МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ

ВСЕРОССИЙСКИЙ ИНСТИТУТ НАУЧНОЙ И ТЕХНИЧЕСКОЙ ИНФОРМАЦИИ
РОССИЙСКОЙ АКАДЕМИИ НАУК
(ВИНИТИ РАН)

НАУЧНО • ТЕХНИЧЕСКАЯ ИНФОРМАЦИЯ

Серия 2. ИНФОРМАЦИОННЫЕ ПРОЦЕССЫ И СИСТЕМЫ

ЕЖЕМЕСЯЧНЫЙ НАУЧНО-ТЕХНИЧЕСКИЙ СБОРНИК

Издается с 1961 г. **№ 11** **Москва 2018**

ОБЩИЙ РАЗДЕЛ

УДК 004.6.021

О.В. Сюнтюренко

Теоретические и прикладные аспекты автоматизации процедур многомерного анализа данных*

Исследованы актуальные теоретические и прикладные аспекты современной информатики в их привязке к вопросам аналитической обработки научно-технической и технико-экономической информации. Представлены основные направления использования автоматизированных непараметрических процедур логико-математической обработки массивов (потоков) цифровых данных. Рассмотрены методические особенности разработки новых технологических подходов и алгоритмов аналитической постобработки, которые позволяют создавать широкий диапазон многошаговых процедур оценивания и многомерного анализа научных и технико-экономических данных на основе полиграммной оценки функционалов. Показано, что процедуры и алгоритмы, реализуемые на основе рассмотренных методов непараметрической статистики и многомерного анализа данных, имеют хорошую перспективу применения в различных приложениях в рамках развития аналитических технологий Big Data (Больших Данных).

Ключевые слова: *информатика, анализ данных, аналитическая постобработка, непараметрическая статистика, кластерный анализ, прогнозирование, технология Больших Данных*

* Статья подготовлена в рамках работ по гранту РФФИ № 17-07-00256

ВВЕДЕНИЕ

Современные мировые тенденции развития сетевых технологий характеризуются переходом от «паутины документов» к «паутине данных». Становление цифровой экономики, развитие онлайновой инфраструктуры науки и цифровая трансформация информационного пространства активизируют разработку новых технологических направлений более глубокой, аналитической постобработки информации. Конвергенция методов наукометрии (понимаемых расширительно) и многомерного анализа данных в цифровой среде позволяет создавать системы и технологии, реализующие формирование качественно новых информационных продуктов и услуг, ориентированных на поддержку научно-технических и технико-экономических исследований и разработок [1, 2], в которых значительное место занимают методы и модели вероятностно-статистического характера.

Использование вероятностно-статистических моделей и методов в работах, связанных с анализом реальных данных, носит далеко не формальный характер и сопряжено с рядом принципиальных трудностей [3]: малые объемы выборок, существенная неоднородность наблюдений, значительные отклонения эмпирических распределений от нормального закона, наличие ошибок в таблицах и плохая формализуемость изучаемых ситуаций и процессов. Во многом существующие методы многокритериальной оценки инноваций, процессов трансфера технологий, оценки научно-технического уровня выпускаемой и разрабатываемой продукции недостаточно достоверны и продуктивны, так как не позволяют эффективно обрабатывать все возрастающие объемы цифровой научной и технико-экономической информации. Это обусловлено, во-первых, огромной трудоемкостью соответствующих процедур и, во-вторых, малой скоростью анализа данных. В идеальных условиях (при строгом выполнении предпосылки о наличии многомерного нормального распределения) наилучшими являются процедуры классической статистики. Однако качество этих процедур резко снижается даже при небольших отклонениях от нормальности. Так как распределения эмпирических данных плохо описываются гауссовой кривой, растет интерес к методам, свободным от распределения (робастным и непараметрическим), которые являются центральным звеном современной прикладной статистики и анализа данных.

Главное направление в развитии статистической методологии связано с концепцией анализа данных, которая была предложена Дж. Тьюки [3, 4] и по сути представляет синтез вероятностного (стохастического), детерминированного и эвристического подходов к обработке выборочной информации об объектах, процессах и системах. В настоящей работе предложен подход к созданию технологии аналитической постобработки информации на основе алгоритмов, реализующих разработанные методы непараметрической статистики и анализа данных, для решения задач: прогнозирования; выявления эмпирических закономерностей, объективно существующих для экономических объектов и систем; сопоставительного анализа уровня технических и экономических объектов; получения интегральных характеристик для совокупностей однородных объектов техники или экономики.

ПРЕДВАРИТЕЛЬНЫЙ АНАЛИЗ ДАННЫХ

Начальный этап научно-технических и технико-экономических прикладных статистических исследований неизбежно связан с необходимостью выборочной оценки параметров положения и рассеяний маргинальных распределений, а также ковариаций изучаемых числовых и, как правило, непрерывных признаков, характеризующих рассматриваемые объекты. Однако классические выборочные оценки типа среднего арифметического нестандартного отклонения оказываются крайне ненадежными при работе с реальными данными [5]. Это обусловлено тем, что эмпирические распределения на практике часто обнаруживают значительные отклонения от нормального. Выборочное среднее весьма чувствительно к отклонениям ассиметрии, в то время как стандартное отклонение чрезвычайно реагирует на ненормальности эксцесса. Выборки реальных наблюдений часто содержат аномальные выбросы (грубые ошибки). Природа выбросов различна: искажения в каналах телекоммуникаций, ошибки экспертной классификации, случайная или умышленная дезинформация, ошибки кодирования при вводе данных в БД. Наличие в выборке не только малого (с числом наблюдений $N \simeq 20$), но и умеренного объема аномальных выбросов часто делает классические оценки бесполезными. Все это ведет к тому, что на этапе предварительного анализа данных используются не классические выборочные оценки, а современные робастные и непараметрические методы оценивания. Отметим, что робастная статистика занимает «промежуточное» положение между классической (параметрической) и непараметрической статистикой [6]. В нашем случае предварительный анализ данных основан на применении процедур полиграммного оценивания характеристик распределений [7]. Полиграмма, как непараметрическая оценка неизвестной функции плотности вероятности (ФПВ) непрерывной случайной величины (НСВ), используется для непараметрического оценивания функционала энтропии. Общий подход к полиграммной оценке интегральных функционалов, зависящих от непрерывной и аналитически неизвестной ФПВ, обоснован в [7-10]. При нем возникает возможность оценивать моменты распределения НСВ с неизвестной ФПВ.

Пусть НСВ x представлена вариационным рядом без связок вида $x_{(1)} < x_{(2)} < ... < x_{(N)}$. Для выборочных квантилей ξ_j уровня j/m введем обозначение:

$$\xi_j = x_{(jk)};\ j = \overline{1,m};\ \mathrm{m} = [(\mathrm{N}+1)/\mathrm{k}];\ \mathrm{k} = \mathrm{N}^{\gamma};\ 0<\gamma<1.$$

Тогда в качестве полиграммной оценки математического ожидания μ НСВ x используем выражение:

$$\hat{\mu} = \frac{1}{m}\sum\nolimits_{j=1}^{m-1}\left(\xi_{j+1} - \xi_j\right)\mathrm{sign}\left(\xi_j\right)\Big/\ln\left|\xi_{j+1}/\xi_j\right|, \qquad (1)$$

а для полиграммного оценивания дисперсии $(\mu)\sigma^2$ можно воспользоваться соотношением:

$$\hat{\sigma}^2 = \frac{1}{m}\sum\nolimits_{j=1}^{m-1}\left(\xi_j\xi_{j+1}\right) - \hat{\mu}^2 . \qquad (2)$$

Как следует из общей теории полиграммного оценивания интегральных функционалов, линейно зависящих от ФПВ НСВ [7], выражения (1) и (2) дают состоятельные асимптотически (по объему выборки N) нормальные оценки: соответственно μ и σ^2 НСВ x, если ее ФПВ такова, что существуют конечные моменты распределения до четвертого порядка включительно.

ВЫЯВЛЕНИЕ ФАКТОРНЫХ ПРИЗНАКОВ

На основе полиграммного подхода можно получить и непараметрическую (причем асимптотически нормальную) оценку ковариации $\mathrm{R_{ij}}$ признаков x_i и x_j. Для этого достаточно воспользоваться известным соотношением:

$$\mathrm{R_{ij}} = \frac{1}{2}\left(\sigma_{ij}^2 - \sigma_i^2 - \sigma_j^2\right), \qquad (3)$$

где σ_{ij}^2 – дисперсия НСВ ($x_i + x_j$), а σ_i^2 и σ_j^2, соответственно, дисперсии x_i и x_j.

Ковариационная матрица R, полученная на основе соотношения (3), обладает высокой устойчивостью и позволяет оценить степени взаимозависимости признаков даже при небольших объемах технико-экономической информации. Матрицу R можно использовать для получения факторного решения в рамках распространенной модели компонентного анализа [11–13]. Известно, что существует единственный ортогональный оператор B с точностью до: а) перестановок строк его матрицы местами; б) умножения любой его матрицы на -1, приводящей матрицу R к каноническому виду:

$$\mathrm{B^*RB} = \mathrm{E} = \mathrm{diag}\left(\varepsilon_i^2\right);\ \ i = \overline{1,m}\ ; \qquad (4)$$

где $\mathrm{B^*}$ – транспонированная матрица B.

Оператор B вызывает изометрическое преобразование пространства признаков $x_i\left(i = \overline{1,m}\right)$:

$$f_j = \sum\nolimits_{j=1}^{m} b_{ij}x_i\ ;\ j = \overline{1,m}; \qquad (5)$$

Для определенности принимают (это просто соответствует определенной нумерации факторных признаков f_i), что соответствующие значения ε_i^2 матрицы R расположены в порядке убывания: $\varepsilon_1^2 \ge \varepsilon_2^2 \ge \ldots \ge \varepsilon_m^2 \ge 0$.

Такая нумерация ε_i^2 делает очевидным то, что заданная часть (например, 0,80 или 0,95 суммарной дисперсии) статистической изменчивости признаков $x_i\left(i = \overline{1,m}\right)$ учитывается обобщенными признаками $f_j\left(j = \overline{1,p}\right)$. При этом заведомо $p < m$, однако на практике часто оказывается, что $p << m$. Так решается задача снижения размерности пространства исследуемых признаков, т.е. с заданной наперед полнотой информативности описания наблюдений изучаемой области вместо m признаков x_i можно использовать p обобщенных признаков f_j.

Факторный анализ (в частности, описанный выше один из его вариантов – метод главных компонент) используется для выявления обобщенных характеристик объектов, в пространстве которых существенно облегчен процесс сопоставительного анализа сложных процессов и систем технического и экономического характера. В приложениях целесообразно работать с предварительно центрированными и нормированными по масштабу, например, на параметры (1) и (2) случайными переменными x_i, что позволяет избежать влияния выбора единиц измерения. Кроме того, для положительно определенных переменных, часто встречающихся в технико-экономических работах, повышению надежности получаемых результатов способствует переход к логарифмическому масштабу [3, 4] из-за того, что для отлогарифмированных переменных без добавления новых параметров модели можно в значительной степени учесть нелинейность реальных взаимосвязей x_i.

НЕЧЕТКАЯ КЛАССИФИКАЦИЯ НАБЛЮДЕНИЙ

Одной из важнейших задач прикладных научно-технических и технико-экономических исследований, связанных с выявлением образцов-аналогов, групп однородных объектов, сопоставлением уровня развития и т.д., является классификация многомерных наблюдений. Существует множество подходов к многомерной таксономии, основу которых составляют различные, вероятностные и детерминированные методы распознавания образов. На практике наиболее часто применяется кластерный анализ наблюдений. Классическая постановка задачи кластерного анализа состоит в том, чтобы наилучшим способом разбить N наблюдений x^k, каждое из которых характеризуется m признаками x_i $p\left(p < \mathrm{N}\right)$ непересекающихся классов [14]. Таким образом, каждое из изучаемых наблюдений должно попасть в один и только один кластер. Однако на практике добиться этого удается далеко не всегда. Как отмечается в [14], в приложениях для данного объекта X и класса Y в большинстве случаев вопрос состоит не в том, принадлежит ли X к Y, а в том, до какой степени X принадлежит Y. Такая трактовка кластеризации соответствует концепции нечетких (размытых) множеств, в рамках которой целесообразно выполнять многие классификационные задачи из области прикладных научно-технических и технико-экономических исследований. Один из таких подходов к решению указанной проблемы описан в [8] и сводится к следующему.

Пусть C – матрица скалярных произведений векторов наблюдений $x^k\left(k=\overline{1,N}\right)$, элементы которой имеют вид:

$$C_{kl}=\sum_{i=1}^{m}x_i^k x_i^l \;\; k,l=\overline{1,N} \;; \qquad (6)$$

A – ортогональный оператор, приводящий C к диагональному виду:

$$A*CA=L\equiv \mathrm{diag}\left(\lambda_i^2\right);\;\; i=\overline{1,N}\;; \qquad (7)$$

Оператор вызывает (теперь уже в пространстве наблюдений) изометрическое преобразование переменных вида:

$$F^k=\sum_{i=1}^{N}a_{lk}x^l \;;\;\; k=\overline{1,N}\;; \qquad (8)$$

Заметим, что каждый из векторов (8) определен лишь с точностью до знака. Все λ_i^2 (как собственные значения матрицы C) имеют смысл дисперсий по некоторым попарно ортогональным направлениям в пространстве наблюдений. Поэтому по аналогии с компонентным анализом можно ограничиться рассмотрением лишь первых p(p<N) линейных многообразий вида (8), подразумевая, что $\lambda_1^2\ge\lambda_2^2\ge...\lambda_p^2\ge...\ge\lambda_N^2\ge 0$.

В силу ортогональности оператора A можно записать:

$$\sum_{i=1}^{N}\left(a_{ij}\right)^2=1=\sum_{j=1}^{N}\left(a_{ij}\right)^2\;;\;\; ij=\overline{1,N}\;; \qquad (9)$$

и

$$x^k=\sum_{j=1}^{N}a_{kj}F^j\simeq\sum_{j=1}^{p}a_{kj}F^j\;;\;\; k=\overline{1,N}\;; \qquad (10)$$

Из соотношений (9) и (10) видно, что в качестве естественной меры принадлежности наблюдений x^k классу F^j следует использовать величины $\alpha_{kj}=\left(a_{kj}\right)^2$; $k=\overline{1,N}$; $j=\overline{1,p}$. Смысл этих величин точно соответствует «лингвистическим переменным» Л.А. Заде [15, 16], применяемым в теории размытых множеств и в нечеткой классификации наблюдений в качестве характеристик принадлежности.

Векторы (8) ортогональны, но не ортонормированны. Поэтому в качестве нового (генерального) базиса пространства наблюдений следует использовать ортонормированную систему векторов $\gamma_i=F_i/\lambda_i$; $i=\overline{1,p}$. Тогда координаты наблюдений x^k в генеральном базисе запишутся в виде:

$$f_i^k=\sum_{i=1}^{m}x_j^k\sum_{i=1}^{N}a_{li}x_j^l/\lambda_i\;;\; i=\overline{1,p}\;;\;\; k=\overline{1,N}\;; \qquad (11)$$

Отсюда следует, что (аналогично (5)) в «терминах исходных признаков» $x_i\left(i=\overline{1,m}\right)$ полученная нечеткая классификация имеет вид:

$$f_i=\sum_{j=1}^{m}Q_{ij}x_j\;;\;\; i=\overline{1,p}\,. \qquad (12)$$

При этом соотношение (4) оказывается верным:

$$a_{ij}=\sum_{k=1}^{m}x_k^i Q_{kj}/\varepsilon_j\;;\;\; ij=\overline{1,N}\;; \qquad (13)$$

Связь оператора перехода с оператором диагонализации матрицы ковариаций, определенным как решение матричного уравнения (4), можно показать в виде:

$$Q_{ij}=B_{ij}\,. \qquad (14)$$

Этот результат позволяет не пользоваться на практике процедурой приведения C к каноническому виду, а исходить при вычислении матрицы A из матрицы B, принимая за основу соотношение (14).

Таким образом, последовательность действий сводится к следующему:

Шаг 1. Вычисляется (на основе полиграммного оценивания) матрица ковариаций R.

Шаг 2. На основе уравнения (4) определяется матрица перехода к главным компонентам B.

Шаг 3. По матрице B на основе соотношения (14) вычисляется матрица Q.

Шаг 4. По матрице Q на основе выражения (13) вычисляется матрица A.

Шаг 5. По формулам $a_{kj}=\left(a_{kj}\right)^2$ вычисляются значения функций нечеткой принадлежности наблюдений к выявляемым размытым кластерам.

Такой алгоритм дает возможность, руководствуясь единой позицией, решать задачи выявления обобщенных признаков (факторный анализ) и классификации наблюдений (кластер-анализ с использованием нечеткости множеств).

АНАЛИЗ ВЗАИМОСВЯЗЕЙ ПРИЗНАКОВ

Процедура многомерного анализа также позволяет решать такую важнейшую задачу, как выявление эмпирических взаимосвязей между числовыми признаками наблюдений рассматриваемой области. Вновь обратимся к факторным характеристикам f_i, определяемым на основе уравнения (4). Поскольку нумерация величин ε_i^2 принята строго по убыванию, постольку с наперед заданной точностью, начиная с некоторого номера, будет выполняться условие: $f_j^k\simeq 0\pm 3\varepsilon_j\;;\;\; j=\overline{(S+1,m)}$ для большинства наблюдений исследуемой области. Оценка погрешности "$3\varepsilon_j$", где ε_j – корень из соответствующего собственного значения матрицы R, использована здесь на основании неравенства Чебышева и правила «трех сигм», являющихся его следствием. Эта оценка соответствует (грубо) уровню доверительной вероятности примерно 0,9. В прикладных работах такая приближенная оценка сверху значительно практичнее и надежнее, чем вычисления границ на основе t – статистик, которые базируются на мало реалистичной предпосылке о нормальности исследуемого эмпирического распределения.

Таким образом, на основе уравнения (4) можно получить систему линейных приближенных уравнений:

$$\sum_{i=1}^{m} b_{ij} x_i \cong 0 + 3\varepsilon_j; \quad j = \overline{(S+1, m)}; \qquad (15)$$

которые в виде неявных функций определяют $(m - S)$ взаимозависимости между изучаемыми признаками. Однако на практике чаще требуется выявить взаимозависимости между признаками в рамках традиционной модели регрессивного анализа, а не в форме системы уравнений (15). Переход от этой системы к стандартной модели множественной регрессии прост и легко реализуется.

Допустим, часть признаков необходимо рассматривать в качестве независимых переменных (регрессоров). Тогда для определенности принимаем, что это признаки $x_j\left(j = \overline{1, S}\right)$. Другую часть переменных требуется рассматривать как зависимые переменные (отклики). В этом случае для определенности будем считать, что это признаки $x_i\left(i = \overline{S+1, m}\right)$. Теперь путем элементарных алгебраических преобразований линейная система (15) может быть легко преобразована к виду:

$$x_i \cong \sum_{j=1}^{S} \Theta_{ij} x_j \pm \delta_i; \quad i = \overline{(S+1, m)}; \qquad (16)$$

где параметры δ_i и Θ_{ij} определены значениями элементов матрицы B и вектора $\vec{\varepsilon}$. Система уравнений (16) непосредственно пригодна для приближенных прогностических, плановых аналогичных подобных расчетов при проведении прикладных исследований.

При работе с положительно определенными переменными в логарифмическом масштабе точный аналог системы (16) имеет вид:

$$x_i = (1 \pm \delta_i) \cdot \eta_i \prod_{j=1}^{S} \left(x_j / \eta_j\right)^{\Theta_{ij}}; \; i = \overline{(S+1, m)}, \qquad (17)$$

где $\vec{\eta}$ – вектор центра распределения.

При прочих равных условиях система (17) дает более надежные и точные результаты в технико-экономических расчетах по сравнению с системой (16).

ОПТИМАЛЬНЫЙ ВЫБОР НАБЛЮДЕНИЙ

В различных работах научно-технического и технико-экономического характера одной из наиболее важных прикладных проблем является сопоставление многомерных наблюдений в отношении их «качества». К этому кругу задач относятся процессы ранжирования изделий техники, научных организаций и т.п. по их уровню, сопоставление экономических стратегий, сравнение эффективности проектов и программ развития и др. В настоящее время известны три пути практического решения указанных задач.

Первый путь связан с экспертными оценками и обычно предусматривает «усреднение» мнений экспертов на основе какой-либо из современных модификаций метода Дельфи. Этот подход наряду с сильными сторонами (в частности, использование неформализуемого личного опыта специалистов-экспертов) обладает и явными недостатками, среди которых высокая трудоемкость, плохая алгоритмизуемость и существенная субъективность оценок (полностью никакими «усреднениями» она не снимается). Кроме того, все процедуры экспертного сравнения базируются на посылке равной компетенции экспертов, что не всегда соответствует действительности. Это обстоятельство негативно влияет на результаты процедур.

Второй путь – «эвристический» – предусматривает математизацию собственных представлений специалистов-разработчиков об «оптимальности» объектов на основе каких-либо, обычно простых, формальных соотношений. Здесь часто пользуются сопоставлением, исходя из весовых функций, обычно линейного характера, причем вес оценочных признаков выбираются произвольно. Множество методик такого рода создано в самых различных областях как в нашей стране, так и за рубежом, но значительное их большинство выглядит неубедительно в силу почти полного отсутствия теоретического обоснования и аналитической проработки предлагаемых моделей.

Третий путь предполагает использование формального аппарата теории выбора и принятия решений. Здесь широко применяются современные компьютерные технологии и методы информатики, и в настоящее время этот подход явно перспективнее двух предыдущих.

Рассмотрим процедуру оптимального выбора, которая программно может быть реализована в рамках работ (проекта) по аналитической (статистической) постобработке научно-технической или технико-экономической информации.

