выход EB) переключился в рабочую фазу при рабочем значении индикаторного выхода;
- вход управления (выход EB) переключился в спейсер при спейсере на индикаторном выходе;
- информационный вход, не являющийся ПФС, изменил свое состояние при рабочем значении входа управления (выхода EB);
- вход СС-установки переключился в активное состояние при рабочем значении информационного ПФС-входа;
- информационный ПФС-вход переключился в рабочее состояние при активном значении входа СС-установки;
- вход СС-установки переключился в активное состояние при рабочем значении входа управления (выхода EB);
- вход управления (выход EB) переключился в рабочее состояние при активном значении входа СС-установки.

Общий вид генератора предупреждения о нарушении, связанном с изменением некоторого сигнала в течение указанного периода, в языке Verilog выглядит следующим образом [6]:

```
$nochange (reference_event, data_event, start_edge_offset,
end_edge_offset, notifier);
```

В случае СС-схем значения «`start_edge_offset`» и «`end_edge_offset`» достаточно выбрать нулевыми — это отвечает требуемой дисциплине формирования информационных, управляющих и установочных сигналов. Например, для триггера R0CE10 в Verilog-модель добавляется код

```
  specify
    $nochange(posedge I, negedge E, 0, 0, notifier);
    $nochange(negedge I, posedge E, 0, 0, notifier);
    $nochange(negedge C, posedge E, 0, 0, notifier);
    $nochange(posedge E, R, 0, 0, notifier);
    $nochange(posedge E, S, 0, 0, notifier);
    $nochange(posedge E, negedge C, 0, 0, notifier);
    $nochange(negedge I, negedge C, 0, 0, notifier);
    $nochange(posedge I, posedge C, 0, 0, notifier);
endspecify
```

Описанная выше методика уточнения процесса характеризации для последовательностных элементов была алгоритмизирована и реализована в версии 3.0 системы характеризации САХИБ.

3 Доработка системы САХИБ

Для решения описанных выше задач система САХИБ была дополнена интерактивной работой с векторами, реализацией характеризации на основе данных векторов и модулем анализа возможных нарушений.

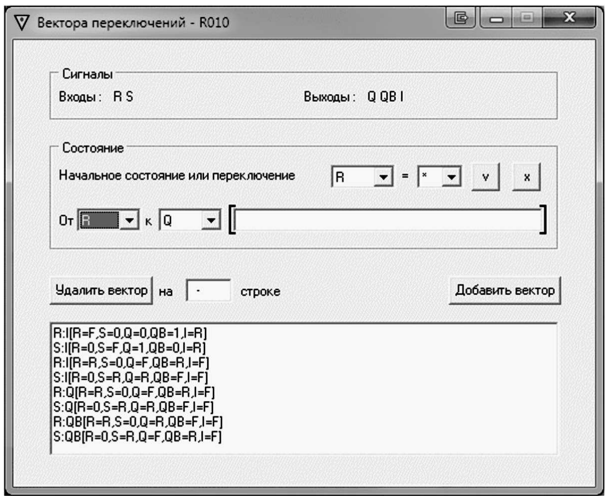


Рис. 2 Окно интерактивного задания векторов

Предусмотрена возможность начального получения всех векторов, формируемых программой характеризации автоматически, в качестве основы (шаблона) для формирования целесообразной совокупности векторов.

Выбор пары «активный вход – активный выход» и задание вектора для нее осуществляется в окне «Вектора переключений» (рис. 2). При наличии готового файла векторов для характеризуемой схемы с именем ⟨имя_схемы⟩.vin его содержимое появится в нижней части окна «Вектора переключений». Перед запуском характеризации пользователь может посмотреть, изменить или пополнить список векторов.

Режим процесса характеризации выбирается пользователем:

– автоматический, программа сама составляет необходимые векторы;

– ручной, используются только заданные пользователем векторы;

– комбинированный.

Модуль анализа нарушений дисциплины формирования внешних и внутренних сигналов в характеризуемой схеме автоматически анализирует алгоритм ее функционирования и добавляет необходимые конструкции «$nochange», описанные выше, в генерируемое системой характеризации Verilog-описание схемы. Это позволяет в процессе моделирования СС-схем локализовать нарушения дисциплины формирования сигналов в СС-схеме. Для сложных СС-схем такая возможность облегчает поиск причин нарушения самосинхронности схемы, выявленных на этапе анализа схемы на самосинхронность.

4 Заключение

Правильно спроектированные СС-схемы являются таковыми только при соблюдении соответствующей дисциплины входов, выходов и внутренних сигналов схемы. Поэтому наличие этой дисциплины должно учитываться при характеризации последовательностных СС-схем.

Дополнительные атрибуты задания на характеризацию последовательностного СС-элемента, задаваемые пользователем, — вектор определения статических значений входов и выходов или направлений переключения — позволяют однозначно определить начальное состояние входов, выходов и внутренних цепей элемента и сделать процедуру характеризации достоверной и полной.

Использование в Verilog-моделях элементов, формируемых по результатам характеризации, операторов анализа взаимного расположения их входных сигналов на временно́й оси обеспечивает дополнительный контроль соблюдения СС-дисциплины внешних и внутренних сигналов в СС-схеме. Это позволяет выявить и исправить нарушения самосинхронности на этапе функционально-логической верификации алгоритма работы схемы.

Литература

1. *Степченков Ю. А., Дьяченко Ю. Г., Горелкин Г. А.* Самосинхронные схемы — будущее микроэлектроники // Вопросы радиоэлектроники, 2011. № 2. С. 153–184.
2. *Бобков С. Г., Горбунов М. С., Дьяченко Ю. Г., Рождественский Ю. В., Степченков Ю. А., Сурков А. В.* Использование самосинхронной логики для снижения потребляемой мощности и повышения надежности микропроцессоров // Проблемы разработки перспективных микро- и наноэлектронных систем: Сб. трудов. — М.: ИППМ РАН, 2014. Ч. I. С. 43–48.
3. Library Characterization, Cadence. https://www.cadence.com/content/cadence-www/global/en_US/home/tools/custom-ic-analog-rf-design/library-characterization.html.
4. CHARISMA: система характеризации библиотек стандартных ячеек // Radix Tools. http://www.radixtools.ru/products-charisma.
5. *Морозов Н. В., Дьяченко Ю. Г., Степченков Д. Ю.* Система характеризации самосинхронных элементов САХИБ. Версия 2. Свидетельство о государственной регистрации программы для ЭВМ № 2016663661 от 13.12.2016.
6. IEEE Standard Verilog Hardware Description Language. IEEE Computer Society. IEEE Std 1364-2001 (Revision of IEEE Std 1364-1995). — New York, NY, USA: Institute of Electrical and Electronics Engineers, Inc., 2001. 778 p.

Поступила в редакцию 07.08.19

SEQUENTIAL SELF-TIMED CELL CHARACTERIZATION

Yu. A. Stepchenkov, Yu. G. Diachenko, N. V. Morozov, D. Yu. Stepchenkov, and D. Yu. Diachenko

Institute of Informatics Problems, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, 44-2 Vavilov Str., Moscow 119133, Russian Federation

Abstract: Functional specificity of the self-timed circuits makes special requirements to their characterization procedure. This procedure should take into account a signal conditioning discipline for information and phase signals on base of user defined attributes of the characterized cell's inputs and outputs. The paper describes a technique of adjusting characterization process for sequential self-timed cells. It is based on using vectors that set static values and transition direction for all inputs and outputs. Algorithmization and implementation of the suggested approach in new SAHIB characterization system version have increased its efficiency and provided the valid characterization of all sequential cell types in the self-timed cell library for 65-nanometer standard CMOS (complementary metal-oxide-semiconductor) process. Automatic introduction of the Verilog constructions analyzing change order of all cell inputs and notifying their invalid sequence into the sequential cell models during characterization procedure accelerates and mitigates self-timed circuit design.

Keywords: self-timed circuit; timing parameters; characterization; simulation; sequential cell; initial state

DOI: 10.14357/08696527190309

Acknowledgments

The research was performed within the project #KP19-260 funded by the Ministry of Education and Science of Russia.

References

1. Stepchenkov, Yu. A., Yu. G. Diachenko, and G. A. Gorelkin. 2011. Samosinkhronnye skhemy — budushchee mikroelektroniki [Self-timed circuits are the future of microelectronics]. *Voprosy radioelektroniki* [Issues of Radio Electronics] 2:153–184.
2. Bobkov, S. G., M. S. Gorbunov, Yu. G. Diachenko, Yu. V. Rozhdestvenskij, Yu. A. Stepchenkov, and A. V. Surkov. 2014. Ispol'zovanie samosinkhronnoy logiki dlya snizheniya potreblyaemoy moshchnosti i povysheniya nadezhnosti mikroprotsessorov [Delay insensitive circuits for low power and highly reliable microprocessors]. *Conference (International) "Problems of Perspective Micro- and Nanoelectronic Systems Development" Proceedings*. Moscow: IPPM RAN. 1:43–48.

3. Library Characterization, Cadence. Available at: https://www.cadence.com/content/cadence-www/global/en_US/home/tools/custom-ic-analog-rf-design/library-characterization.html (accessed July 5, 2019).
4. CHARISMA: sistema kharakterizatsii bibliotek standartnykh yacheek [CHARISMA: Standard cell library characterization system]. *Radix Tools*. Available at: http://www.radixtools.ru/products-charisma (accessed July 5, 2019).
5. Morozov, N. V., Yu. G. Diachenko, and D. Yu. Stepchenkov. 13.12.2016. Sistema kharakterizatsii samosinkhronnykh elementov SAKhIB. Versiya 2 [SAHIB: Self-timed cell characterization system, version 2]. Certificate on official registration of the computer program No. 2016661383.
6. IEEE Standard Verilog Hardware Description Language. 2001. IEEE Computer Society. IEEE Std 1364-2001 (Revision of IEEE Std 1364-1995). New York, NY: The Institute of Electrical and Electronics Engineers, Inc. 778 p.

Received August 7, 2019

Contributors

Stepchenkov Yuri A. (b. 1951) — Candidate of Science (PhD) in technology, leading scientist, Institute of Informatics Problems, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, 44-2 Vavilov Str., Moscow 119133, Russian Federation; YStepchenkov@ipiran.ru

Diachenko Yuri G. (b. 1958) — Candidate of Science (PhD) in technology, senior scientist, Institute of Informatics Problems, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, 44-2 Vavilov Str., Moscow 119133, Russian Federation; diaura@mail.ru

Morozov Nikolai V. (b. 1956) — senior scientist, Institute of Informatics Problems, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, 44-2 Vavilov Str., Moscow 119133, Russian Federation; NMorozov@ipiran.ru

Stepchenkov Dmitri Yu. (b. 1973) — senior scientist, Institute of Informatics Problems, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, 44-2 Vavilov Str., Moscow 119133, Russian Federation; stepchenkov@mail.ru

Diachenko Denis Yu. (b. 1987) — research engineer, Institute of Informatics Problems, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, 44-2 Vavilov Str., Moscow 119133, Russian Federation; diaden87@gmail.com

СИСТЕМЫ И СРЕДСТВА ИНФОРМАТИКИ, 2019. Т. 29. № 3. С. 114–126

МЕТОД ВЫБОРА ВАРИАНТА ПОСТРОЕНИЯ ИНФОРМАЦИОННО-ТЕЛЕКОММУНИКАЦИОННОЙ СИСТЕМЫ*

***А. А. Зацаринный*[1], *Ю. С. Ионенков*[2]**

Аннотация: Статья посвящена описанию метода выбора варианта построения информационно-телекоммуникационной системы (ИТКС). Рассмотрен общий методологический подход к выбору системотехнических решений построения ИТКС, учитывающий их особенности, принципы и условия построения. Предложен метод выбора варианта построения ИТКС, включающий две взаимоувязанные методики: методику оценки эффективности ИТКС и методику выбора варианта построения ИТКС. Дана общая характеристика методики оценки эффективности ИТКС, представленной в предыдущих публикациях. Разработана методика выбора варианта построения ИТКС с учетом вклада в эффективность соответствующей организационной системы, технической реализуемости и рисков разработки и применения. Предложен перечень частных показателей эффективности для каждой из трех групп обобщенных показателей эффективности (вклад в эффективность организационной системы, техническая реализуемость и риски).

Ключевые слова: информационно-телекоммуникационная система; эффективность; показатель; критерий; технология

DOI: 10.14357/08696527190310

1 Введение

Информация в настоящее время превратилась в один из наиболее важных ресурсов, а ИТКС широко применяются во всех сферах деятельности. При этом все чаще вопросы эффективности применения этих ИТКС рассматриваются с позиций вклада в повышение эффективности функционирования соответствующей организационной системы (министерства, ведомства и т. п.). Информационно-телекоммуникационные системы создаются на основе информационных и телекоммуникационных технологий и реализующих эти технологии аппаратно-программных средств. На этапе разработки ИТКС с учетом имеющегося множества технологий и аппаратно-программных средств появляется

*Работа выполнена при частичной финансовой поддержке РФФИ (проект 18-29-03091).

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

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

множество системотехнических решений, способных обеспечить выполнение требований, предъявляемых к системе. При этом возникает задача выбора рациональных системотехнических решений при построении ИТКС. Это может быть сделано на основе разработки соответствующих методов с использованием оценки эффективности предлагаемых решений.

Таким образом, весьма актуальной задачей является создание методологической базы количественной оценки эффективности вариантов построения ИТКС, которая в существующей научно-технической литературе представлена недостаточно. Кроме того, традиционные подходы к оценке эффективности ИТКС учитывают только затраты и наиболее очевидные прямые эффекты, не рассматривая такие элементы, как их вклад в эффективность соответствующей организационной системы, техническую реализуемость и риски.

Целью настоящей статьи является рассмотрение общего методологического подхода к выбору системотехнических решений, описание метода выбора варианта построения ИТКС на основе двух взаимоувязанных методик, а также разработка методики выбора варианта ИТКС, включая обоснование перечня показателей эффективности.

2 Общая характеристика метода

Задача разработки и применения ИТКС делится на несколько этапов, начинающихся с формирования общего организационно-технического замысла и заканчивающегося выведением из эксплуатации этой системы. С самой общей точки зрения весь процесс, связанный с разработкой и применением ИТКС, состоит из следующих шагов.

1. Формирование организационно-технического замысла системы, т. е. перечня условий, целей и задач проекта, основных требований к системе, бюджета, временны́х, технологических, нормативных и других ограничений, а также исполнителей.
2. Разработка системы — решение проблемы выбора системотехнических решений, технологий, аппаратно-программных средств для создания ИТКС с учетом условий и требований соответствующей организационной системы и с помощью заранее определенных критериев эффективности.
3. Производство — обеспечение технических и технологических условий производства, условий, программ и методик испытаний систем.
4. Применение и поддержка применения, подразумевающие перечень необходимых работ, требуемых для обеспечения непрерывной эксплуатации ИТКС в течение срока службы и для необходимых модернизаций.

Общий методологический подход к задаче оценки системотехнических решений для построения ИТКС включает в себя несколько шагов.

1. Определение совокупности исходных условий разработки и применения ИТКС $N = \{n_j | j = 1, \ldots, q\}$, включающих в себя нормативные, экономические, технические и технологические особенности проектирования.
2. Обоснование перечня параметров ИТКС $T = \{t_i | i = 1, \ldots, p\}$ и требований к ним $U = \{u_i | i = 1, \ldots, p\}$. Требования могут быть представлены как в виде заданных диапазонов, так и в виде точных значений.
3. Формирование перечня возможных системотехнических решений для построения ИТКС $S = \{s_n | n = 1, \ldots, v\}$.
4. Выбор основного критерия Э — показателя эффективности, определенного на множестве альтернатив. Основной критерий должен соответствовать обоснованному значению ценности каждого системотехнического решения с учетом выполнения условий N и требований U.

Таким образом, в формализованном виде задача выбора системотехнических решений для построения ИТКС сводится к следующей: *необходимо осуществить выбор альтернативы S_z из множества альтернатив S при такой совокупности параметров ИТКС T, которая при заданных условиях N позволила бы создать систему, удовлетворяющую заданным требованиям U, с обеспечением наилучшего значения показателя критерия эффективности Э*:

$$Э(N, U, S, T) \rightarrow \max .$$

На основании вышесказанного формальная постановка задачи выбора системотехнических решений может быть представлена в виде схемы, отражающей ряд основных шагов (рис. 1).

Разработка ИТКС производится на основе следующих исходных данных:

– требования к системе, предъявляемые соответствующей организационной системой, для которой создается ИТКС;
– действующая нормативно-техническая база, регламентирующая разработку ИТКС;
– технико-экономические факторы.

Виды требований к большинству отечественных ИТКС определены в ГОСТ РВ 15.201-2003 и ГОСТ 34.602-89. Требования к ИТКС позволяют обосновать их основные характеристики, служат основой для выбора аппаратных и программных средств. При этом требования дополняются и конкретизируются в результате обследования объектов автоматизации, в интересах которых осуществляется разработка.

Действующая нормативно-техническая база, регламентирующая разработку ИТКС, определяет требования к разработке, порядок разработки, взаимодействие заказчика и исполнителя работ, состав разрабатываемых документов.

Технико-экономические факторы определяют условия проведения разработок.

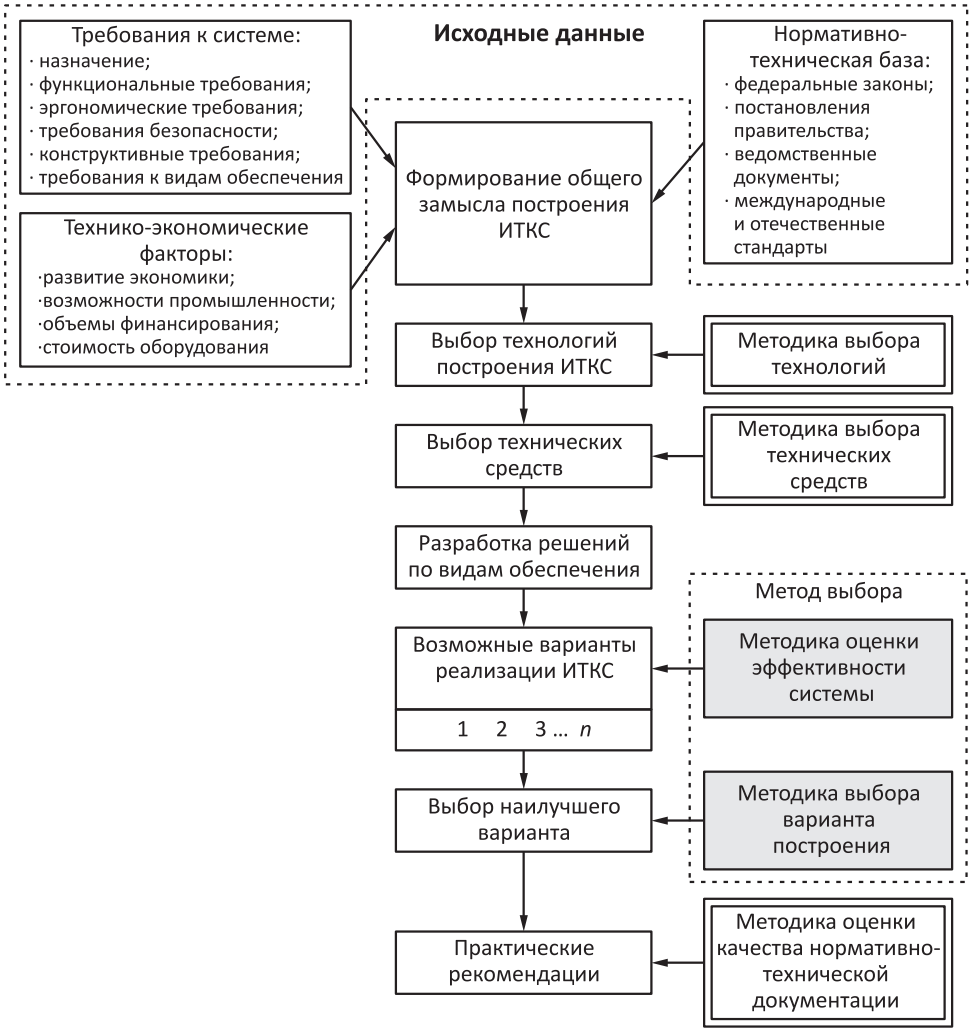


Рис. 1 Общий методический подход к оценке системотехнических решений

Фактор технических и технологических возможностей промышленности особенно актуален для нашей страны в связи с относительной отсталостью отечественной электронной промышленности и промышленности средств связи.

Кроме того, необходимо учитывать тенденции развития информационных и телекоммуникационных технологий, что позволит правильно выбрать технологию и спроектировать систему, позволяющую обеспечить растущие информационные потребности организационной системы с учетом требований информационной безопасности.

Объемы ассигнований, выделяемых на развитие ИТКС, выступают в качестве ограничений, исходя из которых с учетом других исходных данных и стоимостных показателей средств (комплексов) обосновывается номенклатура закупаемых аппаратных и программных средств.

На основании изложенных выше исходных данных осуществляется формирование общего замысла построения ИТКС. Затем осуществляется выбор технологий построения ИТКС, технических средств, а также оценка возможных вариантов ИТКС и выбор наилучшего варианта. Выбор осуществляется с помощью соответствующих методик, учитывающих наиболее важные показатели.

Для достижения указанной цели необходимо разработать комплекс взаимоувязанных методов и методик, которые обеспечили бы решение задачи оценки системотехнических решений на всех стадиях жизненного цикла ИТКС.

Весь комплекс разработанных методик применяется на ранних стадиях жизненного цикла ИТКС — стадиях формирования замысла и разработки систем. На последующих стадиях (производства, применения и поддержки применения) применяются методики оценки эффективности системы с конкретными наборами показателей эффективности, учитывающими специфику соответствующих систем.

Одним из первых этапов в рамках формирования облика ИТКС является выбор технологий, обеспечивающих выполнение функциональных задач системы. Методика выбора технологий достаточно подробно рассмотрена в [1].

Следующий этап — выбор технических средств, реализующих выбранные системотехнические решения и технологии и удовлетворяющих предъявляемым к системе требованиям. Возможная методика выбора технических средств представлена в [2].

Предлагаемый метод выбора варианта ИТКС включает две взаимоувязанные методики: методику оценки эффективности ИТКС и методику выбора варианта построения ИТКС (рис. 2).

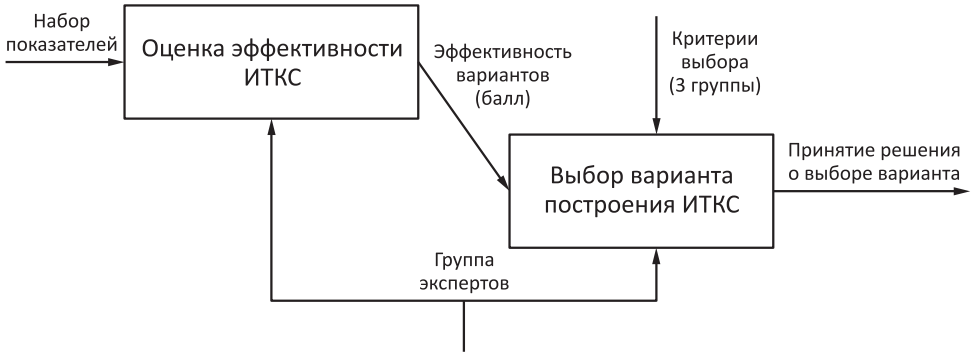


Рис. 2 Общая характеристика метода выбора варианта ИТКС

После выбора по представленным выше методикам применяемых в ИТКС технологий и технических средств формируется множество вариантов системотехнических решений, различающихся техническими, технологическими, экономическими, организационными и другими показателями. Для снижения размерности задачи и сокращения объемов вычислений на первом этапе используется методика оценки эффективности ИТКС, которая обеспечивает выбор наиболее эффективных вариантов, соответствующих требованиям тактико-технического задания на систему, из имеющегося множества. На втором этапе из полученного на первом этапе числа наиболее эффективных вариантов (3–5) с помощью методики выбора варианта ИТКС отбирается наилучший вариант с учетом его вклада в эффективность соответствующей организационной системы, технической реализуемости и рисков разработки и применения.

3 Методика оценки эффективности информационно-телекоммуникационной системы

Под эффективностью системы понимается степень достижения целей, поставленных при ее создании [3].

Для оценки эффективности ИТКС разработана методика на основе средневзвешенной суммы [4, 5], учитывающая совокупность множества количественных данных о функционировании системы в сочетании с применением экспертных оценок.

Одной из важных задач в данной методике является выбор и обоснование номенклатуры показателей системы с точки зрения их влияния на эффективность решения стоящих перед системой задач.

Выбор показателей эффективности целесообразно осуществлять исходя из следующих соображений:

- соответствие показателей целям разработки и назначению системы;
- измеримость с помощью существующих физических (объективных) величин. Желательно выбирать показатели, которые могут быть выражены количественно;
- выбор оптимального числа показателей, так как при их малом числе не в полной мере учитываются целевые функции системы, а с ростом числа показателей возрастает трудоемкость оценки;
- показатели эффективности должны учитывать требования, регламентируемые действующими нормативно-техническими документами в области ИТКС.

В рамках ряда работ ФИЦ ИУ РАН были разработаны перечни показателей эффективности для различных типов ИТКС. В частности, для оценки эффективности ведомственной системы ситуационных центров [4] использовались обобщенные показатели эффективности компонентов ситуационного центра:

функционального, технологического, технического, эксплуатационного и организационного.

Для оценки эффективности системы распределенных ситуационных центров органов государственной власти [6] были предложены такие обобщенные показатели эффективности, как развитие системы, своевременность, полнота, достоверность, организация.

В работе [7] приведены обобщенные и частные показатели эффективности для типовой ИТКС на основе требований ГОСТ РВ 51987-2002, а в [8] обоснованы показатели для ИТКС на основе применения облачных технологий.

В методике оценки эффективности ИТКС для расчета интегрального показателя эффективности используется математическое выражение в виде средневзвешенной суммы обобщенных показателей эффективности их отдельных сторон (факторов).

Обобщенные показатели эффективности для большинства факторов определяются путем аддитивной свертки частных показателей, выбранных для каждого фактора.

Для расчета обобщенного показателя эффективности в части информационной безопасности используется мультипликативная свертка. Применение мультипликативной модели обосновывается тем, что низкая оценка даже по одному показателю в данном случае неприемлема. Аддитивная свертка здесь не подходит, так как частные показатели в данном случае не компенсируют друг друга.

Ряд частных показателей эффективности, имеющих количественное выражение, определяется из технической документации либо расчетным путем по типовым формулам (коэффициент готовности и др.).

Некоторые показатели определяются как отношение имеющегося значения к требуемому.

Другие частные показатели эффективности определяются методом экспертных оценок.

Найденные значения частных показателей переводятся в баллы по десятибалльной шкале и приводятся к значениям от 0 до 1.

Весовые коэффициенты определяются экспертным путем, в том числе с использованием математических методов, например метода попарных сравнений и др.

Данная методика позволяет оценить эффективность ИТКС на разных стадиях их жизненного цикла и выбрать наиболее эффективные варианты их реализации.

4 Методика выбора варианта информационно-телекоммуникационной системы

После того как с помощью методики оценки эффективности ИТКС определены наиболее эффективные варианты реализации системы, встает задача выбрать наилучший из них с учетом его вклада в эффективность соответству-

ющей организационной системы, технической реализуемости и рисков разработки и применения.

С этой целью для каждой из трех групп обобщенных показателей эффективности (вклад в эффективность организационной системы, техническая реализуемость и риски) предлагается несколько частных показателей.

В качестве показателей вклада в эффективность организационной системы предложены:

- уровень реализации задач организационной системы с помощью ИТКС;
- сокращение времени на принятие управленческих решений (определяется как отношение времени принятия решения с помощью конкретной системы к времени принятия аналогичного решения без ее участия);
- сокращение времени на сбор, обработку и доступ к информации;
- повышение уровня доступности информации;
- своевременность предоставления отчетности и трудозатраты на ее подготовку;
- прямые доходы либо экономический эффект, получаемый при внедрении системы.

В качестве показателей технической реализуемости используются:

- уровень развития производственной базы разработчика ИТКС, включая ее технические и технологические характеристики, производственные площади, оборудование и т. п.;
- укомплектованность персоналом;
- наличие отечественного оборудования для ИТКС;
- возможности использования отечественного программного обеспечения;
- наличие у разработчика опыта создания систем аналогичного назначения;
- возможности производственной кооперации.

В качестве показателей рисков применены:

- наличие и степень соблюдения нормативно-правовой базы разработки и применения ИТКС;
- уровень и своевременность финансирования разработки и применения;
- уровень руководства проектом;
- квалификация разработчиков и обслуживающего персонала;
- некомпетентность заказчика или разработчика;
- непредвиденные финансовые расходы.

Источниками информации для получения показателей всех трех групп служат:

- структура предприятий (организационных систем);
- отчеты о деятельности предприятий (организационных систем);
- статистическая и бухгалтерская отчетность;

– положения и должностные инструкции;
– нормативы и стандарты отрасли.

При формировании показателей, естественно, используются и экспертные оценки.

Для расчета обобщенного показателя эффективности и выбора наилучшего варианта ИТКС использован метод анализа иерархий, подробно описанный в [6]. Этот метод позволяет рассматривать иерархию критериев по уровням, проводить сравнение критериев на основе попарных сравнений, а также формализовать как количественную, так и качественную экспертную информацию.

В рамках изложенного выше подхода для выбора варианта построения ИТКС разработана следующая методика с применением метода анализа иерархий.

Шаг 1. Моделирование процесса и математическая постановка задачи. На этом шаге осуществляется построение трехуровневой иерархической структуры:

(1) уровень цели — выбор варианта построения ИТКС;

(2) первый уровень критериев: вклад в эффективность организационной системы, техническая реализуемость и риски разработки и применения;

(3) второй уровень критериев — показатели по каждой группе критериев первого уровня, представленных выше.

Общий вид иерархической структуры представлен на рис. 3.

Шаг 2. Сбор исходных данных и расчет частных показателей эффективности 2-го уровня.

Осуществляется сбор данных, характеризующих представленные выше три группы обобщенных показателей и их обработка и нормирование экспертами.

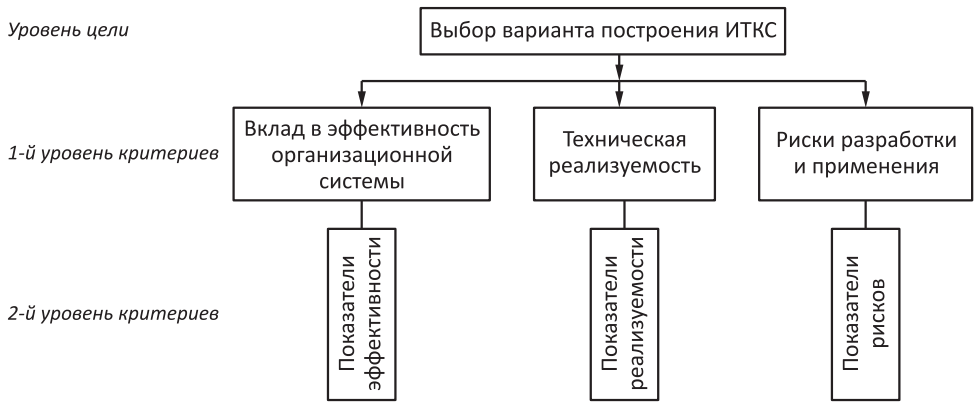


Рис. 3 Иерархия критериев оценки эффективности ИТКС

Ряд показателей определяется экспертным путем, оценивается по десятибалльной шкале и приводится к значениям от 0 до 1.