В основу любого формального выбора объектов (элементов) абстрактного множества Ω лежит понятие отношения на Φ. Отношением Φ на множестве Ω называется заданное подмножество декартова произведения $\Phi \leq \Omega \times \Omega$. Смысл такого определения состоит в том, что Φ означает такие упорядоченные пары элементов (наблюдений) из Ω, которые находятся между собой в соответствии. Однако понятие абстрактного множества слишком общее для построения эффективных прикладных процедур сопоставления и выбора наблюдений. Поэтому на практике оптимальный выбор наблюдений (т.е. элементов Ω) производят, исходя из отношений, заданных в конечномерном векторном (евклидовом) пространстве ε_m. Одним из важнейших в прикладном плане отношений, задаваемых на ε_m, является отношение Парето, представляющее частный случай отношения частичного порядка.

Пусть Ω представляет собой конечный набор, состоящий из N наблюдений рассматриваемой области $x_k\left(k = \overline{1, N}\right)$, каждое из которых характеризуется m оценочными признаками $x_i\left(i = \overline{1, m}\right)$. Оценочным признаком наблюдения (называемым далее для краткости параметром) будем считать любую числовую характеристику, изменение которой меняет наше представление о «качестве» наблюдения. При этом «качество» по своей природе должно определяться

конкретным видом отношения на $\Omega \le \varepsilon_m$. Договоримся также, что возрастание любого параметра $x_i\left(i=\overline{1,m}\right)$ вызывает некоторое увеличение «качества» наблюдения. Разумеется, на практике это не всегда так. Например, параметр «мощность компьютера» удовлетворяет нашему условию, а параметр «масса компьютера» - нет. Но для любого параметра x_i, который не удовлетворяет сделанному предложению, всегда легко подобрать такое преобразование, которое приведет указанный параметр в соответствие требуемой предпосылке. В данном случае вместо массы компьютера M можно ввести в рассмотрение параметры «1/M» или «-M».

Наблюдение x^k находится в отношении Парето к наблюдению x^l (обозначение: $x^k > x^l$) по определению в том случае, если выполнено условие:

$$x^k \succ x^l \Leftrightarrow \begin{cases} \forall_i = 1,m : x_l^k \ge x_i^l \\ \exists_i \in \{1,2,...,m\} : x_j^k > x_j^l \end{cases} \qquad (18)$$

На основе (18) можно легко установить, что $x^k \succ x^l$ в том случае, если x^k не уступает x^l ни по одному параметру и хотя бы по одному из них строго его превосходит. Отметим, что отношение Парето инвариантно к любым монотонным сжатиям и растяжениям масштаба, а также к изменению положения начала координат.

Отношение (18) порождает функцию выбора по Парето, которая отображает множество Ω в множество Парето $C:\Omega \to c(\Omega)$. Функция выбора C имеет смысл блокировки отношения Парето, а множество Парето $c(\Omega)$ – набора мажорант по отношению Парето:

$$C(\Omega) = \left\{ x^k \in \Omega \middle| \forall x^l \in \Omega : x^l \prec x^k \right\}, \qquad (19)$$

где $\succ$ – отношение, являющееся дополнением к отношению Парето:

$$x^l \prec\prec x^k \Leftrightarrow \left[x^k \succ x^l\right] \vee \left[\exists_{ij} : \left(x_l^k > x_i^l\right) \wedge \left(x_j^k < x_j^l\right)\right]. \qquad (20)$$

Суть множества Парето $C(\Omega)$ состоит в том, что любой элемент из $C(\Omega)$ в смысле отношения (18) «не хуже» любого другого наблюдения из Ω.

ВЫБОР ПО ФАКТОРНЫМ ОЦЕНОЧНЫМ ПРИЗНАКАМ

При практическом использовании критерия Парето (18) обычно сталкиваются с двумя затруднениями. Во-первых, оптимальный выбор по Парето явно подразумевает «равноправие» параметров, что не всегда достижимо. Включение же в набор параметров второстепенных и сравнительно малозначимых оценочных характеристик влечет неправомерное (в прикладном отношении) расширение множества Парето. Во-вторых, в прикладном плане нежелательно, чтобы в набор $\left\{x_i\left(i=\overline{1,m}\right)\right\}$ попали хотя бы две характеристики x_i и x_j, которые жестко взаимосвязаны в виде $x_j^c \sim 1/x_i$; $c > 0$. Но известно, что оценочные параметры порой «противоборствуют», так как на практике часто встречаются пары параметров, между которыми существуют жесткие монотонно убывающие взаимозависимости. Очевидно, что наличие даже одной такой пары параметров в списке оценочных характеристик приведет к неравномерному увеличению множества Парето. Следовательно, при практическом использовании процедуры оптимального по Парето выбора требуется «нейтрализовать» эти затруднения. Один из возможных путей решения такой проблемы сводится к использованию выбора не по исходным параметрам x_i, а по факторному решению, полученному на основе компонентного анализа пространства оценочных параметров x_i [14]. По смыслу и природе построения факторы являются, как взаимно (в статистическом плане) независимыми, так и обобщенными (комплексными, наиболее содержательными) характеристиками наблюдений исследуемой области. Таким образом, при реализации оптимального по Парето выбора в пространстве факторных оценочных параметров указанные трудности будут преодолены.

Выбор по Парето позволяет разбить множество наблюдений на непересекающиеся классы эквивалентности. Обозначим $\omega_1 = C(\Omega)$, а Ω_1/ω_1. Произведем выбор $c(\Omega_1)$, результат которого обозначим ω_2. Далее произведем выбор $c(\Omega_2)$, где $\Omega_2 = \Omega_1/\omega_2$. Последовательно продолжая выбор $c(\Omega_k)$; $k = 1,2,...,$ на некотором шаге с номером p будем иметь $c(\Omega_p) = \Omega_p$, что неизбежно в силу конечности наблюдений N. Таким образом, получен набор подмножеств $\{\omega_i\}$, причем $\bigcup_{i=1}^{p} \omega_1 = \Omega$ и $\omega_i \cap \omega_j = \varnothing (i \ne j)$. Приведенную процедуру разбиения Ω на классы эквивалентности (в смысле отношения Парето) можно на практике использовать для анализа уровня изделий техники (инноваций), уровня экономического развития фирм и регионов, сопоставительного анализа отечественных и зарубежных научно-технических и технико-экономических объектов.

Как легко видеть, отношение Парето обладает свойствами транзитивности: $x^k \succ x^l$, $x^l \succ x^p \Rightarrow x^k \succ x^p$. Это обстоятельство позволяет построить графические и иерархические структуры (типа деревьев) для объектов исследуемой области техники или экономики. Такие структуры могут оказаться весьма полезными, например, при номенклатурном анализе научно-технической продукции, в том числе военного назначения, сопоставления уровней ее качества, а также при решении задач прогнозирования развития техники и экономики. Следует отметить, что большинство современных реализуемых алгоритмов прогнозирования носит характер экстраполирования стационарных статистических последовательностей.

АНАЛИТИКА *BIG DATA*

Развитие сетевых технологий, суперкомпьютинга и лавинообразный рост объемов цифровых данных активизируют все более широкое применение технологии *Big Data* (Большие Данные). Термин Большие Данные относится к огромному количеству структурированных и неструктурированных данных, генерирумых в научно-технической и социально-экономической сферах [17]. В зависимости от вида анализа технологии Больших Данных разделяются на три класса [18]: 1) пакетного анализа, когда данные вначале накапливаются, а затем анализируются; 2) потоковой обработки, при которой данные анализируются в реальном масштабе времени; 3) интерактивного анализа, при котором пользователи управляют алгоритмами обработки в зависимости от получаемых результатов и дополнительных данных.

Для аналитической обработки очень больших массивов цифровых данных в последние годы активно разрабатывается технология, называемая *Data Analitics* [19]. Она позволяет для уменьшения Больших данных до приемлемого управляемого размера использовать такие статистические средства и методы, как PCA (Principal Component Analysis) – анализ главных компонентов [13], кластеризация, нечеткая логика и т.д. *Data Analitics* применяет эти средства для выделения значимой информации, создания баз знаний на основе выделенных данных и развивает непараметрическую модель *Big Data*. В качестве примера приведем результаты исследования междисциплинарного характера (медицина, география, метеорология). Анализ за девять лет (2004–2013 гг.) десятков миллионов запросов в интернете о средствах от депрессии показал, что частота таких запросов зависит от температуры в том месте, где находился пользователь интернета. Плохое настроение и хандра чаще бывают не в жару, а при холоде. Особенно проявляется корреляция со средней температурой января. Остальные факторы (продолжительность светового дня, атмосферное давление, количество осадков и др.) влияют значительно меньше [20].

Рассмотренные выше процедуры и алгоритмы могут быть успешно использованы при разработке, например, интерактивного анализа крупномасштабных наборов данных в рамках развития технологии и моделей *Data Analitics*. В обозримом будущем следует ожидать создание и широкое применение автоматизированных систем поддержки и принятия решений на основе Больших Данных с расширенными возможностями многомерного анализа научно-технических и технико-экономических данных. По экспертным оценкам развитие технологий Больших Данных ведет к появлению в скором времени распределенных самообучающихся систем когнитивных вычислений, качественно новых и эффективных методов прогнозирования научно-технических, инновационно-технологических, экономических и социальных процессов. В определенной мере технологии *Big Data* – это ответ на качественно новые задачи в промышленности и науке.

ЗАКЛЮЧЕНИЕ

Показаны подходы к трем проблемам аналитической постобработки научно-технической и технико-экономической информации: предварительному анализу данных, многомерному анализу данных и оптимальному выбору многомерных наблюдений. В современной цифровой среде областью приложения программной реализации рассмотренных процедур постобработки на основе методов непараметрической статистики и многомерного анализа данных могут быть исследования и разработки, связанные с задачами: выявления эмпирических закономерностей, объективно существующих для научно-технических и экономических объектов, систем и процессов; сопоставительного анализа технических и экономических объектов на основе теории выбора (по критерию Парето), а также ранжирования уровня объектов и систем; получения интегральных характеристик совокупностей однородных объектов техники или экономики; прогнозирования динамики изменения системы взаимосвязанных технико-экономических показателей во времени; классификации объектов и систем в целях выявления объектов-аналогов и групп однородных объектов; проектирования систем поддержки принятия решений; компьютерной лингвистики (машинный анализ большого объема текста).

Процедуры и алгоритмы, реализуемые на основе рассмотренных методов непараметрической статистики и многомерного анализа данных, имеют хорошую перспективу применения в различных приложениях в рамках развития аналитических технологий Больших Данных – ведущего тренда экономического и технологического развития.

СПИСОК ЛИТЕРАТУРЫ

1. Борисова Л.Ф., Сюнтюренко О.В. Реферативный банк данных ВИНИТИ РАН: перспективы постобработки информации с использованием методов анализа данных // Научно-техническая информация. Сер. 1 – 2007. – № 11. – С. 6–11.
2. Сюнтюренко О.В. Производство информационно-аналитических продуктов и услуг с использованием методов наукометрии и анализа данных // Материалы Международной конференции к 65-летию ВИНИТИ РАН «Информация в современном мире». – М.: ВИНИТИ, 2017. – С. 317–321.
3. Мостеллер Ф., Тьюки Дж. Анализ данных и регрессия. Вып. 1. – М.: Финансы и статистика, 1982. – 379 с.
4. Тьюки Дж. Анализ результатов наблюдений. Разведочный анализ. – М.: Мир, 1981. – 693 с.
5. Курочкин Е.П. Гарантированное оценивание параметров процессов и систем по ограниченным данным // Техника средств связи. Сер. ТЭУ. – 1986. – Вып.2(19). – С. 35–40.
6. Хампель Ф.Р. Современные тенденции в теории устойчивых статистических процедур // Сб. Математическая статистика и ее приложения. – Томск: Изд-во ТГУ, вып. 6, 1980. – С. 57–59.
7. Тарасенко Ф.П., Черепанов Е.В. Полиграммные оценки линейных функционалов //

Математическая статистика и ее приложения. – Томск: Изд-во ТГУ, вып. 12, 1985. – 279 с.
8. Сюнтюренко О.В., Черепанов Е.В., Щиренко Е.Г. Некоторые вопросы моделирования технико-экономических процессов и систем на основе многомерного анализа фактографической информации по технике и экономике // Материалы IV Всесоюзного симпозиума «Машинные методы обнаружения закономерностей». – Новосибирск: ИМ СО АН СССР, 1983. – С. 19.
9. Сюнтюренко О.В., Симонов О.В., Черепанов Е.В. Некоторые автоматизированные процедуры многомерного анализа технико-экономических данных // Техника средств связи. Сер. ТРПА. – 1985. – Вып. 2. – С. 56–66.
10. Сюнтюренко О.В., Черепанов Е.В. Информатика: анализ данных и эконометрия // Средства связи. – 1986. – № 4. – С. 39–44.
11. Иберла К. Факторный анализ. – М.: Статистика, 1980. – 177 с.
12. Choi S., Ahn J. H., Cichocki A. Constrained projection approximation algorithms for component analysis // Neural Process. Lett. – 2006. – Vol. 24. – P. 53–65.
13. Кривенко М.П. Реконструкция осей главных компонент // Информатика и ее применение. – 2018. – Т. 12, Вып. 1. – С. 71–77.
14. Дюрен Б., Одел П. Кластерный анализ. – М.: Статистика, 1977. – 241 с.
15. Заде Л.А. Размытые множества и их применение в распознавании образов и кластер-анализе // Классификация и кластер. – М.: Мир, 1980. – 47 с.
16. Кофман А. Введение в теорию нечетких множеств. – М.: Радио и связь, 1983. – 297 с.
17. Lynch C. Big Data: How do your data grow? // Nature. – 2008. – Vol.455, 4 September. – P. 28–29.
18. Rodriguez-Mazahua L. et al. A general perspective of Big Data: applications, tools, challenges and trends // Journal of Supercomputing. – 2016. – 72. – P. 3073–3113.
19. Barnabas K. Tannahill, Mo Jamshidi. System of Systems and Big Data analytics – Bridging the gap // Computer and Electricity Engineering. – 2014. – Vol.40, № 1. – P. 2–15.
20. Погода и настроение // Наука и жизнь. – 2014. – № 10. – С. 47.

Материал поступил в редакцию 24.07.18.

Сведения об авторе

СЮНТЮРЕНКО Олег Васильевич – доктор технических наук, профессор, ведущий научный сотрудник Библиотека естественных наук РАН, Москва
e-mail: olegasu@mail.ru

УДК 004.414.2:004.451.25

В.Н. Шведенко, В.В. Шведенко, О.В. Щекочихин

Применение структурного полиморфизма при создании информационных систем процессного управления

Рассматривается создание информационного обеспечения процессного управления производственных систем на основе структурного полиморфизма. Представлен обзор видов полиморфизма, их преимущества и недостатки. Дается новое определение структурного полиморфизма, расширяющее его возможности. Предлагается новый подход к созданию информационных ресурсов и программного обеспечения, где структурный полиморфизм используется как на этапе концептуально-логического проектирования системы, так и на этапе реализации информационного и программного обеспечения.

Ключевые слова: *структурный полиморфизм, процессное управление, проектирование информационных систем, язык манипулирования данными*

ВВЕДЕНИЕ

Процессный подход – один из методов, обеспечивающих управление сложными организационно-техническими системами. Проектирование бизнес-процессов на основе типовых решений не всегда дает возможность решить практические задачи. Для расширения возможностей информационных систем, не заложенных в проект, используются встроенные языки программирования [1].

Информационные системы, как правило, включают в себя несколько встроенных языков программирования для автоматизации задач управления процессами, интеграции, выборки данных из баз [2, 3]. Эти языки программирования классифицируются по:

- парадигме – декларативные, императивные, структурированные, объектно-ориентированные;
- системе типов – динамические, статические, сильно- и слаботипизированные, нетипизированные;
- уровню абстракции – высокого, низкого уровня;
- модели исполнения – компилируемые, интерпретируемые, скриптовые;
- области применения – универсальные, встроенные в прикладные программы, встраиваемые.

Языки для реализации процессного управления строятся на принципах объектно-ориентированной парадигмы. Одним из четырех базовых понятий объектно-ориентированных технологий является полиморфизм, который обеспечивает возможность объектов с одинаковой спецификацией иметь различную реализацию. Общие свойства объектов объединяются в систему, которая носит название «интерфейс» или «класс». Общность имеет внешнее и внутреннее выражение. Внешняя общность проявляется как одинаковый набор методов с одинаковыми именами и сигнатурами (именем методов и типами аргументов, их количеством). Внутренняя общность – одинаковая функциональность методов. Её можно описать интуитивно или выразить в виде закономерностей, правил, которым должны подчиняться методы. В настоящее время полиморфизм совершенствуется как в теоретическом, так и в практическом планах для повышения эффективности исполнения программного кода. Различают такие виды полиморфизма как параметрический, структурный, ограниченный, переопределения, полиморфизм-перегрузка [4–7].

Программирование на основе классов приводит к излишней концентрации внимания на таксономии классов и к росту отношений между ними. При этом в многослойной структуре классов только нижний уровень связан с объектами реального мира. Все вышележащие уровни являются абстрактными и необходимы для увеличения повторяемости использования кода.

В противоположность этому подход на основе прототипов основан на поведении некоторого небольшого количества «образцов», которые затем классифицируются как «базовые» объекты и используются для создания других объектов, что по сути дела является наследованием. Под прототипом понимается фрагмент информационного объекта в многомерном пространстве его состояний. Многие прототип-ориентированные системы поддерживают изменение прототипов во время выполнения программы, т.е. обеспечивают динамический полиморфизм [8].

В настоящей работе предлагаются теоретические основы создания объектно-ориентированного языка на базе прототипов для программирования процессов в информационно-управляющих системах.

ОПИСАНИЕ СТРУКТУРНОГО ПОЛИМОРФИЗМА

Базовыми элементами объектно-ориентированного языка на основе прототипов для программирования процессов в информационно-управляющих системах являются многомерные конструкции информационных ресурсов, представляющие собой набор иерархий прототипов, отражающих объекты реального мира и позволяющие рассматривать эти объекты с различных точек зрения в многомерном пространстве их состояний.

На концептуальном уровне объектно-ориентированного языка рассматриваются свойства объектов через их измерения системой мер и единиц измерения, а на уровне программной реализации такая система представляется в виде типов данных, соответствующих логике вычислений и возможностям аппаратной платформы. Без такой искусственной функции привязки свойств объектов в настоящее время нельзя обойтись. Проблемы типизации данных рассматривались в работах [9–12].

В основе объектно-ориентированного языка при создании моделей управления процессами используются конструкции прототипов, методы работы с процессами, а также устанавливается факт того, что свойства объектов должны описываться мерой, единицами измерений, а тип данных, который может обработать соответствующую меру и единицу измерения на конкретной аппаратной платформе, определяется автоматически в момент исполнения программного кода. Таким образом, язык является нетипизированным.

Набор прототипов, описывающих модель предметной области, образует структуру информационного объекта. На каждом уровне иерархии каждому прототипу соответствует набор объектов реального мира. При этом в зависимости от значений свойств объекта, меры их измерений и наличия определенных событий будет однозначно определяться их информационное наполнение.

События в реальном мире порождают соответствующие события в информационной системе. При наличии связи между реальным объектом и информационной системой через набор событий осуществляется изменение информационного наполнения объектов информационной системы.

Структурный полиморфизм реализуется через набор событий информационной системы, осуществляющих преобразование данных и запись их в информационный объект в соответствии с методами обработки событий с учетом фрагмента структуры иерархии прототипов.

Структурный полиморфизм проявляется во внешней общности методов работы с экземплярами информационных объектов. При этом метод не зависит от внутренней структуры данных. Структурный полиморфизм реализуется как через наследование прототипов, так при перегрузке методов. Он дает возможность представить методы работы со структурами данных в виде одинаковых спецификаций и методов, которые выбираются в зависимости от меры входных параметров.

Структура объекта проявляется в его многомерности. Это означает, что один и тот же объект может рассматриваться в пространстве его состояний. Например, состояние объекта с точки зрения конструктора, с точки зрения технолога, с точки зрения службы снабжения. Каждый из них имеет свою систему мер, которая определяет проекцию объекта в многомерном пространстве: координаты и значения его свойств.

Также система мер позволяет соотносить один объект с другим и оценивать связь между отдельными элементами системы. Мера – это средство формирования измерительной оси для оценки качественных и количественных свойств объектов. Мера может описываться типами данных. Кроме того, мера – это структурная характеристика, потому что она описывает объект на более высоком уровне абстракции.

Например, для производства объекта машиностроения требуются материалы, комплектующие, детали, узлы. В процессе изготовлении изделия задействованы различные службы. Конструктор оценивает изделие с точки зрения его прочности, надежности, мощности, производительности. Технолог анализирует изделие с точки зрения технологии его изготовления, обрабатываемости материалов, а также других технологических параметров. Отдел снабжения рассматривает изделие как заказ металла в тоннах и рублях с учетом логистики.

Существуют базовые меры, определяющие объект в многомерном пространстве его состояний и времени, которые представляются вектором, описывающим траекторию жизненного цикла изделия. Таким образом, в различных точках пространства имеется проекция состояния свойства объекта на определенные меры. Каждая мера характеризуется набором единиц измерения. В ходе изменений состояния объекта каждому значению вектора состояний будет соответствовать свой набор проекций.

Меры могут быть зависимыми и независимыми. Зависимые меры определяются связями между проекциями мер. Например, для движения транспортной единицы – расстояние, скорость, время. Набор мер и соответствующих шкал измерений дает проекцию объекта, которая может соответствовать функциям состояния объекта. Например, функция состояния – обеспечение себестоимости изготовления изделия в заданном диапазоне. Это связано с расходом различных ресурсов – материальных, трудовых, информационных и других.