Шаг 3. Сравнительный анализ элементов процесса.

Для этого с привлечением экспертов производится заполнение обратно-симметричных матриц парных сравнений для критериев первого и второго уровня. После заполнения матриц производится усреднение суждений экспертов по формуле:

$$a_{ij} = \sqrt[n]{a_{ij}^1 a_{ij}^2 \cdots a_{ij}^n},$$

где a_{ij} — элементы матриц, заполненных соответствующими экспертами; n — число экспертов.

Шаг 4. Определение значений весовых коэффициентов для критериев первого и второго уровня из заполненных матриц.

Шаг 5. Оценка степени согласованности экспертов.

Для оценки степени согласованности экспертов используется следующий алгоритм подсчета [9].

1. В матрице парных сравнений суммируются элементы каждого столбца.
2. Сумма элементов каждого столбца умножается на соответствующие нормализованные компоненты вектора весов, определенного из той же матрицы.
3. Полученные числа суммируются, значение суммы определяется как $\lambda_{\max}$.
4. Находится индекс согласованности

$$L = \frac{\lambda_{\max} - n}{n - 1},$$

где n — размер матрицы.

5. Выбирается среднее значение индекса согласованности R для матриц, заполненных случайным образом (см. таблицу).

Значения случайного индекса согласованности [10]

n	R
2	0
3	0,58
4	0,90
5	1,12
6	1,24
7	1,32
8	1,41
9	1,45
10	1,49
11	1,51
12	1,54
13	1,56
14	1,57
15	1,58

6. Вычисляется отношение согласованности $T = L/R$. Если величина $T \leq 0{,}1$, то степень согласованности экспертных данных считается приемлемой [9].

Шаг 6. Определение частных показателей эффективности по критериям второго и первого уровня, а также обобщенного показателя эффективности ИТКС.

Шаг 7. Оценка полученных результатов и выбор наилучшего варианта.

5 Заключение

Задача выбора лучшего из множества вариантов системотехнических решений на стадиях замысла и разработки ИТКС является весьма актуальной. При решении данной задачи необходимо учитывать множество факторов — технических, технологических, экономических, организационных и других, включая возможности разработчиков, риски и т. п.

Для решения указанной задачи в статье предложен метод выбора варианта построения ИТКС, включающий две взаимоувязанные методики: методику оценки эффективности ИТКС и методику выбора варианта построения ИТКС и учитывающий кроме традиционных показателей вклад в эффективность соответствующей организационной системы, техническую реализуемость и степень риска разработки и применения. Разработан перечень частных показателей эффективности для каждой из трех групп обобщенных показателей (вклад в эффективность организационной системы, техническая реализуемость и риски) и методика выбора варианта ИТКС, использующая эти показатели.

Предложенный метод позволяет выбрать оптимальные варианты системотехнических решений для различных типов ИТКС и оценить их вклад в эффективность организационной системы в целом, что имеет большое практическое значение при разработке ИТКС.

Литература

1. *Зацаринный А. А., Ионенков Ю. С.* Некоторые аспекты выбора технологии построения информационно-телекоммуникационных сетей // Системы и средства информатики, 2007. Вып. 17. С. 5–16.
2. *Зацаринный А. А., Ионенков Ю. С.* Методика выбора технических средств для построения телекоммуникационных сетей // Системы и средства информатики, 2009. Доп. выпуск. С. 4–14.
3. ГОСТ 34.003-90. Информационная технология. Автоматизированные системы. Термины и определения. — М.: Стандартинформ, 2005. 14 с.
4. *Зацаринный А. А., Ионенков Ю. С., Шабанов А. П.* К вопросу о сравнительной оценке эффективности ситуационных центров // Системы и средства информатики, 2013. Т. 23. № 2. С. 155–171.
5. *Ионенков Ю. С.* Методический подход к оценке эффективности информационно-телекоммуникационных систем // Математическое моделирование и информационные технологии в инженерных и бизнес-приложениях: Сб. мат-лов Междунар. научн. конф. — Воронеж: ВГУ, 2018. С. 209–217.
6. *Зацаринный А. А., Ионенков Ю. С.* К вопросу оценки эффективности автоматизированных систем с использованием метода анализа иерархий // Системы и средства информатики, 2015. Т. 25. № 3. С. 161–178.
7. *Ионенков Ю. С.* Научно-практические аспекты оценки эффективности информационно-телекоммуникационных систем // Радиолокация, навигация, связь: Сб. трудов XXIV Междунар. научн.-технич. конф. — Воронеж: Вэлборн, 2018. Т. 1. С. 140–149.

8. *Зацаринный А. А., Ионенков Ю. С., Сучков А. П.* Некоторые аспекты оценки эффективности облачных технологий // Системы и средства информатики, 2018. Т. 28. № 3. С. 106–119.
9. *Ларичев О. И.* Теория и методы принятия решений. — М.: Логос, 2007. 392 с.
10. *Саати Т.* Принятие решений. Метод анализа иерархий. — М.: Радио и связь, 1993. 278 с.

Поступила в редакцию 15.08.19

THE METHOD OF SELECTING A VARIANT OF THE CONSTRUCTION OF INFORMATION AND TELECOMMUNICATION SYSTEMS

***A. A. Zatsarinny*[1] *and Yu. S. Ionenkov*[2]**

[1]Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, 44-2 Vavilov Str., Moscow 119333, 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: The article is devoted to the description of the method of choosing the option of building an information and telecommunication system (ITCS). The authors discuss the general methodological approach to the selection of solutions to system integrators of building ITCS taking into account their characteristics, principles, and conditions of the build. The method of choice of variants of building an ITCS includes two interrelated techniques: the methodology to assess the effectiveness of ITCS and the method of selection of variants of ITCS. The authors describe the methodology for assessing the effectiveness of the ITCS developed in previous publications. The method of choosing the option of building ITCS takes into account the contribution to the effectiveness of the relevant organizational system, technical feasibility, and risks of development and application. A list of specific performance indicators for each of the three groups of generalized performance indicators (contribution to the efficiency of the organizational system, technical feasibility, and risks) is proposed.

Keywords: information and telecommunication system; efficiency; indicator; criterion; technology

DOI: 10.14357/08696527190310

Acknowledgments

The work was partly supported by the Russian Foundation for Basic Research (project 18-29-03091).

References

1. Zatsarinnyy, A. A., and Yu. S. Ionenkov. 2007. Nekotorye aspekty vybora tehnologii postroeniya informatsionno-telekommunikatsionnykh setey [Some aspects of the choice

of technology for building information and telecommunication networks]. *Sistemy i Sredstva Informatiki — Systems and Means of Informatics* 17:5–16.
2. Zatsarinnyy, A. A., and Yu. S. Ionenkov. 2009. Metodika vybora tekhnicheskikh sredstv dlya postroeniya telekommunikatsionnykh setey [Method of selection of technical means for construction of telecommunication networks]. *Sistemy i Sredstva Informatiki — Systems and Means of Informatics*. Additional Issue:4–14.
3. GOST 34.003-90. 2005. Informatsionnaya tekhnologiya. Kompleks standartov na avtomatizirovannye sistemy. Avtomatizirovannye sistemy. Terminy i opredeleniya [Information technology. Set of standarts for automated system. Automated system. Terms and definitions]. Collection of State Standarts. Moscow: Standardinform Publs. 14 p.
4. Zatsarinnyy, A. A., Yu. S. Ionenkov, and A. P. Shabanov. 2013. K voprosu o sravnitel'noy otsenke effektivnosti situatsionnykh tsentrov [Regarding comparative evaluation of situational centers efficiency]. *Sistemy i Sredstva Informatiki — Systems and Means of informatics* 23(2):155–171.
5. Ionenkov, Yu. S. 2018. Metodicheskiy podkhod k otsenke effektivnosti informatsionno-telekommunikatsionnykh system [Methodical approach to evaluation of information and telecommunication systems efficiency]. *Mathematical Modeling and Information Technologies in Engineering and Business Applications: Scientific Conference (International) Proceedings*. Voronezh: VSU Publs. 209–217.
6. Zatsarinnyy, A. A., and Yu. S. Ionenkov. 2015. K voprosu otsenki effektivnosti avtomatizirovannykh sistem s ispol'zovaniem metoda analiza ierarkhiy [Regarding automated systems efficiency evaluation using analytic hierarchy process]. *Sistemy i Sredstva Informatiki — Systems and Means of Informatics* 25(3):161–178.
7. Ionenkov, Yu. S. 2018. Nauchno-prakticheskie aspekty otsenki effektivnosti informatsionno-telekommunikatsionnykh system [Scientific and practical aspects of evaluating the effectiveness of information and telecommunication systems]. *Radar, Navigation, Communication: 24th Scientific and Technical Conference (International) Proceedings*. Voronezh: Walburn. 1:140–149.
8. Zatsarinny, A. A., Yu. S. Ionenkov, and A. P. Suchkov. 2018. Nekotorye aspekty otsenki effektivnosti oblachnykh tekhnologiy [Some aspects of cloud computing efficiency estimation]. *Sistemy i Sredstva Informatiki — Systems and Means of Informatics* 28(3):106–119.
9. Larichev, O. I. 2007. *Teoriya i metody prinyatiya resheniy* [Theory and methods of decision making]. Moscow: Logos. 392 p.
10. Saati, T. 1993. *Prinyatie resheniy. Metod analiza ierarkhiy* [Decision making. Analytic hierarchy process]. Moscow: Radio i svyaz'. 278 p.

Received August 15, 2019

Contributors

Zatsarinny Alexander A. (b. 1951) — Doctor of Science in technology, professor, Deputy Director, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, 44-2 Vavilov Str., Moscow 119333, Russian Federation; AZatsarinny@ipiran.ru

Ionenkov Yurij S. (b. 1956) — 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; uionenkov@ipiran.ru

СИСТЕМЫ И СРЕДСТВА ИНФОРМАТИКИ, 2019. Т. 29. № 3. С. 127–138

О ПРОБЛЕМЕ ИНТЕГРАЦИИ ИНФОРМАЦИОННЫХ РЕСУРСОВ*

С. К. Дулин[1], *И. Н. Розенберг*[2], *В. И. Уманский*[3]

Аннотация: Анализируются процессы, характерные для интеграции информации, знаний аналитиков и их совместных действий в информационной среде. Под интеграцией знаний здесь понимается процедура синтеза существующих знаний с целью получения новых. Выделены и рассмотрены три этапа аналитической деятельности и их особенности. Предложено проводить интеграцию информационных ресурсов на основе динамической реструктуризации базы знаний для поддержания ее структурной согласованности и представления ее в виде структурированной совокупности информационных ресурсов в соответствии с требованиями интероперабельности. Для решения подобных задач авторами используется методика, основанная на индуктивно-комбинаторном аппарате сравнения структур связей произвольного множества и одного из типов согласованных множеств. Эта методика выбрана авторами в качестве теоретической базы реализации рассматриваемых задач.

Ключевые слова: информационные ресурсы; реструктуризация базы знаний; интероперабельность; интеграция информации

DOI: 10.14357/08696527190311

1 Введение

Для поддержки принятия решений общепризнана необходимость существования информационной среды, включающей в себя конечных пользователей и содержащей слабосвязанные источники знаний: аналитиков, базы знаний, репозитории, электронные документы.

В настоящее время пользователям открыт доступ к огромным массивам информации и существует возможность объединить аналитических работников в единую информационную среду при тесном взаимодействии друг с другом, если обеспечено решение проблемы интероперабельности на уровне, выше чем техническом [1]. Масштабы современных проблем определяют необходимость

*Работа выполнена при финансовой поддержке РФФИ (проект 17-20-02153 офи_м_РЖД).

[1]Институт проблем информатики Федерального исследовательского центра «Информатика и управление» Российской академии наук; Научно-исследовательский и проектно-конструкторский институт информатизации, автоматизации и связи на железнодорожном транспорте (АО НИИАС), skdulin@mail.ru

[2]Научно-исследовательский и проектно-конструкторский институт информатизации, автоматизации и связи на железнодорожном транспорте (АО НИИАС), I.Rozenberg@vniias.ru

[3]Научно-исследовательский и проектно-конструкторский институт информатизации, автоматизации и связи на железнодорожном транспорте (АО НИИАС), umanvi@yandex.ru

привлечения групп специалистов из одной или разных областей знаний, порождая задачу интеграции их знаний, информации и действий в процессе их деятельности в информационной среде.

При всем богатстве и разнообразии существующих на сегодня инструментальных средств работы с информацией они, к сожалению, не отвечают современным требованиям и оказываются неэффективными в ситуациях, когда необходима реструктуризация базы знаний аналитика в соответствии с динамично изменяющимися условиями и характером решаемых задач. Для решения этой проблемы авторами была разработана методика, опирающаяся на концепцию модели предметной области аналитика, которая представляется в виде сетевой структуры взаимосвязанных объектов. Динамическая реструктуризация базы знаний осуществляется на основе разработанного алгоритма поддержания ее структурной согласованности [2], инвариантной к специфике предметной области и допускающей с определенной степенью точности моделирование семантической согласованности [3]. На базе этой методики авторами разработан макет программно-технического комплекса, предлагаемый в качестве прототипа автоматизированного рабочего места аналитика.

2 Элементы аналитической деятельности

Процесс накопления знаний носит итеративный характер. Схема преобразования неявных (tacit) знаний в явные (explicit) [4] состоит из четырех основных этапов: обобществления, формализации, диссеминации и усвоения (рис. 1). Под обобществлением понимается обобщенный процесс выработки аналитиком неявных знаний. Этот процесс может принимать самые разнообразные формы: участие в семинарах и конференциях, непосредственное общение с другими экспертами, проведение опытов и экспериментов и т. д. Формализация предполагает изложение аналитиком своих знаний в той или иной явной, например наглядной или читаемой, форме. Нужно четко различать две формы представления знаний. Одна форма связана с тем, как и в каких моделях хранятся знания у человека. Другая форма связана с тем, как они будут описаны и представлены в системе знаний. От степени согласованности этих двух форм репрезентации знаний между собой зависит эффективность системы знаний [5]. Чтобы ознакомить с результатами своего труда других специалистов, аналитик преобразует форму представления знания из одной в другую и распространяет ее по доступным ему каналам связи. Этот процесс назван в упомянутой схеме диссеминацией. На четвертом этапе происходит изучение экспертами полученного ими знания, его осознание и усвоение.

Этап диссеминации, по своей сути являющийся процессом приобретения новых знаний, связан с технологическими аспектами обработки, хранения и передачи информации. Следовательно, эффективность соответствующих процедур и программ, реализующих эти функции как составные (базисные) элементы управляющей системы, в значительной степени определяет степень интеропера-

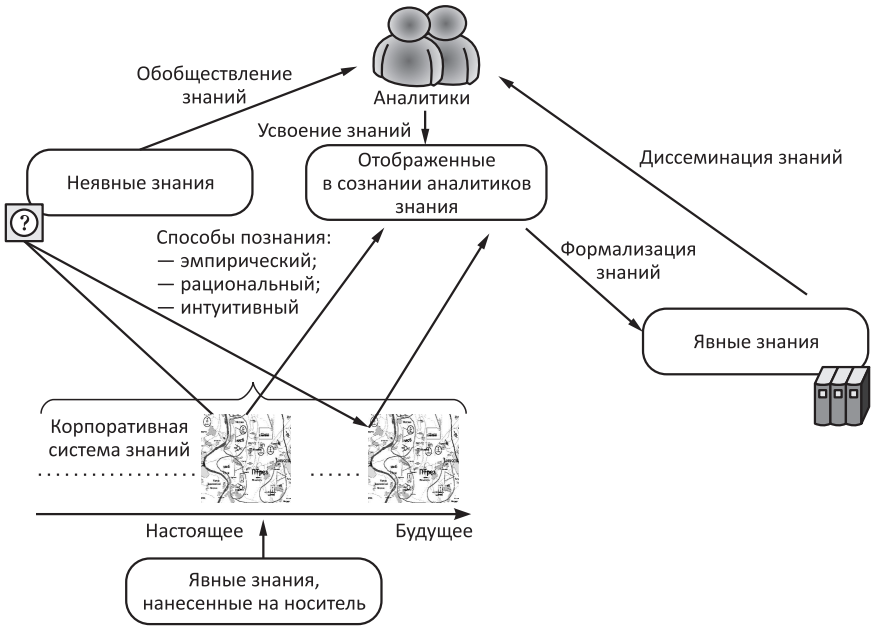


Рис. 1 Схема преобразования неявных знаний в явные

Рис. 2 Этапы создания модели предметной области для управления интеграцией знаний в информационной среде

бельности аналитиков, т. е. степень их взаимодействия на уровне знаний своей предметной области.

Модель предметной области аналитика, предлагаемая в качестве основного инструментария обеспечения его деятельности, должна удовлетворять определенным требованиям, которые диктуются особенностями этой деятельности и ее содержанием. Однако аналитическая деятельность представляет собой сложный процесс, который носит преимущественно творческий характер и практически не поддается какому-либо алгоритмическому описанию. Тем не менее авторам представляется возможным выделить основные элементы этой деятельности, условно представив процесс решения задачи создания модели предметной области поделенным на три этапа (рис. 2). Анализ этих этапов позволит сформулировать основные требования к средствам обеспечения интеграции знаний и определить свойства, которыми должна обладать модель предметной области.

2.1 Постановка задачи

В настоящее время этап постановки задачи носит больше организационный характер, чем исследовательский. Главное, что определяется на этом этапе, — это цель предстоящей работы и условия ее выполнения. Условия могут накладывать определенные ограничения, и их необходимо учитывать при планировании всей работы. Цель важна прежде всего с той точки зрения, что позволяет в определенной степени формализовать стоящую перед аналитиком задачу, конкретизировать ее и определить степень ее соответствия предметной области.

В рамках предлагаемого авторами подхода структурная схема задачи может быть представлена в качестве подмодели общей модели рассматриваемой предметной области, которая фактически является объединением всех подмоделей. Поэтому при постановке новой аналитической задачи ее структура может быть частично определена путем поиска в базе знаний той подмодели, которая соответствует наиболее близкой к ней задаче. Понятие «близости» решаемых аналитиком задач позволяет рассматривать всю совокупность этих задач как множество взаимосвязанных объектов и рассматривать проблему его разбиения на классы. При возникновении новой задачи, т. е. при добавлении к этому множеству объектов нового элемента, возникает необходимость реструктуризации этого множества с целью определения того класса задач, в который попадает новая. Представление об этом классе и том месте, которое занимает в нем стоящая перед аналитиком задача, позволяет ему в значительной мере формализовать основные цели планируемого исследования и адекватно спланировать свою работу.

2.2 Мобилизация знаний

Второй этап решения аналитической задачи, хотя и носит подготовительный характер, является исключительно важным, поскольку его основная задача — сбор всей информации, которой располагает аналитик в данной области и которая имеет непосредственное отношение к стоящей перед ним задаче. Основная проб-

лема здесь состоит в том, что требуется не только локализовать информацию по заданной теме, но и представить ее в виде структурированной совокупности информационных ресурсов в соответствии с требованиями семантической интероперабельности [6], что в значительной степени облегчило бы задачу ее аналитической обработки пользователями и аналитиками.

В настоящее время существуют достаточно мощные средства поиска тематической информации, каждое из которых имеет свои неоспоримые преимущества, но, к сожалению, ни одно из них пока не отвечает всем предъявляемым к ним требованиям. Несмотря на все разнообразие этих средств, условно их можно разделить на три класса: (1) поисковые системы; (2) тематические рубрикаторы; (3) интеллектуальные агенты.

Поисковые системы относятся к наиболее мощным и наиболее популярным средствам поиска информации в сети Интернет. Поиск в них осуществляется на основе логической комбинации ключевых слов, которая в той или иной степени характеризует область интереса пользователя.

Тематические рубрикаторы (классификаторы) представляют собой другой тип поисковых средств, также широко используемый в интернете. Они представляют собой иерархически организованный список рубрик, который просматривается пользователем в поисках интересующей его информации. Рубрикаторы могут иметь различную глубину иерархии, а для каждой рубрики нижнего уровня имеется список документов, отвечающих данной конкретной тематике.

Интеллектуальные агенты представляют собой поисковые средства следующего поколения. Это один из типов программных роботов, которые действуют автономно, осуществляя поиск необходимой информации на основе некоторой информационной модели, задаваемой пользователем и отражающей область его интереса. Они оказываются способными определять информационно значимые части в больших массивах информации и выявлять внутренние связи между различными информационными объектами.

Интеллектуальные агенты осуществляют поиск информации, основываясь не на строго определенных ключевых словах или фразах, а на информационной модели, заложенной в них пользователем.

Данные, которыми располагает аналитик, образуют совокупность взаимосвязанных информационных объектов (документов). Связи между ними могут носить различный характер, и «вес» той или иной связи в значительной мере зависит от решаемой задачи. Это означает, что рассматриваемую совокупность можно представить в виде графа, где между вершинами существуют множественные связи с определенными числовыми коэффициентами, а задача поиска решается определением «значимости» того или иного типа взаимосвязей и соответствующим разбиением графа на подмножества.

Подобные задачи обычно решаются методами кластеризации, которых существует достаточно большое число. Однако при всей эффективности этих методов они не в состоянии полностью решить проблемы, характерные для рассматриваемой области приложения. Существующие в настоящее время алго-

ритмы кластеризации данных [7] представляют собой, как правило, достаточно четкую и жесткую последовательность выполнения шагов и не предполагают вмешательства извне, предлагая пользователю результаты классификации как конечный результат поиска информации. При этом степень соответствия этих результатов реальным потребностям пользователя или то, каким образом и за счет чего они получены, остается неизвестным. Отсутствует и возможность содержательной оценки целесообразности выбора того или иного набора признаков, по которым осуществляется данная классификация. В связи с этим авторам представляется целесообразным предоставить пользователю (аналитику) возможность контролировать ход реализации алгоритма кластеризации [7] (управлять направлением поиска) и при необходимости вмешиваться в этот процесс. Ряд проведенных авторами экспериментов показал, что этим требованиям отвечает алгоритм, основанный на понятии согласованности объектов по набору выбираемых признаков [8], который и лег в основу предлагаемого подхода.

2.3 Анализ и поиск

Этап, связанный с анализом и поиском информации, составляет основу всей аналитической деятельности эксперта. На этом этапе решаются следующие основные задачи: определение недостающих сведений, поиск информации, ввод и реорганизация базы знаний.

Определение недостающих сведений играет ключевую роль в планировании дальнейших действий эксперта, связанных с поиском требуемой информации. Если учитывать, что в интернете информация отличается слабой структурированностью, распределенностью по множеству потенциальных источников и большими объемами, то для эффективной локализации недостающих знаний исключительно важной становится конкретизация объекта поиска (рис. 3).

Предлагаемая авторами концепция модели предметной области способна оказать аналитику определенную помощь в решении этой задачи. Как уже указывалось, каждая решаемая аналитиком задача представляется в виде подмодели, реализованной в виде некоторой сетевой структуры. Изучение взаимосвязей между отдельными элементами этой структуры и реструктуризация ее в соответствии с выбираемыми критериями согласованности позволяют разбивать все множество рассматриваемых в модели информационных объектов на консонансные подмножества. В работе [2] в качестве визуализации подобных структур предлагалась знаковая матрица сходства между различными объектами. Согласно этому подходу, консонансным будет подмножество, все связи между объектами которого «положительны», а связи элементов этого подмножества с остальными — «отрицательны». Изучение подобной знаковой интерпретации рассматриваемой аналитиком подмодели позволяет определять элементы, где эта модель рассогласована, например где элементы одного подмножества имеют отрицательную взаимосвязь. Подобная рассогласованность позволяет выдвинуть гипотезы о недостающих знаниях и тем самым конкретизировать объект поиска.

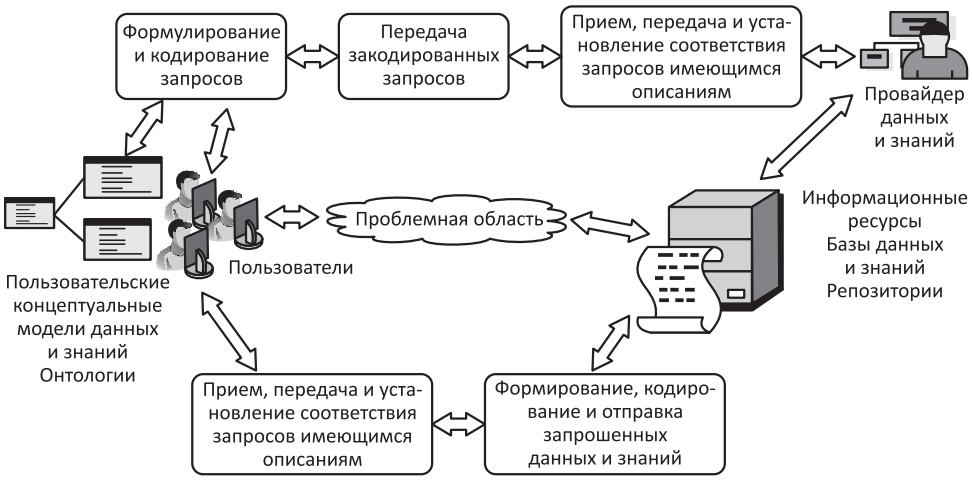


Рис. 3 Локализация, поиск и кодирование знаний

Умение находить подмножество предметной области, адекватное каждой конкретной задаче, стоящей перед аналитиком, является исключительно важным фактором, поскольку способно в значительной степени повлиять на эффективность поиска, сужая его на стадии выбора потенциальных источников информации. Если в качестве одного из параметров накапливаемых аналитиком данных указывать источник их получения, то структурная схема (подмодель) рассматриваемой задачи, о которой говорилось в предыдущих разделах, может быть преобразована в совокупность взаимосвязанных объектов, представляющих собой информационные источники. Взаимосвязь этих объектов основывается на тематической близости соответствующей им информации.

Результаты поиска требуют своей аналитической обработки, которая предполагает генерацию новых знаний на основе полученных экспертом данных. Следовательно, одним из результатов аналитической деятельности является расширение базы знаний, и, соответственно, возникают проблемы, связанные с поддержанием ее согласованности и непротиворечивости.

В базе знаний связи между компонентами знаний носят неоднозначный характер, поэтому всякое изменение в них, связанное с добавлением нового знания или устранением существующего, требует проверки всех связей, в которых участвует это знание. Такая проверка носит глобальный характер и может приводить к реорганизации знаний так, как это, по всей видимости, происходит у аналитика. Но реорганизация структуры знаний у него происходит не только в процессе получения знаний, но и в процессе анализа уже имеющейся информации. Это дает возможность предположить, что система знаний у человека активна, так как в ней можно увидеть «. . .наличие классифицирующих связей с соответствующими

процедурами обобщения, пополнения и выявления противоречивых и несовместимых в рамках одного описания знаний, что вызывает активацию знаний, при которой внутренние потребности самой системы представления знаний становятся потребностями в реализации определенных программ деятельности» [9].

Таким образом, фактически любые результаты поиска требуют проведения определенной реструктуризации имеющихся знаний. Для решения подобных задач в [2] была предложена методика, основанная на индуктивно-комбинаторном аппарате сравнения структур связей произвольного и одного из типов согласованных множеств, которая выбрана авторами в качестве теоретической базы реализации рассматриваемых задач.

3 Вопросы интеграции

Интеграция в единую информационную среду разнообразных по тематике и составу источников данных и аналитических работников, решающих сходные по тематике или характеру задачи, требует развития методов и средств интероперабельности. Наибольшую актуальность это развитие приобретает при построении крупных корпоративных сетей, основанных на тех же самых технологиях, где в качестве источников информации могут выступать как информационные массивы в сети Интернет, так и внутрикорпоративные (локальные) базы данных. Интеграция аналитических работников в единую информационную среду предполагает прежде всего выполнение условий согласованной интероперабельности [6], направленных на решение общих задач, что должно обеспечиваться совместным планированием поиска, интеграцией и синтезом документов и ведением обобщенной базы знаний, где вопросы их согласования приобретают весьма большое значение.

Первая задача, которая возникает при рассмотрении вопросов интероперабельности аналитических работников, связана с доступностью результатов поиска каждого из них для других.

В действительности эта задача может быть поставлена шире. Речь может идти не только о поиске данных и знаний, которыми располагают эксперты, а об интеграции их знаний, что является существенно более сложной, но и более важной проблемой. При решении той или иной общей задачи каждый из участников этого процесса «видит» проблему по-своему в соответствии со своими знаниями, опытом, интуицией или другими факторами. В соответствии с этим аналитики строят свои модели решаемых задач, которые затем интегрируются в единое хранилище информационных ресурсов (рис. 4).

Множественность таких моделей обусловливает возможную рассогласованность интегрированной базы знаний и, как следствие, необходимость ее реструктуризации.

Реструктуризация базы знаний приводит, естественно, и к соответствующим изменениям в индивидуальных моделях решаемой задачи, что ведет к использо-

Рис. 4 Интеграция моделей решаемых задач в хранилище информационных ресурсов

ванию в работе того или иного аналитика результатов коллективного творчества, которые отличаются большей степенью объективности и универсальности.

4 Заключение

Междисциплинарность современных проблем определяет необходимость привлечения групп специалистов из одной или разных областей знания, порождая задачу интеграции их знаний, информации и действий в процессе их деятельности. В работе выделены три этапа процесса решения задачи создания модели предметной области. Обусловлена необходимость динамической реструктуризации базы знаний для поддержания ее структурной согласованности и представления ее в виде структурированной совокупности информационных ресурсов в соответствии с требованиями семантической интероперабельности. Интеграция информационных ресурсов в единой информационной среде предполагает успешное решение задач тематического поиска и синтеза интегрируемых данных и знаний.