При человеко-машинной обработке данных для удобства восприятия информационный объект должен быть представлен на плоскости (экранной форме) в виде набора его свойств, мер, единиц измерения и типов представления данных в компьютерной системе.

При автоматической обработке данных методами искусственного интеллекта, таких как искусственные нейронные сети, генетические алгоритмы и т.п., мерность представления прототипа может ограничиваться только вычислительными возможностями информационной системы. Так, например, регрессионный

анализ позволяет получать прогнозные значения в плоских и линейных системах. А при использовании искусственных нейронных сетей многомерность ограничивается слоями сети и точностью прогноза.

Таким образом, структура прототипа определяется методами обработки данных, а его мерность определяется ограничениями этих методов.

МЕТРИЧЕСКАЯ СИСТЕМА

Рассмотрим использование структурного полиморфизма для автоматизации управления процессами на основе метрической системы. Такая система представляет собой набор показателей, организованных в соответствии со структурой процессов и используемых в них информационных объектов для контроля достижения цели управления во взаимосвязанных подсистемах. Метрическая система предполагает наличие оперативного, тактического и стратегического контуров принимаемых решений, настраиваемых в строгом соответствии с установленным деревом целей, настройка параметров (ключевых ориентиров) которого, в свою очередь, зависит от условий микро и макросреды предприятия, а также долгосрочных интересов бизнеса.

В зависимости от контура принимаемых решений метрическая система активизирует ряд ключевых показателей, необходимых для принятия управленческих решений. В соответствии с метрической системой показатели связываются между собой через процессы и осуществляют мониторинг достижения цели управления.

Таким образом, выбирается цель управления (например, минимизация издержек производства, расширение присутствия на рынке, диверсификация производства и т.п.), задается набор процессов, объектов, свойств объектов, связей между объектами, процессами и контурами управления, а также методы работы с объектами и процессами.

Для повышения эффективности управления процессом необходимо увеличивать его информативность за счет усложнения структур информационных объектов и методов работы с ними. Каждое свойство объекта может рассматриваться как показатель исполнения процесса, для чего анализируются изменения его значений или влияние показателя на достигаемые цели управления. Показатель рассматривается как количественное или качественное отражение состояния свойства элемента системы или связи между элементами, которые измеряются и представляются наборами данных. Каждый показатель может быть получен из материальной системы, из другой информационной системы, рассчитан через группу иных показателей или вычислен собственным методом. Метод рассматривается как процесс преобразования данных об объектах, их свойствах и связях между объектами и свойствами объектов.

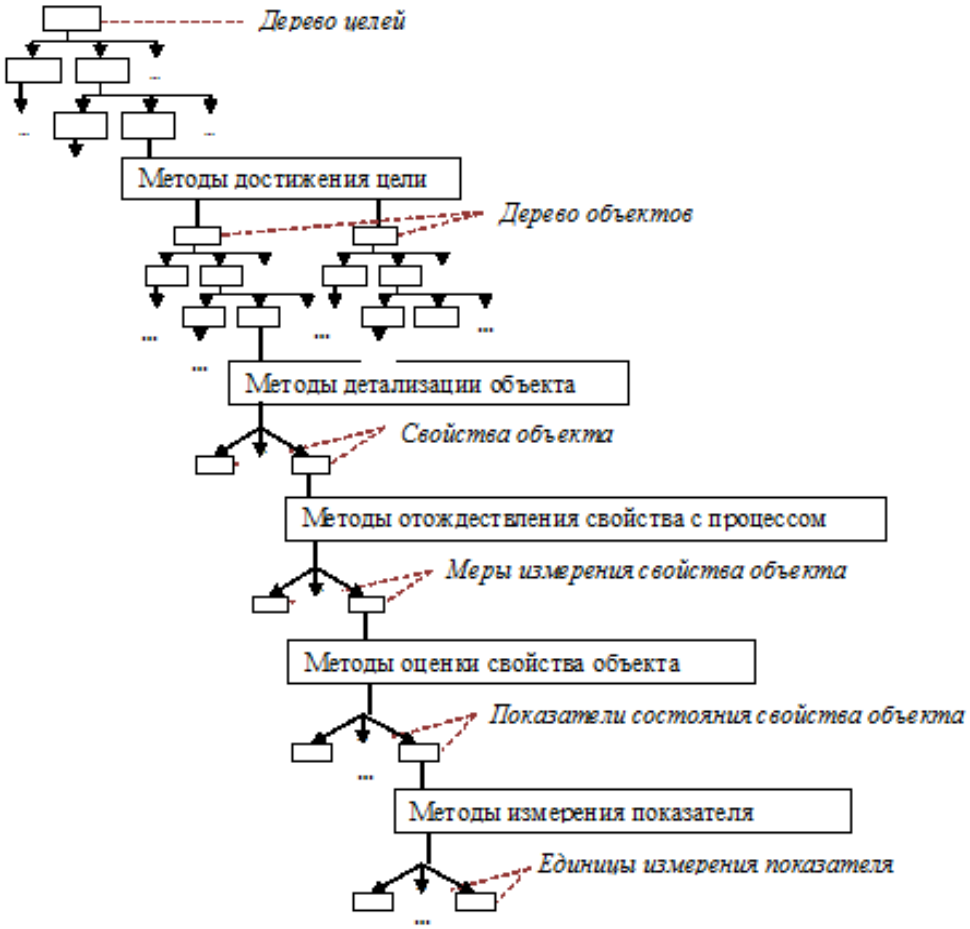

Рис. 1. Схема взаимодействия объектов, процессов и методов в структурном полиморфизме

Информационный объект представляет собой динамическую структуру, развивающуюся во времени. В каждый конкретный момент времени процесс оперирует прототипом – информационным объектом с определенным уровнем детализации.

Полная характеристика объекта и набор его свойств заранее не определены. Поэтому есть версии объекта, содержащие свойства, требующиеся для реализации методов на текущий момент времени. Таким свойствам объекта присваивается статус «Активный». Все прочие свойства, которые могут описывать тот или иной аспект объекта, имеют статус «Латентный». При увеличении количества активных свойств объекта (свойств, состояние которых оценивается, сопоставляется, для которых учитываются изменения их параметров в зависимости от изменения параметров свойств других объектов и т.д.) формируется новый прототип, наследующий свойства предыдущего.

Чем полнее структура прототипа, тем более информативной является модель материального мира, представляющая собой совокупность свойств объектов и связей между ними. Каждое свойство описывается одной или несколькими мерами. Мера определяет группу показателей. Например, мера длины включает набор абсолютных и относительных показателей (протяженность, % преодоленного расстояния от заданного значения, и т.д.). Каждый показатель имеет свою систему единиц измерений, соотносимых по отношению друг к другу. Например, длина может быть измерена по шкалам: микроны, мм, см, дм, м, км.

Значение каждого показателя раскладывается на следующие проекции: фактическое значение (*FZ*), набор нормативных значений *(NZ)*, которые зависят от целей управления, набор абсолютных отклонений значений показателя (*FZ–NZ*), набор относительных отклонений значения показателя (*FZ/NZ*). Пример набора нормативных значений: лучший в мире, лучший в отрасли, нормативное значение, которое должно быть достигнуто через 5, 4, 3, 2, 1 года и т.д.

Показатели связаны между собой через: меры, единицы изменения, процессы передачи и преобразования данных, т.е. значения показателей (рис. 1).

МАТЕМАТИЧЕСКАЯ МОДЕЛЬ СИСТЕМЫ ПРОЦЕССНОГО УПРАВЛЕНИЯ

Для информационного обеспечения логики принятия управленческих решений представлено формальное математическое описание на основе теории множеств и предикатов первого порядка. Для формирования алгебраической системы управления предприятием введены структуры:

- *MK* – множество контуров управления;
- *F* – множество функциональных направлений;
- *PR* – множество лиц или устройств, принимающих решения (ЛПР /УПР);
- *PF* – множество функций УПР/ЛПР;
- *CR* – множество центров ответственности.

Задача управления бизнесом позволяет сформировать дерево целей в соответствии со структурой управления. Структура управления состоит из ЛПР, наделенных функциями и системой показателей. Множество УПР/ЛПР через показатели связано с множеством функций, исполняемых УПР/ЛПР. Множество приложений связано с УПР/ЛПР через исполняемые функции.

Дерево целей может иметь несколько проекций – индикаторов текущего состояния производственной системы в различных аспектах, в которых наряду с нормативными значениями показателей содержатся их фактические значения, величина отклонения и степень значимости отклонения, что является моделью метрической системы.

Проекция дерева целей на каждом контуре управления представляется графом:

$$G_A=(A, T),$$
$$A=\{A^0, A^1, \ldots A^{P-1}\},$$

где A_i – показатель, представляющий кортеж *(ID, Name, NV)*. При этом *ID* – идентификатор показателя; *Name* – наименование показателя; *NV* – нормативное значение показателя; *T* – множество связей между показателями. Каждая связь определяет методы агрегирования показателя.

Индикатор текущего состояния системы представляется графом:

GB = (B, T), где *B* – состояние показателя, представляющее кортеж информации: *(ID, FV, D, SD)*. При этом *ID* – идентификатор показателя из графа *GA; FV* – фактическое значение показателя; *D* – величина отклонения фактического значения от нормативного, *SD* – степень значимости отклонения; *T* – множество связей между показателями.

Представленные алгебраические структуры позволяют описывать взаимодействие контуров метрической системы управления предприятием.

Для описания информационной системы предприятия заданы алгебраические структуры:

A – множество имен элементарных свойств;
B – множество имен объектов,
O – множество объектов системы,
К – множества типов прототипов,
H – множество элементарных свойств,
MR – мера,
BP – множество бизнес-процессов,
L – множество типов связей,
APP – множество приложений, входящих в информационную систему,
P – множество прототипов,
EI – множество единиц измерения,
DT – конечное множество, определяющее абсолютное и относительное значение даты и времени, которые используются в приложениях.

Элементы ранее описанных множеств задают основное множество (Θ) алгебраической системы *U* и множество констант (*C*) сигнатуры (Σ). Описание представленных множеств и констант осуществляется в соответствии с объявленными объектами и процессами предметной области.

Предикат, описывающий уникальный идентификатор прототипа (УИП), имеет вид:

$$uip(O,N,H,O',Y, DT),$$

Предикат определяет свойство из множества *H*, входящее в структуру объекта *O'*, в качестве составного в наборе уникального идентификатора прототипа объекта *O* под порядковым номером *N* из множества констант *Y* в момент времени *DT*.

$Y \subset N$, $y\mu$ – номер свойства, входящего в УИП.

Предикат определен на множестве:

$$T^{u} = O \times N \times H \times O \times Y \times DT =$$
$$\{(o_i, n, h_j, o_\phi, y_\mu, dt) \mid o_i \in$$
$$\in \mathrm{O}:(\exists \mathrm{b}_w \in \mathrm{B}, \mathrm{prop\_name}(o_i, \mathrm{b}_i)),$$
$$\mathrm{n} \in \mathrm{N}, \mathrm{Prototip}(o_i, \mathrm{n}), y_\mu \in N, h_j \in$$
$$\in \mathrm{H}:[(\mathrm{s\_prop}^{*}(o_i, n, h_j) \to \phi = \mathrm{i}) \vee$$
$$\vee \exists \overline{o_r} \in S^{i}:\mathrm{s\_prop}^{*}(o_r, n, h_j)]\}$$

Таким образом, определив модели прототипов, можно рассматривать процессы их взаимодействия посредством приложений в этапах бизнес-процесса (БП).

Далее в модели выделены конструкты, описывающие структуры процессов:

struct_SBP(BP, N, $\overline{O}, \overline{O}, \overline{O}, \ldots$) – предикат, описывающий БП версии *N*.

Предикат определен на множестве:

$$M^{sbp} = BP \times N \times \overline{O} \times \overline{O} \times \overline{O} \times \ldots \times \overline{O} \supset$$
$$\supset \{(bp_j, n, \overline{o_1}, \overline{o_2}, \ldots) \mid bp_j \in BP, n \in N, \overline{o_i} \in \overline{O} \wedge$$
$$\wedge o_i \in OM + F\}$$

Бизнес-процесс представляет собой совокупность прототипов и показателей метрической системы, поэтому предикатом задается соответствие коду bp_j БП набора прототипов $\overline{P_i}$, для которых $p_j \in OM+F$.

Так как процесс имеет этапную структуру, вводится предикат, задающий для *n*-го этапа набор из единственного прототипа, контура управления, лица принимающего решения, соответствующего контура управления, приложений, которые обрабатывают поступающую информацию:

step_SBP(BP,N, N, MK, P, PR, PF, CR, APP, $\overline{O}, \overline{O}, \overline{O}, \ldots$*)*

Множество истинности предиката имеет вид:

$$M^{ssbp} = BP \times N \times N \times MK \times P \times PR \times PF \times CR \times APP \times$$
$$\times \overline{O} \times \overline{O} \times \overline{O} \times \ldots \times \overline{O} \supset \{(bp_j, n, n', \overline{o_1}, \overline{o_2}, \ldots) \mid bp_j \in BP,$$
$$n \in N, n' \in N, \exists! o_i \in OM \wedge o_q \in F, i \neq q\}$$

ТЕОРЕТИЧЕСКИЕ ПОЛОЖЕНИЯ НА ОСНОВЕ ПРОТОТИПОВ ДЛЯ ПРОГРАММИРОВАНИЯ ПРОЦЕССОВ

Полученная многоосновная алгебраическая система позволяет создавать основы языка описания прототипов и методов работы с ними при программировании задач процессного управления. Язык имеет два компонента: язык для описания структуры прототипов и язык для выборки и обновления данных в соответствии с метрической системой.

Первый компонент языка включает функции для построения структуры прототипов, и структуры процессов. Наиболее важные функции языка описания структур данных и определяющие их предикаты, представлены в табл. 1.

Имеется ряд системных объектов, их свойств и процессов, обеспечивающих работу системы. Например, приложение, его версия, элементы управления приложением, и др.

Второй компонент языка содержит описание функций манипулирования данными, которые поддерживают взаимодействие приложений в рамках исполняемого процесса.

Базовые функции добавления, извлечения данных представлены в табл. 2.

На основе предложенных теоретических выкладок разработаны интерпретатор и среда разработки прототипов информационных объектов для создания информационно-управляющих систем (рис. 2).

Таблица 1

Базовые функции языка описания структур данных

Описание функции	Предикат
Создание типа прототипа	*Prot_type(O,K)*
Создание меры свойства объекта	*Prop_м(H,T)*
Назначение имени объекта	*obj_name(O,B)*
Назначение имени свойства	*Prop_name(H,A)*
Создание объекта	*uiio(O,N,H,O,Y, DT)*
Создание прототипа	*prototip(O,N)*
Добавление в структуру прототипа нового свойства	*s_prop(O,N,H)*
Добавление в структуру прототипа другого объекта	*inserted_sd(O, N,* $\overline{O}$*, L)*
Создание этапа бизнес-процесса	*step_SBP(BP,N, N, APP,* $\overline{O}, \overline{O}, \overline{O}, \ldots$*)*
Создание бизнес-процесса	*struct_SBP(BP, N,* $\overline{O}, \overline{O}, \overline{O}, \ldots$*)*

Таблица 2

Базовые функции языка манипулирования данными

Название функции	Аргументы	Значение функции	Обозначение
Создание экземпляра прототипа	Идентификатор прототипа объекта Ключ, определяющий местоположение свойства в структуре прототипа Идентификатор приложения Значение свойства	Идентификатор экземпляра прототипа объекта либо код ошибки	InstanceProt (Prot, PropKey-InTree, APP, Value)
Запуск бизнес-процесса в момент времени *t*	Идентификатор бизнес-процесса Идентификатор экземпляра прототипа объекта Регламентное время	Идентификатор экземпляра бизнес-процесса Идентификатор экземпляра прототипа объекта Фактическое время получения данных либо код ошибки	StartBusi-nessProcess (BP, IP, DT)
Извлечение данных	Идентификатор прототипа объекта Идентификатор экземпляра прототипа объекта Ключ, определяющий местоположение свойства в структуре объекта Временной интервал ограничения	Поток данных либо код ошибки	GetData (Prot, IP, PropKeyInTree, DTbegin, DTend)

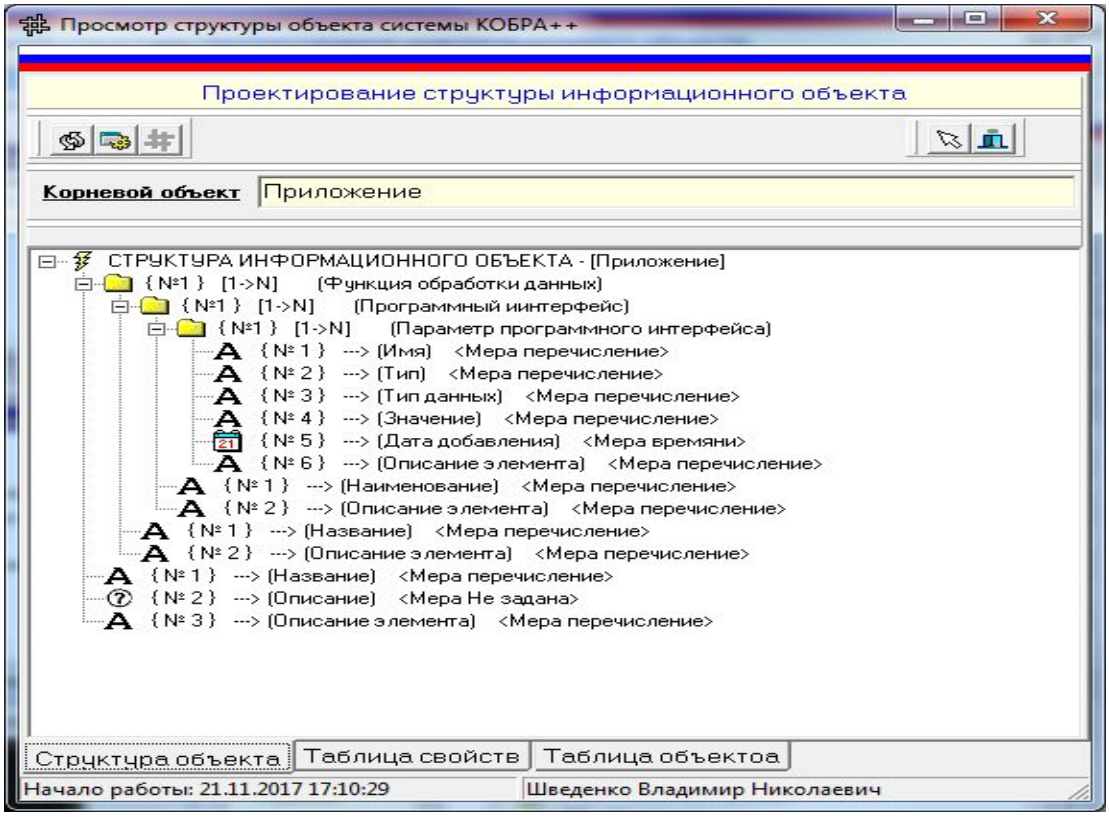

Рис. 2. Окно проектирования прототипа информационного объекта

ЗАКЛЮЧЕНИЕ

В настоящей статье изложен новый подход к созданию языка описания информационных ресурсов информационно-управляющих систем на основе структурного полиморфизма. Структурный полиморфизм позволяет осуществлять динамическое преобразование набора структур иерархий прототипов информационных объектов, интегрированных в систему процессного управления. События и методы сбора, обработки, преобразования, записи и хранения данных определяются процессами, протекающими в организационно-технической системе.

Структурный полиморфизм реализован в метрической системе управления процессами, которая представляет собой набор показателей, организованных в соответствии со структурой процессов и используемых в них информационных объектов для обеспечения контроля достижения цели управления во взаимосвязанных подсистемах. Каждый показатель характеризуется мерой и единицами измерений.

Метрическая система управления создается на этапе концептуального проектирования информационно-управляющей системы. Здесь же задаются методы работы с показателями ее состояния.

Разработана математическая модель процессного управления на основе метрической системы, которая позволяет получить структуру прототипов информационных объектов и связать их со свойствами организационно-технической системы. Представлены фрагменты объектно-ориентированного языка программирования процессов в информационно-управляющей системе для манипулирования данными.

СПИСОК ЛИТЕРАТУРЫ

1. ARIS scripting tutorials. – URL: http://www.aris-community.com/users/eva-klein/2010-04-27-aris-scripting-tutorials (accessed 1 May 2018).
2. Pierce B.C. Types and Programming Languages. – L: MIT Press, 2002. – 645 p.
3. Lammel R., Visser J. Typed Combinators for Generic Traversal // Practical Aspects of Declarative Languages: 4th International Symposium, 2002. – p. 153.
4. Booch G. Object-Oriented Analysis and Design with Applications. 3rd ed. – Upper Saddle River, N.J.: Addison-Wesley, 2007. – 543 p.
5. Strachey C. Fundamental Concepts in Programming Languages // Higher-Order and Symbolic Computation. – 2000. – Vol. 13, Issue 1/2.
6. Garrigue J. Simple type inference for structural polymorphism // The Ninth International Workshop on Foundations of Object-Oriented Languages. – Portland, Oregon, 2002.
7. Garrigue J. A certified implementation of ML with structural polymorphism // Proc. Asian Symposium on Programming Languages and Systems. Vol. 6461 of SpringerVerlag LNCS., Shanghai, 2010. – P. 360–375.
8. Litvinov V. Constraint-Bounded Polymorphism: an Expressive and Practical Type System for Object-Oriented Languages: a dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy. – Washington: University of Washington, 2003.
9. Millstein T. Modular Typechecking for Hierarchically Extensible Datatypes and Functions // Proceedings of the seventh ACM SIGPLAN international conference on Functional programming, Pittsburgh, PA. – 2002. – Vol. 37, Iss. 9. – P 110–122.
10. Neubauer M. Functional Logic Overloading // Conference Record of the 29th Symposium on Principles of Programming Languages, Portland, Oregon. – 2002. – P. 233–244.
11. Pottier F. A Versatile Constraint-Based Type Inference System // Nordic Journal of Computing. – 2007. – № 4. – P. 312–347.
12. Tobin-Hochstadt S. The design and implementation of typed scheme // Symposium on Principles of Programming Languages, San Francisco, California. – 2008. – P. 395-406.