Литература

1. *Lemmens R.* Semantic interoperability of distributed geo-services. Publications on Geodesy 63 Netherlands Geodetic Commission. — Rotterdam, Netherlands: Optima Graphic Communication, 2006. 312 p.
2. *Дулин С. К., Розенберг И. Н.* Об одном подходе к структурной согласованности геоданных // Мир транспорта, 2005. № 3. С. 16–29.
3. *Du Shihong, Qin Qimin, Wang Qiao, Ma Haijian.* Evaluating structural and topological consistency of complex regions with broad boundaries in multi-resolution spatial databases // Inform. Sciences, 2008. Vol. 178. Iss. 1. P. 52–68.
4. *Nonaka Ikujiro, Takeuchi Hirotaka.* Theoretical models of information and knowledge management. Spiral model. http://www.tlu.ee/~sirvir/IKM/Theoretical_models_of_Information_and_Knowledge_Management/the_nonaka_and_takeuchi_knowledge_spiral_model_page_3.html.
5. *Schreiber G.* Knowledge acquisition and the web // Int. J. Hum.-Comput. St., 2013. Vol. 71. P. 206–210.
6. *Дулин С. К., Дулина Н. Г., Никишин Д. А.* О проблемах реализации семантической геоинтероперабельности в SEMANTIC WEB // Системы и средства информатики, 2014. Т. 24. № 2. С. 143–165.
7. *Xu Rui, Wunsch D.* Survey of clustering algorithms // IEEE T. Neural Networ., 2005. Vol. 16. Iss. 3. P. 645–678.
8. *Дулин С. К., Розенберг И. Н., Уманский В. И.* Структуризация проблемы улучшения пространственной согласованности баз геоданных арктической зоны // Системы высокой доступности, 2017. Т. 13. № 3. С. 3–14.
9. Представление знаний в человеко-машинных и робототехнических системах. Том А: Фундаментальные исследования в области представления знаний. — М.: ВЦ АН СССР, 1984. 262 с.

Поступила в редакцию 25.06.19

ABOUT THE PROBLEM OF INFORMATION RESOURCES INTEGRATION

***S. K. Dulin*[1,2], *I. N. Rozenberg*[2], *and V. I. Umanskiy*[2]**

[1]Institute of Informatics Problems, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, 44-2 Vavilov Str., Moscow 119133, Russian Federation

[2]Research & Design Institute for Information Technology, Signalling and Telecommunications on Railway Transport (JSC NIIAS), 27-1 Nizhegorodskaya Str., Moscow 109029, Russian Federation

Abstract: The paper analyzes the processes typical for the integration of information, knowledge of analysts, and their joint actions in the information environment. Here, the integration of knowledge refers to the procedure for the

synthesis of existing knowledge in order to obtain the new one. Three stages of analytical activity were identified and considered as well as their features. It is proposed to integrate information resources based on dynamic restructuring of the knowledge base to maintain its structural consistency and to present it as a structured set of information resources in accordance with the requirements of interoperability. To solve such problems, the authors use a technique based on an inductive-combinatorial apparatus comparing the structures of the connections of an arbitrary set and one of the types of consistent sets. This technique was chosen by the authors as the theoretical basis for the implementation of the tasks under consideration.

Keywords: information resources; knowledge base restructuring; interoperability; information integration

DOI: 10.14357/08696527190311

Acknowledgments

The work was supported by the Russian Foundation for Basic Research (project 17-20-02153 ofi_m_RZhD).

References

1. Lemmens, R. 2006. Semantic interoperability of distributed geo-services. Publications on Geodesy 63 Netherlands Geodetic Commission. Rotterdam, Netherlands: Optima Graphic Communication. 312 p.
2. Dulin, S. K., and I. N. Rozenberg. 2005. Ob odnom podkhode k strukturnoy soglasovannosti geodannykh [On an approach to structural consistency of geodata]. *Mir transporta* [World of Transport and Transportation] 3:16–29.
3. Du, Shihong, Qimin Qin, Qiao Wang, and Haijian Ma. 2008. Evaluating structural and topological consistency of complex regions with broad boundaries in multi-resolution spatial databases. *Inform. Sciences* 178(1):52–68.
4. Nonaka, Ikujiro, and Hirotaka Takeuchi. 2019. Theoretical models of information and knowledge management. Spiral model. Available at: http://www.tlu.ee/~sirvir/IKM/Theoretical_models_of_Information_and_Knowledge_Management/the_nonaka_and_takeuchi_knowledge_spiral_model_page_3.html (accessed March, 2019).
5. Schreiber, G. 2013. Knowledge acquisition and the web. *Int. J. Hum.-Comput. St.* 71:206–210.
6. Dulin, S. K., N. G. Dulina, and D. A. Nikishin. 2014. O problemakh realizatsii semanticheskoy geointeroperabel'nosti v SEMANTIC WEB [On the problems of implementing semantic geo-interoperability in SEMANTIC WEB]. *Sistemy i Sredstva Informatiki — Systems and Means of Informatics* 24(2):143–165.
7. Xu, Rui, and D. Wunsch. 2005. Survey of clustering algorithms. *IEEE T. Neural Networ.* 16(3):645–678.
8. Dulin, S. K., I. N. Rozenberg, and V. I. Umanskij. 2017. Strukturizatsiya problemy uluchsheniya prostranstvennoy soglasovannosti baz geodannykh arkticheskoy zony [Structuring the problem of improving the spatial consistency of geodatabases in the Arctic zone]. *Sistemy vysokoy dostupnosti* [High Availability Systems] 13(3):3–14.

9. Computer Center of the USSR Academy of Sciences. 1984. *Predstavleniye znaniy v cheloveko-mashinnykh i robototekhnicheskikh sistemakh. Tom A. Fundamental'nye issledovaniya v oblasti predstavleniya znaniy* [Representation of knowledge in man-machine and robotic systems. Vol. A. Basic research in the field of knowledge representation]. Moscow. 262 p.

Received June 25, 2019

Contributors

Dulin Sergey K. (b. 1950) — 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 119133, Russian Federation; principal scientist, Research & Design Institute for Information Technology, Signalling and Telecommunications on Railway Transport (JSC NIIAS), 27-1 Nizhegorodskaya Str., Moscow 109029, Russian Federation; skdulin@mail.ru

Rozenberg Igor N. (b. 1965) — Doctor of Science in technology, professor, Director General, Research & Design Institute for Information Technology, Signalling and Telecommunications on Railway Transport (JSC NIIAS), 27-1 Nizhegorodskaya Str., Moscow 109029, Russian Federation; I.Rozenberg@vniias.ru

Umanskiy Vladimir I. (b. 1954) — Doctor of Science in technology, First Deputy Director General, Research & Design Institute for Information Technology, Signalling and Telecommunications on Railway Transport (JSC NIIAS), 27-1 Nizhegorodskaya Str., Moscow 109029, Russian Federation; umanvi@yandex.ru

МОДЕЛИРОВАНИЕ КОНФЛИКТОВ АГЕНТОВ В ГИБРИДНЫХ ИНТЕЛЛЕКТУАЛЬНЫХ МНОГОАГЕНТНЫХ СИСТЕМАХ

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

Аннотация: Управление конфликтами — неотъемлемая часть процесса решения проблем коллектива экспертов за «круглым столом», обеспечивающая поощрение положительно влияющих на ход решения проблемы конфликтов и предотвращение или разрешение всех остальных. Существующие модели гибридных интеллектуальных многоагентных систем (ГиИМАС) обладают существенным недостатком — конфликты агентов в них не моделируются, а итоговые решения принимаются единственным агентом на основе рекомендаций других агентов. Моделирование конфликтов в ГиИМАС позволит управлять ходом «обсуждения», активируя различные типы коллективного мышления в зависимости от характера и интенсивности конфликта, что обеспечит их релевантность малым коллективам экспертов, успешно решающим проблемы, характеризующиеся высокой комбинаторной сложностью, неоднородностью, недоопределенностью и другими НЕ-факторами. С этой целью в работе предлагаются модели проблемно- и процессно-ориентированного конфликта в ГиИМАС.

Ключевые слова: конфликт; гибридная интеллектуальная многоагентная система; коллектив экспертов; круглый стол

DOI: 10.14357/08696527190312

1 Введение

Решение проблем в биопроизводственных, социально-экономических и технических системах, например планирование урожая и агротехнических мероприятий [1], дифференциальная диагностика заболеваний [2], оперативное планирование мелкосерийного производства [3], восстановление региональной электросети после масштабных аварий [4], осложняется их неоднородностью, необходимостью принятия нескольких последовательных взаимосвязанных решений в реальном времени, что делает нерелевантными традиционные абстрактно-математические модели управления. Традиционно для преодоления неоднородности используются методы системного анализа и привлекаются коллективы экспертов со специфическими для каждого специалиста моделями внешнего мира, обусловливающими его понимание и решение проблемы. Взаимодействуя при решении проблем, эксперты не только предлагают свои частные решения,

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

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

но и обмениваются мнениями, опытом, знаниями, обеспечивают эволюцию коллектива и его адаптацию к изменениям среды. В качестве источников изменения и эволюции коллектива выступают, как правило, противоречия и конфликты [5]. Согласно современным представлениям о конфликте, он не просто деструктивен или продуктивен, но одновременно содержит в себе обе стороны, поэтому для обеспечения эффективности работы и развития коллектива требуется не избегать конфликтов, а управлять ими, подавляя деструктивные элементы и стимулируя конструктивные [6].

Концепция интеграции разнородных, дополняющих друг друга знаний в интересах компьютерной поддержки индивидуальных решений воплотилась в гибридных интеллектуальных системах А. В. Колесникова [1], а расширение ее аппаратом многоагентных систем в смысле В. Б. Тарасова [7] позволило моделировать процесс коллективного решения проблем [3]. Однако в существующих ГиИМАС моделируется лишь небольшое число групповых процессов и эффектов, присущих реальным коллективам экспертов, в частности конфликты агентов в них не моделируются, а итоговые решения принимаются единственным агентом на основе рекомендаций других агентов. Цель настоящей работы — расширение существующих моделей ГиИМАС моделями проблемно- и процессно-ориентированного конфликта для повышения их релевантности малым коллективам экспертов, решающих проблемы «за круглым столом».

2 Конфликт в малых коллективах экспертов, решающих проблемы «за круглым столом»

Под коллективом принято понимать малую группу, соответствующую набору высоких требований: успешно справляется с решением поставленных проблем, хорошие отношения между участниками, способность к творчеству и пр. [8]. Психологически развитой как коллектив считается такая малая группа, в которой сложилась дифференцированная система различных деловых и личных взаимоотношений [9]. Именно поэтому очень важен вопрос регулирования конфликтного взаимодействия в малых коллективах.

Понятие конфликта сегодня не принадлежит какой-то одной определенной области науки или практики; выделяются не менее 11 областей научного знания, так или иначе изучающих конфликты [10]. Анализ и сравнение разных определений конфликта [6] позволили выделить в качестве инвариантных такие его характеристики: биполярность; активность, направленная на преодоление противоречий; субъектность (наличие субъекта или субъектов как носителей конфликта). Таким образом, конфликт — биполярное явление (противостояние двух начал), проявляющееся в активности сторон, направленной на преодоление противоречий, причем стороны представлены активными субъектами. В случае конфликта в малых коллективах экспертов субъекты — его участники.

Управление конфликтом — целенаправленное, обусловленное объективными законами воздействие на его динамику в интересах развития или разрушения

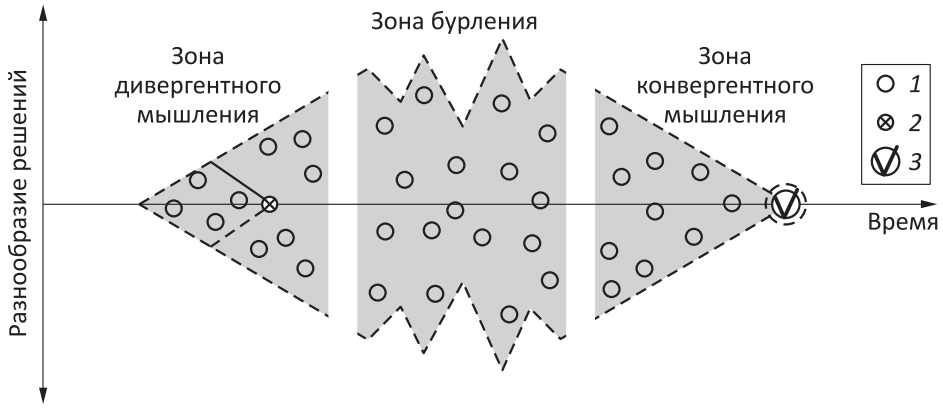
социальной системы, к которой относится конфликт [11]. При управлении конфликтами в малых коллективах экспертов актуален его продуктивный аспект, направленный на развитие коллектива и состоящий в предотвращении деструктивных конфликтов, стимулировании и разрешении конструктивных.

Конструктивный конфликт [5] обеспечивает развитие отношений между субъектами; высвобождает накапливающееся внутреннее напряжение, тем самым обеспечивая сохранение связей; актуализирует разные позиции и мнения по поводу возникающих в коллективе проблем и тем самым способствует поиску оптимальных способов их решения. Такой конфликт проявляется преимущественно, когда цели и потребности конфликтующих экспертов не противоположны целям, ценностям и нормам коллектива, а социально-психологическая структура коллектива относительно гибкая, т. е. для коллектива характерна относительная открытость подгрупп и динамичные связи между ними и не включенными в них членами. Конструктивные конфликты — преимущественно инструментальные конфликты, возникающие по поводу проблемы или процесса ее решения. Открытое обсуждение и споры по поводу проблемы увеличивают результативность деятельности группы, потому что эксперты предлагают и оценивают различные решения, тем самым достигая оптимальных решений и результатов. Процессно-ориентированные конфликты связаны с технологией и способами решения поставленной проблемы, распределением ролей и ответственности между участниками коллектива.

Для деструктивного конфликта характерно разрушение психологической целостности коллектива, снижение его социальной, экономической, социально-психологической эффективности, блокирование развития отношений между субъектами [5]. Подобный конфликт возникает, если цели и потребности конфликтующих субъектов радикально расходятся с целями, ценностями и нормами коллектива, при этом коллектив обладает жесткой структурой, т. е. для него характерна закрытость подгрупп по отношению к коллективу и явное доминирование дезинтегративных связей между ними. Строго говоря, коллектив экспертов, в котором преобладают деструктивные конфликты, представляет собой, скорее, малую группу.

Чаще всего деструктивные конфликты — конфликты по поводу отношений, т. е. разногласия между экспертами по личным вопросам и проблемам, не относящимся к выполняемой работе, которые обусловлены несовместимостью и враждебностью между ними. При компьютерном моделировании коллективного решения проблем малыми коллективами экспертов основная цель — повышение качества предлагаемых автоматизированными системами решений, поэтому конфликты по поводу отношений в них не моделируются и в дальнейшем не рассматриваются.

Мера выраженности инструментальных конфликтов зависит от стадии решения проблемы в коллективе [5]: формирование, дивергентное мышление, бурление, конвергентное мышление, принятие решения и расформирование [12–14] (см. рисунок).



Ромб группового принятия решений С. Кейнера, К. Толди, С. Фиск, Д. Бергера: *1* — альтернатива; *2* — досрочное несогласованное решение; *3* — согласованное решение

На первой стадии члены коллектива знакомятся, обмениваются официальной информацией друг о друге, вносят предложения о работе коллектива, придерживаются общепринятых точек зрения, высказывают очевидные решения [13]. Если задача имеет очевидное решение, дискуссия завершается, иначе возникают процессы дивергентного, расходящегося мышления, в рамках которых поощряется безоценочная дискуссия и генерация большого числа решений [14]. На этой стадии возникает умеренный процессно-ориентированный конфликт, так как решающее значение имеют аргументы тех, кто имеет более высокую квалификацию для решения проблемы и ее частей [5]. На данной стадии фасилитатор, т. е. участник коллектива, отвечающий за эффективную организацию процесса решения проблемы, обеспечивает разрешение процессно-ориентированных конфликтов и применяет методы дивергентного коллективного мышления для стимулирования проблемно-ориентированных конфликтов.

Если коллективу удалось выйти за границы устоявшихся мнений, процесс обсуждения переходит в стадию бурления, когда между членами коллектива возможны конфликты из-за противоречивых решений, т. е. проблемно-ориентированные конфликты. При этом наблюдается низкий уровень процессно-ориентированных конфликтов, так как эксперты все свое внимание фокусируют на проблеме и поиске ее решений. На данной стадии цель фасилитатора — сгладить разногласия между экспертами, сблизить их точки зрения на решаемую проблему, разрешить возникшие конфликты, не допустив перерождения конструктивных конфликтов в деструктивные.

На стадии конвергентного мышления эксперты переформулируют ценные мысли в конкретные предложения и «шлифуют» их, пока все участники дискуссии не придут к конечному решению, воплощающему все разнообразие точек зрения. Эта стадия характеризуется «сходящимся мышлением»: классифика-

цией идей, их обобщением, вынесением оценок. При этом может происходить небольшое усиление процессно-ориентированных конфликтов, так как эксперты активно обсуждают вклад каждого в процесс решения проблемы и полученный результат [5].

В ходе принятия решений и расформирования формулируется коллективное решение, учитывающее мнения всех участников обсуждения.

Для моделирования конфликтологического аспекта решения проблем коллективом экспертов «за круглым столом» предлагается модифицировать модель ГиИМАС [3], расширив ее элементами для управления конфликтами агентов.

3 Управление конфликтами агентов в гибридной интеллектуальной многоагентной системе

Формально ГиИМАС для моделирования проблемно- и процессно-ориентированных конфликтов определяется следующим образом:

$$\text{himas} = \langle \text{AG}^*, \text{env}, \text{INT}, \text{ORG}, \{\text{cnfm}\}\rangle. \tag{1}$$

Здесь $\text{AG}^* = \{\text{ag}_1, \ldots, \text{ag}_n, \text{ag}^{\text{dm}}, \text{ag}^{\text{fc}}\}$ — множество агентов, включающее агентов-экспертов (АЭ) ag_i, $i \in \mathbb{N}$, $1 \le i \le n$, агента, принимающего решения (АПР), — ag^{dm} и агента-фасилитатора (АФ), управляющего взаимодействиями агентов при решении проблемы с учетом возможных конфликтов между ними, — ag^{fc}; n — число АЭ; env — концептуальная модель внешней среды ГиИМАС; INT^* — элементы структурирования взаимодействий агентов, описываемые выражением:

$$\text{INT} = \{\text{prot}_{\text{cnfm}}, \text{lang}, \text{ont}\}\ ,$$

где $\text{prot}_{\text{cnfm}}$ — протокол взаимодействия агентов, позволяющий организовать управление конфликтами агентов, lang — язык передачи сообщений, ont — модель предметной области; ORG — множество архитектур ГиИМАС; $\{\text{cnfm}\}$ — множество концептуальных моделей макроуровневых процессов в ГиИМАС; cnfm — модель процесса управления конфликтами при коллективном решении проблем:

$$\text{cnfm} = \langle \mathbf{CNF}, \text{cnfcl}, \text{cmkb}, \text{act}_{\text{cnfm}}, \text{ACT}_{\text{agcr}}\rangle\ . \tag{2}$$

В (2) **CNF** — матрица, описывающая конфликты между каждой парой агентов кортежем, представленным выражением:

$$\text{cnf}_{i\,j\,\text{cnft}} = \langle \text{ag}_i, \text{ag}_j, \text{cnfin}, \text{cnft}, \text{ACT}_{\text{agcr}\,i}, \text{ACT}_{\text{agcr}\,j}\rangle\ ,$$

где ag_i и ag_j — агенты-субъекты конфликта, $i, j \in \mathbb{N}$, $1 \le i, j \le n$, $i \ne j$, cnfin — напряженность конфликта в виде скалярной величины $\text{cnfin} \in [0, 1]$, вычисляемая классификатором конфликтов cnfcl в соответствии с реализуемой им мерой напряженности конфликта, в зависимости от того, предлагают АЭ

частные решения проблемы или альтернативные решения проблемы в целом, могут использоваться меры на основе совместимости частных решений [15] или на основе ранжирования альтернатив [16, 17], cnft — символьная переменная «тип конфликта», определенная на множестве $\mathrm{CNFT} = \{$«проблемно-ориентированный», «процессно-ориентированный»$\}$, значение которой устанавливается классификатором конфликтов cnfcl в зависимости от предмета конфликта: если он возник из-за противоречий по поводу альтернативных решений проблемы или ее частей, то cnft = «проблемно-ориентированный», если же конфликт обусловлен распределением ролей и ресурсов, то cnft = «процессно-ориентированный», $\mathrm{ACT}_{\mathrm{agcr}\,i}$ $\mathrm{ACT}_{\mathrm{agcr}\,j}$ — множество допустимых действий агентов ag_i и ag_j соответственно по разрешению противоречий, $\mathrm{ACT}_{\mathrm{agcr}\,i}, \mathrm{ACT}_{\mathrm{agcr}\,j} \subseteq \mathrm{ACT}_{\mathrm{agcr}}$; cnfcl — классификатор конфликтов агентов, идентифицирующий их характер и оценивающий напряженность, т. е. формирующий для каждой пары агентов значение элемента матрицы $\mathbf{CNF}$; cmkb — база знаний об эффективности методов управления конфликтами в зависимости от характеристик проблемы и конфликтов между агентами, которая разрабатывается по результатам тестирования алгоритмов, реализующих эти методы; $\mathrm{act}_{\mathrm{cnfm}}$ — функция «управление конфликтом» АФ, обеспечивающая идентификацию конфликтов с помощью классификатора cnfcl и инициализацию методов гетерогенного мышления [14, 15] в соответствии с базой знаний cmkb и протоколом $\mathrm{prot}_{\mathrm{cnfm}}$; $\mathrm{ACT}_{\mathrm{agcr}}$ — множество допустимых действий агентов по разрешению противоречий.

Таким образом, функция ГиИМАС для моделирования проблемно- и процессно-ориентированных конфликтов в целом описывается выражением:

$$\mathrm{act}_{\mathrm{himas}} = \left(\bigcup_{\mathrm{ag} \in \mathrm{AG}^*} \mathrm{act}_{\mathrm{ag}} \right) \cup \mathrm{ACT}_{\mathrm{agcr}} \cup \mathrm{act}_{\mathrm{cnfm}} \cup \mathrm{act}_{\mathrm{col}} \,. \tag{3}$$

Здесь $\mathrm{act}_{\mathrm{ag}}$ — функция АЭ из множества AG^*, описываемая формулой:

$$\mathrm{act}_{\mathrm{ag}} = (\mathrm{met}_{\mathrm{ag}}, \mathrm{it}_{\mathrm{ag}}) \,, \;\; \mathrm{ag} \in \mathrm{AG}^* \,, \;\; \left| \bigcup_{\mathrm{ag} \in \mathrm{AG}^*} \mathrm{it}_{\mathrm{ag}} \right| \geq 2 \,,$$

где $\mathrm{met}_{\mathrm{ag}}$ — метод решения проблемы, $\mathrm{it}_{\mathrm{ag}}$ — интеллектуальная технология, в рамках которой реализован метод $\mathrm{met}_{\mathrm{ag}}$; $\mathrm{act}_{\mathrm{col}}$ — коллективная функция ГиИМАС, конструируемая динамически в соответствии с протоколом $\mathrm{prot}_{\mathrm{cnfm}}$.

В отличие от функции ГиИМАС, представленной в [3], выражение (3) дополнено множеством $\mathrm{ACT}_{\mathrm{agcr}}$, обеспечивающим моделирование конфликта, т. е. активностей агентов, направленных на преодоление противоречий, и элементом $\mathrm{act}_{\mathrm{cnfm}}$, отвечающим за управление конфликтами в ГиИМАС в соответствии с моделью ромба группового принятия решений на рисунке. В результате ГиИМАС,

моделирующая конфликты агентов, способна регулировать интенсивность взаимодействия агентов и активировать релевантные ситуации методы коллективного гетерогенного мышления в зависимости от напряженности и типа конфликтов, т. е. конфликты выполняют сигнальную функцию. Кроме того, моделирование конфликтов обеспечивает развитие ГиИМАС и ее самоорганизацию в сильном смысле [18], т. е. возникающую за счет распределенного взаимодействия агентов без явного централизованного управления этим процессом одним из них. Таким образом, предложенные модели конфликтов и процесса управления ими повысят релевантность ГиИМАС малым коллективам экспертов, решающих проблемы «за круглым столом».

4 Заключение

Рассмотрены вопросы управления конфликтами в малых коллективах экспертов, решающих биопроизводственные, социально-экономические и технические проблемы «за круглым столом». Показано, что конфликты в таких коллективах могут играть как деструктивную, так и конструктивную роль, поэтому требуется не избегать конфликтов, а правильно управлять ими. В этом случае конфликты выступают одним из факторов развития коллектива и повышения его эффективности. Предложены модели проблемно- и процессно-ориентированных конфликтов, применение которых в ГиИМАС позволит смоделировать процесс коллективного решения проблем на основе анализа напряженности конфликта, применять эффективные методы организации коллективных рассуждений, такие как гетерогенное мышление, чтобы вырабатывать решения, сопоставимые с решениями реальных коллективов экспертов.

Литература

1. *Колесников А. В.* Гибридные интеллектуальные системы. Теория и технология разработки. — СПб.: СПбГТУ, 2001. 711 с.
2. *Кириков И. А., Колесников А. В., Листопад С. В., Румовская С. Б.* «Виртуальный консилиум» — инструментальная среда поддержки принятия сложных диагностических решений // Информатика и её применения, 2016. Т. 10. Вып. 3. С. 81–90.
3. *Колесников А. В., Кириков И. А., Листопад С. В.* Гибридные интеллектуальные системы с самоорганизацией: координация, согласованность, спор. — М.: ИПИ РАН, 2014. 189 с.
4. *Колесников А. В., Листопад С. В.* Модель гибридной интеллектуальной многоагентной системы гетерогенного мышления для информационной подготовки оперативных решений в региональных электрических сетях // Системы и средства информатики, 2018. Т. 28. № 4. С. 31–41.
5. *Сидоренков А. В.* Конфликт в малой группе: понятие, функции, виды и модель // Северо-Кавказский психологический вестник, 2008. Т. 6. № 4. С. 22–28.
6. *Гришина Н. В.* Психология конфликта. — 2-е изд. — СПб.: Питер, 2008. 544 с.

7. *Тарасов В. Б.* От многоагентных систем к интеллектуальным организациям: философия, психология, информатика. — М.: Эдиториал УРСС, 2002. 352 с.
8. Понятия малой группы и коллектива. http://azps.ru/articles/soc/soc96.html.
9. *Андреева Г. М.* Социальная психология. — М.: Аспект Пресс, 1996. 342 с.
10. *Анцупов А. Я., Шипилов А. И.* Конфликтология: теория, история, библиография. — М.: Дом Советов, 1996. 143 с.
11. *Емельянов С. М.* Практикум по конфликтологии. — 3-е изд. — СПб.: Питер, 2009. 384 с.
12. *Занковский А. Н.* Организационная психология. — 2-е изд. — М.: Флинта: МПСИ, 2002. 648 с.
13. Организационное поведение / Под ред. Г. Р. Латфуллина, О. Н. Громовой. — СПб: Питер, 2004. 432 с.
14. *Kaner S., Lind L., Toldi C., Fisk S., Beger D.* The facilitator's guide to participatory decision-making. — San Francisco, CA, USA: Jossey-Bass, 2011. 368 p.
15. *Kolesnikov A. V., Listopad S. V.* Hybrid intelligent multiagent system of heterogeneous thinking for solving the problem of restoring the distribution power grid after failures // Open Semantic Technologies for Intelligent Systems: Research Papers Collection. — Minsk: BGUIR, 2019. P. 133–138.
16. *Bana e Costa C.* The use of multi-criteria decision analysis to support the search for less conflicting policy options in a multi-actor context: Case study // J. Multi-Criteria Decision Analysis, 2001. Vol. 10. Iss. 2. P. 111–125.
17. *Fasth T., Larsson A., Ekenberg L., Danielson M.* Measuring conflicts using cardinal ranking: An application to decision analytic conflict evaluations // Advances Oper. Res., 2018. Vol. 2018. Art. ID 8290434. 14 p.
18. *Serugendo G. D. M., Gleizes M.-P., Karageorgos A.* Self-organization in multiagent systems // Knowl. Eng. Rev., 2005. Vol. 20. Iss. 2. P. 165–189.

Поступила в редакцию 31.07.19

MODELING OF AGENT CONFLICTS IN HYBRID INTELLIGENT MULTIAGENT SYSTEMS

S. V. Listopad and I. A. Kirikov

Kaliningrad Branch of the Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, 5 Gostinaya Str, Kaliningrad 236022, Russian Federation

Abstract: Conflict management is an integral part of the problem solving process by expert team at the round table, encouraging of conflicts that positively influence the course of solving the problem and preventing or resolving all others. The existing models of hybrid intelligent multiagent systems have a significant drawback, which is that they do not model agent conflicts and a single agent based on recommendations of other agents makes final decisions. Modeling

of conflicts in hybrid intelligent multiagent systems will make it possible to manage the "discussion" process, activating various types of collective thinking depending on the nature and intensity of the conflict, which will ensure their relevance to small teams of experts who successfully solve problems, which are underdetermined, characterized by a high combinatorial complexity, heterogeneity, and other NON-factors. For this purpose, the paper proposes a model of problem and process oriented conflict in hybrid intelligent multiagent systems.

Keywords: conflict; hybrid intelligent multiagent system; expert team; round table

DOI: 10.14357/08696527190312

References

1. Kolesnikov, A. V. 2001. *Gibridnye intellektual'nye sistemy. Teoriya i tekhnologiya razrabotki* [Hybrid intelligent systems: Theory and technology of development]. St. Petersburg: SPbGTU Publs. 711 p.
2. Kirikov, I. A., A. V. Kolesnikov, S. V. Listopad, and S. B. Rumovskaya. 2016. "Virtual'nyy konsilium" — instrumental'naya sreda podderzhki prinyatiya slozhnykh diagnosticheskikh resheniy ["Virtual council" — source environment supporting complex diagnostic decision making]. *Informatika i ee Primeneniya — Inform. Appl.* 10(3):81–90.
3. Kolesnikov, A. V., I. A. Kirikov, and S. V. Listopad. 2014. *Gibridnye intellektual'nye sistemy s samoorganizatsiey: koordinatsiya, soglasovannost', spor* [Hybrid intelligent systems with self-organization: Coordination, consistency, dispute]. Moscow: IPI RAN. 189 p.
4. Kolesnikov, A. V., and S. V. Listopad. 2018. Model' gibridnoy intellektual'noy mnogoagentnoy sistemy geterogennogo myshleniya dlya informatsionnoy podgotovki operativnykh resheniy v regional'nykh elektricheskikh setyakh [Model of a hybrid intelligent multiagent system of heterogeneous thinking for preparation of information about operational decisions in a regional power system]. *Sistemy i Sredstva Informatiki — Systems and Means of Informatics* 28(4):31–41.
5. Sidorenkov, A. V. 2008. Konflikt v maloy gruppe: ponyatie, funktsii, vidy i model' [Conflict in a small group: Concept, functions, forms and model]. *Severo-Kavkazskiy psikhologicheskiy vestnik* [North-Caucasian Psychological J.] 6(4):22–28.
6. Grishina, N. V. 2008. *Psikhologiya konflikta* [Psychology of conflict]. 2nd ed. St. Petersburg: Piter. 544 p.
7. Tarasov, V. B. 2002. *Ot mnogoagentnykh sistem k intellektualnym organizatsiyam: filosofiya, psikhologiya, informatika* [From multiagent systems to intelligent organizations: Philosophy, psychology, and informatics]. Moscow: Editorial URSS. 352 p.
8. Ponyatiya maloy gruppy i kollektiva [The concepts of a small group and a team]. Available at: http://azps.ru/articles/soc/soc96.html (accessed July 25, 2019).
9. Andreeva, G. M. 1996. *Sotsial'naya psikhologiya* [Social psychology]. Moscow: Aspekt Press. 342 p.
10. Antsupov, A. Ya., and A. I. Shipilov. 1996. *Konfliktologiya: teoriya, istoriya, bibliografiya* [Conflictology: Theory, history, bibliography]. Moscow: Dom Sovetov. 143 p.

11. Emel'yanov, S. M. 2009. *Praktikum po konfliktologii* [Practicum on conflictology]. 3rd ed. St. Petersburg: Piter. 384 p.
12. Zankovskiy, A. N. 2002. *Organizatsionnaya psikhologiya* [Organizational psychology]. Moscow: Flinta: MPSI. 648 p.
13. Latfullin, G. R., and O. N. Gromova, eds. 2004. *Organizatsionnoe povedenie* [Organizational behavior]. St. Petersburg: Piter. 432 p.
14. Kaner, S., L. Lind, C. Toldi, S. Fisk, and D. Beger. 2011. *The facilitator's guide to participatory decision-making.* San Francisco, CA: Jossey-Bass. 368 p.
15. Kolesnikov, A. V., and S. V. Listopad. 2019. Hybrid intelligent multiagent system of heterogeneous thinking for solving the problem of restoring the distribution power grid after failures. *Open Semantic Technologies for Intelligent Systems: Research Papers Collection*. Minsk. 133–138.
16. Bana e Costa, C. 2001. The use of multi-criteria decision analysis to support the search for less conflicting policy options in a multi-actor context: Case study. *J. Multi-Criteria Decision Analysis* 10(2):111–125.
17. Fasth, T., A. Larsson, L. Ekenberg, and M. Danielson. 2018. Measuring vonflicts using cardinal ranking: An application to decision analytic conflict evaluations. *Advances Oper. Res.* Article ID 8290434. 14 p.
18. Serugendo, G. D. M., M.-P. Gleizes, and A. Karageorgos. 2005. Self-organization in multiagent systems. *Knowl. Eng. Rev.* 20(2):165–189.

Received July 31, 2019

Contributors

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 236022, Russian Federation; ser-list-post@yandex.ru

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 236022, Russian Federation; baltbipiran@mail.ru

СИСТЕМЫ И СРЕДСТВА ИНФОРМАТИКИ, 2019. Т. 29. № 3. С. 149–159

АЛГОРИТМ НЕЧЕТКОГО СРАВНЕНИЯ ПРИ ОБРАБОТКЕ ПЕРСОНАЛЬНЫХ ДАННЫХ

***О. В. Бобылева*[1], *И. С. Бекешева*[2], *В. А. Бобылев*[3], *В. В. Чаркова*[4]**

Аннотация: Обоснована необходимость разработки нового метода нечеткого сравнения, ориентированного на сравнение слов в базах данных, содержащих персональные данные. На конкретных примерах из области медицинского страхования указаны плюсы применения данного алгоритма. Приведены результаты работы разработанного и реализованного алгоритма.

Ключевые слова: алгоритм; нечеткий поиск; нечеткое сравнение; матрица

DOI: 10.14357/08696527190313

1 Введение

При проектировании и разработке информационной системы, содержащей большой объем персональных данных, необходимо заранее предусматривать, какой вид поиска в ней будет использоваться. Если объем персональных данных относительно небольшой или представляет собой своеобразный справочник, который обновляется крайне редко и исключает дублирование данных, то можно использовать простые алгоритмы поиска по точному совпадению или по совпадению подстроки. Однако такой подход не всегда применим для информационных систем персональных данных (ИСПД), сведения в которых обновляются периодически и могут поступать из разных источников (подобная ситуация характерна для большинства информационных систем государственных служб и учреждений). Число обновляемых и дополняемых записей в таких информационных системах может составлять до нескольких миллионов за один год. В этом случае возможны ошибки ввода персональных данных, для определения которых простые алгоритмы поиска не подходят, потому что заранее не известно, в каком месте оператор мог допустить ошибку. Поэтому при необходимости сопоставить данные, исключив ошибки, в таких ИСПД используют алгоритмы нечеткого поиска. Данные алгоритмы могут применяться, например, в сфере здравоохранения для сбора статистических данных о некотором заболевании, когда важно знать не только, сколько всего раз обращались за медицинской помощью, но

[1]Хакасский государственный университет им. Н. Ф. Катанова, nimdar@bk.ru

[2]Хакасский государственный университет им. Н. Ф. Катанова, irrisskay@mail.ru

[3]Территориальный фонд обязательного медицинского страхования Республики Хакасия, vadimbobylev@yandex.ru

[4]Хакасский государственный университет им. Н. Ф. Катанова, charkova_97@mail.ru

и сколько раз обращался за помощью один и тот же человек с данным диагнозом. Точность таких расчетов вместе с расчетом суммы средств, потраченных гражданином или государством при оказании услуги, имеет большое значение при планировании бюджетов различных уровней [1].

2 Особенности нечеткого поиска в информационных системах персональных данных

На данный момент существующие алгоритмы нечеткого поиска сравнивают между собой строки, однако персональные данные можно отнести к особой категории строк, сравнение которых существующими алгоритмами дает верный результат далеко не всегда. Основными персональными данными являются фамилия, имя, отчество — основная триада персональных данных. Но в этой триаде один из элементов может отсутствовать или, наоборот, может появиться четвертый составной элемент имени, который меняет свое местоположение в полном имени. В этих случаях алгоритм нечеткого поиска начинает давать неверный результат. Для России с большим разнообразием национального состава такие случаи далеко не редкость, и число ошибок становится больше, если вводятся персональные данные, имеющие свои особенности, в регионе, где таких особенностей нет.

В любой ИСПД первичный ввод данных осуществляется оператором, и чем больше объем данных, тем выше вероятность появления как ошибочных записей, так и дублирующих уже существующие. При проектировании информационных систем одним из основных критериев их функционирования служит способ хранения данных. Например, для информационных систем записи актов гражданского состояния (ЗАГС), Федеральной миграционной службы (ФМС) и т. д. персональные данные должны исключать дублирование (у гражданина может быть только один паспорт или свидетельство о рождении). Другие информационные системы предполагают накопительное хранение персональных данных. Например, информационная система Министерства внутренних дел (МВД) хранит все факты обращения гражданина. Существуют также и комбинированные информационные системы. Например, информационные системы обязательного медицинского страхования (ОМС), хранящие как уникальные записи (регистр застрахованных), так и накопительные (данные об оказанной медицинской помощи). Независимо от способа хранения персональных данных (уникальные, накопительные, комбинированные), разрабатываемый алгоритм нечеткого сравнения ориентирован именно на персональные данные и может применяться как для задач дедупликации (уникальные), так и для задач ассоциации разных записей с одним субъектом персональных данных (накопительные).

Приведем пример из медицинской информационной системы. В одном из дальневосточных регионов живет уроженец Северного Кавказа. Следовательно, в информационной системе медицинской организации по месту его постоянного

жительства есть данные об этом человеке. Но для данного региона не характерны национальные имена, используемые на Кавказе, поэтому велика вероятность ошибок при вводе данных о посещении врача: кроме того, что оператор может допустить простую ошибку в виде опечатки, он может неверно понимать четвертую составляющую имени, например «оглы», и эта четвертая составляющая может как отсутствовать, так и идти после имени или отчества. Перед информационной системой встает непростая задача ассоциировать новую запись с существующими данными, и алгоритм нечеткого поиска здесь чаще всего дает неверный результат. Еще более сложная задача будет стоять перед информационной системой фонда ОМС, ведь после обращения в поликлинику человека могли направить на другой уровень оказания медицинской помощи, на консультативный прием в другую медицинскую организацию, в которой сведений о пациенте нет и оператор введет данные не в той форме, что оператор поликлиники. После консультации данного человека могли направить на лечение в другой регион, где оператор также мог ввести персональные данные с ошибкой. В результате ИСПД фонда ОМС столкнется с тремя различными формами написания персональных данных, и ее задачей будет определить, что все три записи принадлежат одному и тому же гражданину. Сфера ОМС получит точные сведения, необходимые для оплаты медицинской помощи, а система здравоохранения региона получит точные сведения о маршруте движения пациента и результатах лечения на разных уровнях оказания медицинской помощи.

В медицинских информационных системах накопление информации происходит с различной периодичностью (ежедневно небольшими объемами). В информационных системах фондов ОМС и информационных системах страховых медицинских организаций эти сведения обновляются преимущественно один раз в месяц передачей сведений об оказанной медицинской помощи полным объемом за месяц. В обоих случаях в ИСПД и медицинских организаций, и фондов ОМС должны использоваться алгоритмы нечеткого поиска данных, предназначенные для исключения ошибок при вводе данных в ИСПД. Оператор в медицинской организации при вводе сведений о случае оказания медицинской помощи может допустить ошибку при вводе персональных данных пациента, задача информационной системы определить, вводятся сведения о существующем в ИСПД пациенте, данные нового пациента или данные о существующем пациенте введены с ошибкой. В случае обнаружения ошибки ИСПД должна уведомить оператора и внести коррективы во введенные данные. В ИСПД фондов ОМС и страховых медицинских организаций алгоритмы нечеткого поиска должны сопоставить существующие данные с вновь вводимыми и в случае обнаружения соответствий при полном или частичном совпадении (когда не превышен порог ошибок) загрузить новые данные, установив их связь с уже имеющимися для определенного субъекта персональными данными. Таким образом, во всех ИСПД на каждого субъекта персональных данных должна быть выстроена цепочка ассоциаций, позволяющая проследить все случаи его обращения за медицинской помощью, даже если когда-то персональные данные были введены с ошибкой. Установление

таких ассоциаций имеет большое значение для принятия решений в сфере охраны здоровья граждан и финансировании программы государственных гарантий оказания медицинской помощи. Ассоциации помогают ответить на следующие вопросы:

- соотношение числа обращений за медицинской помощью и численности пациентов;
- средняя стоимость оказания медицинской помощи одному пациенту;
- средняя стоимость оказания медицинской помощи одному пациенту по медицинской организации в целом и по отделениям;
- обнаружение случаев повторных обращений за медицинской помощью в определенные периоды.

Ответы на эти вопросы помогают решить, например, такую задачу. В ИСПД есть следующие сведения о случаях укуса клеща: за первый год — 5000 случаев; за второй год — 4500 случаев; за третий год — 5500 случаев. Но это только случаи обращения за медицинской помощью, а не количество пострадавших человек. При укусе клеща пациент будет приходить на прием к врачу несколько раз: если результаты анализов будут хорошими, все закончится двумя посещениями, а если будет обнаружен клещевой энцефалит, предстоит длительное лечение. Перед органом управления здравоохранением стоит задача: на основе сведений о выявленных случаях укуса клеща определить потребность в количестве вакцины и иммуноглобулина, необходимых для лечения больных. Зная число пациентов и количество случаев выявления клещевого энцефалита у этих пациентов, можно сделать следующие выводы: насколько часто укус клеща приводит к клещевому энцефалиту, сколько пациентов перенесло заболевание в легкой (тяжелой) форме, сколько из заболевших было вакцинировано и как протекало течение болезни у данной группы пациентов. Таким образом, если есть точные данные о каждом субъекте и течении его болезни, можно максимально точно определить потребность данного региона в вакцине и иммуноглобулине.

Примеры подобных применений ассоциаций между персональными данными в ИСПД можно найти не только в государственных информационных системах, но и в информационных системах бизнеса.

Для определения соответствия введенной информации определенному субъекту персональных данных в ИСПД используются различные сведения. Однако в Российской Федерации первичными всегда остаются фамилия, имя, отчество (три составляющих полного имени). Независимо от того, какой из алгоритмов нечеткого поиска был выбран для определения соответствия между существующими и вводимыми данными, будет получен ответ, что данные полностью соответствуют друг другу или показана степень различия между двумя наборами данных. На основании этого ответа и принимается решение об идентичности данных и возможности установления между ними связи.

3 Математическая модель алгоритма нечеткого сравнения в информационных системах персональных данных

При необходимости сопоставить персональные данные, исключив ошибки, используют алгоритмы нечеткого поиска. Любой из существующих алгоритмов в той или иной форме указывает на степень похожести сравниваемых строк. При этом некоторые из них, например метод n-грамм, просты для реализации в любой информационной системе, другие, такие как алгоритм «Расстояние Левенштейна», дают более быстрый результат.

Под нечетким поиском понимается поиск по ключевым словам с учетом возможных произвольных ошибок в написании ключевого слова или, напротив, ошибок написания слова в целевом запросе. Ключевым элементом организации нечеткого поиска выступает выбор меры сходства слов или обратной функции — функции расстояния между словами, называемой метрикой. В качестве метрик используют расстояния Хэмминга, Левенштейна, Дамерау–Левенштейна.

Большинство современных алгоритмов текстового поиска основываются на модификациях одного из следующих методов.

Алгоритм Вагнера–Фишера позволяет для двух строк найти расстояние. Оптимизировать нечеткий поиск позволяет наиболее удобная организация данных — использование метрических деревьев. Самый быстрый поиск обеспечивают деревья Бернхарда Келлера (БК-деревья), однако они усложняют сам алгоритм поиска. Алгоритм расширения выборки базируется на сведении задачи о нечетком поиске к задаче о точном поиске. Метод n-грамм основывается на следующем свойстве: если слово u получается из слова w в результате не более чем k операций редактирования (кроме транспозиции символов), то при любом представлении в виде конкатенации из $(k+1)$-й строки одна из строк такого представления будет точной подстрокой w. Таким образом, задача поиска сводится к задаче выборки слов, содержащих заданную подстроку.

Другим вариантом реализации нечеткого поиска является метод хеширования по сигнатуре. Поиск с использованием хеширования состоит в подборе отображения (хеш-функции) слова, например в множество чисел или строк, сохраняющего основные характеристики исходного слова и устойчивого к наиболее распространенным ошибкам.

Широко используемый в настоящее время алгоритм Soundex, разработанный Р. Расселом и М. К. Оделл, использует сравнение двух строк по их звучанию с помощью специально введенных индексов. Недостаток алгоритма заключается в привязке к языку, на котором написаны анализируемые строки.

При проверке правильности написания слов используется алгоритм расширения выборки. Алгоритм сводит задачу о нечетком поиске к задаче о точном поиске.

Кроме того, для создания алгоритмов нечеткого поиска широко применяются линейные коды. Однако их эффективность при сравнении слов с частыми, но небольшими ошибками достаточно низкая.

В настоящее время, хотя и разработано немало методов и алгоритмов нечеткого поиска, результаты сравнения, выдаваемые алгоритмами нечеткого поиска применительно к персональным данным, можно назвать лишь условно точными. Кроме того, ключевыми недостатками этих алгоритмов остаются либо время выполнения поиска, либо значительные затраты оперативной памяти.

Для получения максимально точных результатов необходима разработка нового алгоритма нечеткого поиска, применяемого для сравнения персональных данных с учетом отсутствия или появления дополнительных элементов. Кроме того, не стоит забывать и простые, но не всегда используемые процедуры очистки персональных данных, удаления лишних символов, возможной замены одних букв другими и т. д. Максимально точный результат сравнения существующих и вводимых персональных данных может быть получен в результате использования алгоритма нечеткого поиска, ориентированного не на простое сравнение строк, абстрагированное от контекста, а на нечеткое сравнение персональных данных.

Ошибки, которые возникают при использовании существующих алгоритмов нечеткого поиска в базах данных, не позволяют использовать эти алгоритмы в полной мере. Алгоритмы нечеткого поиска в качестве результата чаще всего выдают определенное числовое значение: это или коэффициент подобия двух строк, или число символов, на которое отличаются две строки. Персональные данные при сравнении следует рассматривать как единое целое, а не по отдельным элементам, так как триада персональных данных может быть заполнена не в том порядке или может быть пропущена одна из составляющих этой триады. При этом цена ошибки в ИСПД может быть очень значима. Например, сравниваются три набора персональных данных с использованием одного из существующих алгоритмов нечеткого поиска или их комбинации, при этом все персональные данные принадлежат одному и тому же субъекту персональных данных. Первая и вторая пары идентичны, но при вводе третьей записи в отчество была добавлена четвертая именная составляющая «оглы», в результате сравнения будет выдан ответ: первая и третья, вторая и третья различаются на четыре знака, что находится выше «максимального порога» различия двух наборов персональных данных.

Математическая модель данного алгоритма разработана на основе матричного исчисления, точнее на основе клеточного строения матриц.

Представим теоретические основы, разработанные для сопоставления основной триады персональных данных.

Пусть в базе данных имеется строка-образец S длины N, состоящая из Q строк-компонентов $S_1, S_2, \ldots, S_x, \ldots, S_Q$ (где $x \in \{1, \ldots, Q\}$, $Q \geq 1$), длины которых равны $N_1, N_2, \ldots, N_x, \ldots, N_Q$ соответственно.

Образуем новую «ошибочную» строку C длины M, состоящую из P строк-компонентов. С помощью возможных операций редактирования из строк-компонентов $S_1, S_2, \ldots, S_x, \ldots, S_Q$ образуем «ошибочные» строки-компоненты $C_1, C_2, \ldots, C_x, \ldots, C_Q$ длиной $M_1, M_2, \ldots, M_x, \ldots, M_P$ соответственно так,

чтобы общее количество несовпадений этих строк не превышало двух элементов (иначе персональные данные считаются принадлежащими разным людям).

По условию исследования предполагается, что строка C может быть образована в результате удаления одного из компонентов строки S или вставки одного произвольного компонента, а также перестановки «ошибочных» строк-компонентов произвольным образом.

Таким образом, получается строка C длины M, состоящая из P строк-компонентов $C_1, C_2, \ldots, C_y, \ldots, C_P$ (где $|Q - P| \in \{0, 1\}$).

Требуется: установить сходство строк S и C.

Строки S и C считаем *совпадающими* только в том случае, если общее число несовпадений соответствующих строк-компонентов не превышает двух элементов и число строк-компонентов в строках отличается на ноль или на один компонент.

Для сравнения строк была разработана следующая математическая модель алгоритма.

Первый шаг — предварительная обработка строк: пробелы заменяются символом тире («—»).

Полагаем, что изначально все компоненты ФИО разделены между собой одним пробелом или символом тире.

Второй шаг — построение блочной матрицы $\mathbf{B}$ размером $\min(P, Q) \times \times \max(P, Q))$.

Блок матрицы $\mathbf{B}$ будем обозначать прописной буквой с двумя индексами $\mathbf{B}_{u,z}$. Первый индекс $u \in \{1, \ldots, \min(P, Q)\}$ указывает номер строки, а второй индекс $z \in \{1, \ldots, \max(P, Q)\}$ — номер столбца, в котором располагается соответствующий блок.

Построение блочной матрицы осуществляется следующим образом.

Символы меньшей строки записываем в столбец, большей — в строку. Для разбиения матрицы $\mathbf{B}$ на блоки используются одноразмерные матрицы (матрицы-столбцы и матрицы-строки), элементами которых является символ «∞». Такое разбиение производится в позициях строк, где $S[i] =$ «—», $C[j] =$ «—».

Далее вычисляем элементы блоков матрицы по следующим правилам.

1. Если длины строк-компонентов отличаются не более чем на два символа ($|N_x - M_y| \leq 2$), то вычисляются элементы бинарной матрицы $\mathbf{B}_{x,y}$, т. е. блок $\mathbf{B}_{u,z} = \mathbf{B}_{x,y}$. При этом если символы большей строки-компонента расположены в столбце, а меньшей — в строке, то сначала производится транспонирование данной матрицы, затем вычисляются элементы бинарной матрицы $\mathbf{B}_{x,y}$.
2. Если длины строк-компонентов отличаются более чем на два символа (т. е. $|N_x - M_y| > 2$), то элементы матрицы $\mathbf{B}_{u,z}$ не вычисляются (пустой блок), так как допущено более двух ошибок.

Таким образом, каждый блок $\mathbf{B}_{u,z}$ отражает сходство соответствующих строк-компонентов S_x и C_y.

Третий шаг — подсчет числа несовпадений в каждом блоке.

Используя алгоритм приблизительного сравнения строк-компонентов, вычислим число несовпадений $K_0(\mathbf{B}_{uz})$ в каждом блоке.

Четвертый шаг — отображение числа несовпадений $K_0(\mathbf{B}_{u,z})$ в элементы матрицы $\mathbf{A}$:

$$f\left(K_0\left(\mathbf{B}_{u,z}\right)\right) \to \begin{cases} \times\,, & \text{если } K_0 > 2\,; \\ K_0(\mathbf{B}_{u,z})\,, & \text{если } K_0 \le 2\,. \end{cases}$$

Матрицу $\mathbf{A}$ будем называть *матрицей отображения ошибок*, так как матричные элементы $\mathbf{A}_{u,z}$ являются отображением числа несовпадений соответствующих строк-компонентов. Матрица $\mathbf{A}$ задана над множеством $\mathbf{T} = \{0, 1, 2, \times\}$, где элемент $\times = \max(T)$.

Матрица позволяет сравнивать все имеющиеся строки-компоненты между собой.

Пятый шаг — установление соответствия между $\min(P, Q)$-компонентами наименьшей строки с $\min(P, Q)$-компонентами наибольшей строки.

Строка-компонент C_y будет соответствовать компоненту S_x, если число несовпадений $K_0(\mathbf{B}_{x,y}) = \mathbf{A}_{u,z}$ для пары S_x и C_y принимает минимальное значение.

Для установления соответствий строк-компонентов разработана следующая рекуррентная формула:

$$\mathbf{K} = \begin{cases} \mathbf{I} = \{1, 2, \ldots, \min(P, Q)\}\,; \\ \mathbf{J} = \{1, 2, \ldots, \max(P, Q)\}\,; \\ \mathbf{I}_i = \mathbf{I}_{i-1} \cup \{u\}\,; \\ \mathbf{J}_j = \mathbf{J}_{j-1} \cup \{z\}\,; \\ \mathbf{K}_{u,z} = \min\limits_{z \in \mathbf{J} \setminus \mathbf{J}_j} \min\limits_{u \in \mathbf{I} \setminus \mathbf{I}_j} \mathbf{A}_{u,z}\,; \\ \mathbf{K}_t = \mathbf{K}_{u,z}\,, \text{ где } (u + z) \text{ — минимальная.} \end{cases}$$

Шестой шаг — установление сходства строк.

Сумму всех элементов множества $\mathbf{K}$ определим как общее число несовпадений строк-компонентов K_0, т. е. $K_0 = \sum \mathbf{K}$.

Пусть $\mathbf{I} = \{1, 2, \ldots, \max(P, Q)\}$; $\mathbf{J} = \{1, 2, \ldots, \max(P, Q)\}$.

Строки S и C считаем совпадающими в следующих случаях:

(1) если $(\sum K \le 2)$ и $(\max(P, Q) - \min(P, q) = 0)$ и $(\mathbf{I}_i = \mathbf{I})$ и $(\mathbf{J}_j = \mathbf{J})$;

(2) если $(\sum K \le 2)$ и $(\max(P, Q) - \min(P, q) = 1)$ и $(\mathbf{I}_i = \mathbf{I})$.

Разработанный алгоритм дает достаточно точные результаты сравнения, что подтверждается программной реализацией алгоритма. Реализация была произведена на языке Delphi 7 в связке с системой управления базами данных Access. Полученные результаты представлены на рис. 1 и 2.

Разработанный алгоритм имеет ограничения на строки. Сравниваемые строки должны быть образованы в результате следующих преобразований: не более

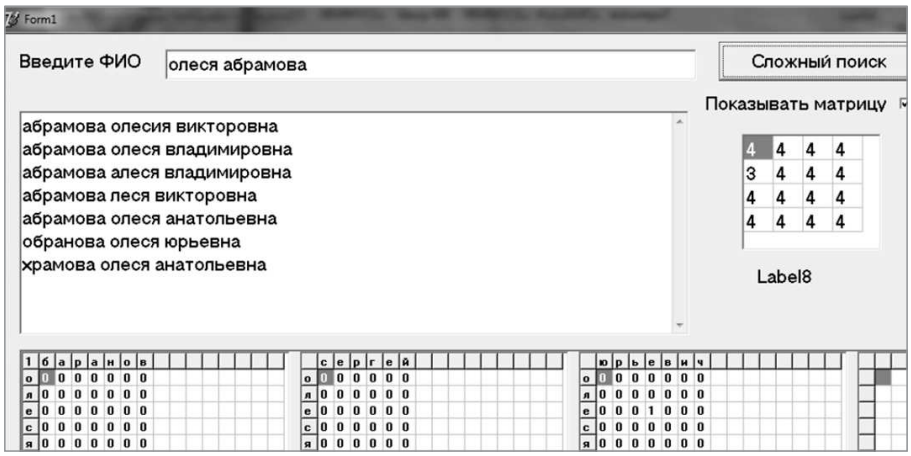


Рис. 1 Фамилии без дополнительных элементов

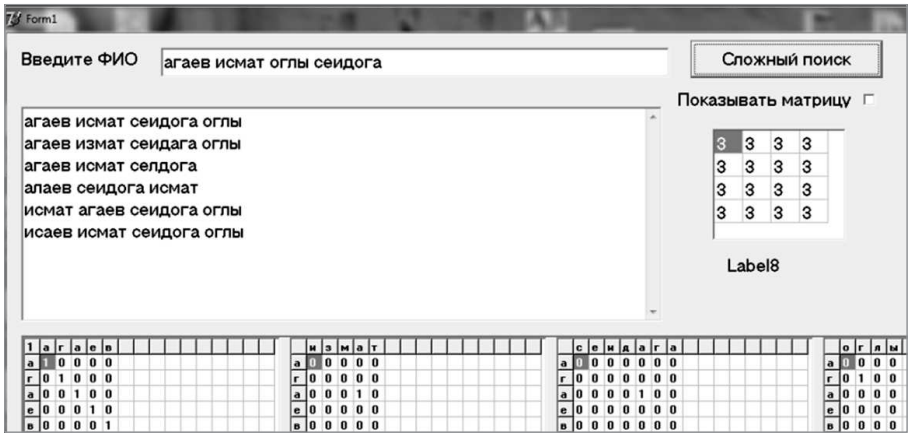


Рис. 2 Фамилии с дополнительными элементами

двух опечаток в компонентах ФИО; удаление или появление одного компонента ФИО; смена местоположения компонентов. Однако если строка образована, например, в результате склеивания (конкатенации) строк-компонентов, то алгоритм будет давать неверные результаты сравнения.

4 Заключение

Выбор применяемого алгоритма в информационной системе зависит от предъявляемых условий точности. В статистических системах всегда существует

определенный коэффициент погрешности, что позволяет использовать быстрые «поверхностные» алгоритмы. В других же системах должна достигаться максимальная точность, т. е. полное исключение дублирований, так как за каждой записью в том или ином виде стоит государственное финансирование. Например, в системе ОМС — страховые взносы на застрахованного. В таких информационных системах на первое место выходит точность полученных данных, а не время, потраченное на их обработку.

Существующие алгоритмы нечеткого поиска дают не всегда точный результат в ИСПД. Разработанный алгоритм нечеткого сравнения основан на особенностях структуры матриц и дает точные результаты при работе с персональными данными в базах данных, исключая случай «склеивания» данных. В алгоритме учтены возможности неправильного написания приставок к фамилиям или двойных фамилий, имен, а также допущения ошибок в записи персональных данных.

Литература

1. *Бобылева О. В., Бобылев В. А.* Нечеткий поиск персональных данных в информационных системах // Динамика развития современной науки: Сб. статей Междунар. научн.-практич. конф. — Уфа: Омега Сайнс, 2015. С. 18–20.

Поступила в редакцию 03.08.18

FUZZY STRING COMPARISON WHEN PROCESSING PERSONAL DATA

***O. V. Bobyleva*[1], *I. S. Bekesheva*[1], *V. A. Bobylev*[2], *and V. V. Charkova*[1]**

[1]Khakas State University named after N. F. Katanov, 90 Lenina Str., Abakan 655017, Russian Federation

[2]Territorial Fund for Mandatory Medical Insurance of the Republic of Khakassia, 199 Pushkina Str., Abakan 655017, Russian Federation

Abstract: The article substantiates the necessity to develop a new fuzzy search method targeted at word comparison within the databases containing personal data. The advantages of this algorithm application are demonstrated through the specific examples from the sphere of health insurance. The development of mathematical algorithm model is carried out on the basis of cell texture of matrices.

Keywords: algorithm; fuzzy search; fuzzy comparison; matrix

DOI: 10.14357/08696527190313

References

1. Bobyleva, O. V., and V. A. Bobylev. 2015. Nechetkiy poisk personal'nykh dannykh v informatsionnykh sistemakh [Fuzzy search for personal data in information systems].

Dynamics of the Development of Modern Science: Scientific-Practical Conference (International) Proceedings. Ufa: Omega Sines Publs. 18–20.

Received August 3, 2018

Contributors

Bobyleva Oksana V. (b. 1982) — Candidate of Science (PhD) in physics and mathematics, assistant professor, Department of Mathematics and Mathematics Teaching Methods, Khakas State University named after N. F. Katanov, 90 Lenina Str., Abakan 655017, Russian Federation; nimdar@bk.ru

Bekesheva Irina S. (b. 1987) — Candidate of Science (PhD) in education, assistant professor, Department of Mathematics and Mathematics Teaching Methods, Khakas State University named after N. F. Katanov, 90 Lenina Str., Abakan 655017, Russian Federation; irrisskay@mail.ru

Bobylev Vadim A. (b. 1981) — Head of Information Support Department, Territorial Fund for Mandatory Medical Insurance of the Republic of Khakassia, 199 Pushkina Str., Abakan 655017, Russian Federation; vadimbobylev@yandex.ru

Charkova Viktoriya V. (b. 1997) — student, Khakas State University named after N. F. Katanov, 90 Lenina Str., Abakan 655017, Russian Federation; charkova_97@mail.ru

СИСТЕМЫ И СРЕДСТВА ИНФОРМАТИКИ, 2019. Т. 29. № 3. С. 160–168

ПРОЦЕСС КОРРЕКЦИИ ОШИБОК СЕМАНТИЧЕСКОЙ СЕТИ КАК НЕЛИНЕЙНАЯ ДИНАМИЧЕСКАЯ СИСТЕМА

И. М. Адамович[1]***, О. И. Волков***[2]

Аннотация: Статья продолжает серию работ, посвященных моделированию ошибок независимых пользователей при формировании семантической сети, лежащей в основе распределенной технологии поддержки конкретно-исторических исследований. Данная статья посвящена описанию и обоснованию подхода к моделированию организационных мер поиска и исправления ошибок подсети экземпляров семантической сети технологии. Описана специфика данного вида ошибок и обоснована необходимость их изучения. Предложенный подход заключается в анализе процессов изменения числа ошибок семантической сети и усилий пользователей, противостоящих их росту, как нелинейной динамической системы. В рамках данных усилий выделяется и описывается отдельный подкласс — волонтерство, характеризующийся добровольными и целенаправленными акциями пользователей по коррекции ошибок. С помощью данного подхода была количественно оценена эффективность действий волонтеров, и на основании этой оценки были сформулированы рекомендации для сообщества пользователей технологии.