Материал поступил в редакцию 08.06.18.

Сведения об авторах

ШВЕДЕНКО Владимир Николаевич – доктор технических наук, профессор, ведущий специалист ВИНИТИ РАН, Москва
e-mail: vv_shved@mail.ru

ШВЕДЕНКО Валерия Валериевна – кандидат экономических наук, системный аналитик ООО "РЕГУЛ+", Москва
e-mail: vv_shved@mail.ru

ЩЕКОЧИХИН Олег Владимирович – кандидат технических наук, доцент, инженер-программист ООО "РЕГУЛ+"
e-mail: slim700@yandex.ru

УДК 004.75:336.743–028.27

А.А. Рязанова

Технология распределенных реестров: накопленный опыт и потенциал

Рассматривается применение технологии распределенных реестров, положительные и отрицательные эффекты, опыт использования в различных областях.

Ключевые слова: *распределенные реестры, блокчейн, хеш-функция, алгоритм консенсуса, платформы, криптовалюты, биткоин*

ВВЕДЕНИЕ

Уровень развития науки и образования в области информационных технологий на сегодняшний день не только не отвечает требованиям современного общества, но и ощутимо не соответствует его перспективам и трендам. Это связано с совокупностью факторов – начиная от стабильного технологического отставания и заканчивая отсутствием осознания того факта, что современное информационное общество, включая все информационные процессы, происходящие в нем, стало приобретать квантомеханическую природу, следовательно, регулируется принципом неопределенности Гейзенберга. В отношении информационных систем и процессов он может быть сформулирован как активное влияние участников информационного обмена на сами информационные процессы.

Это положение в явном виде мы наблюдаем при становлении децентрализованных систем, зависящих от функционирования распределенных реестров. Влияние немногочисленных участников порождает лавинообразный саморегулирующийся процесс, охватывающий все области деятельности и постепенно вовлекающий крупнейших игроков, которые не только регулируют и направляют процессы, но также сами становятся заложниками созданных и запущенных ими процессов. Ярким примером может служить проект *IBM – HyperLedger Fabric*.

Это можно назвать фактором непроизвольного рефлексивного воздействия участников системы на процессы, происходящие в ней. Такой фактор существенно усложняет оценку и прогнозирование развития современных информационных технологий, требуя включения в анализ новых аспектов, поскольку участники системы также представляют собой системные целостности, имеющие собственные структуры и процессы и вступающие во взаимодействие с другими системными целостностями. Адекватно оценить все перспективы перехода на децентрализованные системы не представляется возможным, однако накопленный эмпирический опыт позволяет выявить тенденцию к смене парадигмы с централизованной модели на контролируемую децентрализованную.

Особенностью системы блокчейна криптовалюты биткоин как первой крупнейшей децентрализованной системы является то, что управление валютой производится через распределенную компьютерную сеть без участия центрального органа [1]. Поэтому лежащая в основе криптовалюты биткоин технология блокчейн фактически рассматривается как инновационная технология, которая благодаря своим свойствам сможет существенно трансформировать область цифровых валют, а также выйти за ее пределы.

Согласно прогнозу Всемирного экономического форума, сделанному в 2017 г., фундаментальной технологией для Четвертой индустриальной революции станет технология распределенных реестров, а за следующие одиннадцать лет доля транзакций, проводимых через системы блокчейн, будет составлять 10% мирового ВВП. Но, несмотря на множество проектов и практических инициатив в области блокчейна, а также на растущее внимание общественности издано крайне мало научных работ на эту актуальную и требующую системного подхода тему.

Технология распределенных реестров уже утвердилась в области финансов. Изучены технические и возможные правовые аспекты ее применения. Однако системные исследования практически отсутствуют. В существующих публикациях авторы, как правило, ограничиваются рассмотрением частных случаев применения, отказываясь от анализа системных категорий более высокого уровня.

Все эти факторы вместе с неопределенностью в перспективах и технологическим отставанием, а также недостаточно активным участием в мировых процессах дает России возможность найти собственный путь. Осмысление совершенных ошибок и достаточно мощный теоретический потенциал российской науки неоднократно давали импульс научно-технологической революции глобального масштаба, подобно системному прорыву в области термоядерного оружия, совершенному в ситуации объективного отставания по атомному проекту.

В области криптовалют и распределенных реестров российские ученые уже сейчас предложили целый ряд технологий, недоступных крупнейшим игрокам и реализующих ожидания как участников

процесса, так и потенциальных его регуляторов. Это изолированные криптовалютные сети, авторизованные и надежно защищенные криптовалюты, развитие идей и технологий распределенных реестров (блокчейнов) в двумерные блок-матрицы.

Эти положительные тенденции нуждаются в организационной поддержке и институализации. Однако современная академическая наука практически не содействует проведение работ в области технологий распределенных реестров. Исключением является Всероссийский институт научно-технической информации РАН (ВИНИТИ РАН), создавший Центр развития криптовалют и цифровых финансовых активов.

Все это свидетельствует о том, что для современной науки важны изучение и анализ накопленного опыта с целью представить перспективы и потенциальные области применения технологии распределенных реестров. В ходе проводимых исследований осуществляется комплексный анализ основ технологии, структурных шансов и рисков, связанных с ее применением, создаются различные модели, описываются принципы работы блокчейн-систем. При этом блокчейн рассматривается также в качестве информационно-инфраструктурного решения.

ОСНОВЫ ТЕХНОЛОГИИ РАСПРЕДЕЛЕННЫХ РЕЕСТРОВ

Распределенные реестры, в частности, блокчейн определяется как тип базы данных, в которой записи группируются в блоки. Эти блоки соединяются в неразрывную цепь в хронологическом порядке, который обеспечивается хеш-значением каждого последующего блока, зависимого от хеш-значения предыдущего блока. Каждый блок содержит записи, созданные с момента его добавления [2]. В случае криптовалюты блок биткойна включает в себя историю всех выполненных транзакций с момента его добавления в сеть. Принцип формирования записей в блоки отличает блокчейн от иных распределенных реестров, в которых записи вносятся непрерывно.

Системы управления любыми распределенными реестрами описываются как системы консенсуса, так как алгоритмы консенсуса обеспечивают согласованность между узлами сети. Такие системы основаны на криптографии и принципах *Peer-to-Peer* (*P2P*) для обеспечения консенсуса всей сети.

Из приведенных определений видно, что системы блокчейн являются распределенными системами, их можно охарактеризовать несколькими свойствами: во-первых, узлы таких систем взаимодействуют и синхронизируются друг с другом; во-вторых, отказ отдельного узла не сказывается негативно на работе системы в целом. Кроме того, каждый сетевой узел содержит в себе информацию всей системы, поэтому отказ одного или нескольких узлов не может привести к полной или частичной потере данных. В системах, основанных на технологии блокчейн, вся цепочка блоков хранится в каждом узле [2].

Важнейший принцип *Peer-to-Peer* – прямой обмен между узлами в пиринговой (децентрализованной) сети, т.е. отсутствие центрального узла для координации (каждый узел является и клиентом, и сервером). Это обусловливает необходимость применения криптографии для цепочек блоков. Как криптографические методы, так и алгоритмы консенсуса, с помощью которых сетевые узлы координируют состояние системы, являются основой технологии блокчейн.

Принципы технологии блокчейн могут быть рассмотрены на примере платежной системы биткойн, которая является ее первым практическим применением [3]. Криптовалюты уже существовали на протяжении двадцати лет до появления биткойна. Однако их существенным недостатком было наличие посредника в платежных операциях, поэтому первые попытки были обречены на провал [4]. Этот недостаток был устранен с началом применения технологии блокчейн, так как криптовалюты на базе этой технологии являются цифровыми валютами, при обращении и управлении которыми отсутствует посредник. При этом децентрализованная пиринговая компьютерная сеть выполняет и проверяет транзакции [5].

Блокчейн биткойна является реестром всех проведенных транзакций, сформированным в хронологическом порядке. Копия этого децентрализованного реестра хранится у каждого участника сети. Криптографические алгоритмы и децентрализованное управление позволяют безопасно проводить транзакции. Функционирование системы биткойн-блокчейна основано на двух принципах криптография с открытым ключом и цифровые подписи, а также криптографические хеш-функции.

Концепция криптографии с открытым ключом была разработана в 1976 г. [6]. Алгоритм предполагает генерацию пары математически связанных ключей – открытого и закрытого ключа. Ключ парной связи позволяет шифровать информацию таким образом, чтобы она была доступна только двум пользователям. Для этого отправитель снабжает сообщение закрытым ключом, который известен только ему, и отправляет подписанное сообщение получателю. С помощью открытого ключа получатель устанавливает авторство документа и его неизменность после подписания.

Цифровая подпись позволяет достичь трех целей: только отправитель знает секретный ключ, поэтому он может доказать свое авторство; отправитель не может отрицать, что именно он подписал сообщение; благодаря асимметричному шифрованию не остаются незамеченными изменения, что обеспечивает целостность содержимого блока. Блокчейн биткойна использует криптографическую хеш-функцию. Алгоритм хеширования – это алгоритм, который преобразует строку любой длины в строку фиксированной длины (хеш-значение, или хеш-функция).

Хеш-функция является детерминированной функцией, т. е. одни и те же входные данные всегда будут иметь одно и то же хеш-значение, а любое изменение входных данных приводит к его изменению. Особенностью криптографических хеш-функций является невозможность как восстановить зашифрованные данные по хеш-значению, так и найти другие данные, которые имеют то же хеш-значение.

Как уже указывалось, одним из главных достижений в системе биткоин является механизм консенсуса, который позволяет принять в сеть новый блок. Все узлы сети достигают консенсуса о принятии нового блока посредством алгоритма *Proof-of-Work* (доказательство работы), представляющего собой набор требований к сложным компьютерным вычислениям, которые необходимо провести, чтобы создать блок и добавить его в блокчейн. Цель алгоритма – как верификация сделки, которая позволяет избежать двойного расходования, так и создание новой единицы биткойна.

Ключевой особенностью является то, что вычисления должны быть умеренно сложными для узла сети, но достаточно простыми для сети в целом. Это достигается с помощью криптографических алгоритмов. Каждый узел сети пытается найти решение задачи первым; при этом найти его фактически можно лишь методом прямого перебора, поэтому для успешного решения требуется множество попыток. Вновь сгенерированный блок является вознаграждением за трудоемкие вычисления. При этом транзакции случайным образом группируются в блок, а участники сети (узлы) подтверждают легитимность транзакций в блоке. Транзакции добавляются к блоку, который теперь доступен любому участнику системы (узлу, внешнему пользователю и т.д.)

Сложность создания блоков в сети биткоин обусловлена требованием высокой распределенной вычислительной мощности сети. Чем выше мощность, тем больше число вычислений, которые нужно провести, чтобы создать новый блок [1]. Метод также повышает стоимость создания блока, побуждая участников наращивать эффективность своих систем, чтобы сохранить положительную динамику.

Алгоритм *Proof-of-Work* используется не только в системе биткоина, но и во многих других криптовалютных системах, основанных на блокчейне. В каждом случае конкретные особенности *Proof-of-Work* могут немного отличаться, поскольку создаются индивидуально для каждого блокчейна.

ПРЕИМУЩЕСТВА И НЕДОСТАТКИ КРИПТОВАЛЮТ, БАЗИРУЮЩИХСЯ НА ТЕХНОЛОГИИ БЛОКЧЕЙН

Отметим следующие преимущества и недостатки криптовалют по сравнению с государственными валютами.

К преимуществам относятся:

- делимость на достаточно малые единицы;
- повышенный уровень приватности, так как отсутствует привязка к банковскому счету и не требуется удостоверение личности;
- минимальные транзакционные издержки и временные затраты благодаря отсутствию посредников;
- невозможность фальсификации криптомонеты, которая следует из свойств блокчейна;
- более высокий уровень надежности благодаря отсутствию посредников.

Среди недостатков можно выделить:

- значительные колебания обменного курса криптовалют; однако, волатильность криптовалют имеет тенденцию к снижению;
- необратимый характер транзакций;
- анонимность незаконных операций;
- возможные хакерские атаки, так как цифровые активы находятся в электронных онлайн-кошельках.

КАТЕГОРИЗАЦИЯ БЛОКЧЕЙНА

Устойчивая категоризация по областям применения начала появляться в последние годы в результате того, что технологии блокчейн стали использоваться не только в сфере цифровых валют, но и в других областях. При этом часто выделяют три фазы категоризации [1]. Первая фаза включает в себя валюты и связанные с ними приложения в сфере финансовых услуг. Ко второй фазе относятся приложения в области экономики и финансов, такие как, например, смарт-контракты, которые описаны ниже. В третьей фазе разрабатываются приложения, выходящие за рамки финансов и рынков, например, блокчейн-приложения в государственном секторе.

Существенный вклад в развитие категоризации блокчейна внес Уильям Мугаяр [7], согласно которому среда применения технологии блокчейн включает в себя инфраструктуру и платформы, связующее программное обеспечение, приложения и сопутствующие услуги. Это категории верхнего уровня, которые» и далее по тексту, в свою очередь, могут быть разделены на несколько подкатегорий. При этом разделение на подкатегории строится, исходя из функциональных характеристик.

1. Инфраструктура и платформы. Инфраструктура информационных технологий (ИТ-инфраструктура) представляет собой сочетание определенных непрерывно используемых ИТ-ресурсов, которые формируют основу для применения технологии. Указанные ИТ-ресурсы включают в себя как аппаратные ресурсы, так и операционные системы, сети и информационные технологии, а также основную информацию и приложения для обработки данных. Применительно к блокчейну в эту категорию входят, в том числе, собственно блокчейн-системы, мультиплатформы для разработки и внедрения различных приложений, которые базируются на блокчейн-системах, а также специализированное аппаратное обеспечение для реализации различных алгоритмов.

Мультиплатформенность обеспечивает возможность разработки инновационных производных финансовых инструментов, которые могут найти применение в экономическом или общественном кон-тексте [1]. Такие платформы представляют собой приложения, связанные для выполнения определенной задачи. В технологии блокчейн они предназначены для разработки и внедрения децентрализованных блокчейн-приложений.

2. Связующее программное обеспечение – это универсальные сервисы, являющиеся объединяющим звеном между платформами и приложениями. Они выполняют требования большого числа приложений. Например, к связующему ПО относятся смарт-контракты, реализуемые на мультиплатформах криптовалют, прикладные программные интерфейсы для реализации приложений в блокчейн-системах.

3. Приложения исполняются на базе существующей инфраструктуры и связующего ПО. В то время как концепция смарт-контрактов относится к сервисам связующего ПО, специализированные частные реализации смарт-контрактов являются приложениями. Примером применения может служить регистрация цифровых и материальных объектов в распределенном реестре. В этом случае система формирует цифровые документы, содержащие идентификационные данные объекта и владельца и хранит их в зашифрованном виде, т.е. в системе блокчейн хранятся не документы, а хеш-функции от содержимого блоков.

Проблема, которая негативно сказывается на реализации таких приложений, заключается в ограниченной масштабируемости открытых блокчейн-систем, что осложняет их применение в областях с большим количеством объектов, подлежащих учету. Кроме того, затруднена точная идентификация объекта, поэтому для включения объекта в блокчейн необходимо участие доверенной стороны.

4. Вспомогательные услуги. В качестве четвертой категории в среде применения технологии блокчейн представляются вспомогательные услуги. Сюда относятся поставщики данных, сспециализированные средства массовой информации или отраслевые инвесторы. Вспомогательные услуги предоставляются, например, интернет-сайтом *coingecko.com*, который представляет обзор, сравнительный анализ, показатели рыночной капитализации и прогноз развития различных криптовалют.

Категории блокчейна приведены на рисунке.

ОБЛАСТИ ПРИМЕНЕНИЯ ТЕХНОЛОГИИ РАСПРЕДЕЛЕННЫХ РЕЕСТРОВ

Области применения технологии распределенных реестров не ограничиваются только сферой финансов. Технология получает широкое распространение во многих других сферах деятельности человека.

Блокчейн служит средством реализации смарт-контрактов (протоколов транзакций, отвечающих за исполнение условий договора). В качестве входящих данных используется информация из внешнего источника, которая запускает исполнение протокола, исходя из установленных в смарт-контракте правил.

Соответствующие детали договора в режиме реального времени записываются и сохраняются в цепи блокчейн по определенному адресу. Смарт-контракты являются средством автоматизации действий сторон договора. При этом контракты при помощи соответствующих алгоритмов могут исполняться, верифицироваться, прекращаться, а переводы средств, передача прав, как и управление правами, осуществляются прозрачно, безопасно, с сохранением в цепи блокчейн в неизменном виде. Невозможность отменить действие при определенных условиях может расцениваться как преимущество. Автономное исполнение контракта исключает атаки и вмешательство извне, но при этом не исключает возможную зависимость его исполнения от наступления внешних событий. Даже если стороны договора не доверяют друг другу, смарт-контракты могут обеспечить честный обмен между ними без участия доверенного посредника.

Категории блокчейна

Однако использование смарт-контрактов сопряжено с определенными рисками. Такие контракты исполняются программой, а не традиционным способом, поэтому к ним не применимо договорное право и законодательство о защите прав потребителей. Следовательно, не определены случаи наступления правовой ответственности при нарушении условий этих контрактов. Кроме того, возможна нехватка специалистов для внедрения технологии смарт-контрактов, что может свести на нет преимущества от использования технологии в масштабах целых отраслей. Еще одна проблема заключается в том, что пока смарт-контракт оперирует данными и другими элементами внутри блокчейн-платформы (например, мы говорим только об операциях с криптовалютой), можно говорить об определенной безопасности. Но если попытаться расширить сферу применения распределенных реестров, то неизбежное взаимодействие с внешней средой приводит к возникновению угроз безопасности различного рода. Формально утверждение о небезопасности взаимодействия с внешними источниками данных сформулировано в работе [8].

Распределенные реестры применяются в сфере управления данными для ведения записей и их фиксации [9]. Такие записи обладают метками времени и хранятся в распределенных реестрах, что позволяет снизить стоимость и сложность управления ими.

Технологию распределенного реестра также используют в решениях, предназначенных для идентификации и подтверждения прав доступа [9]. Принцип идентичности в системе блокчейн может предоставить пользователям возможность более надежно контролировать доступ к их персональным данным и управлять степенью их открытости для других.

Блокчейн позволяет подтверждать и фиксировать право авторства и научный приоритет с возможностью их обнародования через заданный промежуток времени [10]. С помощью уникальных идентификаторов и цифровых сертификатов для подтверждения авторства и подлинности можно получить и сохранить приоритет своего изобретения. Кроме того, возможно создание механизма передачи права владения от правообладателя к покупателю.

Технология блокчейн применяется для совершения транзакций и хранения любых форм денег, информации о товарах или сырье [2]. Юридические и физические лица могут перевести средства на свой аккаунт с помощью банковского перевода, дебетовой, кредитной карт, или со своих биткойн-кошельков.

Применительно к сфере энергетики [11] технология позволяет в реальном времени измерять уровень выработки и потребления электроэнергии и других видов энергоресурсов.

Технологию блокчейн со встроенной криптографией целесообразно использовать для анонимного голосования [10], что гарантирует точность и достоверность результатов, а также невозможность их подтасовки.

Блокчейн может быть применен для повышения прозрачности и целостности политических систем [9]. В частности, существует целая международная виртуальная нация, в которой есть свои граждане, послы, партнеры и присутственные места по всему миру. Присоединиться к ней может каждый желающий без каких-либо ограничений.

Что касается решений для эффективного управления внутри организаций, то для этой цели создаются специализированные сервисы, основанные на технологии блокчей, которые автоматизируют процесс управления компанией.

Существуют блокчейн-платформы в области интернета вещей [2], нацеленные на улучшение практики потребления. В этом случае система хранит идентификационные данные физических предметов со встроенными микрочипами. Это позволяет создавать безопасные и совместимые со множеством других систем цифровые идентификаторы, что открывает возможности для новых механизмов взаимодействия с потребителем, основанные на его близости к предмету. Существует ряд приложений для крупномасштабного умного управления промышленными системами и оборудованием. В основе разработок лежат принципы децентрализации, криптографической защиты и автономности.

Одной из первых стран, использующих с 2013 г технологию блокчейн для хранения данных в базах электронного правительства, является Эстония. При этом предпосылки к использованию технологии создавались с начала 2000-х гг., когда была официально введена цифровая подпись и децентрализованная система хранения данных (*X-Road*), в которой все административно-территориальные единицы имеют доступ ко всем данным.

Правительство Эстонии использует инфраструктуру безключевой подписи (*Keyless Signature Infrastructure – KSI*) на основе блокчейна *Guardtime* для проверки подлинности данных в электронных реестрах. Данные сохраняются в распределенном реестре, что обеспечивает их целостность.

Эстония предлагает своим гражданам сотни инновационных электронных услуг, в том числе электронную подачу налоговой декларации, оформление завещания, подачу заявки на получение государственных субсидий, оформление медицинских рецептов.

Компании имеют возможность отправлять ежегодные отчеты в режиме онлайн, подавать заявки на лицензирование. Создание и регистрация в госреестре юридического лица может производиться в течение нескольких минут. Работники Администрации используют систему для шифрования документов, безопасного обмена данными, подписания разрешений. В кабинете министров для организации заседаний используется децентрализованная база данных [12].