Ключевые слова: семантическая сеть; модель; ошибки пользователей; динамическая система; исправление ошибок

DOI: 10.14357/08696527190314

1 Введение

В статье [1] была описана новая распределенная технология поддержки историко-биографических исследований, для которой была обоснована форма организации информации в виде семантической сети. В статье [2] была поставлена и обоснована задача оценки качества этой семантической сети, формируемой одновременно множеством не связанных между собой исследователей. Была описана и обоснована модель, построенная на базе сочетания принципов графодинамики [3] с принципом предпочтительного присоединения [4], позволяющая изучить свойства информации, организованной в форме семантической сети, в динамике. В статье [5] было описано дальнейшее развитие модели семантической сети, предусматривающее включение в нее механизмов параллельной фиксации ошибочных и соответствующих им безошибочных (идеальных) действий

[1]Институт проблем информатики Федерального исследовательского центра «Информатика и управление» Российской академии наук, Adam@amsd.com

[2]Институт проблем информатики Федерального исследовательского центра «Информатика и управление» Российской академии наук, Volkov@amsd.com

пользователей, а также механизма имитации поисковых запросов, выполняемых параллельно в искаженной и неискаженной подсетях. С помощью данной модели были проведены экспериментальные проверки влияния ошибок пользователей на качество семантической сети в динамике. В статье [6] были описаны возможные организационные меры поиска и исправления ошибок, а также приведены результаты экспериментальной проверки их эффективности.

Под ошибками пользователей в указанных работах понимался определенный класс возможных ошибок, а именно:

- ошибки, влияющие на структуру данных (ОСД), — это, прежде всего, ошибки формирования подсети понятий, а также ошибки связей типа «часть»;
- ошибки, влияющие на формирование информационного запроса, поскольку они, как и ОСД, влияют на возможность получения информации от семантической сети.

При этом ошибки формирования экземпляра (ОФЭ) не рассматривались в принципе, поскольку они могут привести к затруднению корректного поиска только этого экземпляра и не являются ОСД. Взаимовлияние ОФЭ при их накоплении в семантической сети невозможно, а следовательно, доля ошибочных экземпляров, даже при отсутствии каких-либо мер по их выявлению и исправлению, ограничена вероятностью ошибки оператора, значение которой для нестрессовых ситуаций мало [7].

Но, поскольку в процессе эксплуатации реальной семантической сети пользователями неизбежно будут предприниматься попытки выявления и исправления ошибок, в том числе и ОФЭ, представляется целесообразным заранее оценить эффективность таких усилий, с тем чтобы иметь возможность сформулировать для пользователей сети соответствующие методические рекомендации. Так, неограниченный рост объема документов в сочетании с лимитированным объемом ресурсов, выделяемых сообществом исследователей на проверку данных, могут привести к постепенному снижению эффективности таких проверок вплоть до их полной бесполезности.

Целью настоящей статьи является исследование эффективности усилий не связанных между собой пользователей распределенной технологии поддержки историко-биографических исследований по выявлению и исправлению ошибок подсети экземпляров семантической сети технологии.

2 Описание подхода

Опишем динамику изучаемой семантической сети с точки зрения изменения числа ОФЭ и усилий пользователей, противостоящих их росту, как динамическую систему.

Далее под ошибками будем понимать исключительно ОФЭ, под документами — элементы подсети экземпляров, под пользователями — пользователей

технологии поддержки конкретно-исторических исследований, под волонтерами — пользователей, добровольно осуществляющих в данный момент помимо исследовательской деятельности целенаправленную деятельность по поиску и исправлению ошибок.

Так же, как и в [5, 6], полагаем, что технология находится на стационарном этапе развития сообщества [8] и число пользователей можно считать неизменным.

Изменение числа ошибок является результатом трех происходящих в системе процессов:

(1) накопления ошибок, происходящего с постоянной скоростью, пропорциональной скорости добавления документов в систему и вероятности для пользователя сделать ошибку;

(2) исправления пользователями случайно обнаруженных в процессе исследовательской деятельности ошибок, происходящего со скоростью, пропорциональной плотности ошибок в системе;

(3) целенаправленного исправления ошибок волонтерами, происходящего со скоростью, пропорциональной текущему числу волонтеров, плотности ошибок, среднему объему документов, просматриваемых волонтером за один сеанс исправления, и вероятности заметить ошибку.

Принимаем, что вероятность события, когда пользователь временно берет на себя функции волонтера, пропорциональна числу ошибок, замеченных им в процессе основной деятельности в последнее время. Соответственно, скорость роста числа волонтеров будет пропорциональна плотности ошибок, но с учетом ограничения на предельное значение числа волонтеров, равного общему числу пользователей.

Ограничение числа волонтеров предельным значением может быть описано уравнением Ферхюльста (логистическим уравнением) [9]:

$$\dot{N} = rN\left(1 - \frac{N}{K}\right),$$

обеспечивающим ограничение роста популяции $N(t)$ предельным значением K (потенциальная емкость системы). Параметр r характеризует скорость роста популяции.

Введем обозначения:

t — время в днях от начала эксперимента;

ε — текущее число ошибок;

ω — текущее число волонтеров;

m_0 — исходное число документов;

n — число пользователей (константа);

w — средняя скорость добавления документов одним участником;

$m = nw$ — скорость роста числа документов;

$\mu = m_0 + mt$ — текущее число документов;

$\rho = \varepsilon/\mu$ — текущая плотность ошибок;

v — вероятность для участника сделать ошибку;

p — средняя скорость просмотра документов одним участником;

p_v — среднее число документов, проверяемых волонтером за один раз;

v_0 — вероятность заметить ошибку в ошибочном документе;

v_ω — скорость роста числа волонтеров при увеличении плотности ошибок.

В данных обозначениях поведение системы описывается с помощью следующей системы уравнений:

$$\left.\begin{aligned} \frac{d\varepsilon}{dt} &= mv - pv_0 n\rho - p_v v_0 \omega\rho\,; \\ \rho &= \frac{\varepsilon}{\mu}\,; \\ \mu &= m_0 + mt\,; \\ \frac{d\omega}{dt} &= -r\omega\left(1 + \frac{\omega}{n}\right) + v_\omega\rho\,. \end{aligned}\right\} \qquad (1)$$

Первое уравнение системы описывает динамику числа ошибок. Первый член правой части этого уравнения соответствует линейному характеру роста числа ошибок при отсутствии каких-либо исправлений. Второй член соответствует исправлению пользователями случайно обнаруженных в процессе исследовательской деятельности ошибок, происходящему со скоростью, пропорциональной плотности ошибок в системе. Третий член соответствует целенаправленному исправлению ошибок волонтерами, происходящему со скоростью, пропорциональной текущему числу волонтеров и плотности ошибок.

Второе уравнение системы описывает связь между плотностью ошибок и их числом.

Третье уравнение описывает линейный характер роста числа документов в системе на стационарном этапе развития сообщества.

Четвертое уравнение описывает динамику числа волонтеров. Первый член правой части этого уравнения представляет собой логистическое уравнение, ограничивающее их рост общим числом пользователей, а также описывающее экспоненциальный спад их числа при отсутствии ошибок. Второй член соответствует линейной зависимости скорости роста числа волонтеров от плотности ошибок.

Перегруппировав в системе (1) члены и введя обозначения $a = mv$, $b = pv_0n$, $c = p_v v_0$, $d = r$ и $e = v_\omega$, преобразуем ее к следующему виду:

$$\left.\begin{aligned} \varepsilon' &= a - \frac{b\varepsilon + c\omega\varepsilon}{m_0 + mt}; \\ \omega' &= -d\omega\left(1 + \frac{\omega}{n}\right) + \frac{e\varepsilon}{m_0 + mt}, \end{aligned}\right\} \tag{2}$$

где a, b, c, d и e — константы.

В каждый момент времени состояние динамической системы, соответствующей системе (2), описывается точкой фазового пространства (ε, ω). Интерес представляет не общее число ошибок, которое, безусловно, будет расти по мере роста числа документов, а их плотность, которая характеризует степень снижения поисковой доступности информации [10]. Но поскольку ρ и ε связаны простой зависимостью, переход от пространства (ε, ω) к пространству (ρ, ω) не вызывает затруднений.

3 Анализ системы

Система (2) является нелинейной неавтономной системой дифференциальных уравнений, что затрудняет ее аналитический анализ. Поэтому анализ был осуществлен численно с помощью пакета MatLab. На рис. 1 приведен вид решений системы. Из графика видно, что решения довольно быстро асимптотически приближаются к некоторым стационарным значениям, что говорит о наличии положения равновесия системы.

Для поиска положения равновесия в фазовом пространстве (ρ, ω) следует учесть, что

$$\rho' = \frac{\varepsilon'\mu - \mu'\varepsilon}{\mu^2} = \frac{(m_0 + mt)(a - b\rho - c\omega\rho) - \varepsilon m}{(m_0 + mt)^2} = \frac{a - (b + m)\rho - c\omega\rho}{m_0 + mt}.$$

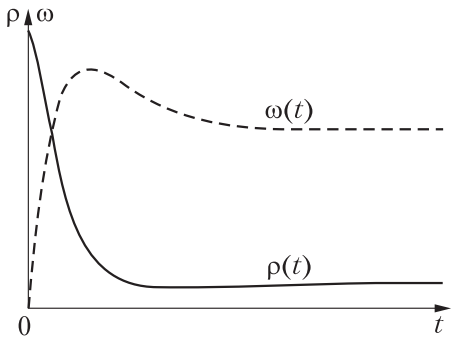


Рис. 1 Вид решений системы

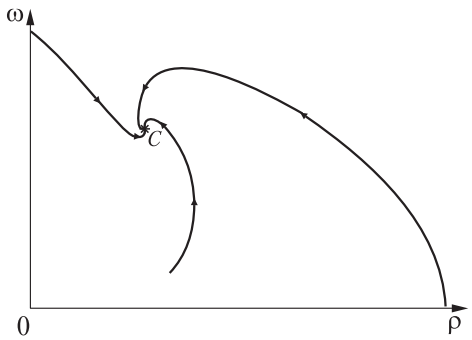


Рис. 2 Фазовый портрет системы

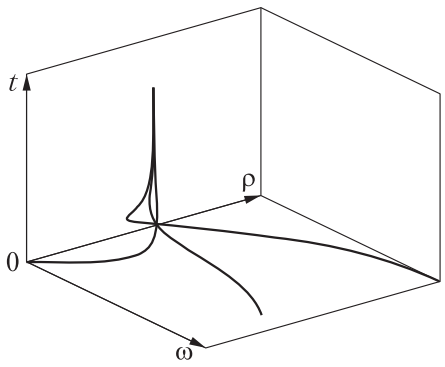


Рис. 3 Интегральные кривые системы

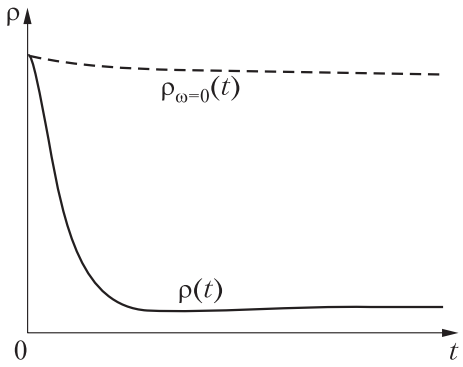


Рис. 4 Сравнение изменения плотности ошибок для случаев нулевой $\rho_{\omega=0}(t)$ и ненулевой $\rho(t)$ активности волонтеров

Из этого следует, что

$$\begin{cases} \rho' = 0\,; \\ \omega' = 0\,; \end{cases} \Leftrightarrow \begin{cases} a - (b+m)\rho - c\omega\rho & = 0\,; \\ -d\omega\left(1 + \dfrac{\omega}{n}\right) + e\rho & = 0\,. \end{cases}$$

Данная система имеет единственное решение, и положение C этого решения на фазовой плоскости не зависит от t, несмотря на неавтономность системы (2).

На рис. 2 приведен фазовый портрет системы, из которого видно, что точка C является устойчивым фокусом.

На рис. 3 приведен вид соответствующих интегральных кривых, демонстрирующий быстрый характер сходимости для различных решений системы (2).

4 Анализ эффективности усилий волонтеров

Для анализа эффективности усилий волонтеров было произведено сравнение решений исходной системы с решением системы модифицированной, в которой коэффициент c установлен равным нулю. Такая модификация позволяет смоделировать полное отсутствие влияние волонтеров на снижение числа ошибок. Сравнение вида кривых $\rho(t)$ приведено на рис. 4.

Расчет показывает, что для значений параметров, соответствующих параметрам моделей, приведенных и обоснованных в [5, 6], соотношение ρ-координат точек равновесия C для модифицированной и немодифицированной систем находится в диапазоне 0,803–0,894, что соответствует диапазону 11,877%–24,513% для уровня эффективности волонтеров.

5 Выводы

Проведенное исследование показало, что предположение о постепенном снижении эффективности усилий не связанных между собой пользователей распределенной технологии поддержки историко-биографических исследований по выявлению и исправлению ошибок подсети экземпляров семантической сети технологии по мере роста объема документов не подтвердилось.

Эффективность при различных значениях параметров модели не опускается ниже 12%, что является довольно высоким показателем.

Это позволяет рекомендовать сообществу пользователей технологии осуществление проверок на вышеописанных принципах.

Литература

1. *Адамович И. М., Волков О. И.* Технология распределенного автоматизированного анализа исторических текстов // Системы и средства информатики, 2016. Т. 26. № 3. С. 148–161.
2. *Адамович И. М., Волков О. И.* Об одном подходе к моделированию процесса развития семантической сети // Системы и средства информатики, 2017. Т. 27. № 2. С. 143–154.
3. *Юдицкий С. А.* Графодинамическое имитационное моделирование развития сетевых структур // Управление большими системами, 2011. Вып. 33. С. 21–34.
4. *Бадрызлов В. А.* Принципы генерации случайных графов для моделирования сети Интернет // Омский научный вестник, 2014. № 3(133). С. 204–208.
5. *Адамович И. М., Волков О. И.* Влияние ошибок пользователей на динамику качества семантической сети // Системы и средства информатики, 2017. Т. 27. № 4. С. 150–163.
6. *Адамович И. М., Волков О. И.* Модель процесса коррекции ошибок в семантической сети // Системы и средства информатики, 2018. Т. 28. № 1. С. 65–76.
7. *Белов С. В., Ильницкая А. В., Козьяков А. Ф. и др.* Безопасность жизнедеятельности / Под общ. ред. С. В. Белова. — 6-е изд. — М.: Высш. шк., 2006. 616 с.
8. *Сазанов В. М.* Сравнительный анализ социально-сетевых проектов // Социально-ориентированные сети и технологии: Исследовательский ресурс социального интернета, 2009. http://v-school.narod.ru/PAPERS/analiz.doc.
9. *Братусь А. С., Новожилов А. С., Платонов А. П.* Динамические системы и модели биологии. — М.: Физматлит, 2010. 400 с.
10. *Морвиль П.* Тотальная видимость / Пер. с англ. — СПб.: Символ-Плюс, 2008. 272 с. (*Morville P.* Ambient findability. — Beijing: O'Reilly Media, 2005. 208 p.)

Поступила в редакцию 12.02.19

THE ERROR CORRECTION PROCESS IN THE SEMANTIC NETWORK AS A NONLINEAR DYNAMIC SYSTEM

I. M. Adamovich and O. I. Volkov

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: This article continues the series of works devoted to modeling the errors of independent users in the formation of a semantic network which is the basis for the distributed technology to support specific historical investigations. This article is devoted to the description and substantiation of the approach to modeling of organizational measures of search and correction of errors of subnetworks of the semantic network of technology copies. The specificity of this type of errors was described and the necessity of their study was substantiated. The proposed approach is to analyze the processes of the number of semantic network errors changing and the efforts of users countering its increase as a nonlinear dynamic system. As a part of these efforts, a separate subclass — volunteerism characterized by voluntary and targeted actions of users to correct the errors — was highlighted and described. Using this approach, the effectiveness of volunteers' actions was quantified and on the basis of this estimate, the recommendations for the community of technology users were formulated.

Keywords: semantic net; model; user errors; dynamic systems; error correction

DOI: 10.14357/08696527190314

References

1. Adamovich, I. M., and O. I. Volkov. 2016. Tekhnologiya raspredelennogo avtomatizirovannogo analiza istoricheskikh tekstov [Distributed automated technology of historical texts analysis]. *Sistemy i Sredstva Informatiki — Systems and Means of Informatics* 3(26):148–161.
2. Adamovich, I. M., and O. I. Volkov. 2017. Ob odnom podkhode k modelirovaniyu protsessa razvitiya semanticheskoy seti [An approach to modeling the semantic net evolution]. *Sistemy i Sredstva Informatiki — Systems and Means of Informatics* 2(27):143–154.
3. Yuditskiy, S. A. 2011. Grafodinamicheskoe imitatsionnoe modelirovanie razvitiya setevykh struktur [Graphodynamic simulation modeling of network structures evolution]. *Upravlenie bol'shimi sistemami* [Large-Scale Systems Control] 33:21–34.
4. Badryzlov, V. A. 2014. Printsipy generatsii sluchaynykh grafov dlya modelirovaniya seti Internet [The principles of generation of random graphs for simulation of the Internet]. *Omskiy nauchnyy vestnik* [Omsk Scientific Bulletin] 3(133):204–208.
5. Adamovich, I. M., and O. I. Volkov. 2017. Vliyanie oshibok pol'zovateley na dinamiku kachestva semanticheskoy seti [The influence of user errors on semantic net quality

dynamics]. *Sistemy i Sredstva Informatiki — Systems and Means of Informatics* 4(27):150–163.

6. Adamovich, I. M., and O. I. Volkov. 2018. Model' protsessa korrektsii oshibok v semanticheskoy seti [The model of semantic net error correction process]. *Sistemy i Sredstva Informatiki — Systems and Means of Informatics* 1(28):65–76.
7. Belov, S. V., A. V. Ilnitskaya, A. F. Koziakov, *et al.* 2006. *Bezopasnost' zhiznedeyatel'nosti* [Life safety]. Ed. S. V. Belov. 6th ed. Moscow: Vysshaya Shkola Publs. 616 p.
8. Sazanov, V. M. 2009. Sravnitel'nyy analiz sotsial'no-setevykh proektov [Comparative analysis of social-network projects]. *Sotsial'no-orientirovannye seti i tekhnologii. Issledovatel'skiy resurs sotsial'nogo interneta* [Socially-oriented networks and technologies. Research resource of social Internet]. Available at: http://v-school.narod.ru/PAPERS/analiz.doc (accessed February 11, 2019).
9. Bratus, A. S., A. S. Novozhilov, and A. P. Platonov. 2010. *Dinamicheskie sistemy i modeli biologii* [Dynamical systems and models in biology]. Moscow: Fizmatlit. 400 p.
10. Morville, P. 2005. *Ambient findability*. Beijing: O'Reilly Media. 208 p.

Received February 12, 2019

Contributors

Adamovich Igor M. (b. 1934) — Candidate of Science (PhD) in technology, leading scientist, Institute of Informatics Problems, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, 44-2 Vavilov Str., Moscow 119133, Russian Federation; Adam@amsd.com

Volkov Oleg I. (b. 1964) — leading programmer, Institute of Informatics Problems, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, 44-2 Vavilov Str., Moscow 119133, Russian Federation; Volkov@amsd.com

СИСТЕМЫ И СРЕДСТВА ИНФОРМАТИКИ, 2019. Т. 29. № 3. С. 169–179

ФОРМИРОВАНИЕ СИТУАЦИОННО ЗАВИСИМЫХ СИСТЕМ ТРЕБОВАНИЙ К РЕШЕНИЯМ ЗАДАЧ ПЛАНИРОВАНИЯ РАСХОДОВ

***А. В. Ильин*[1], *В. Д. Ильин*[2]**

Аннотация: Предложен подход к экспертному формированию ситуационно зависимых систем требований к решениям задач планирования расходов. Приведена постановка и методы решения линейной задачи ситуационного планирования расходов. В зависимости от набора требований задача решается либо методом приоритетного интервального распределения, либо методом целевого перемещения решения. Каждый из методов позволяет найти план расходов, всегда удовлетворяющий обязательным требованиям и максимально удовлетворяющий ориентирующим требованиям. На каждом шаге поиска плана в режиме вычислительного эксперимента постановка задачи определяется системой обязательных и ориентирующих требований, которая формируется экспертом-планировщиком на основе анализа портретов ситуаций. Предусмотрена возможность задать несколько показателей качества решения. Представление данных и результата планирования в виде числовых отрезков позволяет учесть точность прогнозирования величины распределяемого ресурса и ожидаемых расходов. Формируемые с помощью цифровых двойников портреты ситуаций (целевой, стартовой и достигнутой) представлены формализованным описанием ключевых параметров, характеризующих состояние источников расходуемого ресурса, его потребителей и условия планирования. Приведена характеристика действующего интернет-сервиса планирования расходов.

Ключевые слова: ситуационно зависимые системы требований; ситуационное планирование расходов; метод приоритетного интервального распределения; метод целевого перемещения решения; портреты ситуаций; интернет-сервис планирования расходов

DOI: 10.14357/08696527190315

1 Введение

Неуклонно усиливаются основания оценивать уровень социально-экономического развития страны и ее регионов в тесной связи с уровнем продуктивного *применения цифровых информационных технологий* (далее — *цифровизации*) для совершенствования ключевых видов деятельности (научной, образовательной,

[1]Государственный научно-исследовательский институт авиационных систем, ilyin@res-plan.com

[2]Вычислительный центр им. А. А. Дородницына Федерального исследовательского центра «Информатика и управление» Российской академии наук, vdilyin@yandex.ru

производственной и др.) [1–3]. Повышение экономической эффективности реализации социально-экономических проектов предполагает совершенствование методологии планирования расходов на федеральные и региональные проекты на основе передовых информационных технологий. План расходов, как правило, обновляется на всех этапах жизненного цикла проекта. Конкретизируются детали проекта: уточняются задачи, источники и объемы располагаемых ресурсов, статьи и величины расходов, изменение цен и другие факторы. При этом невозможность точного прогноза доходной части проекта остается аксиомой, справедливой и для расходов по составляющим проекта. Чтобы учесть реальную *точность данных и реализации решений* [4], при планировании расходов целесообразно *данные и результат представлять в виде числовых отрезков* (их левые границы соответствуют наименьшим ожидаемым значениям, а правые — наибольшим) [5].

Рассматриваемые результаты. В статье представлены результаты развития методологии *ситуационного планирования расходов* при изменениях требований к решению в условиях информационной неполноты [6]. Результаты получены при выполнении научно-исследовательской работы «Моделирование социальных, экономических и экологических процессов» (№ 0063-2016-0005), выполняемой в соответствии с государственным заданием ФАНО России для Федерального исследовательского центра Информатика и управление РАН.

Запись формул и выделение фрагментов текста. Для записи формул, выделения определений и замечаний далее используются средства языка TSM-комплекса (TSM: Textual Symbolic Modeling), разработанного для формализованного описания текстовых моделей[1].

В статье применены следующие средства выделения фрагментов текста:

□ ⟨фрагмент описания⟩ □ ≈ утверждение (определение, аксиома и др.) (здесь и далее символ ≈ заменяет слово «означает»);

◊ ⟨фрагмент описания⟩ ◊ ≈ замечание;

◯ ⟨фрагмент описания⟩ ◯ ≈ пример.

Курсивом выделены первые вхождения названий понятий и фрагменты описания, к которым авторы хотят привлечь внимание.

Приоритеты требований записываются в форме

⟨требование⟩ ← ⟨приоритет⟩.

Инфраструктурное ядро цифровизации. Современное инфраструктурное ядро цифровизации включает технологии цифровых двойников (*англ.* digital twins) [7], облачных вычислений (*англ.* cloud computing) [8–11], электронных сервисов [12–14], Интернета вещей (*англ.* Internet of Things, IoT) [15, 16] и M2M-технологии (*англ.* Machine-to-Machine, M2M) [17, 18]. Темпы развития этого ядра во многом определяют темпы *нормализации экономического механизма* [19].

[1] *Ильин В. Д.* Символьное моделирование // Большая российская энциклопедия, 2019. [Электронный ресурс]. URL: http://dev.bigenc.ru/technology and technique/text/4010980 (дата обращения: 17.07.2019).

□ *Цифровой двойник* (DT — Digital Twin) — реализованная в s-среде обучаемая символьная модель наблюдаемого объекта, предназначенная для анализа и совершенствования его поведения. □ Для построения DT используют значения параметров, характеризующих поведение моделируемого объекта (◯ для технических устройств — данные, поступающие от соответствующих датчиков ◯). С момента создания DT (как образа объекта) он накапливает знания о своем прообразе, обмениваясь с ним сообщениями и обновляясь. Различают DT-прототипы (DT Prototype, DTP), DT-экземпляры (DT Instance, DTI) и агрегированные DT (DT Aggregate, DTA). DTP используется при создании физической версии моделируемого объекта; DTI — модель существующего объекта, с которым DTI предназначено взаимодействовать (DTI собирают данные о своих прообразах с помощью IoT-датчиков); DTA — комплекс средств обработки данных, предназначенный для взаимодействия с заданной совокупностью DTI. ◊ DT-технология построения и функционирования цифровой *символьной модели* произвольного объекта предполагает возможность получения данных о его поведении (путем измерения значений некоторой совокупности параметров объекта). ◊

2 Постановка задачи

Для задающего величину распределяемого ресурса числового отрезка $\left[a^{\min}, a^{\max}\right]$ ($a^{\min} \geq 0$, $a^{\max} > 0$), отрезков $\left[b_i^{\min}, b_i^{\max}\right]$ ($b_i^{\min} \geq 0$, $b_i^{\max} > 0$, $i = 1, \ldots, n$), задающих запросы по расходным статьям, и весовых коэффициентов (приоритетов) расходных статей $p_i > 0$ ($i = 1, \ldots, n$) требуется найти план расходов по статьям:

$$\left[x_i^{\min}, x_i^{\max}\right] : \left\{0 \leq x_i^{\min} \leq b_i^{\min},\ x_i^{\max} \leq b_i^{\max},\ \sum x_i^{\min} \leq a^{\min},\right.$$
$$\left.\sum x_i^{\max} \leq a^{\max},\ i = 1, \ldots, n\right\}. \quad (1)$$

Для совокупного вектора искомого плана $\boldsymbol{x} = (x_1^{\min}, \ldots, x_n^{\min}, x_1^{\max}, \ldots, x_n^{\max})$ может быть также задан набор требований

$$\boldsymbol{C}\boldsymbol{x} \leq \boldsymbol{d} \leftarrow \boldsymbol{q}\,, \quad (2)$$

где $\boldsymbol{C}$ — матрица вещественных коэффициентов размера $k \times 2n$ ($k \geq 1$); $\boldsymbol{d}$ — вектор-столбец вещественных констант размера k; $\boldsymbol{q}$ — вектор-столбец весовых коэффициентов (приоритетов) требований ($0 < q_i \leq +\infty$, $i = 1, \ldots, k$).

Требования к решению. *Обязательные требования* имеют приоритет $+\infty$. Приоритеты *ориентирующих требований* задаются положительными вещественными числами. ◊ Это делает эксперт-планировщик, учитывая относительную важность выполнения рассматриваемых требований. ◊ Требования (1) являются

обязательными. Требования (2) могут быть как обязательными, так и ориентирующими. ◊ Возможность задания набора требований (2) позволяет эксперту-планировщику ввести отражающие специфику ситуации ограничения в виде линейных соотношений между компонентами совокупного вектора плана. ◊

Прикладная точность. ◊ *Прикладная точность*, задаваемая экспертом-планировщиком, определяет в выбранных единицах измерения минимальную значимую сумму (◯ целые степени 10 от -4 до 9 ◯). Данные округляются в соответствии с заданной прикладной точностью. Результаты расчетов округляются по специальному алгоритму — с сохранением требуемой суммы, учетом запросов и приоритетов. *Возможность задания прикладной точности позволяет*, в частности, *решать целочисленные задачи*. ◊

Портреты ситуаций. □ *Ситуация* — состояние доходной и расходной части проекта, представленное совокупностью параметров, описывающих источники поступлений ресурсов, расходные статьи и др. *Целевой* называем ситуацию, которую планируется создать; *стартовой* — существующую в начале поиска плана расходов; *достигнутой* — созданную в результате выполнения этапного плана. □

□ *Портрет ситуации* — формализованное описание состояния доходной и расходной части проекта. □ Совокупность правил, применяемых для построения портрета ситуации, порядок их срабатывания и истинность описаний определяются с помощью *механизма интерпретации* запросов на заданном множестве правил [5, 6]. Мониторинг доходов и расходов, построение и анализ портретов ситуаций, формирование систем требований, поиск планов и оценка их эффективности выполняются с помощью онлайн-сервисов, функционирующих в среде DTI, управление поведением которых осуществляется с помощью специального DTA. Цифровые двойники представляют собой dt-инфы [20], предназначенные для совершенствования поведения онлайн-сервисов.

3 Наборы требований и методы решения

В трех следующих случаях задача решается итеративным *методом приоритетного интервального распределения* [6], реализованным в действующем *интернет-сервисе планирования расходов* [13]:

(1) $\sum b_i^{\min} > a^{\min}$, $\sum b_i^{\max} > a^{\max}$ $(i = 1, \ldots, n)$
и набор требований (2) имеет вид:

$$\Big\{\sum x_i^{\min} = a^{\min} \leftarrow +\infty,\ \sum x_i^{\max} = a^{\max} \leftarrow +\infty\ (i = 1, \ldots, n),$$
$$p_j b_j^{\min} x_i^{\min} - p_i b_i^{\min} x_j^{\min} = 0 \leftarrow 1$$
$$\text{для каждой пары } i, j\{1 \le i \le n,\ 1 \le j \le n\},$$
$$p_j \left(b_j^{\max} - b_j^{\min}\right)\left(x_i^{\max} - x_i^{\min}\right) - p_i \left(b_i^{\max} - b_i^{\min}\right)\left(x_j^{\max} - x_j^{\min}\right) = 0 \leftarrow 1$$

$$\text{для каждой пары } i,j:\left\{b_i^{\max}>b_i^{\min},\ b_j^{\max}>b_j^{\min},\ 1\le i\le n,\ 1\le j\le n\right\},$$
$$p_j b_j^{\max} x_i^{\max} - p_i b_i^{\max} x_j^{\max} = 0 \leftarrow 1$$
$$\text{для каждой пары } i,j:\left\{b_i^{\max}=b_i^{\min},\ b_j^{\max}=b_j^{\min},\ 1\le i\le n,\ 1\le j\le n\right\}\Big\}; \quad (3)$$

(2) $\sum b_i^{\min} \le a^{\min}$, $\sum b_i^{\max} > a^{\max}$ $(i=1,\ldots,n)$ и набор требований (2) имеет вид:

$$\Big\{x_i^{\min}=b_i^{\min}\leftarrow+\infty, \sum x_i^{\max}=a^{\max}\leftarrow+\infty\ (i=1,\ldots,n),$$
$$p_j\left(b_j^{\max}-b_j^{\min}\right)\left(x_i^{\max}-x_i^{\min}\right)-p_i\left(b_i^{\max}-b_i^{\min}\right)\left(x_j^{\max}-x_j^{\min}\right)=0\leftarrow 1$$
$$\text{для каждой пары } i,j:\left\{b_i^{\max}>b_i^{\min},\ b_j^{\max}>b_j^{\min},\ 1\le i\le n,\ 1\le j\le n\right\},$$
$$p_j b_j^{\max} x_i^{\max} - p_i b_i^{\max} x_j^{\max} = 0 \leftarrow 1$$
$$\text{для каждой пары } i,j:\left\{b_i^{\max}=b_i^{\min},\ b_j^{\max}=b_j^{\min},\ 1\le i\le n,\ 1\le j\le n\right\}\Big\}; \quad (4)$$

(3) $\sum b_i^{\min} > a^{\min}$, $\sum b_i^{\max} \le a^{\max}$ $(i=1,\ldots,n)$ и набор требований (2) имеет вид:

$$\Big\{\sum x_i^{\min}=a^{\min}\leftarrow+\infty,\ x_i^{\max}=b_i^{\max}\leftarrow+\infty\ (i=1,\ldots,n),$$
$$p_j b_j^{\min} x_i^{\min} - p_i b_i^{\min} x_j^{\min} = 0 \leftarrow 1$$
$$\text{для каждой пары } i,j\ \{1\le i\le n,\ 1\le j\le n\}\Big\}. \quad (5)$$

В случаях когда набор требований (2) отличается от вариантов (3)–(5), задача решается *методом целевого перемещения решения* [21]. Для удобства читателя далее приведено схематическое описание идеи этого метода.