Таким образом, опыт использования технологии распределенных реестров, в частности блокчейна, позволяет выделить следующие свойства информационной блокчейн-инфраструктуры:

- открытость: в открытых блокчейн-системах любой участник может реализовать собственный продукт или приложение;
- гетерогенность: технология блокчейн применима в различных приложениях и областях. Для различных целей применяются свои блокчейн-структуры;
- адаптируемость: дальнейшее развитие технологии предполагает ее применение для множества приложений;

- рекурсивность: вокруг блокчейнов сложились определенные рекурсивные связи, в частности, в виде так называемых сайдчейнов, сопряженных с родительскими блокчейнами, а также в виде иных приложений, работающих на криптовалютных платформах;
- неконтролируемость: блокчейн-системы – по определению распределенные системы, в них нет доверенного центра, а текущий статус системы актуализируется участниками при помощи механизма консенсуса;
- распределенность: блокчейн содержится во всех узлах, входящих в сеть.

ПРЕИМУЩЕСТВА И НЕДОСТАТКИ СТРУКТУРЫ БЛОКЧЕЙН-СИСТЕМ

Структура блокчейн-систем имеет преимущества и недостатки, которые зависят от области применения технологии. Необходимость шифрования данных обусловила применения процедуры детального контроля доступа. Блокчейн-системы считаются анонимными и открытыми, так как адрес представлен открытым ключом, однако на практике их можно назвать псевдоанонимными – открытый ключ служит псевдонимом [13]. Хеширование также является неотъемлемой частью функционирования блокчейн-систем и обеспечивает неизменность содержащихся в блоках данных.

Свойства распределенной пиринговой сети в комбинации с механизмом консенсуса для подтверждения статуса сети гарантируют отказоустойчивость системы и доступность данных, т.к. во всех узлах действует механизм верификации электронных подписей. В блокчейне криптовалюты с помощью алгоритма консенсуса также решается проблема двойного расходования средств. В каждом блоке проходит проверка легитимности транзакций, поэтому не требуется доверие между узлами сети.

Процессы в сети проходят по соответствующему программному коду, что обеспечивает технологическую целостность. История транзакций открыта для всех узлов и это предопределяет высокий уровень прозрачности системы.

Кроме того, технология блокчейн имеет следующие преимущества:

- защита больших объемов данных с помощью шифрования и управления доступом;
- возможность сбора и анализа больших объемов данных от многих предприятий;
- более простая верификация точек доступа к данным;
- автоматическое обнаружение уязвимостей в цепочках поставок, платежных операциях и других бизнес-процессах;
- сокращение затрат на ИТ-инфраструктуру;
- снижение расходов на внутренние и внешние финансовые транзакции и управление;
- возможность оперативного представления финансовой отчетности.

Вместе с тем технология блокчейн обладает существенными недостатками. Помимо высоких затрат на электроэнергию при использовании алгоритма *Proof-of-Work*, существуют ограничения пропускной способности и вычислительных ресурсов некоторых узлов, что является критичной проблемой при значительном увеличении числа транзакций и участников [14]. Несовместимость блокчейн-моделей приводит к определенным проблемам, связанным со взаимодействием с открытой системой, которые частично решаются созданием сайдчейнов.

Также существует риск утери или кражи закрытого ключа. В этом случае внесение несанкционированных изменений может иметь серьезные последствия, так как отменить транзакцию невозможно. При успешно проведенной атаке возможны разного рода манипуляции, позволяющие осуществить повторный запуск алгоритма *Proof-of-Work*, начиная с первого измененного блока, что означает двойное расходование средств и создание дополнительной единицы криптовалюты.

Среди недостатков также следует указать:

- недостаточную масштабируемость;
- низкую скорость передачи данных;
- ограниченное дисковое пространство;
- непростое управление правами доступа;
- необходимость наличия высокой скорости Интернет-соединения;
- сложная интеграция с существующими в организации системами.

Подводя итог, можно выделить следующие значительные препятствия на пути внедрения технологии блокчейн:

- недостаточно высокий уровень современной техники ограничивает масштабируемость блокчейн-систем и обуславливает невозможность минимизировать транзакционные риски (внесение записей в распределенный реестр необратимо), а также не позволяет успешно внедрять блокчейн-системы;
- отсутствует законодательное регулирование применения смарт-контрактов;
- не разработаны необходимые стандарты применения технологии;

При создании блокчейн-систем должны учитываться всевозможные факторы, оказывающие влияние на успех реализации, такие как возможность создания пиринговой сети, прозрачная и воспроизводимая история действий. При этом преимущества смарт-контрактов для автоматизации процессов могут в полной мере проявиться в том случае, если каждое действие процесса и его следствия имеют четкое описание.

Несмотря на оптимистичные взгляды, связанные с применением технологии блокчейн в различных областях, потребуется достаточно много времени, чтобы в полной мере оценить эффект от ее применения. На сегодняшний день можно утверждать, что в силу своей универсальности технология заслуживает пристального внимания и проведения системных исследований с последующим созданием блокчейн-лаборатории для более детального и системного изучения ее практической реализации.

Кроме того, целесообразно создание федеральной службы по регулированию обращения цифровых активов и использования распределенных реестров, в функции которой будет входить утверждение нормативов (правил) в указанной области, мониторинг работы организаций в сфере использования цифровых

активов, аналитика в области движение криптовалют, доведение ее до заинтересованных ведомств, курирование работы операторов распределенных реестров.

Основной аргумент в пользу создания такой службы – потребность в решении принципиально новых вопросов и проблем, возникающих при обращении цифровых активов и использовании распределенных реестров, которые невозможно локализовать в рамках одного или нескольких ведомств, а также указанные выше причины методологического плана, которые затрудняют работу существующих ведомств, находящихся в рамках сложившихся принципов и технологий принятия решений. Подконтрольность государству указанной деятельности позволит не только не допустить полулегальные операции, но и избежать дисбаланса в сфере финансов.

В связи с этим необходимо создание технологической платформы для цифровой трансформации, удовлетворяющей следующим фундаментальным требованиям:

- доверенность (универсальная открытая структура и открытый программный код) [4];
- использование национальных стандартов криптографической защиты информации;
- универсальность (возможность использования технологии в различных областях деятельности: в сфере финансов, смарт-контрактов, для фиксации научного приоритета и подтверждения авторского права, для обеспечения функционирования информационных систем в государственном секторе и др.).

ВЫВОДЫ

Для современной науки крайне важно изучение накопленного опыта для того, чтобы представлять перспективы и потенциальные области применения технологии распределенных реестров. В статье был проведен анализ основ технологии, структурных шансов и рисков, связанных с ее применением, описаны принципы работы блокчейн-систем. При этом блокчейн рассматривается также и в качестве информационно-инфраструктурного решения.

Блокчейн как технология распределенного хранения данных способен стабилизировать информационные системы в областях государственного управления и науки, но может дестабилизировать сферу финансов и экономику в целом. В связи с этим необходимо создание национальной технологической платформы для цифровой трансформации, которая удовлетворяла бы описанным в работе требованиям.

СПИСОК ЛИТЕРАТУРЫ

1. Nakamoto S. Bitcoin: A Peer-to-Peer Electronic Cash System. – 2008. – URL: bitcoin.org.
2. Schlatt V., Schweizer A., Urbach N., Fridgen G. Blockchain: Grundlagen, Anwendungen und Potenziale // Projektgruppe Wirtschaftsinformatik des Fraunhofer-Instituts für Angewandte Informationstechnik FIT. – Bayreuth, 2016. – P. 1–51.
3. Peters G.W., Panayi E., Chapelle A. Trends in crypto-currencies and blockchain technologies: A monetary theory and regulation perspective. – 2015. – URL: http://arxiv.org/pdf/1508.04364.pdf.
4. Romano D., Schmid G. Beyond Bitcoin: a critical look at blockchain-based systems. –2017. – URL: http://www.mdpi.com/2410-387X/1/2/15/htm
5. Ahamad S., Nair M., Varghese B. A survey on crypto currencies // Proceedings of the 4th International Conference on Advances in Computer Science (ACS 2013), December 13–14, Delhi, India. – P. 42–48.
6. Diffie W., Hellman M. E. New Directions in Cryptography // IEEE Transactions on Information Theory. – 1976. – Vol. 22, № 6. – P. 644–654.
7. Mougayar W. The business blockchain. Promise, practice, and application of the next internet technology // John Wiley & Sons, Inc. – Hoboken, New Jersey, 2016. – 208 p.
8. Правиков Д.И., Щербаков А.Ю. К вопросу об изменении парадигмы информационной безопасности // Системы высокой доступности. – 2018. – Т. 14, № 2. – С. 35–39.
9. Voshmgir Sh. Blockchains, smart contracts und das dezentrale Web // Technologie Stiftung Berlin. – Berlin, 2016. – P. 1–35.
10. Рязанова А.А. Технология блокчейн в научно-информационной деятельности // Научно-техническая информация. Сер.1. – 2018. – № 4. – С. 8–12.
11. Johannes Scherk B.Sc., Mag. Gerlinde Pöchhacker-Tröscher. Die Blockchain – Technologiefeld und wirtschaftliche Anwendungsbereiche // Pöchhacker Innovation Consulting GmbH. – Linz, 2017. – P. 1–63.
12. No author's name. Distributed Ledger Technology: beyond block chain // A report by the UK Government Chief Scientific Adviser. – UK Government Office for Science, 2016. – P. 1–87.
13. Mainelli M., von Gunten C. Chain Of A Lifetime: How Blockchain Technology Might Transform Personal Insurance // Z/Yen Group, Long Finance. – London, 2014. – P. 1–51.

Материал поступил в редакцию 24.07.18

Сведения об авторе

РЯЗАНОВА Алина Александровна – научный сотрудник ВИНИТИ РАН, Москва
e-mail: int.co-op@ras.ru

ИНФОРМАЦИОННЫЙ ПОИСК

УДК 004.65:54

С.В. Трепалин, Ю.Е. Бессонов, Б.С. Фельдман, Е.В. Кочетова, Н.И. Чуракова, Л.М. Королева

База структурных данных по химии ВИНИТИ РАН. Автономная система структурного поиска

Описывается система, позволяющая в автономном режиме выполнять поиск информации о различных характеристиках химических соединений в Базе структурных данных по химии ВИНИТИ РАН. Приведены примеры поиска и рассмотрены перспективы дальнейшего развития системы.

Ключевые слова: *база структурных данных по химии, системы информационного поиска химической информации*

ВВЕДЕНИЕ

Неотъемлемой составляющей системы информационного обеспечения специалистов-химиков являются специализированные базы данных структурной химической информации. Понимание механизма химических процессов и разработка синтеза новых материалов невозможны без использования такой информации. По некоторым данным [1] сведения о строении и свойствах химических соединений составляют около 85% общего потока химической информации.

В ВИНИТИ РАН с 1975 г. началась работа по созданию специализированной базы данных структурной химической информации (База СД) – крупнейшей в России и третьего в мире хранилища структурных химических данных. Сейчас База СД содержит информацию о более чем 7 млн химических структур, около 4 млн химических реакций и 15 млн свойств химических соединений. Накопленный с 1975 г. массив информации включает разноформатные данные о структуре химических соединений, их физико-химических свойствах, о химических реакциях, методах синтеза, анализа, биологической активности, токсичности, областях применения в химии, сельском хозяйстве, фармакологии, медицине, металлургии, электротехнике, физике, экологии и др.. База СД состоит из двух составляющих – базы структур химических соединений и базы реакций, в которых эти соединения принимают участие. Наполнение Базы СД в настоящее время производится с помощью программного комплекса *CBASE32*, который является развитием оболочки *CBASE* [2]. Указанный программный комплекс обеспечивает ввод и обработку структурных, фактографических и библиографических данных. Графическая информация о химических соединениях и реакциях сопровождается разнообразной текстовой информацией, представленной в Базе СД в виде предметных характеристик (термов), система которых была разработана в ВИНИТИ и в настоящее время постоянно пополняется [3]. Предметные характеристики представляют собой зашифрованный заглавными буквами латинского алфавита по иерархическому принципу список различных сведений о соединении: физико-химические свойства соединения, особенности его получения, проявляемая активность, области применения и т.д. Таким образом, в Базе СД содержится релевантная, тщательно отобранная, критически оцененная и обработанная высококвалифицированными специалистами-химиками информация, что является одним из серьезных преимуществ Базы СД среди многообразного потока химической информации, предоставляемой пользователям в современном информационном пространстве.

Неоспоримым достоинством любой базы данных является наличие разнообразных способов поиска необходимой пользователю информации. Каждая база данных имеет свой поисковый инструмент. Программно-технологический комплекс Базы СД включает программные системы, реализующие два способа предоставления информации пользователям: поиск в интерактивном режиме и поиск в автономном режиме. Система поиска в интерактивном режиме подробно изложена в [4]. В настоящей статье рассматривается автономная система поиска структурных данных, позволяющая выполнять поиск информации в режиме *off-line* в локальной базе данных системы, сформированной из базы данных структурной химической информации.

ТЕХНИЧЕСКИЕ ТРЕБОВАНИЯ К АВТОНОМНОЙ СИСТЕМЕ ПОИСКА

Современная поисковая система должна обладать комфортным интуитивным интерфейсом и реализовывать возможность быстрого релевантного поиска

по всем элементам информационного массива базы данных. При разработке поисковой системы Базы СД были определены следующие требования:

- удобный графический интерфейс;
- быстрый поиск с наглядными результатами;
- поиск по таким характеристикам химических соединений, как:
  - молекулярная формула;
  - название или фрагмент названия;
  - точная структура и фрагмент структуры;
  - предметные характеристики;
- поиск дополнительной информации в Интернете;
- показ библиографических данных из каталога ВИНИТИ РАН;
- отображение в результатах поиска следующей информации:
  - название химического соединения и его структурная формула;
  - библиографические ссылки;
  - предметные характеристики с комментариями;
  - данные о химических реакциях, в которых участвует найденное химическое соединение.

КОНВЕРТОР ДАННЫХ. ОСНОВНЫЕ ПРИНЦИПЫ ОРГАНИЗАЦИИ ПОИСКА

База СД представляет собой набор файлов одинаковой структуры в бинарном формате (формат *CBASE32*), каждый из которых содержит плоские таблицы и помимо описанных выше данных включает дополнительную информацию, в том числе и служебную. Файлы сгруппированы в каталогах, соответствующих годам и номерам реферативной базы данных (БД) и реферативного журнала (РЖ) «Химия». Такая организация не позволяет вести эффективный поиск необходимой информации. Требованию быстрого поиска информации отвечает база данных, построенная по иерархическому принципу. Для ее создания была разработана программа, называемая **Конвертор данных**. Основная задача конвертора создание иерархии данных, ориентированной на быстрое выполнение запросов в соответствии с техническими требованиями, перечисленными выше. Поскольку иерархическая база данных, создаваемая **Конвертором данных,** предназначена для предоставления химической информации пользователям, она будет в дальнейшем называться *пользовательской базой данных***.**

Конвертор данных работает следующим образом. Данные из Базы СД нормализуются и приобретают четырехуровневую иерархию. Самый верхний уровень – молекулярная формула, далее идет химическая структура (стереоизомеры считаются разными химическими структурами) и ее название. На третьем уровне иерархии находится список рефератов из БД «Химия», в которых упоминается данная химическая структура, список предметных характеристик и библиография. И, наконец, на четвертом уровне к каждому номеру реферата приписывается список реакций из соответствующей этому реферату статьи. Молекулярные формулы (первый уровень иерархии) формируются по правилу Хилла [5] и сортируются по алфавиту. На втором уровне иерархии названия вместе с химическими структурами также сортируются по алфавиту. Сортировка используется и для списка рефератов. Такой способ представления данных позволяет пользователю быстро найти интересующую информацию в ручном режиме без создания поисковых запросов.

При построении иерархии создаются списки уникальных молекулярных формул и внутри каждого списка формируются списки химических структур, для чего используется строка *InChIkey* [6], которая однозначно характеризует химическую структуру с учетом стереохимической информации. Номера рефератов применяются для формирования списков рефератов.

Программа вместе с информацией оо химических соединениях устанавливается на локальном компьютере. Данные для работы программы хранятся в локальном файле. Такой способ хранения позволяет избежать инсталляции и администрирование *SQL*-серверных приложений. Важное преимущество *SQL*-серверных приложений – возможность многопользовательской работы – в данном случае не играет роли, так как данные не редактируются клиентом и это позволяет ему обращаться к ним из нескольких локальных копий приложений без возникновения конфликтов.

Вышеперечисленные уровни иерархии содержат строки в качестве уникального идентификатора – номера рефератов для списка рефератов, молекулярная формула для списка формул и *InChIKey* для списка химических структур. При конвертации плоских таблиц необходима проверка присутствия данной строки в уже сформированных списках. Если строка присутствует, данные добавляются в существующий список, а если отсутствует создается новый список. Время поиска строки в списке пропорционально log_2N для отсортированного списка и N для неотсортированного, где N – число элементов в списке. Очевидно, что поиск по отсортированным спискам является более эффективным. Но при создании нового списка строку необходимо включать в поиск, а для этого требуется повторная сортировка и временные затраты пропорционально $Nlog_2N$. Если выполнять повторную сортировку после добавления каждого нового элемента, то исчезают преимущества поиска по отсортированному списку. Поэтому для формирования списков используются два массива – отсортированный и неотсортированный. Поиск осуществляется бисекциями в отсортированном массиве (скорость сходимости log_2N) и перебором (скорость пропорциональна N) в неотсортированном массиве. При создании нового списка строка-идентификатор добавляется в неотсортированный массив. Когда число элементов в неотсортированном массиве становится равным 64, они переносятся в отсортированный массив, который повторно сортируется, а неотсортированный массив обнуляется. Такая процедура заметно уменьшает время генерации иерархических списков.

Нами изучалась зависимость времени создания финального отсортированного списка от числа элементов в неотсортированном массиве, после достижения которого происходит генерация единого массива. Кривая, отражающая такую зависимость, имеет минимум, но оптимальное число элементов, при котором необходимо создавать единый массив и выполнять повторную сортировку, зависит от длины

списка и частоты встречаемости новых элементов списка и является предметом отдельного исследования.

Вследствие особенности хранения данных в ВИНИТИ РАН, заключающейся в том, что библиографическая информация и номера рефератов, присвоенные им в РЖ и БД «Химия», хранятся отдельно от массива структурной химической информации, программа-конвертор читает эти данные из внешних источников и заносит в плоские таблицы формата *CBASE32*.

Программа-конвертор принимает на вход списки таблиц и генерирует двоичный файл локальной базы данных для автономной системы структурного поиска. **Конвертор данных** реализован в 64 битовой архитектуре, что позволяет конвертировать очень большой объем данных. На практике вся база структурно-химической информации, содержащая более 8 Мб записей, была успешно конвертирована в пользовательскую БД.

Кроме того, **Конвертор данных** может быть успешно использован для формирования тематических баз. Тематические база – это пользовательская БД, которая содержит информацию о соединениях с заданными предметными характеристикам. Их можно указать в программе конвертирования перед формированием базы. В результате получается отфильтрованная база, в которой содержатся химические структуры с заданными термами и соответствующие библиографические данные. Химические реакции не включаются в тематические базы, так как для их показа, в общем виде, необходимы все химические структуры данной публикации, а часть их фильтруется.

ОСОБЕННОСТИ РЕАЛИЗАЦИИ ПОИСКОВЫХ ФУНКЦИЙ СИСТЕМЫ

Автономная система структурного поиска состоит из двух исполняемых модулей – **Конвертор данных** и **Модуль поиска**. Последний реализует функции поиска химических структур в локальной базе данных, сформированной из файлов Базы СД с помощью программы-конвертора.

Главное окно интерфейса **Модуля поиска** системы включает область отображения дерева молекулярных формул и панель инструментов. Дерево молекулярных формул отражает четырехуровневую иерархию базы данных, сформированную программой-конвертором, как это было описано выше.

Просмотр дерева выполняется с помощью полос прокрутки или колёсиком компьютерной мыши. Узлы дерева раскрываются или сворачиваются щелчком мыши. Дерево имеет до четырех уровней. Переход на уровни ниже первого сопровождается показом информации в правой части окна. На первом уровне дерева – молекулярные формулы, упорядоченные по правилу Хилла. На втором уровне – названия химических соединений и их структура (рис. 1). На третьем уровне – библиография и предметная информация. Узлам третьего уровня приписан номер реферата в БД «Химия» и в скобках – год издания (см. рис. 11). На четвертом уровне (если он есть) – уравнения химических реакций, в которых участвует данное соединение (рис. 2).

Панель инструментов состоит из управляющих элементов для реализации различных видов поиска и выполнения вспомогательных действий (рис. 3).

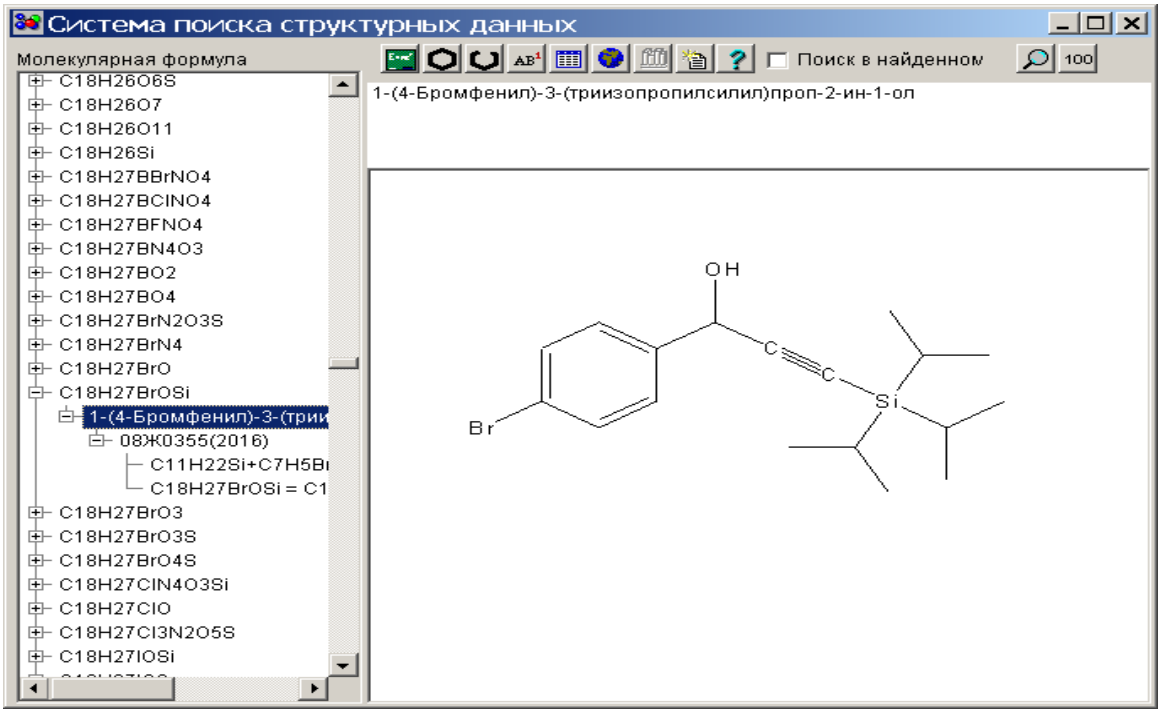

Рис. 1. Главное окно интерфейса Модуля поиска.
Отображение названия и структуры химического соединения.

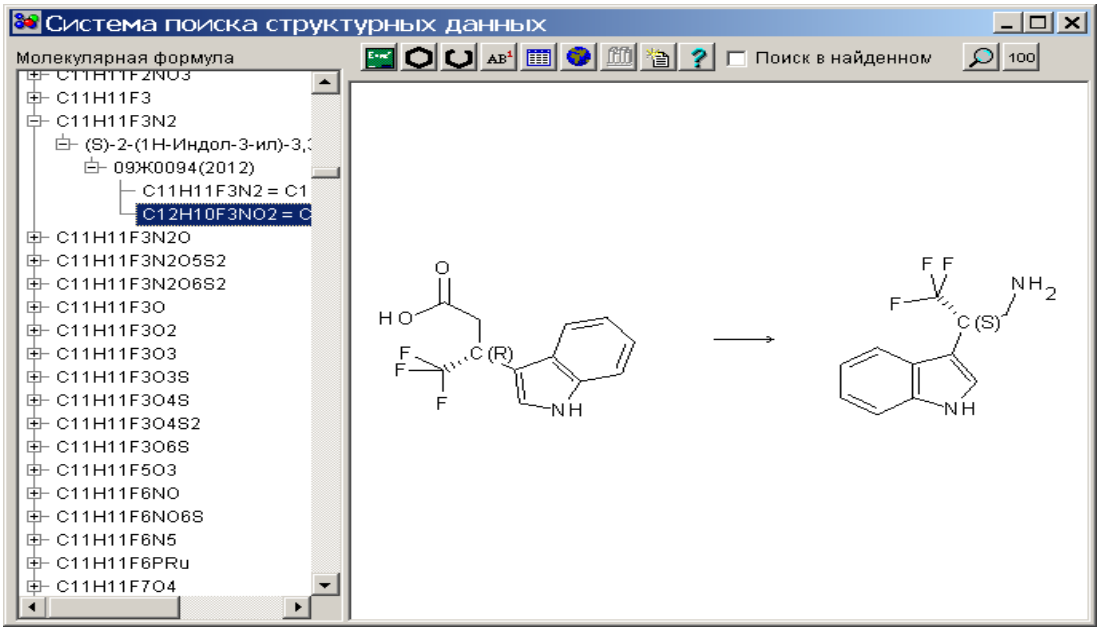

Рис. 2. Пример отображения информации о реакциях.

Рис. 3. Вид панели инструментов.

Поиск по молекулярной формуле и по названию химического соединения выполняется путем задания запроса в текстовом виде в поле специальных окон, вызываемых, соответственно, кнопками  и  . Например, если нажать первую слева кнопку и задать в текстовом поле окна запроса строку *C6H4* (рис. 4), будет получен результат, показанный на рис. 5.

Видно, что найдено не только соединение Циклогекса-1,3-диенин с молекулярной формулой *C6H4*, но и 289 химических соединений, молекулярная формула которых содержит *C6H4* в качестве фрагмента.

Аналогичным образом выполняется поиск по систематическому или тривиальному названию химического соединения. В запросе можно указывать полное название, либо его фрагмент на русском языке.

Поиск по точной структуре выполняется с помощью кнопки  панели инструментов. Щелчком по ней вызывается окно графического редактора [7], в котором можно сформировать структурный запрос в виде рисунка химической структуры. Пример такого запроса представлен на рис. 6.

Когда рисунок будет полностью сформирован, нужно нажать кнопку ОК. Программа выполнит поиск, результаты которого отобразятся в главном окне интерфейса. Алгоритм поиска основан на использовании *InChIKey*, который с большой степенью однозначности идентифицирует химическое соединение.

Строка *InChI* рекомендована *IUPAC* как линейное представление химической структуры [8]. Она содержит данные о связях между атомами, а также стереохимическую информацию и является уникальной для одинаковых структур независимо от того, в каком порядке пронумерованы атомы (каноническое представление). Для простых таутомеров, возникающих при 1,3-миграциях атома водорода, а также цепочек таких миграций генерируются одинаковые строки. Таким образом, сравнивая строки *InChI*, можно определить, являются ли химические структуры одинаковыми или разными.

Недостатком *InChI*-строки является то, что она имеет варьируемую длину и для больших соединений может достигать несколько тысяч символов. Этот недостаток был преодолен в технологии *InChIkey*. *InChIkey* представляет собой хэш *InChI*. Его длина 27 символов. Первые 14 символов содержат информацию о связности, далее после разделителя идет стереохимическая информация (10 символов) и контрольная сумма/версия *InChI* (один символ). Сравнивая первые 14 символов, можно найти химические структуры с одинаковой топологией. Сравнивая 10 стереохимических символов, можно найти одинаковые стереоизомеры.

Имеется ненулевая вероятность, что две разные химические структуры генерируют одинаковый хэш-код. Но она крайне низка, и в настоящее время ни для одной пары химических структур не удалось достоверно обнаружить такое событие. Обсуждаемые в литературе совпадения объясняются различной реализацией генерации *InChI*-строки вследствие ошибок в программах [9].

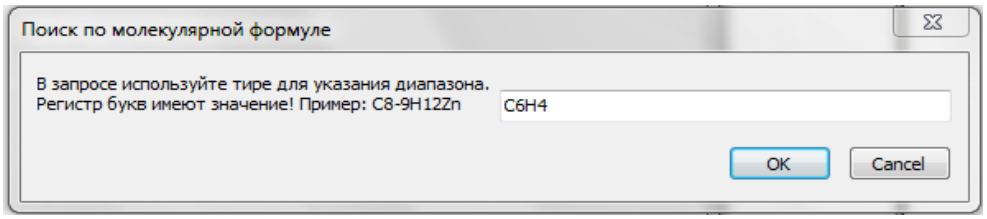

Рис. 4. Окно запроса поиска по молекулярной формуле.

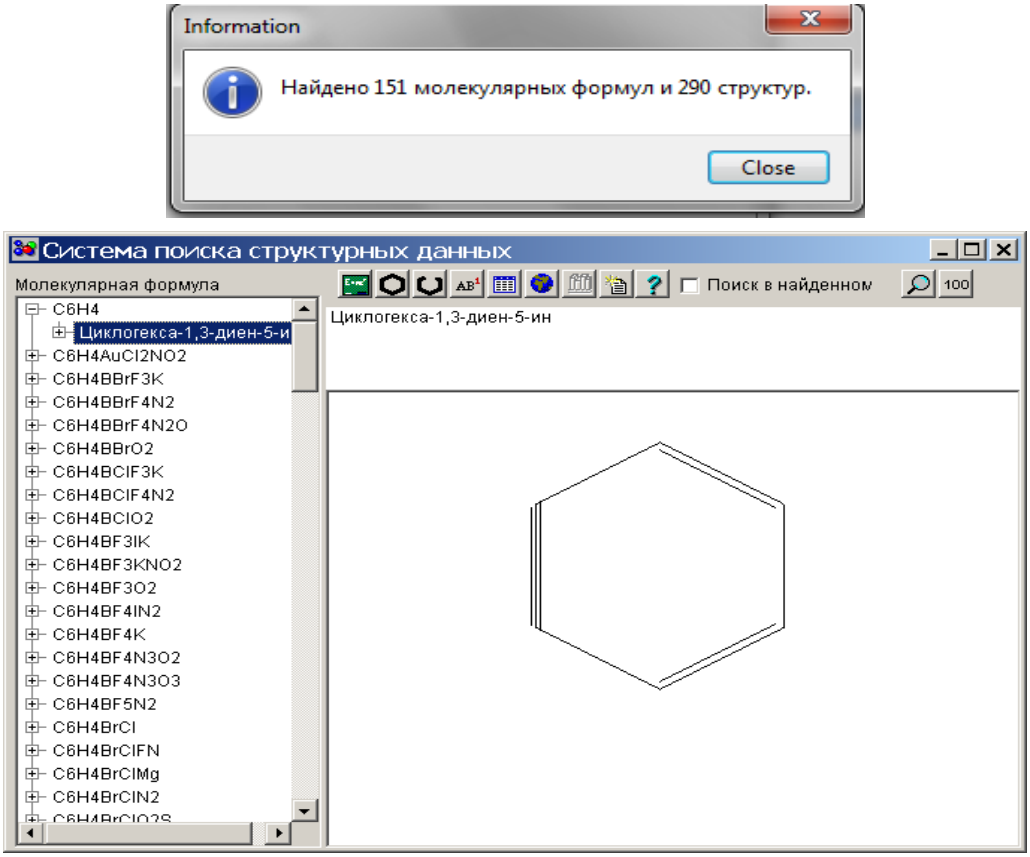

Рис. 5. Пример результата поиска по молекулярной формуле.

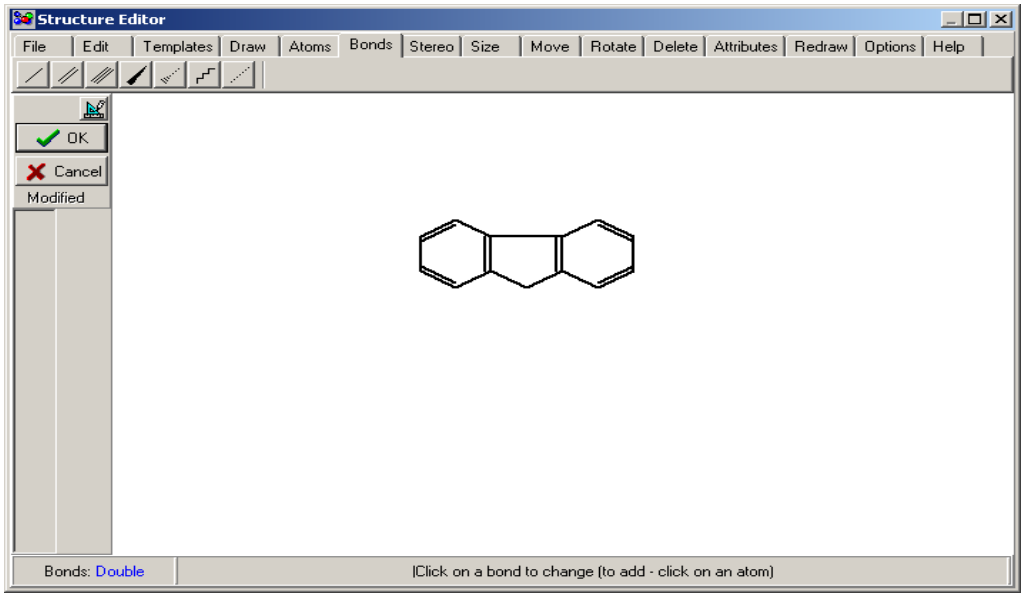

Рис. 6. Окно графического редактора.

При старте программы поиска *InChIkey*-строки сортируются и загружаются в ОЗУ. Для структурного запроса пользователя также генерируется *InChIkey* и поиск осуществляется в массиве с помощью алгоритма бисекций. Для поиска по очень большим данным требуется ничтожное время и результаты сразу представляются пользователю. Подробно описание системы поиска в Базе СД на основе *InChIkey* изложено в работе [10]

Поиск по фрагменту структуры выполняется кнопкой  . Данный вид поиска имеет большое значение при определении связи структуры и активности химических соединений. Алгоритм поиска основан на переборе с возвратом в сочетании с методом релаксации и подробно описан в [11]. Следует отметить, что в терминах теории графов здесь решается задача определения изоморфизма частичному подграфу [12]. Например, по запросу, показанному на рис. 7, может быть найдена структура молекулы циклогексана.

Запрос на поиск по фрагменту структуры задается с использованием всех имеющихся графических средств изображения точной структуры. Кроме этого, для задания поиска по фрагменту существуют специальные средства: *поисковые атомы* и *поисковые связи.* Они изображаются группой команд *Query* графического редактора. Запрос должен содержать минимум два неводородных атома (это естественное и общепринятое требование), и не должен содержать больше 1000 атомов и 1000 связей в связном фрагменте (это ограничение автономной системы структурного поиска). *Поисковыми атомами* являются атомы-заместители, используемые для рисования запроса для поиска по подструктуре. Следующие атомы определены как поисковые:

1. *Any* – любой атом за исключением атома водорода.
2. *Hetero – N, O, Si, P, S, Ge, As, Se, Sb, Te, Po*.
3. *Exactly coordinated* используется в комбинации с любым атомом и означает, что координация атома не может быть другой по сравнению с указанной.
4. *Halogens – F, Cl ,Br, I, At.*
5. *Metal – Li, Be ,Na, Mg, Al, K, Ca, Sc, Ti, V, Cr, Mn, Fe, Co, Ni, Cu, Zn, Ga, Rb, Sr, Y, Zr, Nb, Mo, Tc, Ru, Rh, Pd, Ag, Cd, In, Sn, Cs, Ba, La, Ce, Pr, Nd, Pm, Sm, Eu, Gd, Tb, Dy, Ho, Er, Tm, Yb, Lu, Hf, Ta, W, Re, Os, Ir, Pt, Au, Hg, Tl, Pb, Bi, Fr, Ra, Ac, Th, Pa, U, Np, Pu, Am, Cm, Bk, Cf, Es, Fm, Md, No, Lr.*

Поисковые атомы типов 1, 2, 4, 5 используются как обычные физические атомы. Поисковые атомы типа 3 должны быть нарисованы в комбинации с другими (физическим и поисковыми). *Поисковыми связями* являются связи-заместители, используемые для изображения задачи для поиска по фрагменту структуры. В редакторе определены следующие поисковые связи:

1. *Any* – означает, что два атома связаны произвольной связью.
2. *Aromatic* – означает, что пара атомов является частью ароматического цикла.
3. *S/D* – Одинарная или Двойная. Означает, что атомы связаны либо одинарной либо двойной связью.
4. *Ch* – Цепь. Используется в комбинации с другими связями (физическими и поисковыми) и означает, что связь не включена в цикл.

5) *Rn* – Цикл. Используется в комбинации с другими связями (физическими и поисковыми) и означает, что связь является частью цикла.

6) *Cis/trans*. Используется в комбинации с другими связями и означает тип (*Cis/trans*) помеченной связи.

Проиллюстрируем использование поисковых атомов и связей на примере поискового запроса, в котором *Rn* обозначает циклическую связь, а *X* – атом галогена (рис. 8). Один из результатов этого поиска показан на рис. 9.

Поиск по предметным характеристикам выполняется с помощью кнопки  . Термы обозначаются одним символом или их последовательностью, длиной до четырех символов, представленных в правой части окна (рис. 10)

В нижней части окна имеются инструменты для поиска термов по заданному фрагменту строки терма или текста комментария. После того, как терм выбран и отмечен, нажатие кнопки ОК запустит процедуру поиска. Пример возможных результатов показан на рис. 11.

Поиск выбранной структуры в Интернете выполняется посредством кнопки  . При нажатии на нее формируется символьная строка *InChiKey* данного химического соединения, которая автоматически помещается в поле поискового запроса Google интернет-браузера.

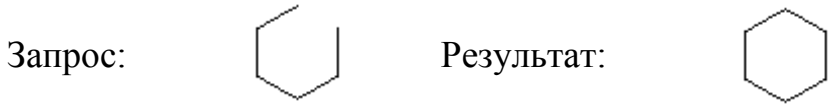

Рис. 7. Пример запроса по фрагменту и найденной структуры.

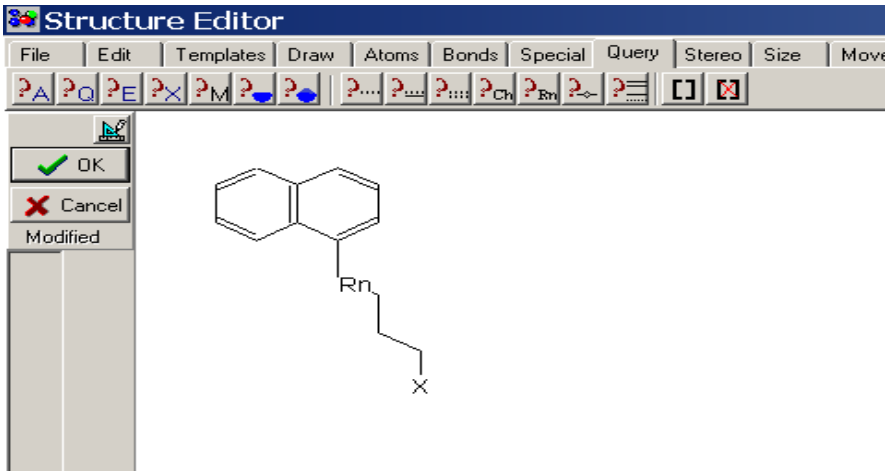

Рис. 8. Пример запроса для поиска по фрагменту структуры.

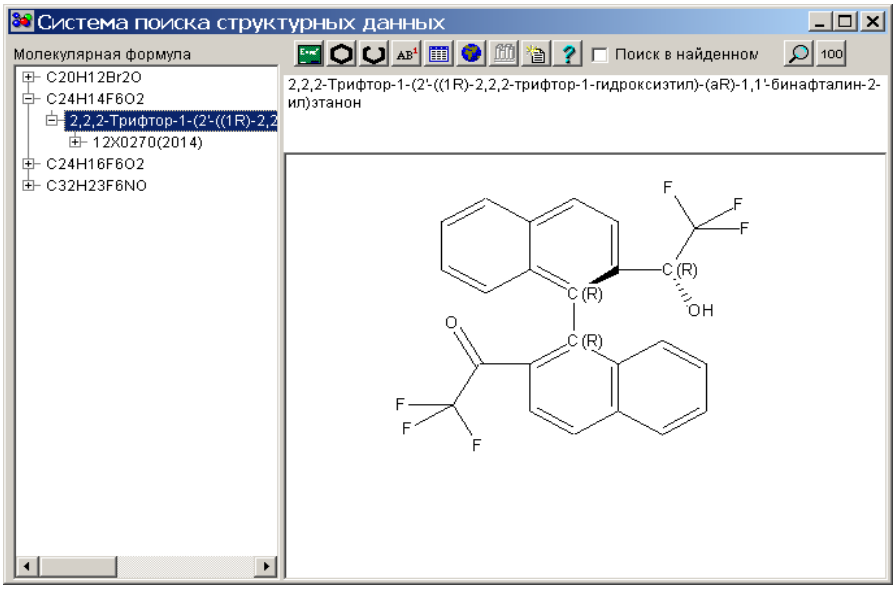

Рис. 9. Один из результатов поиска по фрагменту на рис. 8.

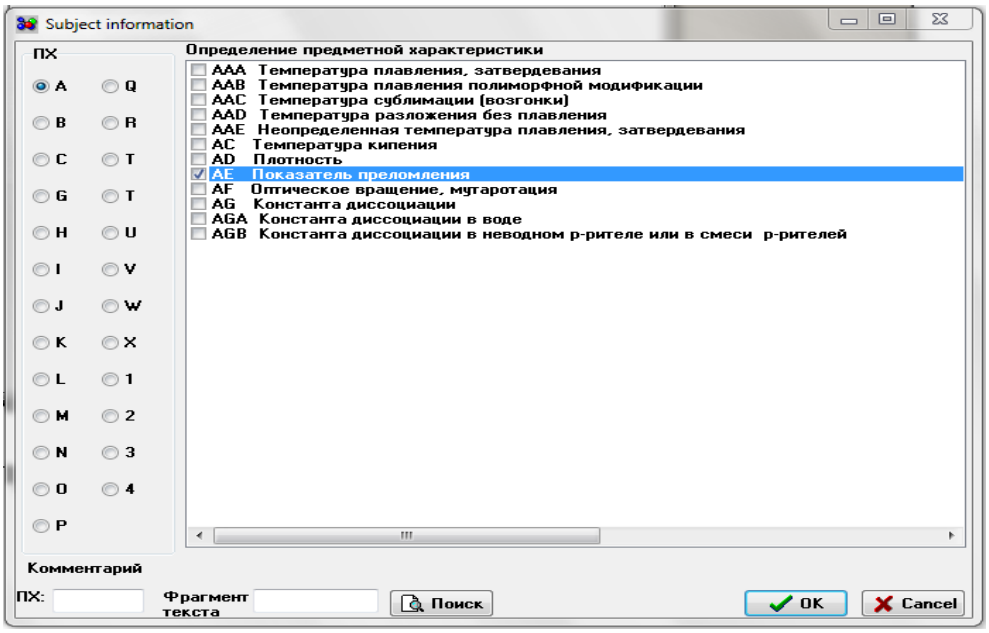

Рис. 10. Окно для выбора предметной характеристики.