При несовместности системы ограничений в качестве начального плана рассматривается чебышёвская точка [21], минимизирующая максимальный дефицит: $\min_x \max_i (c_{i1}x_1+\cdots+c_{i,2n}x_{2n}-d_i)$, $i=1,\ldots,k$. Здесь и далее для удобства записи

$$x_1=x_1^{\min},\ldots,x_n=x_n^{\min},\ x_{n+1}=x_1^{\max},\ldots,x_{2n}=x_n^{\max}.$$

На последующих шагах диалогового поиска плана эксперт-планировщик, сопоставляя значения левых и правых частей набора неравенств $\boldsymbol{Cx}\le\boldsymbol{d}$, оценивает реализуемость и эффективность полученных решений. Текущий план, получивший оценку эксперта, может быть определен как окончательный или могут быть изменены требования, направляющие перемещение решения. Каждый план (с оценкой) может быть сохранен в базе данных для сравнительного анализа и возможного возврата.

Если $\boldsymbol{x}^0$ — текущий план и эксперт определил ориентирующие требования для перемещения от $\boldsymbol{x}^0$ к целевому плану x в виде набора

$$\boldsymbol{C}^*\boldsymbol{x} = \boldsymbol{C}^*\boldsymbol{x}^0 + \boldsymbol{h}\,,$$

где $\boldsymbol{C}^*$ — подмножество из m строк матрицы $\boldsymbol{C}$, а $\boldsymbol{h} = (h_1, \ldots, h_m)$ — вектор ненулевых констант, направляющих перемещение решения, то кандидатом на улучшенное решение будет

$$\boldsymbol{x} = \left(x_1^0 + \frac{\sum_{i=1}^m q_i \Delta_i x_1}{\sum_{i=1}^m q_i} \;\cdots\; x_{2n}^0 + \frac{\sum_{i=1}^m q_i \Delta_i x_{2n}}{\sum_{i=1}^m q_i} \right),$$

где

$$\Delta_i x_j = \frac{c_{ij} h_i}{c_{i1}^2 + \cdots + c_{i,2n}^2},\;\; j = 1, \ldots, 2n\,,\;\; i = 1, \ldots, m\,.$$

Такое решение «ближе» к гиперплоскостям, определяемым требованиями с более высокими приоритетами. Детальное обоснование метода и изложение алгоритма приведено в [21]. Задача решается с помощью специальных программных средств [21].

4 Отличия предложенного подхода от традиционного

Методологические истоки традиционного подхода, основанного на методах линейного программирования, определены в работах [22–24].

Существование, реализуемость и прикладная точность решения. В традиционном подходе при несовместности системы ограничений решения не существует, а если оно отыскивается как чебышевская точка, то не гарантирована реализуемость. При совместности системы ограничений решение не всегда может быть реализовано, так как находится на границе области допустимых решений. Не для любой прикладной точности может быть найдено реализуемое решение. Предложенный подход позволяет найти реализуемое решение и в случае несовместности системы ограничений. Кроме того, реализуемость решения может быть увеличена перемещением от границы области допустимых решений. Реализуемое решение может быть найдено для любой прикладной точности.

Показатели эффективности решения. В традиционном подходе оптимизируется значение одной целевой функции. В предложенном подходе могут контролироваться одновременно несколько показателей эффективности.

5 О действующем интернет-сервисе «Планирование расходов»

Сравнение с известными программными продуктами для неинтервальных расчетов бюджета (PlanGuru, BizBudg Online, Budget Cruncher 3.10, Questica

Budget) подтвердило преимущества интернет-сервиса «Планирование расходов» [13].

Гибкость и эффективность. При осмотрительном задании границ отрезков для ресурса и запросов и следовании плану, рассчитанному сервисом, вероятность выхода за рамки бюджета радикально снижается.

Без сценариев «Если... то...». За счет задания ожидаемых доходов и расходов отрезками нет необходимости просчитывать различные сценарии в отдельных задачах. На любом этапе выполнения плана можно смоделировать планируемый расход (задать минимальный запрос равным максимальному, пометить его как «обязательный», выполнить команду «Распределить» и, получив результат, посмотреть, как изменились границы плана по остальным расходным статьям). Любую расходную статью можно временно исключить из рассмотрения, просто поставив «галочку» в соответствующей клетке таблицы.

Детализация и прикладная точность. Число уровней детализации не ограничено (таким способом может быть задана произвольная иерархическая система распределения). Для любой таблицы расходных статей предусмотрена возможность задания своей прикладной точности (не грубее точности детализируемой статьи) и признака «Использовать весовые коэффициенты» (приоритеты).

6 Заключение

1. Методология планирования расходов на основе ситуационно зависимых систем требований рассчитана на реализацию в виде онлайн-сервисов, работающих в среде цифровых двойников.
2. Основные преимущества предложенных интервальных методов решения линейных задач планирования расходов в сравнении с известными неинтервальными методами: регулируемая набором требований реализуемость и эффективность плана, учет точности прогнозирования распределяемой суммы и ожидаемых расходов, гарантированное получение реализуемого плана с максимально достижимой эффективностью и возможность задания требуемой прикладной точности. Важной особенностью метода целевого перемещения решения является возможность получения решения при несовместности системы ограничений.
3. Методологическая состоятельность предложенного подхода подтверждена сравнением результатов работы действующего интернет-сервиса планирования расходов и известных программных продуктов, предназначенных для бюджетирования.

Литература

1. *Lasi H., Fettke P., Kemper H. G., et al.* Application-pull and technology-push as driving forces for the Fourth Industrial Revolution // Bus. Inf. Syst. Eng., 2014. Vol. 6. No. 4. P. 239–242.

2. *Li Lan-bing, Liu Bing-lian, Liu Wei-lin, et al.* Efficiency evaluation of the regional high-tech industry in China: A new framework based on meta-frontier dynamic DEA analysis // Socio Econ. Plan. Sci., 2017. Vol. 60. P. 26–33.
3. *Skog D. A., Wimelius H., Sandberg J.* Digital disruption // Bus. Inf. Syst. Eng., 2018. Vol. 60. No. 5. P. 431–437.
4. *Куров Б. Н.* Сравнение эффективности алгоритмов управления с учетом точности данных и реализации решений // Управление большими системами, 2011. Вып. 34. С. 279–291.
5. *Ильин В. Д.* Основания ситуационной информатизации. — М.: Наука, Физматлит, 1996. 180 с.
6. *Ilyin A. V., Ilyin V. D.* Situational online resource planning in accordance with mandatory and orienting rules // Системы и средства информатики, 2018. Т. 28. № 1. С. 177–191. doi: 10.14357/08696527180113.
7. The Digital Twin. — General Electric, 2018. https://www.ge.com/digital/sites/default/files/The-Digital-Twin_Compressing-Time-to-Value-for-Digital-Industrial-Companies.pdf.
8. *Armbrust M., Fox A., Griffith R., et al.* A view of cloud computing // Commun. ACM, 2010. Vol. 53. No. 4. P. 50–58. doi: 10.1145/1721654.1721672.
9. *Wang L., Laszewski G., Younge A., et al.* Cloud computing: A perspective study // New Generat. Comput., 2010. No. 28. P. 137–146.
10. *Rogers O., Cliff D.* A financial brokerage model for cloud computing // J. Cloud Computing, 2012. Vol. 1. No. 1. P. 1–12.
11. *Jamsa K. A.* Cloud computing. — Burlington: Jones & Bartlett Learning, 2013. 322 p.
12. *Wei Y., Blake M. B.* Service-oriented computing and cloud computing: Challenges and opportunities // IEEE Internet Comput., 2010. No. 14. P. 72–75.
13. *Ильин А. В.* Интернет-сервис планирования расходов // Системы и средства информатики, 2015. Т. 25. № 2. С. 111–122. doi: 10.14357/08696527150207.
14. *Jede A., Teuteberg F.* Understanding socio-technical impacts arising from software-as-a-service usage in companies // Bus. Inf. Syst. Eng., 2016. Vol. 58. No. 3. P. 161–176.
15. *Perera C., Liu C. H., Jayawardena S.* The emerging Internet of Things marketplace from an industrial perspective: A survey // IEEE T. Emerging Topics Computing, 2015. Vol. 3. No. 4. P. 585–598.
16. «Интернет вещей» (IoT) в России. Технология будущего, доступная уже сейчас. — PWC, 2017. https://www.pwc.ru/ru/publications/IoT.html.
17. *Kim R. Y.* Efficient wireless communications schemes for machine to machine communications // Comm. Com. Inf. Sc., 2011. Vol. 181. No. 3. P. 313–323.
18. *Lien S. Y., Liau T. H., Kao C. Y., et al.* Cooperative access class barring for machine- to-machine communications // IEEE T. Wirel. Commun., 2012. Vol. 11. No. 1. P. 27–32.
19. *Ilyin A. V., Ilyin V. D.* Towards a normalized economic mechanism based on E-services // Agris on-line Papers Economics Informatics, 2014. Vol. 6. No. 3. P. 39–49. http://online.agris.cz/files/2014/agris_on-line_2014_3_ilyin_ilyin.pdf.
20. *Ильин В. Д.* Модель кооперативного решателя задач на основе цифровых двойников // Системы и средства информатики, 2019. Т. 29. № 2. С. 172–179. 10.14357/08696527190215.

21. *Ильин А. В.* Экспертное планирование ресурсов. — М.: ИПИ РАН, 2013. 58 с.
22. *Канторович Л. В.* Математические методы организации и планирования производства. — Л.: ЛГУ, 1939. 67 с.
23. *Dantzig G.* Programming of interdependent activities. Mathematical model // Econometrica, 1949. Vol. 17. No. 3/4. P. 200–211.
24. *Karmarkar N.* A new polynomial-time algorithm for linear programming // Combinatorica, 1984. No. 4. P. 373–395.

Поступила в редакцию 13.07.19

FORMATION OF SITUATIONALLY DEPENDENT SYSTEMS OF REQUIREMENTS FOR SOLVING THE PROBLEMS OF COST PLANNING

***A. V. Ilyin*[1] *and V. D. Ilyin*[2]**

[1]State Research Institute of Aviation Systems, 7 Viktorenko Str., Moscow 125319, Russian Federation
[2]Dorodnicyn Computing Center of the Russian Academy of Sciences, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, 40 Vavilov Str., Moscow 119333, Russian Federation

Abstract: An approach to the expert formation of situationally dependent systems of requirements for solving cost planning problems is proposed. The statement and methods for solving linear problem of situational cost planning are presented. Depending on the set of requirements, the problem is solved either by the method of prioritized interval allocation, or by the method of target displacement of solution. Both methods allow finding a cost plan that always meets the mandatory requirements and satisfy the orienting requirements as much as possible. At each step of the plan search in the computational experiment mode, the problem formulation is determined by the system of mandatory and orienting requirements, which is generated by the expert planner on the base of the situation portraits analysis. It is possible to set several indicators of the solution quality. The presentation of input data and the planning result in the form of numerical segments allows to take into account the accuracy of forecasting the amount of the allocated resource and the expected costs. The portraits of situations (target, starting, and achieved) formed with the help of digital twins are presented by a formalized description of the key parameters characterizing the state of the sources of the consumed resource, its consumers, and the planning conditions. The characteristic of the active online cost planning service is given.

Keywords: situationally dependent systems of requirements; situational cost planning; method of prioritized interval allocation; method of target displacement of solution; situation portraits; online cost planning service

DOI: 10.14357/08696527190315

References

1. Lasi, H., P. Fettke, H. G. Kemper, *et al.* 2014. Application-pull and technology-push as driving forces for the Fourth Industrial Revolution. *Bus. Inf. Syst. Eng.* 6(4):239–242. doi: 10.1007/s12599-014-0334-4.
2. Li, Lan-bing, Bing-lian Liu, Wei-lin Liu, *et al.* 2017. Efficiency evaluation of the regional high-tech industry in China: A new framework based on meta-frontier dynamic DEA analysis. *Socio Econ. Plan. Sci.* 60:26–33. doi: 10.1016/j.seps.2017.02.001.
3. Skog, D. A., H. Wimelius, and J. Sandberg. 2018. Digital disruption. *Bus. Inf. Syst. Eng.* 60(5):431–437. doi: 10.1007/s12599-018-0550-4.
4. Kurov, B. N. 2011. Sravnenie effektivnosti algoritmov upravleniya s uchetom tochnosti dannykh i realizatsii resheniy [Comparison of control algorithms efficiency taking into account data accuracy and solutions implementation]. *Upravlenie bol'shimi sistemami* [Management of Large Systems] 34:279–291.
5. Ilyin, V. D. 1996. *Osnovaniya situatsionnoy informatizatsii* [Foundations of situational informatization]. Moscow: Nauka, Fizmatlit. 180 p.
6. Ilyin, A. V., and V. D. Ilyin. 2018. Situational online resource planning in accordance with mandatory and orienting rules. *Sistemy i Sredstva Informatiki — Systems and Means of Informatics* 28(1):177–191. doi: 10.14357/08696527180113.
7. General Electric. 2018. The Digital Twin. Available at: https://www.ge.com/digital/sites/default/files/The-Digital-Twin_Compressing-Time-to-Value-for-Digital-Industrial-Companies.pdf (accessed July 14, 2019).
8. Armbrust, M., A. Fox, R. Griffith, *et al.* 2010. A view of cloud computing. *Commun. ACM* 53(4):50–58. doi: 10.1145/1721654.1721672.
9. Wang, L., G. Laszewski, A. Younge, *et al.* 2010. Cloud computing: A perspective study. *New Generat. Comput.* 28:137–146. doi: 10.1007/s00354-008-0081-5.
10. Rogers, O., and D. Cliff. 2012. A financial brokerage model for cloud computing. *J. Cloud Computing* 1(1):1–12. doi: 10.1186/2192-113X-1-2.
11. Jamsa, K. A. 2013. *Cloud computing*. Burlington: Jones & Bartlett Learning. 322 p.
12. Wei, Y., and M. B. Blake. 2010. Service-oriented computing and cloud computing: challenges and opportunities. *IEEE Internet Comput.* 14:72–75. doi: 10.1109/MIC.2010.147.
13. Ilyin, A. V. 2015. Internet-servis planirovaniya raskhodov [The online service for cost planning]. *Sistemy i Sredstva Informatiki — Systems and Means of Informatics* 25(2):111–122. doi: 10.14357/08696527150207.
14. Jede, A., and F. Teuteberg. 2016. Understanding socio-technical impacts arising from software-as-a-service usage in companies. *Bus. Inf. Syst. Eng.* 58(3):161-176. doi: 10.1007/s12599-016-0429-1.
15. Perera, C., C. H. Liu, and S. Jayawardena. 2015. The emerging Internet of Things marketplace from an industrial perspective: A survey. *IEEE T. Emerging Topics Computing* 3(4):585–598. doi: 10.1109/TETC.2015.2390034.
16. *"Internet veshchey" (IoT) v Rossii. Tekhnologiya budushchego, dostupnaya uzhe seychas* [Internet of things (IoT) in Russia. Technology of the future, available now]. 2017. PWC. Available at: https://www.pwc.ru/ru/publications/IoT.html (accessed July 14, 2019).

17. Kim, R. Y. 2011. Efficient wireless communications schemes for machine to machine communications. *Comm. Com. Inf. Sc.* 181(3):313–323. doi: 10.1007/978-3-642-22203-0_28.
18. Lien, S. Y., T. H. Liau, C. Y. Kao, *et al.* 2012. Cooperative access class barring for machine-to-machine communications. *IEEE T. Wirel. Commun.* 11(1):27–32. doi: 10.1109/TWC.2011.111611.110350.
19. Ilyin, A. V., and V. D. Ilyin. 2014. Towards a normalized economic mechanism based on E-services. *Agris on-line Papers Economics Informatics* 6(3):39–49. Available at: http://online.agris.cz/files/2014/agris_on-line_2014_3_ilyin_ilyin.pdf (accessed July 14, 2019).
20. Ilyin, V. D. 2019. Model' kooperativnogo reshatelya zadach na osnove tsifrovykh dvoynikov [The model of the cooperative problem solver based on digital twins]. *Sistemy i Sredstva Informatiki — Systems and Means of Informatics* 29(2):172–179. doi: 10.14357/08696527190215.
21. Ilyin, A. V. 2013. *Ekspertnoe planirovanie resursov* [Expert resource planning]. Moscow: IPI RAN. 58 p.
22. Kantorovich, L. V. 1939. *Matematicheskie metody organizatsii i planirovaniya proizvodstva* [Mathematical methods of organizing and planning production]. Leningrad: Leningrad State University. 67 p.
23. Dantzig, G. 1949. Programming of interdependent activities. Mathematical model. *Econometrica* 17(3/4):200–211.
24. Karmarkar, N. 1984. A new polynomial-time algorithm for linear programming. *Combinatorica* 4:373–395.

Received July 13, 2019

Contributors

Ilyin Alexander V. (b. 1975) — Candidate of Science (PhD) in technology, leading scientist, State Research Institute of Aviation Systems, 7 Viktorenko Str., Moscow 125319, Russian Federation; ilyin@res-plan.com

Ilyin Vladimir D. (b. 1937) — Doctor of Science in technology, professor, leading scientist, Dorodnicyn Computing Center of the Russian Academy of Sciences, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, 40 Vavilov Str., Moscow 119333, Russian Federation; vdilyin@yandex.ru

СИСТЕМЫ И СРЕДСТВА ИНФОРМАТИКИ, 2019. Т. 29. № 3. С. 180–193

СПОСОБ ВКРАПЛЕНИЯ ДАННЫХ НА ОСНОВЕ ОДНОЙ СХЕМЫ РАЗДЕЛЕНИЯ СЕКРЕТА

Ю. В. Косолапов[1]

Аннотация: Важными характеристиками стегосистем являются относительная длина α вкладываемого сообщения и относительная эффективность вложения e. Дополнительными существенными характеристиками таких систем представляются степень свободы при выборе модифицируемых битов контейнера и возможность противостоять утрате части блоков стегоконтейнера. Настоящая работа посвящена разработке стегосистемы, которая, с одной стороны, позволяет восстанавливать частично утраченные данные, а с другой стороны, дает возможность выбирать позиции модифицируемых битов. На базе этой системы строятся и исследуются стегоконструкции, для которых вычисляются характеристики α и e, а также оценивается степень свободы при выборе модифицируемых битов и максимальное число стираний блоков, не приводящих к искажениям вкрапленных данных.

Keywords: сокрытие информации; схема разделения секрета

DOI: 10.14357/08696527190316

1 Введение и постановка задачи

Рассмотрим характерную для стеганографии задачу вкрапления сообщения $\mathbf{m} \in \{0,1\}^k$ в файл или цифровой сигнал Σ над конечным алфавитом $\mathcal{X}$ ($\Sigma \in \in \mathcal{X}^*$). Перед выполнением вкрапления по сигналу Σ формируется двоичная последовательность $\mathbf{x} \in \{0,1\}^N$, $N \geq k$, называемая контейнером, в которую вкладывается $\mathbf{m}$. Здесь и далее полагается, что $\mathbf{x}$ формируется с помощью применения некоторой общеизвестной функции $P : \mathcal{X} \to \{0,1\}$ к символам сигнала. Для $\mathcal{X} = \{0,\ldots,255\}$ примером такой функции может быть функция, возвращающая наименее значащий бит байта. Результатом вкрапления является двоичная последовательность (стегоконтейнер) $\mathbf{x}^{(m)}$ длины N. Для удобства сигнал Σ будем называть носителем контейнера, а символом $\Sigma^{(m)}$ обозначим носитель $\mathbf{x}^{(m)}$. Алгоритм вкрапления должен позволять по $\mathbf{x}^{(m)}$ однозначно восстановить $\mathbf{m}$. Будем говорить, что отображения $E : \{0,1\}^k \times \{0,1\}^N \to \to \{0,1\}^N$ и $f : \{0,1\}^N \to \{0,1\}^k$ задают стегосистему, если для любых $\mathbf{m} \in \{0,1\}^k$ и $\mathbf{x} \in \{0,1\}^N$ выполняется условие: $f(E(\mathbf{m},\mathbf{x})) = \mathbf{m}$. Отношение $\alpha = k/N$ называется относительной длиной вкладываемого сообщения, а отношение $e = k/\mathbf{E}[\mathrm{wt}(\mathbf{x} \oplus \mathbf{x}^{(m)})]$ — относительной эффективностью вкрапления, где

[1]Институт математики, механики и компьютерных наук им. И. И. Воровича, Южный федеральный университет, Ростов-на-Дону, itaim@mail.ru

$\mathrm{wt}(\mathbf{a})$ — вес Хэмминга вектора $\mathbf{a}$, а $\mathbf{E}[\mathrm{wt}(\mathbf{x} \oplus \mathbf{x}^{(m)})]$ — среднее число вносимых изменений при вложении. Известно, что $e \leq \alpha / h^{-1}(\alpha)$, где $h(p)$ — функция двоичной энтропии [1].

Один из глубоко изученных методов вложения данных — метод матричного вкрапления, эффективность которого может быть описана в терминах покрывающих кодов. Напомним, что линейный код длины n размерности k с кодовым расстоянием d называется $[n, k, d]$-кодом. Согласно [2], пара отображений (E, f) может быть построена на основе линейного бинарного $[N, N-k, d]$-кода C с радиусом покрытия R, где

$$R = \max \left\{ \min_{\mathbf{c} \in C} \{\mathrm{wt}(\mathbf{c} \oplus \mathbf{x})\} : \mathbf{x} \in \{0,1\}^N \right\}. \tag{1}$$

В этом случае $\alpha = k/N$, $E(\mathbf{m}, \mathbf{x}) = \mathbf{e} \oplus \mathbf{x} = \mathbf{x}^{(m)}$, где $\mathbf{e}$ — вектор минимального веса в смежном классе с синдромом $\mathbf{m} \oplus \mathbf{x}H^\top$, а $H^\top$ — транспонированная проверочная матрица кода C. Правило извлечения сообщения имеет вид:

$$f(\mathbf{x}^{(m)}) = \mathbf{x}^{(m)} H^\top .$$

Из (1) следует, что $\mathrm{wt}(\mathbf{e}) \leq R$ и $e \geq k/R$. Например, для $p \in \mathbb{N}$ двоичный $[2^p - 1, 2^p - 1 - p, 3]$-код Хэмминга имеет радиус покрытия $R = 1$, поэтому с помощью него можно выполнить вкрапление p битов сообщения в контейнер длины $2^p - 1$, изменив при этом не более одного бита.

В работе [3] построена, а в [4] обобщена конструкция отображений E и f (названная впоследствии ZZW-конструкцией), для которых относительная эффективность близка к теоретической границе. В этой конструкции используется представление $\mathbf{x}$ в виде конкатенации блоков меньшей длины, в каждый из которых вносятся необходимые изменения.

Отметим два ограничения конструкции ZZW. *Первое* ограничение связано с тем, что декодирующее отображение f для ZZW-конструкции не позволяет из стегоконтейнера с частично стертыми блоками извлечь $\mathbf{m}$. В то же время, например, появление стираний характерно для случая, когда блоки передаются по ненадежным каналам связи, а также в случае использования пакетной стеганографии [5] (когда теряется часть из набора стегоконтейнеров, в которые внедрено сообщение). *Второе* ограничение связано с отсутствием свободы при внесении изменений в контейнер, а именно: с каждым битом контейнера $\mathbf{x}$ часто связывают величину $0 \leq \rho_i \leq \infty$, обозначающую стоимость изменения i-го бита. При $\rho_i = \infty$ изменение бита x_i запрещено. На практике равенство $\rho_i = \infty$ означает, что изменение i-го бита в пустом контейнере может привести к обнаружению факта вложения, когда наблюдатель (противник) располагает носителем стегоконтейнера. Часто рассматривается модель «влажная» бумага [6], когда $\rho_i \in \{1, \infty\}$. Контейнер состоит из «влажных» битов ($\rho_i = \infty$) и «сухих» битов ($\rho_i = 1$). Для такой модели используются специальные коды (см., например, [7]), позволяющие

избежать модификации «влажных». Говорят, что отправитель обладает свободой при выборе модифицируемых битов [6]. Несмотря на то что в конструкции ZZW используются коды для «влажной» бумаги, сама ZZW-конструкция не позволяет выбирать позиции изменяемых битов контейнера (так как коды для «влажной» бумаги используются для увеличения относительной эффективности вкрапления).

Поэтому актуальна разработка таких методов вкрапления, при которых, во-первых, отправитель имеет возможность не менять биты, стоимость изменения которых равна ∞, во-вторых, которые были бы устойчивы к стиранию части блоков, при этом число вносимых изменений было бы небольшим. В настоящей работе предлагается метод вкрапления данных, который, с одной стороны, устойчив к стиранию одного блока в наборе блоков, а с другой стороны, некоторые модификации этого метода дополнительно обеспечивают свободу при выборе мест модифицируемых битов. Предлагаемый в работе метод основан на схеме разделения секрета (СРС), построенной в [8].

2 Схема разделения секрета

Множество $\{1,\dots,n\}$ обозначим символом $[n]$. Пусть H — $(k\times n)$-матрица над полем $\mathbb{F}_2$ ранга k, $k<n$, $\tau=\{t_1,\dots,t_\mu\}(\subseteq[n])$ — упорядоченное по возрастанию подмножество номеров координат, $\pi_\tau:\{0,1\}^n\to\{0,1\}^\mu$ — оператор проекции, ставящий в соответствие вектору $\mathbf{c}=(c_1,\dots,c_n)$ вектор $\pi_\tau(\mathbf{c})=$ $=(c_{t_1},\dots,c_{t_\mu})$. Символом $\hat\pi_\tau(H)$ обозначим множество столбцов матрицы H с номерами из множества τ, а линейную оболочку, натянутую на эти столбцы обозначим $\mathcal{L}(\hat\pi_\tau(H))$. Пусть $\sigma_1,\dots,\sigma_S$ — такой набор подмножеств, что для любого $A\subset[S]$ мощности s выполняется равенство

$$\dim\bigcap_{i\in A}\mathcal{L}(\hat\pi_{\tau_i}(H)=0,\ \tau_i=[n]\setminus\sigma_i, \tag{2}$$

а для всх подмножеств A меньшей мощности это равенство не выполняется. Для вектора $\mathbf{m}\in\{0,1\}^k$ рассмотрим S векторов $\mathbf{c}_1,\dots,\mathbf{c}_S$, необязательно различных, полученных случайным выбором из множества решений $\Gamma(\mathbf{m})$ системы уравнений $H\mathbf{y}^\top=\mathbf{m}^\top$ относительно вектора неизвестных $\mathbf{y}=(y_1,\dots,y_n)$. Условие (2) позволяет построить (S,s)-схему разделения секрета $\mathbf{m}$, в которой секрет разделяется на S долей, а однозначное восстановление секрета возможно при наличии любых s долей секрета. Дóлей секрета i-го участника служит вектор $\pi_{\sigma_i}(\mathbf{c}_i)$. Для восстановления секрета по s долям с номерами из A, $|A|=s$, достаточно решить систему

$$\hat\pi_{\sigma_i}(H)\pi_{\sigma_i}(\mathbf{c}_i)^\top\oplus\hat\pi_{\tau_i}(H)\pi_{\tau_i}(\mathbf{c}_i)^\top=\mathbf{m}^\top \tag{3}$$

для $i\in A$ относительно векторов неизвестных $\pi_{\tau_i}(\mathbf{c}_i)$. Условие (2) гарантирует, что любое найденное решение однозначно определяет секрет $\mathbf{m}$ (см. [8]).

Описанную схему разделения секрета на основе $(k \times n)$-матрицы H назовем $(H, s, \sigma_1, \ldots, \sigma_S)$-схемой.

Рассмотрим пример. Пусть $\mathrm{supp}(\mathbf{a}) = \{i : a_i \neq 0\}$ — носитель вектора $\mathbf{a} = (a_1, \ldots, a_l)$. Рассмотрим такое семейство кодов $\mathcal{K} = \{K(s)\}_{s \in \mathbb{N}}$, $s \geq 2$, что проверочная матрица кода $K(s)$ может быть получена удалением последней строки из матрицы вида

$$H_s = \left(\begin{array}{c|c} H_{s-1} & I_s \\ \hline \mathbf{0}_{s-1} & \mathbf{1}_s \end{array}\right),$$

где $\mathbf{0}_{s-1}$ — нулевой вектор длины $s-1$; $\mathbf{1}_s$ — вектор из единиц длины s; $H_1 = (1\ 1)^\top$. Заметим, что матрица H_s ранга s имеет $s+1$ строку; матрицу, полученную из H_s удалением последней строки, обозначим H_s^0. Пусть σ_i — носитель i-й строки матрицы H_s. В [8] показано, что для любого подмножества $A \subset [s+1]$ мощности s и матрицы $H = H_s^0$ выполняется равенство (2). При этом для любого подмножества A мощности $s-1$ и менее это равенство не выполняется. Отметим, что $|\sigma_i| = s$ для всех $i \in [s+1]$. Таким образом, код $K(s)$ из $\mathcal{K}$ задает $(H_s^0, s, \sigma_1, \ldots, \sigma_{s+1})$-схему разделения секрета.