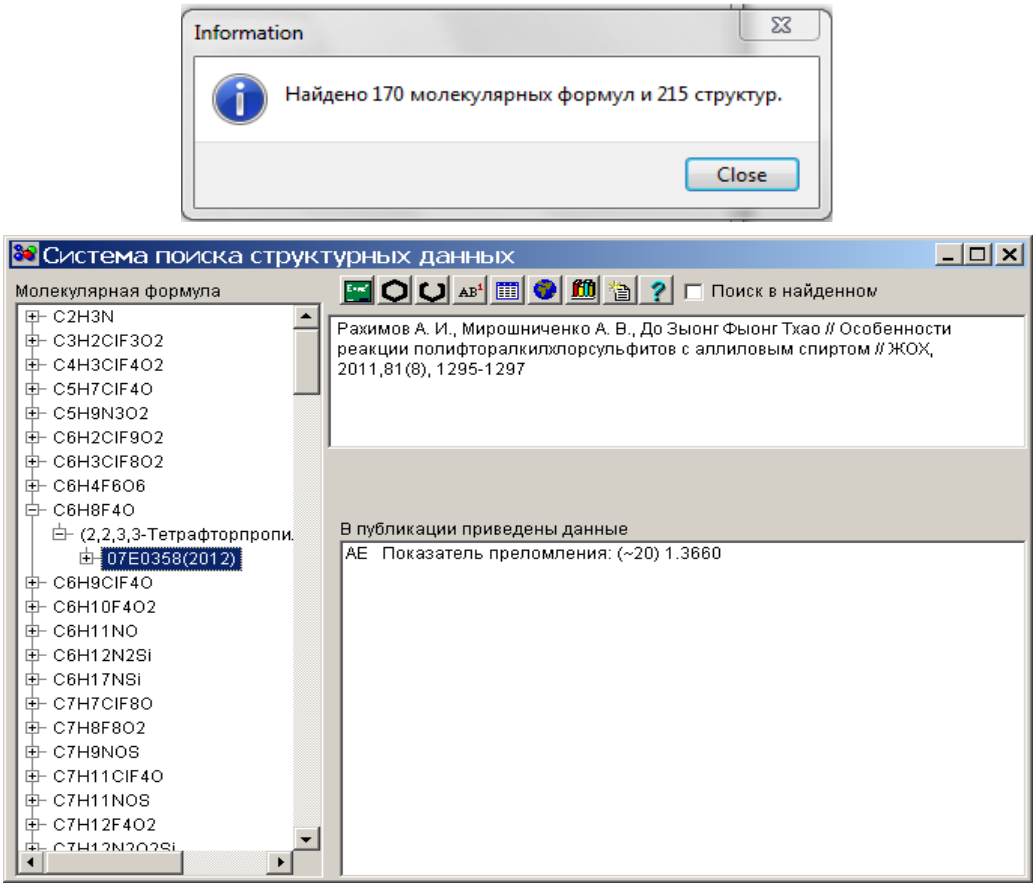

Рис. 11. Пример результата поиска по предметной характеристике.

Для просмотра данных из электронного каталога ВИНИТИ РАН необходимо выделить узел дерева с номером реферата в БД «Химия» (рис. 12).

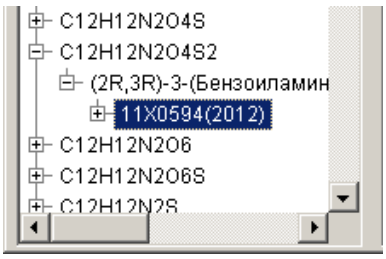

Рис. 12. Выделенный узел с номером реферата.

Тогда кнопка  становится активной и при нажатии на нее устанавливается связь с Электронным каталогом ВИНИТИ РАН. В интернет-браузере отображается библиографическая информация первоисточника и обеспечивается доступ к Электронному каталогу [13], где можно получить дополнительную информацию:

- перевод названия первоисточника на русский язык;
- дата регистрации в ВИНИТИ;
- место хранения в ВИНИТИ;
- аннотация;
- ключевые слова;
- ссылки на все публикации каждого из авторов данной статьи.

Сервис Электронного каталога на вкладке «Поиск» дает возможность найти библиографическую информацию по ключевым словам и другим параметрам.

ЗАКЛЮЧЕНИЕ

В настоящее время задача предоставления пользователям информации из базы данных структурной химической информации решается передачей им пользовательской базы данных и программы **Модуль поиска**, описанной выше. Пользовательская БД создается с помощью специальной программы-конвертора данных, систематизированных по периодам, в течение которых собиралась информация. Так, например, свободно распространяемая демо-версия системы автономного поиска [14] содержит массив данных из Базы СД (порядка 200 тыс. химических структур), накопленный за 2012 г.

Дальнейшее развитие системы может происходить по двум направлениям.

Первое направление – это увеличение объема предоставляемой пользователю информации за счет ретро-данных. Следует отметить, что задача конвертации данных за период 1996–2009 гг. (27% от общего числа химических соединений)решена, но конвертация данных 1975-1995 гг. (62% от общего числа химических соединений) представляет собой достаточно трудоемкую задачу. Проблема состоит в том, что соединения в указанный период времени вводились без использования графического редактора, по-

этому файлы базы данных не хранят информацию о рисунке структурной формулы. Задача автоматического построения рисунка структуры по таблице связей (визуализация молекулы) в настоящее время для большей части химических соединений Базы СД успешно решается [3].

Второе направление предполагает предоставление тематических выборок из Базы СД (например, биологически активных соединений, комплексных соединений, соединений, применяемых в определенной технологии, и т.п.) по требованию пользователей. В настоящее время частично решена задача создания на основе Базы СД специализированных локальных баз данных в автономной системе поиска – разработан конвертор для формирования тематических баз, которые содержат информацию о соединениях с заданными предметными характеристиками.

Рассмотренная система автономного поиска информации в пользовательских базах данных может применяться в следующих областях:

- фундаментальные и прикладные исследования по химии, биохимии, фармацевтике, фармакологии, токсикологии, экологии, медицине;
- химическая промышленность;
- образование (помощь преподавателям, студентам, аспирантам, специалистам, повышающим квалификацию);
- информационное обеспечение (базы данных, научные библиотеки, издательства).

СПИСОК ЛИТЕРАТУРЫ

1. Некоторые наукометрические данные по химии. – URL: http://chem21.info/article/ 388814/
2. Алфимов М.В., Авакян В.Г., Трепалин С.В., Воронежева Н.И., Чуракова Н.И. Универсальная программная оболочка для создания баз данных химических соединений и реакций // Доклады РАН. – 1999. – Т. 366, № 5. – С. 639–642.
3. База структурных данных по химии ВИНИТИ. Индексирование и ввод сведений о химических соединениях при подготовке базы структурных данных по химии с использованием программного комплекса CBASE32. Инструкция ВИНИТИ РАН И81-2010. – М., 2010. – 103 с.
4. Нефедов О.М., Королева Л.М., Трепалин С.В., Бессонов Ю.Е.. Чуракова Н.И. Разработка интегрированной системы структурной химической информации // Научно-техническая информация. Сер. 2. – 2015. – №12. – С. 17-25; Nefedov O.M., Koroleva L.M., Trepalin S.V., Bessonov Yu.E., Churakova N.I. The Development of an Integrated System for Structural Chemical Information. Automatic Documentation and Mathematical Linguistics, 2015, Vol. 49, N 6, P. 213–220.
5. Hill E.A. On a system of indexing chemical literature; Adopted by the Classification Division of the U.S. Patent Office // J. Am. Chem. Soc. – 1900. – Vol. 22 (8). – P. 478–494.
6. Heller S., McNaught A., Stein S., Tchekhovskoi D., Pletnev I. InChI – the worldwide chemical structure identifier standard // Journal of Cheminformatics. – 2013. – Vol. 5(1). – P. 7.
7. Описание графического редактора CBASE32. – URL: https://yadi.sk/i/qMWnyrNFzjkH5.
8. The IUPAC international chemical identifier. – URL: https://iupac.org/who-we-are/divisions/division-details/inchi/.
9. An InChIkey Collision is Discovered and NOT Based on Stereochemistry. – URL: http://www.chemconnector.com/2011/09/01/an-inchikey-collision-is-discovered-and-not-based-on-stereochemistry/.
10. Нефедов О.М., Трепалин С.В., Королева Л.М., Бессонов Ю.Е. Быстрый поиск точных химических структур в больших базах данных с использованием InChI key кодировки структур // Научно-техническая информация. Сер. 2. – 2013. – № 12. – С. 27 – 33
11. Нефедов О.М., Трепалин С.В., Королева Л.М., Бессонов Ю.Е., Чуракова Н.И. База структурных данных по химии ВИНИТИ РАН: проблемы поиска по фрагменту структуры // Научно-техническая информация. Сер. 2. – 2014. – № 12. – С. 19-29.
12. Берж К. Теория графов и ее применение. – М.: Изд-во Иностранная литература, 1962. – 320 с.
13. Электронный каталог ВИНИТИ РАН. – URL: http://catalog.viniti.ru/
14. Демо-версия автономного структурного поиска в базе данных химических соединений. – URL: https://yadi.sk/d/JWUaGwPMvb6Mq.

Материал поступил в редакцию 02.07.18.

Сведения об авторах

ТРЕПАЛИН Сергей Владимирович – кандидат химических наук, ведущий научный сотрудник ФГБУН Институт физиологически активных веществ РАН, Москва
e-mail: trep@chemical-block.com

БЕССОНОВ Юрий Ефимович – кандидат технических наук, ведущий научный сотрудник ВИНИТИ РАН, Москва
e-mail: bessonov-ye@rambler.ru

ФЕЛЬДМАН Борис Семенович – старший научный сотрудник ВИНИТИ РАН
e-mail: bsf@inbox.ru

КОЧЕТОВА Елена Вениаминовна – кандидат химических наук, научный сотрудник ВИНИТИ РАН
e-mail: kelena63@gmail.com

ЧУРАКОВА Наталия Исааковна – кандидат химических наук, зав. Отделом исследований и обработки структурной химической информации ВИНИТИ РАН
e-mail: nichurak@rambler.ru

КОРОЛЕВА Любовь Михайловна – кандидат химических наук, зав. Отделением научной информации по проблемам химии и наук о материалах ВИНИТИ РАН
e-mail: lkorol@viniti.ru

АВТОМАТИЗАЦИЯ ОБРАБОТКИ ТЕКСТА

УДК 81’322.2’37

Н.В. Максимов, А.С. Гаврилкина, В.В. Андронова, И.А. Тазиева

Систематизация и идентификация семантических отношений в онтологиях научно-технических предметных областей

Рассматривается построенная на основе концептуального каркаса таксономия отношений, а также методы идентификации семантических отношений, используемые для автоматического построения онтологий предметных областей. Путём автоматической обработки коллекции текстов научно-технических документов выявлена предикатная лексика, соответствующая классам отношений. Построены шаблоны лингвистических конструкций семантических отношений, учитывающие морфологические признаки слов и морфемы. Исследована статистика классов отношений и использования шаблонов в зависимости от вида документов.

Ключевые слова: *онтологии, текстовые документы, семантические отношения, концептуальный каркас, метод идентификации семантических отношений*

ВВЕДЕНИЕ

Представление семантической информации в большинстве случаев в слабоструктурированной текстовой форме обуславливает актуальность извлечения и формализации знаний из текстов. Для задач информационного поиска, экспертизы проектов, управления знаниями и т.п. перспективным инструментом являются онтологии, как хорошо формализованные семантические структуры.

На сегодняшний день наиболее полный и конструктивный подход в области информационного моделирования сложных систем на основе документальных описаний сложных объектов и процессов – стандарт *ISO* 15926, включающий *Gellish* [1], где основой для семантической интеграция информации является система классов эталонной модели (в том числе 201 класс отношений), в отношении которых определены синтаксические спецификации для английского языка.

Задача построения хорошо формализованного семантического образа, учитывающего отношения, является далеко не новой. Уже в начале 1960-х гг. были разработаны информационно-поисковые языки, включающие не только понятия, но и имманентные и ситуативные связи между ними. Примерами являются языки *RX*-кодов и СИНТОЛ.

Информационно-поисковый язык СИНТОЛ [2] создан в 1960-1962 гг. в Национальном центре научных исследований Франции. Минимальной синтаксической единицей является *синтагма* – двуместный предикат вида *xRy,* где *x* и *y* – лексические единицы, каждая из которых относится к одной из четырех квазиграмматических категорий этого языка, а *R* – одно из четырех главных синтагматических отношений. В основе языка *RX*-кодов [3], разработанного на аналогичных принципах в Институте кибернетики АН УССР, лежит понятие производности значения. Язык содержит некоторое фиксированное (хотя и допускающее изменения) множество лексических единиц, представляющих ряд базовых предметов и отношений, и грамматику. Остальные предметы и отношения, а также все ситуации обозначаются с помощью производных единиц, образуемых с использованием грамматики.

Следует отметить, что особенностью большинства практических реализаций онтологического подхода, как одного из инструментов моделирования и поиска, является то, что определение онтологии не налагает ограничений на типы отношений: при создании онтологий в разных проектах используются разные наборы и типы отношений. Кроме того, семантика естественного языка многогранна и допускает различные способы формализации. И, очевидно, что для того, чтобы было возможно производить вычисления на онтологиях, способы представления семантики и операции должны быть непротиворечивы и эффективны, по крайней мере, для конкретного класса задач. В практическом плане задачи сводятся к

созданию «преобразователя, который может работать с синтаксическим анализатором и выдавать данные, приемлемые для логической дедуктивной системы» [4].

Таким образом, задача систематизации и идентификации семантических отношений становится определяющей для автоматического построения по текстам документов сетевой понятийной семантической структуры, допускающей формализованное построение выводов и оценок смысловой близости текстов.

Онтологии, как семантические структуры, здесь будут рассматриваться в контексте задач класса информационного моделирования и поиска, где они являются инструментом представления смысла документов (в пределе – как «сборочный чертеж» образа объекта исследования или алгоритм функционирования [5]), а основной практической целью будет являться задача формального отождествления смыслов [6]. При этом к современному логико-лингвистическому аппарату предъявляются более широкие требования, в частности:

1) семантическая сеть должна адекватно отображать сколь угодно большой фрагмент текста в целом и, в то же время, учитывать роль каждого отдельного элемента;

2) поскольку текст является формой информации (т.е. текст в зависимости от контекста может иметь разные смыслы), модель онтологии должна обеспечивать возможность построения разных семантических сетей в зависимости от аспекта рассмотрения или области приложения.

Под отношением будем понимать *логическую* (предикатно-ориентированную) форму спецификации взаимосвязи предметов, явлений и процессов в природе, обществе, мышлении. Естественной основой для возможности формальной спецификации является тот факт, что взаимосвязь обнаруживается и потому может быть зафиксирована через свое проявление.

В настоящей работе представлены: 1) таксономия отношений, построенная на основе концептуального каркаса семантических отношений верхнего уровня, 2) метод лингвистического анализа текстов для идентификации семантических отношений, основанный на характеристической предикатной лексике и типизированных шаблонах лингвистических конструкций.

ОБЗОР ПОДХОДОВ К КЛАССИФИКАЦИИ СЕМАНТИЧЕСКИХ ОТНОШЕНИЙ

Упомянутый выше информационно-поисковый язык СИНТОЛ включает четыре класса главных синтагматических отношений: *предикативные* – ориентированная связь между парами слов; *ассоциативные* – отношение зависимости между двумя понятиями (субъекта к действию, действия к его объекту или обстоятельствам, отношение принадлежности, включения и т.п.); *консекутивные* – несимметричные отношения динамического типа, которые существуют между двумя понятиями в тех случаях, когда присутствие одного из них влияет на состояние или положение другого (отношения «причина–следствие», «субъект–объект» и т.п.); *координативные* – неориентированные отношения (эквивалентности, сравнения, дифференциации и т.п.). Кроме того, есть семь синтаксических операторов, которые необходимы для уточнения логической роли отношения. Четыре из них используются с терминами, которые связаны ассоциативными отношениями (*инструментальные*, *места, цели* и *признака)*, три – с терминами, которые связаны координативными отношениями (*сравнения*, *идентификации и дифференциации)*.

Одним из первых практических решений в части формальной спецификации отношений был язык *RX*-кодов. Его базовые отношения включают девять видов и три подвида, для задания которых использовалось около 30 релятем.

Более углубленный подход был применим в системе ТАПАЗ (например, [7]). Таблица семантических элементов (макропроцессов) ТАПАЗ-2 явно подразделяется на физические и другие процессы. Выделяются группы *активизации/эксплуатации/ трансформации/нормализации*, которые, в свою очередь, подразделяются по признакам *среда/оболочка/ядро*, а также по характеру процесса – *инициализация/аккумуляция/амплификация/генерация*.

Аналогичные задачи решаются и при разработке разнообразных онтологических проектов. Однако анализ онтологий, в том числе, *SUMO, DOLCE* и *YATO*, показал, что таксономиям отношений недостаёт обоснованности, в частности, практически нигде не удалось обнаружить явных классификационных признаков.

Структурно и технологически обоснованным в рамках стандартов серии *ISO*-15926 является *Gellish* – формальный язык, предназначенный для выражения фактов (утверждений), представленных на английском языке. Подход основан на предположении, что вещи существуют только в связи с другими вещами и понятиями, т.е. *Gellish* – это попытка построения классов отношений. Однако надо отметить, что приведенный в нем набор из 201 класса отношений формально не является классификацией.

Фреймовый подход, как формализованное описание деятельности в контексте определенной ситуации, наиболее разработан в рамках проекта *FrameNet* [8]. Инструментарий фреймов позволяет определять семантические типы дефиниций, элементов фрейма и самих фреймов. Однако наборы и типы фреймов, как и свойств, в общем случае могут быть различными и определяются предметной областью и характером решаемых задач.

Разные наборы отношений используются и в системах класса лингвистических процессоров и факт-экстракторов. На основе данных из доступных, хотя и разнотипных источников (от демонстрационных роликов до эксплуатационной документации), авторами были рассмотрены подходы, используемые в таких системах, как АОТ, МетаФраз, RCO Fact Extractor, ABBYY Compreno. Эти системы в принципе имеют двуединое назначение – извлечение фактов из текстов и их представление в виде графа типизированных сущностей с одной или несколькими связями. Здесь типовыми классами отношений являются: родовидовые отношения, отношения «целое–часть», синонимия и антонимия, логические и функциональные отношения, атрибутивные и количественные отношения, пространственные и временные отношения, а также лингвистические отношения. Но типы связей, как и

словари сущностей, обычно специально разрабатываются для конкретной предметной области (при отсутствии явного указания на схему классификации), что не позволяет унифицировать «вычисления» на получаемых графах.

Вопросу классификации отношений посвящено немалое количество научных публикаций. В частности, в [9] приводится обстоятельный обзор и обобщающая классификация физических действий и связей. Также конструктивные, но взаимно различающиеся решения представлены в [10–14]. Но при этом отдельные принципы или конкретные классификации, приведенные в этих публикациях, так или иначе, соответствуют рассмотренным выше.

Приведенные выше подходы и решения адекватны задачам идентификации смысла документов в научных и технических областях, где используются преимущественно традиционные схемы определения новых понятий через род и видовые отличия. Используются, в основном, бинарные отношения, а в тех случаях, когда встречаются многоместные отношения, они приводятся к бинарным. Однако, несмотря на логичность и семантическую силу, ни язык *RX*-кодов, ни более современные разработки не нашли пока заметного применения на практике. Причиной этого является очевидная предкоординированность, и, в частности, отсутствие системно обоснованного унифицированного упорядоченного набора отношений.

Иные подходы представлены в лингвистике. Это достаточно значимая составляющая, поскольку лингвистика имеет конечной целью разработку предельно точного (по смыслу) автоматического перевода, причем это справедливо не только для случая перевода текста на другой естественный язык, но и для перевода на формальный или формализованный язык (например, задача автоматического реферирования). Ниже приведен краткий обзор наиболее значимых публикаций, в которых представлены те или иные классификации.

Фундаментальная классификация предикатов Ю.Д. Апресяна [15] содержит 16 классов предикатной лексики, которые пересекаются между собой из-за неоднозначности каждого из глаголов. Поэтому границы между ними размыты.

Семантическая классификация предикатной лексики Л.М. Васильева [16] осуществляется по трём основаниям предиката: по ядерным компонентам, выделяемым по денотативному принципу, по доминирующим компонентам их лексических значений и по синтагматическим компонентам.

В [17] рассматриваются тематические классы глаголов: глаголы восприятия, глаголы знания, глаголы эмоций, глаголы принятия решения, речевых действий, движения, глаголы звука, бытийные глаголы и др. Основной в данном подходе является идея разделять понятия языка на некоторые семантические группы с учетом того, что эти понятия имеют нетривиальный общий смысловой компонент (элементы таких групп склонны иметь один и тот же набор зависимых понятий). Однако главная проблема такого подхода заключается в том, что выделение тематических классов по столь нетривиальному общему смысловому окружению – это достаточно трудоемкий и субъективный процесс, сильно зависящий от индивидуального восприятия.

Достаточно четко определена классификация семантических отношений для научных текстов в [18], где на верхнем уровне выделяются два класса: квалитативные и квантитативные отношения. *Квалитативные отношения* включают четыре группы: 1) «Иерархия» (род–вид; признак–значение признака; инвариант–вариант); 2) «Агрегация» (целое–часть; объект–пространство локализации; объект–свойства/ признак; уровень–единица уровня); 3) «Функциональные отношения» (причина–следствие; условие–действие; событие–действие; объект действия–состояние; состояние–действие; инструмент–действие; данные–действие; данные–величина); 4) отношения соответствия понятий знаковых систем (термин–способ выражения; термин–способ представления термина; термин–способ метаязыкового представления). *Квантитативные отношения* лежат в сфере тождества и противопоставления. К ним относятся отношения тождества (синонимии) и оппозиции (корреляции). Отношение корреляции выражает различия между понятиями и включает: 1) отношение противоположности или контрарное противопоставление; 2) отношение противоречия или контрадикторное противопоставление; 3) противопоставление грамматических форм; 4) противопоставление соподчиненных видовых понятий, имеющих общее родовое понятие; 5) противопоставление частей, образующих единое целое; 6) противопоставление членов отношения соподчинения.

Означенные подходы ориентированы на специфику естественных языков, а в классификациях отсутствуют четкие границы между классами.

Следует отметить, что значительное влияние на представления авторов оказали также фундаментальные положения работ В.В. Виноградова, Ю.С. Маслова, З. Вендлера, У. Чейфа, Ч. Филлмора, У. Кука, Б. Левина.

ПОСТРОЕНИЕ КЛАССИФИКАЦИИ И ТАКСОНОМИИ СЕМАНТИЧЕСКИХ ОТНОШЕНИЙ

Основой (минимальной) для построения исчисления семантики является иерархия классов сущностей, отношений и свойств, базирующаяся на общей системе характеристических признаков. Однако обзор существующих решений показывает, что нет строгой и всеобщей классификации, хотя есть ряд вполне самодостаточных решений[1]. Другая проблема связана с тем, что лексических конструкций, представляющих отношения в тексте, довольно много, и их значение зачастую доопределяется или изменяется контекстом их употребления. Вследствие этого при практическом наполнении классов лингвистическими конструкциями возможны ошибки. Поэтому автоматически выделяемые в тексте лингвистические конструкции отношений на первом шаге экспертом вносились в классификационную таблицу, в которой явно заданы признаки классифи-

[1] Этот факт отражает действие принципа конструктивности: уровень и пути решения должны соответствовать решаемой задаче.

кации. После чего формировалась собственно таксономия отношений. Такая технология позволила выявить некоторые ошибки отнесения к классу, поскольку при группировании и упорядочении классов противоречия могут стать явно заметными.

Учитывая и обобщая упомянутые выше решения, авторами были разработаны классификационная таблица и таксономия семантических отношений. При этом классификация строилась на следующих принципах:

- определяющим является системность уровня «логоса», а не «лексиса» (согласно Д.С. Лотте);
- соотношения свойств (признаков классификации) имеют приоритет перед отдельными свойствами;
- полнота охвата классификации обеспечивается дедуктивностью (от общего к частному) процесса её построения;
- система отношений моделирует соответствующие способы и средства, используемые человеком при выполнении практических задач;
- система отношений должна обладать алгебраическими свойствами для получения количественных оценок силы семантических связей в лексике и тексте;
- система отношений и методы их идентификации (автоматический перевод с естественного языка на информационный) должны быть открытыми и адаптивными, что позволит проследить, как изменяется система семантических связей в процессе функционирования языка.

В качестве основы принят предложенный в [5] концептуальный каркас семантических отношений верхнего уровня, использующий следующие характеристические признаки.

Признак *Реальное/Абстрактно-конкретное/Абстрактное* отражает соотношение «реальность/модель». Абстрактно-конкретные отношения, в отличие от группы реальных, обусловлены некоторой абстрактной моделью соотношения. Определение данного типа отношений позволяет выделить аналитико-синтетические отношения, обусловленные некоторой функциональной моделью деятельности. Группа абстрактных отношений определяется абстрактностью (искусственностью) природы одного из взаимодействующих объектов.

Признак *Отдельное/Целое/Все части целого* отражает комбинации соотношений отдельного *(часть)* и агрегатного *(целое)*. Подкласс *Отдельное–Отдельное* представляет взаимодействие двух отдельных объектов. Подкласс *Все части–Целое* представляет взаимодействия, направленные на соединение/отсоединение некоторого множества частей в единую композицию. Подкласс *Целое–Целое* соответствует взаимодействиям сложных объектов (систем). Но, поскольку с точки зрения «механики» отношения сложность взаимодействующих объектов не принципиальна, то эти подклассы отношений можно считать эквивалентными подклассам *Отдельное-Отдельное*.

По признаку формы проявления связи отношения подразделяются на: *действие-ориентированные* отношения (аналог силы, энергии); *объект–ориентированные* отношения (проявление через изменения состояния целевого объекта); *результат–ориентированные* отношения (аналог события)[2].

В табл. 1 приводится фрагмент классификационной таблицы отношений, построенной на основе концептуального каркаса семантических отношений верхнего уровня.

Кроме того, отношения могут иметь два типа модальности:

- *Достоверное (Актуальное) / Предполагаемое (возможное) / Невозможное*;
- *Выполняющееся / Состоявшееся / Ожидаемое.*

В табл. 2 приведена таксономия отношений, построенная на основе классификационной таблицы путем линейно-иерархического упорядочения[3]. В колонке «Значение» приведены примеры отношений, задаваемых конкретными лексическими конструкциями либо названием связи или характерного действия. В колонке «Признак» приводятся некоторые характеристики отношения или краткий комментарий его характера.

Таблица 1

Классификация отношений на основе концептуального каркаса семантических отношений верхнего уровня

Признак			*Реальное (действие конкретное)*	*Абстрактно-конкретное*	*Абстрактное (операция)*
Креативные	*Отдельное*	**F**	**Создать / Уничтожить** Генерировать, образовать	**Произвести / Разрушить** / **Сохранить** Изготовить, установить	**Утверждение (постулирование)** Утверждать, заявлять,
		O	**Появление / Исчезновение** Быть, являться, жить	**Обнаружение / Потеря / Существование** Найти, обнаружить	**Определение** Дать определение, задать,
		R	**Явление, факт** Земля, вода	**Артефакт / Усложнение / Упрощение** Книга, здание	**Правило, Истина** Значить, означать

[2] В табл. 1 подкласс *действие-ориентированных* отношений обозначен символом «**F**»; *объект–ориентированных* – «**O**»; *результат–ориентированных* – «**R**».

[3] Отметим, что такая форма более удобна для восприятия и эксплуатации.

Признак			*Реальное (действие конкретное)*	*Абстрактно-конкретное*	*Абстрактное (операция)*
Трансформативные/Локативные	*Отдельное–Отдельное*	**F**	**Воздействие / Ограничение** Повредить, оказать воздействие	**Влияние / Защита / Изменение** Смять, сжать, разбавить	**Присвоение, Выполнение** Присвоить значение, сложить
		O	**Передача** переместить, излучение	**Связь/ Запись / Чтение / Индикация** Кодирование, дублирование, копирование	**Обозначение** Обозначать, задать, именовать
		R	**Взаимодействие / Сопоставление** Связать, противопоставлять	**Зависимость** Зависит от, обеспечивать, учитывать	**Сравнение** *(рез-т)* Быть первым, больше, лучше
	Отдельное – Целое	**F**	**Присоединить / Отсоединить** Подключить, прикрепить	**Включение / Исключение Классифицирование**	**Интегрирование / Разложение** Степень, дифференцировать
		O	**Соединение / Фрагменты** Склеить, связать	**Локативность** (*пространство/ время)* Расположить, поместить, изъять	**Принадлежность** Принадлежать, относиться
		R	**Соотношение** - часть/целое, входит в состав Включать, исключать	**Разметка Агрегация / детализация (род/вид)** Обладать свойством, владеть,	**Вычисление** (рез-та) Вычислить, рассчитать,
	Все части – Целое	**F**	**Слияние / Распределение** Рассыпать, раздробить, слить	**Упорядочить** Сортировать, ранжировать	**Объединить / Разделить** Отличать, различать
		O	**Объединение** Объединить, агрегировать	**Собирание / Разборка** Сконструировать, собрать, разобрать	**Распределение, Следование** Рассортировать, ранжировать
		R	**Единение, блокирование** Блокировать, противопоставлять	**Сборка, агрегат, структура** Структурировать	**Порядок, композиция** Упорядочить, комбинировать

Таблица 2

Таксономия отношений, построенная на основе классификационной таблицы путем линейно-иерархического упорядочения

		Тип отношения	*Значение*	*Признак*
Трансформативные	*Конкретные*	***Действие***		
		Создание / Уничтожение / Сохранение	создать/уничтожить	
		Воздействие	ударить, дать, двинуть	*объекта на другой*
		Взаимодействие		*двух объектов*
		Движение (Передача)		
		Перемещение	импорт/экспорт	*целое*
		Излучение / Прием [сигнала]	измерение	*часть*
		Композиция		
		Присоединение / Отсоединение	прикрепить/открепить	*отдельное*
		Слияние / Распределение	соединить/рассортировать	*поток*
		Синтез / Анализ		*сложное*
	Абстрактно-конкретные	***Влияние***		
		Изменение состава, структуры, состояния, формы, положения, поведения	искривить, выровнять, стабилизировать	*внутри. во вне.*
		Преобразование		
		Отображение	имитировать, копировать	*содержание*
		Трансформация	преобразовать	*форма*
		Зависимость	*«быть», «являться»*: входом/выходом, целью/ результатом, причиной/ следствием, ресурсом/ограничением/ средством/инструментом/ управлением/условием/ помехой, основанием/ назначением	*функционально-ролевое*

		Тип отношения	*Значение*	*Признак*
Информационно-аналитические		***Локативность / Соотношение***		
		Размещения / Разграничения *время / пространство / область / порядок*		
		Включения (принадлежности)	род–вид, часть–целое, входит в состав	
		Компаративные отношения		
		сравнения		*в одном пр-ве*
		количественные		
		для мер	<, >, и т.д.	
		для величин	объемнее/выше,	
		качественные	лучше/хуже, …	
		порядковые	слева/справа, до/после, …	
		сходство / различие	равенство, эквивал., подобие, аналогия, толерантность,	*разные пр-ва*
		Квалификативные отношения		
		обладания	«имеет» свойство, признак, признак–значение, «быть владельцем» и т.п.	
		классификации / спецификации	принадлежит, выделяется	*отношение к / выделение из*
		идентификации		
		обозначения	код, «обозначается»	
		описания	именования, разметки	
		содержания	предметизации, индексирования	*представление внутреннее*
		объекта	автор, форма	*внешнее*

ВЫДЕЛЕНИЕ И ИДЕНТИФИКАЦИЯ ОТНОШЕНИЙ В ТЕКСТЕ

Построение типизированных шаблонов основывается на предпосылке, что при всей вариантности естественного языка существуют закономерные конструкции, а также некоторые устойчивые формы и стереотипы, отвечающие грамматике языка. Шаблоны могут учитывать порядок слов, их морфологические и синтаксические характеристики, а также набор конкретных значений конструкций и их частей. Обзор некоторых подходов приводится ниже.

Синтаксический анализ текста, позволяющий выделять тройки, где в качестве отношения между понятиями используется структура, соответствующая сказуемому, рассматривается в [19]. Шаблоны содержат основы глаголов, а также морфологические характеристики слов.

В [20] описывается модель «*деятель-действие-объект*», при этом термин, указывающий на «деятеля», будет являться существительным и иметь именительный падеж, а глагол, представляющий действие, – находиться между «деятелем» и «объектом», на который направлено действие. Отношения представлены глаголами, распределенными по 25 группам.

В [21] предлагается метод автоматического формирования шаблонов для описания событий, который основывается на нахождении в новостном кластере нескольких предложений с одинаковыми участниками, в одном из которых удалось обнаружить извлекаемое событие. При обработке текстов используется морфологический анализатор и учитывается порядок слов в предложении, шаблоны могут содержать несколько участников событий, отношения представлены глаголами.

При анализе способов формализации рассматривались также языки лексико-синтаксических шаблонов *LSPL* [22], *RCO Pattern Extractor*, *Alex*, Томита-парсер.

В настоящей работе шаблоны построены на основе результатов анализа закономерностей состава и структуры лингвистических конструкций отношений того или иного класса. В качестве основы для идентификации семантических отношений принята характеристическая предикатная лексика и типизированные шаблоны соответствующих лингвистических конструкций, представленные морфологическими признаками слов, а также морфемами.

Наборы предикатной лексики для каждого класса отношений формировались экспертным путем на основе автоматически выделенных в корпусе текстов научно-технических документов лингвистических конструкций, представляющих элементарный факт (триплет – пара понятий и лингвистическая конструкция связи между ними), а также словарей глаголов, причастий, предлогов. В качестве понятий использовались именные субстантивные словосочетания. При этом лингвистическая конструкция отношения может быть представлена отдельным словом или комбинацией слов следующих частей речи: глагол, причастие, деепричастие, краткое прилагательное, краткое причастие, наречие, частица, предлог. Глагол и отглагольные части речи при классификации приводятся к форме инфинитива.

В табл. 3 приведены построенные шаблоны. Каждому шаблону отношения соответствует набор лингвистических конструкций, которых в рамках экспериментального исследования всего было выявлено более 800. В таблице приняты следующие обозначения: **pr** – все формы предлогов и существительные, указывающие на класс отношения (неглагольные конструкции); **inf –** отношение, выраженное в тексте глаголом/причастием/деепричастием, представленное для обработки в инфинитивной форме (глагольные конструкции); в скобках могут указываться характерные черты инфинитива, чаще всего это приставка (представлено *курсивом*); знак "+" – иные слова, выражающие свойства, характеризующие отношения (наречия, частицы, модальные глаголы); знак "|" используется для перечисления возможных шаблонов. В скобках курсивом приведены значения морфем, характерных для этого вида отношения.

Проведенный анализ показал, что каждый класс отношений имеет характерные признаки и шаблоны лингвистических конструкций.

Так, для подкласса «Перемещение» класса «Передача» характерно наличие приставки «пере» в глаголах лингвистической конструкции: *передавать, передвигать, переслать, перебазировать, перетащить, перевозить, перенести, перегнать и др.*

Таблица 3

Таблица шаблонов для классификации отношений

	Отношение (класс)	*Шаблон*
Конкретные	*Креативные*: **Создание / Уничтожение**	inf
	Трансформативные (непосредственное взаимодействие):	
	Отдельное	
	Передача	
	Перемещение (импортирование, экспортирование)	inf(*пере-, с-, вы-*) | inf
	Излучение (запись, чтение, информирование)	inf | inf+pr
	Взаимодействие	inf (*взаимо-*) | inf
	Воздействие (одного объекта на другой)	inf | inf+pr
	Композиция	
	Присоединение	inf(*со-, при-*) | inf | inf+pr | inf(*со-, при-*)+pr
	Отделенение	inf (*от-, раз-*) | inf | inf(*от-*)+pr
	Соединение	
	Слияние	inf (*с-, со-*) | inf
	Распределение	inf (*рас-, раз-*) | inf | inf+pr or inf (*рас-, раз-*)+pr
	Сопоставление (непосредственное)	inf(*со-*) | inf | inf+pr | pr
	Локативность	inf | inf + pr | pr
	Время	inf+pr | pr
	Пространство	inf | inf+pr | pr
Абстрактно-конкретные	**Изменение (воздействие на свойства)**	inf | inf+pr
	состава	inf
	структуры	inf
	формы	inf
	состояния	inf | inf+pr
	положения	inf
	поведения	inf
	Зависимость	inf | inf+pr | pr
	Соотношение функционально-ролевое «быть», «являться»	
	источником/приемником, потребителем/поставщиком, входом/выходом	inf | pr
	результатом	inf | inf+pr | pr
	ресурсом	inf | inf+pr | pr
	средством/инструментом	inf | inf+pr | pr
	основанием	inf | inf+pr | pr
	управлением	inf
	помехой	inf | pr
	условием	inf | inf+pr | pr
	целью, предназначением	inf | inf + pr | pr
	причиной/следствием	inf | inf + pr | pr
	ограничением	inf | inf + pr | pr
	Идентификация	
	Утверждения (определения)	inf
	Персонификация	inf

	Отношение (класс)	*Шаблон*
Абстрактные	**Математические операции**	
	Вычисление	inf | inf+pr | pr
	Присвоение	inf+pr
	Сравнение:	
	сравнение (как процесс)	inf | inf+pr | pr
	Тождество	inf
	Различие	inf
	Интегрирование/дифференцирование	inf | inf+pr
	Соотношение	
	Классификации (принадлежность классу)	inf | inf+pr | pr
	Спецификации (выделение из класса) Обладание (признак, свойство, значение, имя, …)	inf | inf+pr
	Отображение	inf
	Упорядочивание	inf

Таблица 4

Примеры типизации связей в триплетах, выделенных согласно шаблонам из конструкторской документации

Понятие 1	Лингвистическая конструкция	Понятие 2	Шаблон	Класс отношений
проектные режимы	реализоваться в	определенная последовательность	inf+pr	Локативность [реализовать в]
проектные режимы	реализоваться по принципу	их действие	inf pr	Создание [реализовать] Быть/являться основанием [по принципу]
баки системы пассивного отвода тепла систем СПОТ ПГ	размещать на	внешняя оболочка	inf+pr	Локативность в пространстве [размещать на]
собственные частоты	находиться в	диапазон 106-119 Гц	inf+pr pr	Локативность в пространстве [находиться в] Быть|являться ограничением [в диапазоне]
подземное сооружение	разделить на	секции	inf(рас-, раз-)+pr	Распределение [разделить на]
патрубки корпуса	приваривать к	трубопроводы главного циркуляционного контура	inf(со-, при-)+pr	Присоединение [приваривать к]

Подкласс «Присоединение» включает в себя отношения присоединения части (отдельного) к целому. Лингвистические конструкции этого подкласса определяется не только глаголом, но и предлогами «с/со» и «к»: *сочетать с, привести к, примыкать/примкнуть к, присоединить к, связать с/со, приварить к, прикрепить к, соприкасаться с.*

Для лингвистических конструкций подкласса «Отсоединение» характерны глаголы с приставками «от» или предлогом «от»: *отсоединить (от), отделить (от), отломить (от), отклеить (от), отмотать от, отрезать (от), отсыпать (от), отгородить (от).*

Конструкциям класса «Локативность» свойственны предлоги «в», «на», «из». Например, для обозначения места в пространстве или времени: *находить на/в, учитывать в, расположить на/в, производить на/в, убрать из, применить в, представить в, установить в/на, возникнуть в/на.* Очевидно, что такие производные предлоги, как «вначале», «задолго до», «в течение», «в период», «во время» характерны для локативности во времени; а такие, как «вокруг», «вдоль», «поперек», «внутри», «снаружи», «слева (от)», «справа (от)», для локативности в пространстве. Для однозначной идентификации отношений из класса «Локативность» как одного из подклассов, таких конструкций как «находиться в», «установить в/на» и т.д., можно, в частности, использовать таксономию сущностей, позволяющую уточнить характер связи за счет свойств объекта.

Отношения могут быть представлены одним или несколькими словами. Например, в триплете «*стандарт – устанавливает – общие требования*» – это одно слово «*устанавливает*», а в триплете «*стадии разработки конструкторской документации – устанавливаются в – техническом задании*» – два. Причем слово «*устанавливаться*» указывает на класс «Создание», а предлог *«в»* – на «Локатив-

ность». И поскольку лексическая конструкция может быть отнесена к двум классам, то класс отношения должен доопределяться на основе контекста триплета. В некоторых случаях, когда конструкция из одного класса является частью лингвистической конструкции другого класса, будет присвоен класс, соответствующий более длинной из них. Например, триплет «*пожарную технику – подразделяют на – группы*» будет отнесен только к классу «Распределение», так как конструкция «*подразделять на*» соответствует шаблону «**inf+pr**» этого класса, хотя отдельно предлог «на» определён в классе «Локативность». В табл. 4 приведены примеры типизации связей в триплетах, выделенных согласно шаблонам из конструкторской документации.

Рассмотрим характер идентификации. Возьмем два примера словесных конструкций: *«Отделить от остальных элементов сернистое зерно»* и *«Раздробить вдребезги камень»*. Согласно признакам классификации первый случай относится к подклассу «Отделение», а второй – подклассу «Распределение». При этом обе лингвистические конструкции и «Отделить», и «Раздробить» относятся к общему классу «Композиция». То есть иерархическая структура классификации позволяет количественно определить семантическую близость пары связей.

ЭКСПЕРИМЕНТАЛЬНЫЕ ИССЛЕДОВАНИЯ

Предложенная таксономия отношений использовалась для типизации связей в триплетах, извлеченных из текстов стандартов и проектно-конструкторской документации: для стандартов из 5457 триплетов были типизированы 4678; для проектно-конструкторской документации из 2323 триплетов типизировались 2069, что составляет 85,7% и 89,1% соответственно.

Распределение типизированных отношений согласно шаблонам приведено на рис. 1 и 2. Классы, к которым было отнесено менее 1% типизированных отношений, объединены под названием «Другие классы».

Как видно из диаграмм на рис. 1 и 2, для обоих типов документов самыми многочисленными являются классы «Зависимость», «Локативность» и «Быть/являться условием». Ко всем этим классам большинство отношений было отнесено за счет предлогов. Однако для случая глагольных конструкций для стандартов превалирует класс «Создание/Уничтожение» (наиболее характерные конструкции: *изготовлять, определять, производить, устанавливать*), в то время как в проектно-конструкторской документации чаще встречаются триплеты с отношением «Обладание» (наиболее часто употребляемые конструкции: *обеспечивать, включать, исключать, включать в, входить в, иметь, состоять из)*.