3 Метод вкрапления на основе схемы разделения секрета

3.1 Правила вложения и извлечения сообщения

Построим стегосистему на основе $(H, s, \sigma_1, \ldots, \sigma_S)$-схемы разделения секрета для случая, когда $\rho_i \in \{1, \infty\}$. Пусть $\mathbf{m}(\in \{0,1\}^k)$ — вкладываемое сообщение, а длина N контейнера $\mathbf{x}$ такая, что $N = \sum_{i=1}^{S} |\sigma_i|$. Представим $\mathbf{x}$ в виде конкатенации S блоков:

$$\mathbf{x} = \mathbf{x}_1 \parallel \cdots \parallel \mathbf{x}_S,\ \mathbf{x}_i \in \{0,1\}^{|\sigma_i|}. \tag{4}$$

Пусть $H = [P\ I_k]$, где P — $(k \times n-k)$-матрица; I_k — единичная $(k \times k)$-матрица. Тогда множество решений $\Gamma(\mathbf{m})$ уравнения $H\mathbf{y}^\top = \mathbf{m}^\top$ может быть представлено в виде:

$$\Gamma(\mathbf{m}) = \left\{(\mathbf{m} \parallel \mathbf{v})\widetilde{G} : \mathbf{v} \in \{0,1\}^{n-k}\right\},\ \widetilde{G} = \begin{bmatrix} O & I_k \\ I_{n-k} & P^\top \end{bmatrix},$$

где O — нулевая матрица. Матрицу $[I_{n-k}\ P^\top]$ обозначим через G, а матрицу $[O\ I_k]$ — через G^*. Для каждого σ_i и вектора $\mathbf{x}_i \in \{0,1\}^{|\sigma_i|}$ из (4) рассмотрим систему $(\mathbf{m} \parallel \mathbf{v}_i)\hat{\pi}_{\sigma_i}(\widetilde{G}) = \mathbf{x}_i$ относительно неизвестного $\mathbf{v}_i$, которая имеет решение, если имеет решение система $\mathbf{v}_i\hat{\pi}_{\sigma_i}(G) = \mathbf{x}_i \oplus \mathbf{m}\hat{\pi}_{\sigma_i}(G^*)$. Пусть $r_i = \mathrm{rank}(\hat{\pi}_{\sigma_i}(G))$, $\mathcal{G}(\hat{\pi}_{\sigma_i}(G))$ — $(r_i \times |\sigma_i|)$-матрица полного ранга, полученная

из матрицы $\hat{\pi}_{\sigma_i}(G)$ исключением линейно зависимых строк. Рассмотрим систему уравнений относительно вектора неизвестных $\tilde{\mathbf{v}}_i \in \{0,1\}^{r_i}$:

$$\tilde{\mathbf{v}}_i \mathcal{G}(\hat{\pi}_{\sigma_i}(G)) = \mathbf{x}_i \oplus \mathbf{m}\hat{\pi}_{\sigma_i}(G^*) =: \mathbf{z}_i. \tag{5}$$

В случае когда $\mathbf{x}_i$ — реализация случайного равновероятного вектора, система (5) имеет решение с вероятностью $2^{r_i-|\sigma_i|}$. Если система (5) не имеет решения, то, так как $\rho_i \in \{1,\infty\}$, необходимо найти такой вектор $\mathbf{e}_i = (e_{i,1},\ldots,e_{i,|\sigma_i|})$, чтобы система

$$\tilde{\mathbf{v}}_i \mathcal{G}(\hat{\pi}_{\sigma_i}(G)) = \mathbf{z}_i \oplus \mathbf{e}_i \tag{6}$$

имела решение, при этом носитель вектора $\mathbf{e}_i$ не должен содержать номера «влажных» координат, а вес вектора $\mathbf{e}_i$ должен быть минимальным среди векторов, обладающих этими свойствами. Правило вложения имеет вид:

$$E(\mathbf{m},\mathbf{x}) = \mathbf{x} \oplus \mathbf{e} = \mathbf{x}^{(m)}, \;\; \mathbf{e} = \mathbf{e}_1 \parallel \cdots \parallel \mathbf{e}_S.$$

Декодирующее отображение f — это решение системы уравнений (3), где $\pi_{\sigma_i}(\mathbf{c}_i) = \mathbf{x}_i^{(m)}$, $i \in A(\subset [S], |A| = s)$. Общее число изменений в контейнере равно

$$C = \sum_{j=1}^{S} \sum_{u=1}^{|\sigma_i|} e_{j,u}.$$

Если $\mathbf{E}[C]$ — среднее число изменяемых битов, то $\alpha = k/N$, $e = k/(\mathbf{E}[C])$.

Лемма 3.1. *Стегосистема (E,f) на основе $(H,s,\sigma_1,\ldots,\sigma_S)$-схемы разделения секрета позволяет восстанавливать сообщение при утрате не более $B = S - s$ блоков.*

Доказательство. Доказательство вытекает из того, что в $(H,s,\sigma_1,\ldots,\sigma_S)$-схеме для любого A, $|A| = s$, выполняется условие (2), достаточное для восстановления секрета. □

Лемма 3.2. *Пусть G — $(n-k \times n)$-матрица ранга $n-k$, $\sigma_i \subseteq [n]$, $\operatorname{rank}(\hat{\pi}_{\sigma_i}(G)) = |\sigma_i| - 1$, $\mathcal{K}_i(\subset \{0,1\}^{|\sigma_i|})$ — $[|\sigma_i|, |\sigma_i|-1, 2]$-код с порождающей матрицей $\mathcal{G}(\hat{\pi}_{\sigma_i}(G))$. Тогда*

(1) *если вектор $\mathbf{x}_j$ — реализация случайного и равновероятного вектора, то система (5) имеет решение с вероятностью 0,5;*

(2) *если (5) не имеет решения, то для любого $\mathbf{e}_j$ веса 1 система (6) имеет решение.*

Доказательство. Первое утверждение следует из того, что векторов $\mathbf{z}_i$ в пространстве $\mathbb{F}_2^{|\sigma_i|}$, для которых система (5) имеет решение, в точности $2^{|\sigma_i|-1}$. При этом, по условию, среди этих векторов нет ненулевых векторов веса 1.

Поэтому если система (5) не имеет решение для $\mathbf{z}_i$, то векторы $\mathbf{z}_i$ и $\mathbf{e}_i$, где $\mathrm{wt}(\mathbf{e}_i) = 1$, принадлежат одному смежному классу кода $\mathcal{K}_i$. По определению смежного класса получаем: $\mathbf{z}_i - \mathbf{e}_i \in \mathcal{K}_i$. □

Следствие 3.1 Если $\mathbf{x}$ вида (4) является реализацией случайного равновероятного вектора и для каждого $i \in [S]$ выполняются условия из леммы 3.2, то для стегосистемы на основе $(H, s, \sigma_1, \ldots, \sigma_S)$-схемы имеем: $e = 2k/S$, $\alpha = k/N$.

3.2 Предварительная рандомизация контейнера

На практике условие случайности (см. лемму 3.1) реальных контейнеров может не выполняться [9], поэтому в среднем может потребоваться модификация более $S/2$ блоков контейнера. Кроме того, возможны пакеты «влажных» битов. Поэтому перед вкраплением данных предлагается выполнять предварительную рандомизацию контейнера: шифровать контейнер шифром типа гаммирования и перемешивать биты. Пусть $\psi_s : \{0,1\}^K \to \{0,1\}^N$ — псевдослучайный генератор, который для ключа $\mathbf{k} \in \{0,1\}^K$ генерирует псевдослучайный вектор длины N; P — перестановочная $(N \times N)$-матрица; (E, f) — стегосистема на основе $(H, s, \sigma_1, \ldots, \sigma_S)$-схемы, $\mathbf{m}(\in \{0,1\}^k)$ — вкладываемое сообщение, $\mathbf{x}$ — контейнер. Построим стегосистему (E', f'), в которой перед вкраплением выполняется предварительная рандомизация контейнера. Пусть $\mathbf{C} = \psi_s(\mathbf{k}) \oplus \mathbf{x}P$, $\mathbf{C}^{(\mathbf{m})} = E(\mathbf{m}, \mathbf{C})$. Тогда

$$\mathbf{x}^{(m)} = E'(\mathbf{m}, \mathbf{x}) = \mathbf{x} \oplus (\mathbf{C}^{(\mathbf{m})} \oplus \mathbf{C})P^{-1}.$$

Так как $\mathbf{x}^{(m)}P \oplus \psi_s(\mathbf{k}) = \mathbf{C}^{(m)}$, то

$$f'(\mathbf{x}^{(m)}) = f(\mathbf{x}^{(m)}P \oplus \psi_s(\mathbf{k})) = \mathbf{m}.$$

В следующем разделе для удобства предполагается, что контейнер является реализацией случайного и равновероятного вектора.

4 Метод вкрапления на основе $(H_s^0, s, \sigma_1, \ldots, \sigma_{s+1})$-схемы

4.1 Базовые конструкции

Далее исследуем стегосистему на основе $(H_s^0, s, \sigma_1, \ldots, \sigma_{s+1})$-схемы. Пусть $G_{s,\mathrm{pt}} = (I_{s-1}\mathbf{1}_{s-1}^\top)$. Отметим, что $G_{s,\mathrm{pt}}$ — порождающая матрица $[s, s-1, 2]$-кода проверки четности.

Лемма 4.1. *Пусть $s \in \mathbb{N}$, $G_s^0 = [I_{n-s}\ H_{s-1}^\top]$, $n = s(s+1)/2$. Для всех i код, порождаемый строками матрицы $\hat{\pi}_{\sigma_i}(G_s^0)$, служит $[s, s-1, 2]$-кодом проверки четности, где σ_i — носитель i-й строки матрицы H_s.*

Доказательство. Выполняется непосредственной проверкой. □

Теорема 4.1. *Стегосистема на основе* $(H_s^0, s, \sigma_1, \ldots, \sigma_{s+1})$*-схемы позволяет восстановить сообщения при стирании не более* $B = 1$ *блоков, при этом* $\alpha = 1/(s+1)$, $e = 2s/(s+1)$.

Доказательство. Из леммы 4.1 следует, что для каждого i выполняются условия леммы 3.2. Так как $k = s$ и $N = s(s+1)$, то равенства $\alpha = 1/(s+1)$ и $e = 2s/(s+1)$ вытекают из следствия 3.1. Оценка числа стертых блоков следует из леммы 3.1. □

Если нет необходимости в защите от потери части блоков, то вместо $(H_s^0, s, \sigma_1, \ldots, \sigma_{s+1})$-схемы может использоваться, например, $(H_s^0, s, \sigma_1, \ldots, \sigma_s)$-схема, так как по любым s долям из $s+1$ штук может быть восстановлен секрет. Отсюда вытекает

Следствие 4.1. Пусть $\{i_1, \ldots, i_s\} \subset [s+1]$. Для стегосистемы на основе $(H_s^0, s, \sigma_{i_1}, \ldots, \sigma_{i_s})$-схемы справедливы равенства: $\alpha = 1/s$; $e = 2$; $B = 0$.

Стегосистему на основе $(H_s^0, s, \sigma_1, \ldots, \sigma_S)$-схемы назовем *базовой конструкцией*, $S \in \{s, s+1\}$. Заметим, что базовая конструкция на основе $(H_s^0, s, \sigma_1, \ldots, \sigma_s)$-схемы эквивалентна по характеристикам α и e простой стегосистеме, в которой четность веса $\mathrm{wt}(\mathbf{x}_i^{(m)})$ подблока $\mathbf{x}_i^{(m)} (\in \{0,1\}^s)$ стеганоконтейнера длины s^2 определяет значение внедренного в этот блок бита m_i сообщения $\mathbf{m} \in \{0,1\}^s$: при $m_i = 1$ число $\mathrm{wt}(\mathbf{x}_i^{(m)})$ нечетное, а при $m_i = 0$ — четное. Если $\mathbf{x}$ — реализация случайного и равновероятного вектора длины s^2, то при внедрении $\mathbf{m}$ длины s потребуется внести одиночные изменения в среднем в половину блоков $\mathbf{x}$ (при этом изменения в рамках подблоков могут производиться в любом месте). Если $\mathbf{m}$ — кодовое слово $[s+1, s, 2]$-кода проверки четности, то стегосистема, в которой значение бита m_i определяется четностью веса $\mathrm{wt}(\mathbf{x}_i^{(m)})(\mathbf{x}_i^{(m)} \in \{0,1\}^s)$, эквивалентна по характеристикам α и e и числу исправляемых стертых блоков базовой конструкции на основе $(H_s^0, s, \sigma_1, \ldots, \sigma_{s+1})$-схемы. Однако использование СРС в основе базовых конструкций позволяет легко получить некоторые новые конструкции с лучшими характеристиками.

Пусть $S \in \{s, s+1\}$. В базовой конструкции на основе $(H_s^0, s, \sigma_1, \ldots, \sigma_S)$-схемы результатом вложения $\mathbf{m} \in \{0,1\}^s$ в $\mathbf{x}$ длины $s \cdot S$ является $\mathbf{x}^{(m)} = \mathbf{x} \oplus \mathbf{e}$. Причем среди подблоков вектора $\mathbf{e} = \mathbf{e}_1 \parallel \cdots \parallel \mathbf{e}_S$, как следует из леммы 3.2, в среднем половина подблоков веса один, а остальные веса ноль. Пусть $g = S/2$. Так как в базовых конструкциях в каждом из g модифицируемых блоков для изменения может быть выбран любой из битов (также следует из леммы 3.2), то будем говорить, что такая конструкция обладает степенью свободы порядка g; степень свободы обозначим буквой F.

4.2 Производные конструкции

В рассматриваемом случае матрица G в выражении (6) имеет вид G_s^0 (см. лемму 4.1) и для каждого $i = 1, \ldots, S$ матрица $\hat{\pi}_{\sigma_i}(G)$ имеет ранг $s-1$, так как $\mathcal{G}(\hat{\pi}_{\sigma_i}(G)) = G_{s,\mathrm{pt}}$. Пусть каждый из векторов набора

$$\tilde{\mathbf{v}}_j (\in \{0,1\}^{s-1}), \;\; i = 1, \ldots, S, \tag{7}$$

является решением системы вида (6). Производные конструкции позволяют увеличить относительную эффективность базовых конструкций, при этом в первых трех конструкциях дополнительная информация $\mathbf{m}^0$ вкладывается в вектор $\tilde{\mathbf{v}} = \tilde{\mathbf{v}}_1 \oplus \cdots \oplus \tilde{\mathbf{v}}_S$, $S \in \{s, s+1\}$.

4.2.1 Конструкция 1

Рассмотрим базовую конструкцию на основе $(H_s^0, s, \sigma_1, \ldots, \sigma_{s+1})$-схемы. Отправитель может дополнительно к s битам вложить в $s(s+1)$ бит еще один бит, контролируя четность числа битов в векторе $\tilde{\mathbf{v}}$ (полагая, например, что четное число битов соответствует нулевому биту, а нечетное — единичному). Если необходимо вложить, например, единичный бит, а в векторе $\tilde{\mathbf{v}}$ четное число единиц, то изменить число единиц можно модификацией одного бита в любом из векторов набора (7). Действительно, так как $G_{s,\mathrm{pt}}$ — порождающая матрица кода проверки четности, то модификация одного бита в векторе $\tilde{\mathbf{v}}_i$ приведет к изменению двух битов в соответствующем векторе $\mathbf{x}^{(m)}$. Поэтому модифицировать следует тот вектор $\tilde{\mathbf{v}}_i$, для которого $\mathrm{wt}(\mathbf{e}_i) = 1$. При этом в векторе $\tilde{\mathbf{v}}_i = (v_{i,1}, \ldots, v_{i,s-1})$ следует модифицировать (инвертировать) бит с номером w, для которого $\mathrm{supp}(1_w G_{s,\mathrm{pt}}) \cap \mathrm{supp}(\mathbf{e}_i) \neq \varnothing$, где 1_w — вектор веса один с единицей в координате с номером w. В этом случае вектор $\mathbf{e}_i$ модифицируется в вектор $\tilde{\mathbf{e}}_i = \mathbf{e}_i \oplus 1_w G_{s,\mathrm{pt}}$, причем $\mathrm{wt}(\tilde{\mathbf{e}}_i) = \mathrm{wt}(\mathbf{e}_i) = 1$. Таким образом, $\alpha = (s+1)/(s(s+1)) = 1/s$ и $e = 2$. Отметим, что в этом случае свобода выбора позиций для модификации по сравнению с базовой конструкцией уменьшается: в одном из блоков должно быть как минимум два «сухих» бита, позиции которых фиксированы. В остальных $g-1$ блоках должно быть не менее одного «сухого» бита в произвольном месте. Поэтому такая конструкция обладает степенью свободы $g-1$. Отметим, что $B = 0$, так как стирание одного из блоков не позволяет восстановить добавленный бит. В связи с этим целесообразно для этой конструкции использовать $(H_s^0, s, \sigma_1, \ldots, \sigma_s)$-схему. Тогда

$$\alpha = \frac{s+1}{s^2}; \quad e = \frac{2(s+1)}{s}.$$

Далее под конструкцией 1 понимается стегосистема на основе $(H_s^0, s, \sigma_1, \ldots, \sigma_s)$-схемы.

4.2.2 Конструкция 2

Пусть $s = 2^p$, $p \geq 3$. Тогда в вектор $\tilde{\mathbf{v}}$ при $S = s + 1$ можно вложить p бит информации ($\mathbf{m}^0 \in \{0,1\}^p$), используя метод матричного вложение из [10]. Пусть $\mathcal{H}$ — проверочная матрица $[2^p - 1, 2^p - 1 - p, 3]$-кода Хэмминга. Найдем такой вектор $\boldsymbol{\epsilon} \in \{0,1\}^{s-1}$ веса 1, что

$$\mathcal{H}\boldsymbol{\epsilon}^\top \oplus \mathcal{H}\tilde{\mathbf{v}}^\top = \mathcal{H}(\boldsymbol{\epsilon} \oplus \tilde{\mathbf{v}})^\top = \mathbf{m}^0 .$$

Как и в конструкции 1, необходимо найти подходящий вектор $\mathbf{e}_i$, соответствующий одному из векторов набора (7). В этом случае $\alpha = (p + 2^p)/(2^p(2^p + 1))$, число изменяемых битов равно $(s + 1)/2$, $e = (2(p + 2^p))/(2^p + 1)$. Степень свободы F такая же, как и в конструкции 1, и $B = 0$. Так как в этой конструкции нет возможности бороться со стираниями блоков, то характеристики e и α этой конструкции могут быть улучшены при использовании $(H_s^0, s, \sigma_1, \ldots, \sigma_s)$-схемы:

$$g = \frac{s}{2}; \quad \alpha = \frac{p + 2^p}{2^{2p}}; \quad e = \frac{2(p + 2^p)}{2^p} .$$

Стегосистему на основе $(H_s^0, s, \sigma_1, \ldots, \sigma_s)$-схемы будем назвать конструкцией 2.

4.2.3 Конструкция 3

Следующая конструкция основана на $(H_s^0, s, \sigma_1, \ldots, \sigma_s)$-схеме. Пусть s — нечетное число, $\mathbf{m}^0 = (m_1^0, \ldots, m_{s-1}^0) \in \{0,1\}^{s-1}$. Каждой паре (m_b^0, m_{b+1}^0), $b \in \{1; 3; \ldots; s-2\}$, поставим в соответствие множество $E_{b,b+1} = \{b, b+1, s\}$, которое представляет собой объединение носителей строк матрицы $G_{s,\mathrm{pt}}$ с номерами b и $b+1$. В силу нечетности s среднее число g модифицируемых блоков с равной вероятностью может быть равно $(s-1)/2$ или $(s+1)/2$. Рассмотрим сначала случай $g = (s-1)/2$. Номера соответствующих подвекторов набора (4) обозначим $i_1, \ldots, i_g$. Вектор $\tilde{\mathbf{v}}_{i_b}$ решения системы (6) при $\mathcal{G}(\hat{\pi}_{\sigma_{i_b}}(G)) = G_{s,\mathrm{pt}}$ будем искать путем модификации в векторе $\mathbf{x}_{i_b}$ любой одной координаты с номером из множества $E_{b,b+1}$; для ясности изложения будем полагать, что меняется бит с номером b.

Пусть $\Delta = (\delta_1, \ldots, \delta_{s-1}) = \mathbf{m}^0 \oplus \tilde{\mathbf{v}}$ — вектор, который необходимо построить для вложения $\mathbf{m}^0$ в $\tilde{\mathbf{v}}$. Способ построения Δ поясним на примере построения первой пары битов (δ_1, δ_2). (Заметим, что таких пар, по предположению, $g = (s-1)/2$.) Этой паре поставим в соответствие вектор $\tilde{\mathbf{v}}_{i_1}$, который был найден после изменения бита с номером $1 (\in E_{1,2})$ в векторе $\mathbf{x}_{i_1}$. Если $(\delta_1, \delta_2) = (0,0)$, то модификация вектора $\tilde{\mathbf{v}}_{i_1}$ не производится. В случае $(\delta_1, \delta_2) = (1,0)$ в векторе $\mathbf{x}_{i_1}$ инвертируются биты с номерами 1 и s, что соответствует прибавлению к вектору $\mathbf{x}_{i_1}$ первой строки матрицы $G_{s,\mathrm{pt}}$. Заметим, что в этом случае общее число изменений битов контейнера равно 1, так как

бит с номером один ранее инвертировался при построении вектора $\tilde{\mathbf{v}}_{i_1}$. При $(\delta_1, \delta_2) = (1, 1)$ меняются биты с номерами 1 и 2, что соответствует прибавлению к вектору $\mathbf{x}_{i_1}$ суммы двух первых строк матрицы $G_{s,\mathrm{pt}}$; общее число изменений также равно 1. В случае $(\delta_1, \delta_2) = (0, 1)$ меняются биты с номерами 2 и s (что соответствует прибавлению к вектору $\mathbf{x}_{i_1}$ второй строки матрицы $G_{s,\mathrm{pt}}$), поэтому общее число изменений равно трем, так как первый бит второй строки нулевой, а вес строки равен двум. При случайном и равновероятном распределении битов сообщения (например, когда вкладывают зашифрованное сообщение) получим, что с вероятностью $1/4$ изменений не потребуется, с вероятностью $1/2$ потребуется изменение одного бита контейнера и с вероятностью $1/4$ потребуется изменение трех битов контейнера. Таким образом, среднее число измененных битов равно $5/4$. Общее число измененных битов при передаче $2s - 1$ битов сообщения в контейнере длины s^2 битов в среднем равно $(5(s-1))/8$. Отсюда $e = (8(2s-1))/(5(s-1)) = (16s-8)/(5s-5)$ при $g = (s-1)/2$. При $g = (s+1)/2$ среднее число измененных битов равно $(5(s-1))/8 + 1$. Отсюда $e = (8(2s-1))/(5(s-1)) = (16s-8)/(5s-5)$ при $g = (s+1)/2$. Так как события $g = (s-1)/2$ и $g = (s+1)/2$ равновероятны, то в среднем

$$e = \frac{(8s-4)(10s-2)}{(5s+3)(5s-5)},$$

при этом $F = 0$, так как в модифицируемых блоках зафиксированы места «сухих» битов; $\alpha = (2s-1)/s^2$, $B = 0$.

4.2.4 Конструкция 4

Пусть s — нечетное число. Воспользуемся $(H_s^0, s, \sigma_1, \ldots, \sigma_{s+1})$-схемой. С помощью специального размещения единиц в векторах ошибок $\mathbf{e}_{i_1}, \ldots, \mathbf{e}_{i_g}$, где $g = (s+1)/2$, можно добиться того, что в векторе вида $\bigoplus_{i=1}^{s+1}(\mathbf{x}_i \oplus \mathbf{e}_i)$ на первых g координатах будет любое значение из $\{0,1\}^g$. Такая конструкция обеспечивает относительную эффективность $(3s+1)/(s+1)$ и относительную длину сообщения $(3s+1)/(2(s+1)s)$. Особенность этой конструкции в том, что на первых g позициях может быть сформирован любой вектор, например кодовый вектор $[g, g-1, 2]$-кода проверки четности, который позволяет исправить одно стирание. В этом случае

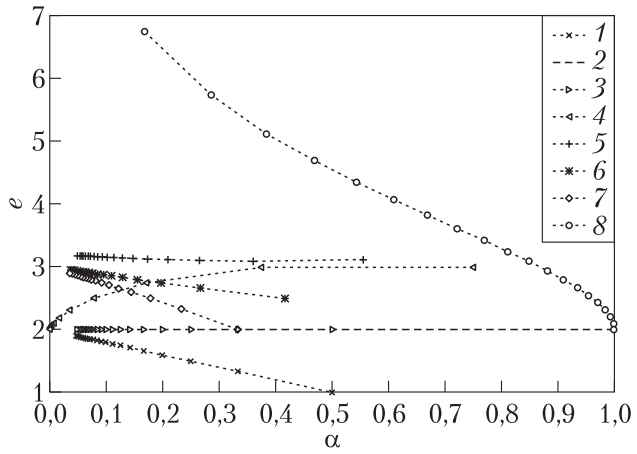
$$\alpha = \frac{3s-1}{2(s+1)s}; \quad e = \frac{3s-1}{s+1},$$

при этом допускается потеря не более одного блока. Степень свободы в этой конструкции отсутствует, так как зафиксированы места «сухих» битов.

Характеристики построенных конструкций обобщены в таблице, а зависимость e от α изображена на рисунке. Как видно на рисунке, построенные

Характеристики стегосистем на основе схемы разделения секрета

№	Конструкция	α	e	B	F	Ограничение
1	Базовая конструкция с исправлением стирания	$\frac{1}{(s+1)}$	$\frac{2s}{(s+1)}$	1	$\left[\left\lceil\frac{s+1}{2}\right\rceil, \left\lfloor\frac{s+1}{2}\right\rfloor\right]$	$s \in \mathbb{N}$
2	Базовая конструкция без исправления стирания	$\frac{1}{s}$	2	0	$\left[\left\lceil\frac{s}{2}\right\rceil, \left\lfloor\frac{s}{2}\right\rfloor\right]$	$s \in \mathbb{N}$
3	Конструкция 1	$\frac{s+1}{s^2}$	$\frac{2(s+1)}{s}$	0	$\left[\left\lceil\frac{s}{2}\right\rceil - 1, \left\lfloor\frac{s}{2}\right\rfloor - 1\right]$	$s \in \mathbb{N}$
4	Конструкция 2	$\frac{s+\log_2 s}{s^2}$	$\frac{2(\log_2 s+s)}{s}$	0	$\left[\left\lceil\frac{s}{2}\right\rceil - 1, \left\lfloor\frac{s}{2}\right\rfloor - 1\right]$	$s = 2^p, p \in \mathbb{N}$
5	Конструкция 3	$\frac{2s-1}{(s^2}$	$\frac{(8s-4)(10s-2)}{(5s+3)(5s-5)}$	0	0	$s \in \mathbb{N} \setminus 2\mathbb{N}$, $s > 1$
6	Конструкция 4 без исправления стирания	$\frac{3s+1}{2(s+1)s}$	$\frac{3s+1}{s+1}$	0	0	$s \in \mathbb{N} \setminus 2\mathbb{N}$
7	Конструкция 4 с исправлением стирания	$\frac{3s-1}{2(s+1)s}$	$\frac{3s-1}{s+1}$	1	0	$s \in \mathbb{N} \setminus 2\mathbb{N}$



Зависимость e от α для построенных конструкций: *1–7* — номера рассматриваемых вариантов из таблицы; *8* — верхняя граница

характеристики далеки от достижения верхней границы. Отметим, что конструкция ZZW, как следует из [3], достигает верхней границы. В то же время, как видно из таблицы, многие построенные в настоящей работе конструкции обладают ненулевой степенью свободы при модификации битов контейнера, а некоторые позволяют бороться с утратой одного из блоков.

Литература

1. *Fridrich J., Soukal D.* Matrix embedding for large payloads // IEEE T. Inf. Foren. Sec., 2006. Vol. 1. No. 3. P. 390–395.
2. *Galand F., Kabatiansky G.* Information hiding by coverings // IEEE Information Theory Workshop Proceedings. — IEEE, 2003. P. 151–154.
3. *Zhang W., Zhang X., Wang S.* Maximizing steganographic embedding efficiency by combining Hamming codes and wet paper codes // Information hiding / Eds. K. Solanki, K. Sullivan, U. Madhow. — Lecture notes in computer science ser. — Springer, 2008. Vol. 5284. P. 60–71.
4. *Zhang W., Wang X.* Generalization of the ZZW embedding construction for steganography // IEEE T. Inf. Foren. Sec., 2009. Vol. 4. No. 3. P. 564–569.
5. *Ker A. D.* Batch steganography and pooled steganalysis // Information hiding / Eds. J. L. Camenisch, C. S. Collberg, N. F. Johnson, P. Sallee. — Lecture notes in computer science ser. — Springer, 2006. Vol. 4437. P. 265–281.
6. *Fridrich J., Goljan M, Lisonek P., Soukal D.* Writing on wet paper // IEEE T. Signal Proces., 2005. Vol. 53. No. 10. P. 3923–3935.
7. *Fridrich J., Goljan M, Soukal D.* Efficient wet paper codes // Information hiding / Eds. M. Barni, J. Herrera-Joancomarti, S. Katzenbeisser, F. Pérez-González. — Lecture notes in computer science ser. — Springer, 2005. Vol. 3727. P. 204–218.
8. *Косолапов Ю. В.* Схема разделения секрета типа схемы Блэкли, основанная на пересечении подпространств // Математические вопросы криптографии, 2017. Т. 8. Вып. 1. С. 13–30.
9. *Ker A. D., Bas P., Bohme R., Cogranne R., Craver S., Filler T., Fridrich J., Pevny T.* Moving steganography and steganalysis from the laboratory into the real world // 1st ACM Workshop on Information Hiding and Multimedia Security Proceedings. — ACM, 2013. P. 45–58.
10. *Fridrich J.* Minimizing the embedding impact in steganography // 8th Workshop on Multimedia and Security Proceedings. — Geneva, Switzerland: ACM, 2006. P. 2–10.

Поступила в редакцию 22.10.18

THE DATA EMBEDDING METHOD BASED ON THE SECRET SHARING SCHEME

Yu. V. Kosolapov

Institute for Mathematics, Mechanics, and Computer Science named after I. I. Vorovich, Southern Federal University, 8a Milchakova Str. Rostov-on-Don 344090, Russian Federation

Abstract: Important characteristics of stegosystems are the relative length α of the message to be inserted and the relative effectiveness of the embedding e. Additional essential characteristics of such systems are the degree of freedom in choosing modifiable bits of the container and the ability to resist loss of part of the stegocontainer blocks. This paper is devoted to the development of the stegosystem which, on the one hand, allows recovering partially lost data and, on the other hand, gives the opportunity to choose modifiable bits. On the basis of this system, stegosystems are constructed and investigated, for which the characteristics of α and e are calculated and the degree of freedom and the number of erasures of blocks that do not distort the disseminated data are estimated.

Keywords: information embedding; a secret sharing scheme

DOI: 10.14357/08696527190316

References

1. Fridrich, J., and D. Soukal. 2006. Matrix embedding for large payloads. *IEEE T. Inf. Foren. Sec.* 1(3):390–395.
2. Galand, F., and G. Kabatiansky. 2003. Information hiding by coverings. *IEEE Information Theory Workshop Proceedings.* IEEE. 151–154.
3. Zhang, W., X. Zhang, and S. Wang. 2008. Maximizing steganographic embedding efficiency by combining Hamming codes and wet paper codes. *Information hiding*. Eds. K. Solanki, K. Sullivan, and U. Madhow. Lecture notes in computer science ser. Springer. 5284:60–71.
4. Zhang, W., and X. Wang. 2009. Generalization of the ZZW embedding construction for steganography. *IEEE T. Inf. Foren. Sec.* 4(3):564–569.
5. Ker, A. D. 2006. Batch steganography and pooled steganalysis. *Information hiding*. Eds. J. L. Camenisch, C. S. Collberg, N. F. Johnson, and P. Sallee. Lecture notes in computer science ser. Springer, 4437:265–281.
6. Fridrich, J., M. Goljan, P. Lisonek, and D. Soukal. 2005. Writing on wet paper. *IEEE T. Signal Proces.* 53(10):3923–3935.
7. Fridrich, J., M. Goljan, and D. Soukal. 2005. Efficient wet paper codes. *Information hiding*. Eds. M. Barni, J. Herrera-Joancomarti, S. Katzenbeisser, and F. Pérez-González. Lecture notes in computer science ser. Springer. 3727:204–218.

8. Kosolapov, Yu. V. 2017. Skhema razdeleniya sekreta tipa skhemy Blekli, osnovannaya na peresechenii podprostranstv [Blakley type secret sharing scheme based on the intersection of subspaces]. *Matematicheskie voprosy kriptografii* [Mathematical Aspects of Cryptography] 8(1):13–30.
9. Ker, A. D., P. Bas, R. Bohme, R. Cogranne, S. Craver, T. Filler, J. Fridrich, and T. Pevny. 2013. Moving steganography and steganalysis from the laboratory into the real world. *1st ACM Workshop on Information Hiding and Multimedia Security Proceedings.* ACM. 45–58.
10. Fridrich, J. 2006. Minimizing the embedding impact in steganography. *8th Workshop on Multimedia and Security Proceedings*. Geneva, Switzerland: ACM. 2–10.

Received October 22, 2018

Contributor

Kosolapov Yury V. (b. 1982) — Candidate of Science (PhD) in technology, associate professor, Institute for Mathematics, Mechanics, and Computer Science named after I. I. Vorovich, Southern Federal University, 8a Milchakova Str. Rostov-on-Don 344090, Russian Federation; itaim@mail.ru

ПОИСК ПУТЕЙ ДИНАМИЧЕСКОЙ РЕКОНФИГУРАЦИИ РАСПРЕДЕЛЕННОЙ ИНФОРМАЦИОННО-ВЫЧИСЛИТЕЛЬНОЙ СИСТЕМЫ В СЛУЧАЕ ЗАХВАТА ХОСТА ПРОТИВНИКОМ*

Н. А. Грушо[1]

Аннотация: Для обеспечения информационной безопасности информационных технологий (ИТ) в распределенных информационно-вычислительных системах (РИВС) ранее был предложен механизм метаданных, реализующий разрешительную систему установления соединений в сети. В случае захвата хоста противником существует стратегия организации атак, которые не выявляются на уровне традиционных метаданных. Ряд ошибок в данных, которые могут быть сгенерированы противником в ходе реализации ИТ, требует построение цепочек причинно-следственных связей, предшествующих ошибке, с целью выявления причины ошибки, при этом метаданные реализуют упрощенную модель причинно-следственных связей при решении задач в ходе выполнения ИТ. Этой моделью можно воспользоваться для поиска указанных ошибок. В работе построена синергетическая связь между решением указанной проблемы информационной безопасности и работой опытного системного администратора (СА) по определению причин неявных ошибок. Эта связь позволяет использовать опыт СА для упрощения поиска захваченного хоста и некоторых стратегий противника по внедрению ошибок в реализацию ИТ. Также эта связь позволяет минимизировать требования по реконфигурации сетей для обхода захваченного хоста.

Ключевые слова: информационная безопасность; метаданные; причинно-следственные связи; системное администрирование; неявные сбои и ошибки

DOI: 10.14357/08696527190317

1 Введение

Метаданные для управления соединениями в РИВС были впервые определены и исследованы в работах [1, 2] как новый механизм обеспечения информационной безопасности. Как было отмечено в [3], метаданные — это упрощения математических моделей бизнес-процессов, достаточные для того, чтобы быстро определять причинно-следственные связи задач в ходе выполнения ИТ. Соединение хостов считается разрешенным, если задачи, решаемые на этих хостах, соответствуют причинно-следственным связям, описанным в метаданных.

*Работа частично поддержана РФФИ (проект 18-07-00274).

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

Возникновение нарушений причинно-следственных связей равносильно нарушению информационной безопасности. При изменении порядка решения задач метаданные могут значительно меняться, что существенно усложняет задачу управления соединениями в сетях с помощью метаданных. В этом случае:

– либо метаданные строятся на основе инвариантов относительно изменений порядка решения задач в ИТ, что снижает защищенность ИТ, достижимую с помощью метаданных;
– либо строится множество локальных траекторий вычислительного процесса (допустимые ограничения соединений, т. е. несколько вариантов метаданных [4]).

В последнем случае строятся цепочки допустимых соединений, куда входят те последовательности задач, которые обеспечивают выполнение ИТ.

Также ранее в [5] было замечено, что в случае захвата хоста и накопления противником данных о возможных соединениях этого хоста (информация о метаданных) порождаются угрозы, не выявляемые на уровне метаданных. Эти угрозы связаны с малыми изменениями значений выходных данных задачи, решаемой на захваченном хосте. Такие изменения могут быть выявлены, как правило, через несколько шагов ИТ. Это выявление возможно в случае отклонения конечных или промежуточных результатов ИТ от ожидаемых. При этом тестирование функционала приложений, операционной и компьютерной систем подтверждает, что все компоненты работают правильно. В работе [5] были предложены некоторые способы модернизации модели метаданных, позволивших решать проблему поиска причин таких ошибок в некоторых частных случаях. В работах [6, 7] выявлены ошибки в значениях параметров конфигураций взаимодействующих программ, которые дают описанный выше эффект. Выявление источника (причины) искажения требует построения модели цепочки предыдущих взаимодействий, на которой выявляется исходное искажение или ошибка. Чем сложнее модель, тем сложнее такой поиск.

Отсюда возникла идея проведения анализа действий опытных СА по выявлению причин подобных ошибок. Формальный подход к поиску ошибок в данных и значениях параметров конфигураций приводит к экспоненциальной сложности таких задач и практически неприемлем, поэтому опытный СА в поиске причин подобных ошибок стремится сократить область поиска, т. е. построить метаданные влияния на эффект, порожденный сбоем. По сути, действия опытного СА восстанавливают причинно-следственные связи конкретных следствий с потенциальными причинами этих следствий по аналогии с метаданными.

Прежде чем строить модели этого подхода, была проведена работа по сбору данных о действиях опытных СА при поиске причин ошибок в данных и в значениях параметров конфигураций. Дальнейшая часть работы посвящена описанию примера построения упрощений вычислительного процесса при поиске неявной ошибки опытным СА. Это и другие подобные описания в дальнейшем предпола-

гается использовать для построения модернизированных метаданных, с помощью которых могут быстрее находиться указанные выше искажения.

2 Выявление неявных сбоев в компьютерных системах

Системный администратор в своей работе по устранению сбоев в компьютерных системах сталкивается со сложным взаимодействием множества программно-аппаратных компонентов. Если аппаратное или программное обеспечение не предоставляет достаточно данных для выявления сбоя, то СА приходится проверять каждый компонент компьютерной системы, связанный с решением конкретной задачи.

Если каждый компонент компьютерной системы может предоставлять подробную информацию о том, правильно ли он выполняет свои функции и насколько хорошо согласованы данные, то СА этого было бы достаточно для оперативного выявления любого сбоя. Некоторые компоненты способны предоставлять подробную диагностическую информацию (журналы работы, диагностическая печать на принтерах), другим же требуется специальное программно-аппаратное обеспечение для отображения такой информации, например специализированные тестеры, адаптеры, отладчики, программы и т. д. Но чем больше компонентов в компьютерной системе, тем больше диагностических средств требуется для получения нужной СА информации. Например, рассмотрим один офисный персональный компьютер. Примерный инструментарий для диагностики каждого компонента:

(1) тестирование материнской платы;

(2) тестирование оперативной памяти;

(3) тестирование процессора;

(4) тестирование жесткого диска;

(5) тестирование блока питания;

(6) тестирование монитора;

(7) тестирование принтера;

(8) тестирование операционной системы;

(9) тестирование драйверов оборудования;

(10) тестирование прикладного программного обеспечения.

Таким образом, СА должен иметь 10 инструментов для тестирования только одного персонального компьютера с одним принтером и монитором. Если же учитывать, что в компьютерных системах используются десятки различных узлов, сетевых устройств и т. д., то число компонентов для тестирования будет исчисляться сотнями. Для более крупных организаций количество инструментов для тестирования компонентов может составлять уже тысячи. Естественно, что

обладать таким огромным инструментарием и уметь его применять способны лишь высококлассные профессионалы. В большинстве же случаев рядовой СА не обладает достаточной квалификацией и инструментарием. В случае, когда СА не в состоянии выявить причину сбоя, привлекают СА-эксперта, который своим опытом заменяет обширный требуемый инструментарий.

Приведем пример выявления сбоя СА-экспертом.

Пример. Принтер этикеток, подключенный через принт-сервер к локальной сети, должен печатать этикетки с удаленного сервера, где установлена товароучетная система, в которой формируется текст этикетки. В результате сбоя принтер осуществляет печать этикеток с пропуском одной этикетки. В данной схеме для сокращения анализа выделена цепочка устройств (метаданные 1), уменьшающая перебор возможных воздействий на ИТ с точки зрения выявленной ошибки. Для данной цепочки выделены следующие компоненты:

- принтер этикеток;
- принт-сервер;
- локальная сеть;
- роутер в локальной сети;
- роутер удаленной сети;
- удаленная локальная сеть;
- соединение через глобальную сеть;
- сервер с операционной системой и драйверами принтера;
- товароучетная система.

Каждый из перечисленных компонентов подвергается проверке системным администратором следующим образом:

- принтер этикеток — проверяется правильность установки рулона этикеток, позиционирование датчика бумаги, проводится тестовая печать с самого принтера, наличие подключения к принт-серверу;
- принт-сервер — осуществляется проверка подключения принтера этикеток к принт-серверу, проводится проверка подключения принт-сервера к локальной сети и его доступность для других узлов локальной сети;
- локальная сеть — проводится проверка связи роутера и принт-сервера по локальной сети;
- роутер локальной сети — проводится проверка установления VPN (virtual private network) подключения к удаленному роутеру;
- роутер удаленной сети, подключение по глобальной сети, удаленная локальная сеть — проверяется установление VPN-соединения и доступность удаленного сервера;

– удаленный сервер — осуществляется проверка работоспособности сервера, установленные на него драйверы оборудования, настройки оборудования;
– товароучетная система — проверяется возможность формирования текста этикетки через предварительный просмотр при печати.

Системный администратор выполнил все эти проверки. Единственным отклонением от нормального поведения компонентов оказалось то, что принтер этикеток при самодиагностике печатал также с пропуском одной этикетки. Но отсутствие понятной СА диагностической информации не позволило ему принять правильное решение по устранению сбоя. Опытный СА, привлеченный в качестве эксперта, ознакомился с результатами проведенной диагностики. Опыт подсказал эксперту, что для поиска причины надо идти от следствия к причине, т. е. необходимо перечислить все параметры конфигурации принтера этикеток вместе с множествами значений, которые могут принимать эти параметры. Каждое значение этих параметров надо попытаться связать с возникающей ошибкой. Построенное множество цепочек определяют метаданные 2. Эксперт обратил внимание на особенность тестовой печати принтером этикеток. Опыт эксперта указывал на то, что перенос на новую страницу при печати возникает только в том случае, когда выводимый на печать макет больше, чем формат бумаги, с которым работает принтер. Если принтер при автономной печати диагностической этикетки совершает такую же ошибку, т. е. тестовый макет больше формата бумаги, значит, формат бумаги, заданный в принтере, оказался меньше необходимого. Эксперт делает заключение о том, что формат бумаги, заданный в настройках принтера, не верен. Последним шагом в рассуждении является тот факт, что сам принтер не имеет возможности настройки размера бумаги. Настройка размера бумаги передается принтеру вместе с заданием на печать от драйвера принтера этикеток. Выводом этих рассуждений эксперта является предположение о том, что принтер получил с заданием на печать неверный размер бумаги. Таким образом, построена цепочка причинно-следственных связей и выделена исходная причина появившейся ошибки. Для того чтобы сбросить полученную и сохраненную принтером настройку, необходимо перезагрузить принтер этикеток. После этого печать этикеток производилась правильно.

В рассмотренном примере наблюдение эксперта за поведением компонента (принтера этикеток) позволило выявить причину, при которой могло возникнуть такое поведение. Эксперт ранее часто сталкивался с аналогичной проблемой на различных видах принтеров, что позволило построить цепочку рассуждений от следствия к причине. При изменении одного параметра происходит изменение получаемого результата. Чтобы приобрести такой опыт, необходимо проведение экспериментов со множеством параметров настроек драйвера принтера и наблюдение за влиянием этих изменений на получаемый результат.

Изменение параметров в компьютерной системе может повлечь за собой изменения как внутри самой системы, так и вне ее, т. е. при изменении размера бумаги принтера в настройках драйвера результат оказывается на бумаге, что находится вне компьютерной системы. В данном случае необходимо наличие

«наблюдателя» — пользователя или СА, который сможет принять решение об удовлетворительном или отрицательном результате.

3 Заключение

Найденная параллель между работой СА-эксперта и причинно-следственными ограничениями, которые описываются метаданными, позволяет наметить дальнейшие пути развития подхода к обеспечению информационной безопасности с помощью метаданных. В частности, развитие метаданных для выявления захвата хоста при описанной стратегии противника должно содержать следующие этапы.

1. На выделенных шагах ИТ необходимо иметь некоторые контрольные функции, вычисление которых позволяет выявлять ошибки (запреты) в данных или в значениях конфигурационных параметров при реализации ИТ.
2. Необходимо уметь строить минимальные цепочки причинно-следственных связей, максимально сокращая поиск причин наблюденной ошибки.
3. Реализация п. 2, возможно, потребует осуществить реконфигурацию вычислительных процессов, реализующих ИТ, т. е. построить минимально возможное число цепочек причинно-следственных связей, чтобы при помощи хотя бы одной цепочки можно было обойти захваченный хост и правильно реализовать ИТ.

Литература

1. *Grusho A. A., Timonina E. E., Shorgin S. Ya.* Modelling for ensuring information security of the distributed information systems // 31st European Conference on Modelling and Simulation Proceedings. — Dudweiler, Germany: Digitaldruck Pirrot GmbH, 2017. P. 656–660. http://www.scs-europe.net/dlib/2017/ecms2017acceptedpapers/0656-probstat_ECMS2017_0026.pdf.
2. *Grusho A., Grusho N., Zabezhailo M., Zatsarinny A., Timonina E.* Information security of SDN on the basis of metadata // Computer network security / Eds. J. Rak, J. Bay, I. V. Kotenko, *et al.* — Lecture notes in computer science ser. — Springer, 2017. Vol. 10446. P. 339–347. https://link.springer.com/chapter/10.1007/978-3-319-65127-9_27.
3. *Грушо А. А., Тимонина Е. Е., Шоргин С. Я.* Иерархический метод порождения метаданных для управления сетевыми соединениями // Информатика и её применения, 2018. Т. 12. Вып. 2. С. 44–49.
4. *Grusho A., Timonina E., Shorgin S.* Security models based on stochastic meta data // Analytical and computational methods in theory probability / Eds. V. Rykov, N. Singpurwalla, A. Zubkov. — Lecture notes in computer science ser. — Springer, 2017. Vol. 10684. P. 388–400. https://link.springer.com/chapter/10.1007/978-3-319-71504-9_32.
5. *Грушо А. А., Грушо Н. А., Левыкин М. В., Тимонина Е. Е.* Методы идентификации захвата хоста в распределенной информационно-вычислительной системе, защищенной с помощью метаданных // Информатика и её применения, 2018. Т. 12. Вып. 4. С. 41–45.

6. *Грушо А. А., Забежайло М. И., Зацаринный А. А., Николаев А. В., Писковский В. О., Тимонина Е. Е.* Классификация ошибочных состояний в распределенных вычислительных системах и источники их возникновения // Системы и средства информатики, 2017. Т. 27. № 2. С. 30–41.
7. *Грушо А. А., Забежайло М. И., Зацаринный А. А., Николаев А. В., Писковский В. О., Сенчило В. В., Судариков И. В., Тимонина Е. Е.* Об анализе ошибочных состояний в распределенных вычислительных системах // Системы и средства информатики, 2018. Т. 28. № 1. С. 99–109.

Поступила в редакцию 15.08.19

METHODS OF IDENTIFICATION OF "WEAK" SIGNS OF VIOLATIONS OF INFORMATION SECURITY

N. A. Grusho

Institute of Informatics Problems, Federal Research Center "Computer Sciences and Control" of the Russian Academy of Sciences; 44-2 Vavilov Str., Moscow 119133, Russian Federation

Abstract: To ensure information security of information technologies in distributed information computing systems, a metadata mechanism implementing a permit system for establishing connections in a network has previously been proposed. If a host is captured by an adversary, there is a strategy for organizing attacks that are not detected at the traditional metadata level. A number of errors in data that can be generated by an adversary during the implementation of information technology require the construction of cause-and-effect chains preceding the error in order to identify the cause of the error. At the same time, metadata implement a simplified model of cause-and-effect relations when solving problems during implementation of information technology. This model can be used to find the specified errors. The author constructs a synergistic relationship between the solution of the mentioned problem of information security and the work of an experienced system administrator to determine the causes of implicit errors. This relationship allows leveraging the expertise of system administrators to make it easier to find a captured host and some strategies of an adversary to incorporate errors into the implementation of information technology. It also minimizes network reconfiguration requirements to bypass the captured host.

Keywords: information security; metadata; cause-and-effect relationships; system administration; implicit failures and errors

DOI: 10.14357/08696527190317

Acknowledgments

The paper was partially supported by the Russian Foundation for Basic Research (project 18-07-00274-a).

References

1. Grusho, A. A., E. E. Timonina, and S. Ya. Shorgin. 2017. Modelling for ensuring information security of the distributed information systems. *31st European Conference on Modelling and Simulation Proceedings*. Digitaldruck Pirrot GmbHP Dudweiler, Germany. 656–660. Available at: http://www.scs-europe.net/dlib/2017/ecms2017acceptedpapers/0656-probstat_ECMS2017_0026.pdf (accessed August 16, 2019).
2. Grusho, A., N. Grusho, M. Zabezhailo, A. Zatsarinny, and E. Timonina. 2017. Information security of SDN on the basis of metadata. *Computer network security*. Eds. J. Rak, J. Bay, I. V. Kotenko, *et al.* Lecture notes in computer science ser. Springer. 10446:339–347. Available at: https://link.springer.com/chapter/10.1007/978-3-319-65127-9_27 (accessed August 16, 2019).
3. Grusho, A. A., E. E. Timonina, and S. Ya. Shorgin. 2018. Ierarkhicheskiy metod porozhdeniya metadannykh dlya upravleniya setevymi soedineniyami [Hierarchical method of meta data generation for control of network connections]. *Informatika i ee Primeneniya — Inform. Appl.* 12(2):44–49.
4. Grusho, A., E. Timonina, and S. Shorgin. 2017. Security models based on stochastic metadata. *Analytical and computational methods in theory probability*. Eds. V. Rykov, N. Singpurwalla, and A. Zubkov. Lecture notes in computer science ser. Springer. 10684:388–400. Available at: https://link.springer.com/chapter/10.1007/978-3-319-71504-9_32 (accessed August 16, 2019).
5. Grusho, A. A., N. A. Grusho, M. V. Levykin, and E. E. Timonina. 2018. Metody identifikatsii zakhvata khosta v raspredelennoy informatsionno-vychislitel'noy sisteme, zashchishchennoy s pomoshch'yu metadannykh [Methods of identification of host capture in the distributed information system which is protected on the base of meta data]. *Informatika i ee Primeneniya — Inform. Appl.* 12(4):41–45.
6. Grusho, A. A., M. I. Zabezhailo, A. A. Zatsarinnyy, A. V. Nikolaev, V. O. Piskovski, and E. E. Timonina. 2017. Klassifikatsiya oshibochnykh sostoyaniy v raspredelennykh vychislitel'nykh sistemakh i istochniki ikh vozniknoveniya [Erroneous states classification in distributed computing systems and sources of their occurrence]. *Sistemy i Sredstva Informatiki — Systems and Means of Informatics* 27(3):30–41.
7. Grusho, A. A., M. I. Zabezhailo, A. A. Zatsarinnyy, A. V. Nikolaev, V. O. Piskovski, V. V. Senchilo, I. V. Sudarikov, and E. E. Timonina. 2018. Ob analize oshibochnykh sostoyaniy v raspredelennykh vychislitel'nykh sistemakh [About the analysis of erratic statuses in the distributed computing systems]. *Sistemy i Sredstva Informatiki — Systems and Means of Informatics* 28(1):99–109.

Received August 15, 2019

Contributor

Grusho Nikolai A. (b. 1982) — Candidate of Science (PhD) in physics and mathematics, senior scientist, Institute of Informatics Problems, Federal Research Center "Computer Sciences and Control" of the Russian Academy of Sciences; 44-2 Vavilov Str., Moscow 119133, Russian Federation; info@itake.ru

ОБ АВТОРАХ

Адамович Игорь Михайлович (р. 1934) — кандидат технических наук, ведущий научный сотрудник Института проблем информатики Федерального исследовательского центра «Информатика и управление» Российской академии наук

Бекешева Ирина Сергеевна (р. 1987) — кандидат педагогических наук, доцент кафедры математики и методики преподавания математики Хакасского государственного университета им. Н. Ф. Катанова

Бобылев Вадим Александрович (р. 1981) — начальник отдела информационного обеспечения Территориального фонда обязательного медицинского страхования Республики Хакасия

Бобылева Оксана Владимировна (р. 1982) — кандидат физико-математических наук, доцент кафедры математики и методики преподавания математики Хакасского государственного университета им. Н. Ф. Катанова

Бунтман Надежда Валентиновна (р. 1957) — кандидат филологических наук, доцент Московского государственного университета им. М. В. Ломоносова

Волков Олег Игоревич (р. 1964) — ведущий программист Института проблем информатики Федерального исследовательского центра «Информатика и управление» Российской академии наук

Гайдамака Юлия Васильевна (р. 1971) — доктор физико-математических наук, профессор Российского университета дружбы народов; старший научный сотрудник Федерального исследовательского центра «Информатика и управление» Российской академии наук

Галина Ирина Владимировна (р. 1965) — ведущий инженер Института проблем информатики Федерального исследовательского центра «Информатика и управление» Российской академии наук

Гончаров Александр Анатольевич (р. 1994) — младший научный сотрудник Института проблем информатики Федерального исследовательского центра «Информатика и управление» Российской академии наук

Грушо Николай Александрович (р. 1982) — кандидат физико-математических наук, старший научный сотрудник Института проблем информатики Феде-

ального исследовательского центра «Информатика и управление» Российской академии наук

Дулин Сергей Константинович (р. 1950) — доктор технических наук, профессор, ведущий научный сотрудник Института проблем информатики Федерального исследовательского центра «Информатика и управление» Российской академии наук; главный научный сотрудник Научно-исследовательского и проектно-конструкторского института информатизации, автоматизации и связи на железнодорожном транспорте (ОАО «НИИАС»)

Дьяченко Денис Юрьевич (р. 1987) — инженер-исследователь Института проблем информатики Федерального исследовательского центра «Информатика и управление» Российской академии наук

Дьяченко Юрий Георгиевич (р. 1958) — кандидат технических наук, старший научный сотрудник Института проблем информатики Федерального исследовательского центра «Информатика и управление» Российской академии наук

Егорова Анна Юрьевна (р. 1991) — младший научный сотрудник Института проблем информатики Федерального исследовательского центра «Информатика и управление» Российской академии наук

Захаров Виктор Николаевич (р. 1948) — доктор технических наук, доцент, ученый секретарь Федерального исследовательского центра «Информатика и управление» Российской академии наук

Зацаринный Александр Алексеевич (р. 1951) — доктор технических наук, заместитель директора 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

Ильин Александр Владимирович (р. 1975) — кандидат технических наук, ведущий научный сотрудник Государственного научно-исследовательского института авиационных систем

Ильин Владимир Дмитриевич (р. 1937) — доктор технических наук, профессор, ведущий научный сотрудник Вычислительного центра им. А. А. Дородницына Российской академии наук Федерального исследовательского центра «Информатика и управление» Российской академии наук

Ионенков Юрий Сергеевич (р. 1956) — 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

Козловская Яна Дмитриевна (р. 1998) — студент Московского авиационного института (национальный исследовательский университет)

Косолапов Юрий Владимирович (р. 1982) — кандидат технических наук, доцент Института математики, механики и компьютерных наук им. И. И. Воровича Южного федерального университета, г. Ростов-на-Дону

Красовицкий Александр Михайлович (р. 1976) — кандидат технических наук, ведущий научный сотрудник Института информационных и вычислительных технологий, Казахстан

Кривенко Михаил Петрович (р. 1946) — доктор технических наук, профессор, ведущий научный сотрудник Института проблем информатики Федерального исследовательского центра «Информатика и управление» Российской академии наук

Кудрявцев Алексей Андреевич (р. 1978) — кандидат физико-математических наук, доцент кафедры математической статистики факультета вычислительной математики и кибернетики Московского государственного университета им. М. В. Ломоносова

Листопад Сергей Викторович (р. 1984) — кандидат технических наук, старший научный сотрудник Калининградского филиала Федерального исследовательского центра «Информатика и управление» Российской академии наук

Мамонова Оксана Сергеевна (р. 1998) — студентка факультета иностранных языков и регионоведения Московского государственного университета им. М. В. Ломоносова

Морозов Николай Викторович (р. 1956) — старший научный сотрудник Института проблем информатики Федерального исследовательского центра «Информатика и управление» Российской академии наук

Мусабаев Рустам Рафикович (р. 1979) — руководитель лаборатории Института информационных и вычислительных технологий, Казахстан

Нуриев Виталий Александрович (р. 1980) — кандидат филологических наук, ведущий научный сотрудник Института проблем информатики Федерального исследовательского центра «Информатика и управление» Российской академии наук

Палионная Софья Игоревна (р. 1995) — студентка факультета вычислительной математики и кибернетики Московского государственного университета им. М. В. Ломоносова

Розенберг Игорь Наумович (р. 1965) — доктор технических наук, профессор, Генеральный директор Научно-исследовательского и проектно-конструкторского института информатизации, автоматизации и связи на железнодорожном транспорте (ОАО «НИИАС»)

Синицын Владимир Игоревич (р. 1968) — доктор физико-математических наук, доцент, главный научный сотрудник Института проблем информатики Федерального исследовательского центра «Информатика и управление» Российской академии наук

Синицын Игорь Николаевич (р. 1940) — доктор технических наук, профессор, заслуженный деятель науки РФ, главный научный сотрудник Института проблем информатики Федерального исследовательского центра «Информатика и управление» Российской академии наук

Степченков Дмитрий Юрьевич (р. 1973) — старший научный сотрудник Института проблем информатики Федерального исследовательского центра «Информатика и управление» Российской академии наук

Степченков Юрий Афанасьевич (р. 1951) — кандидат технических наук, ведущий научный сотрудник Института проблем информатики Федерального исследовательского центра «Информатика и управление» Российской академии наук

Уманский Владимир Ильич (р. 1954) — доктор технических наук, заместитель Генерального директора Научно-исследовательского и проектно-конструкторского института информатизации, автоматизации и связи на железнодорожном транспорте (ОАО «НИИАС»)

Хорошилов Александр Алексеевич (р. 1952) — доктор технических наук, профессор, ведущий научный сотрудник Института проблем информатики Федерального исследовательского центра «Информатика и управление» Российской академии наук

Хорошилов Алексей Алексеевич (р. 1958) — кандидат технических наук, старший научный сотрудник 27-го Центрального научно-исследовательского института Министерства обороны России

Чаркова Виктория Вячеславовна (р. 1997) — студентка Хакасского государственного университета им. Н. Ф. Катанова

Чухно Надежда Викторовна (р. 1995) — студентка магистратуры кафедры прикладной информатики и теории вероятностей Российского университета дружбы народов

Чухно Ольга Викторовна (р. 1995) — студентка магистратуры кафедры прикладной информатики и теории вероятностей Российского университета дружбы народов

Шарнин Михаил Михайлович (р. 1959) — кандидат технических наук, старший научный сотрудник Института проблем информатики Федерального исследовательского центра «Информатика и управление» Российской академии наук

Шестаков Олег Владимирович (р. 1976) — доктор физико-математических, 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

Шоргин Сергей Яковлевич (р. 1952) — доктор физико-математических наук, профессор, главный научный сотрудник Института проблем информатики Федерального исследовательского центра «Информатика и управление» Российской академии наук

Правила подготовки рукописей статей для публикации в журнале «Системы и средства информатики»

Журнал «Системы и средства информатики» публикует теоретические, обзорные и дискуссионные статьи, посвященные научным исследованиям и разработкам в области информационных технологий.

Журнал издается на русском языке. По специальному решению редколлегии отдельные статьи могут печататься на английском языке.

Тематика журнала охватывает следующие направления:

- информационно-телекоммуникационные системы и средства их построения;
- архитектура и программное обеспечение вычислительных машин, комплексов и сетей;
- методы и средства защиты информации.

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 рукопись 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 и 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/publications/collected/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 города и страны и электронного адреса каждого автора;
   - сведения об авторах, в соответствии с форматом, образцы которого представлены на страницах:
     http://www.ipiran.ru/journal/collected/2012_22_02_rus/authors.asp и
     http://www.ipiran.ru/journal/collected/2012_22_02_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 (ссылкм 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".

Примеры ссылок на различные виды публикаций в списке "References":

Описание статьи из журнала:

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 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.

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 «Системы и средства информатики» является некоммерческим изданием. Плата за публикацию не взимается, гонорар авторам не выплачивается.

Адрес редакции журнала «Системы и средства информатики»:

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/collected

Requirements for manuscripts submitted to Journal "Systems and Means of Informatics"

Journal "Systems and Means of Informatics" publishes theoretical, review, and discussion articles on the research and development in the field of information technology.

The journal is published in Russian. By a special decision of the editorial board, some articles can be published in English.

Topics covered include the following areas:

- information and communication systems and tools of their design;
- architecture and software of computational complexes and networks; and
- methods and tools of information protection.

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 "Systems and Means of Informatics" 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).

   The Editorial Board has the right to request from the authors an official expert conclusion that the submitted article has no classified data prohibited for publication.
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 authors receive their article for correction after reviewing, it does not mean that the article is approved to be published. 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/publication/collected/template.doc

8. 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/collected/2012_22_02_rus/authors.asp and
     http://www.ipiran.ru/journal/collected/2012_22_02_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 in 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).

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. "System and Means of Informatics" 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_systems.asp

SYSTEMS AND MEANS OF INFORMATICS (SISTEMY I SREDSTVA INFORMATIKI)

SCIENTIFIC JOURNAL

Volume 29 No.3 Year 2019

I N T H I S I S S U E: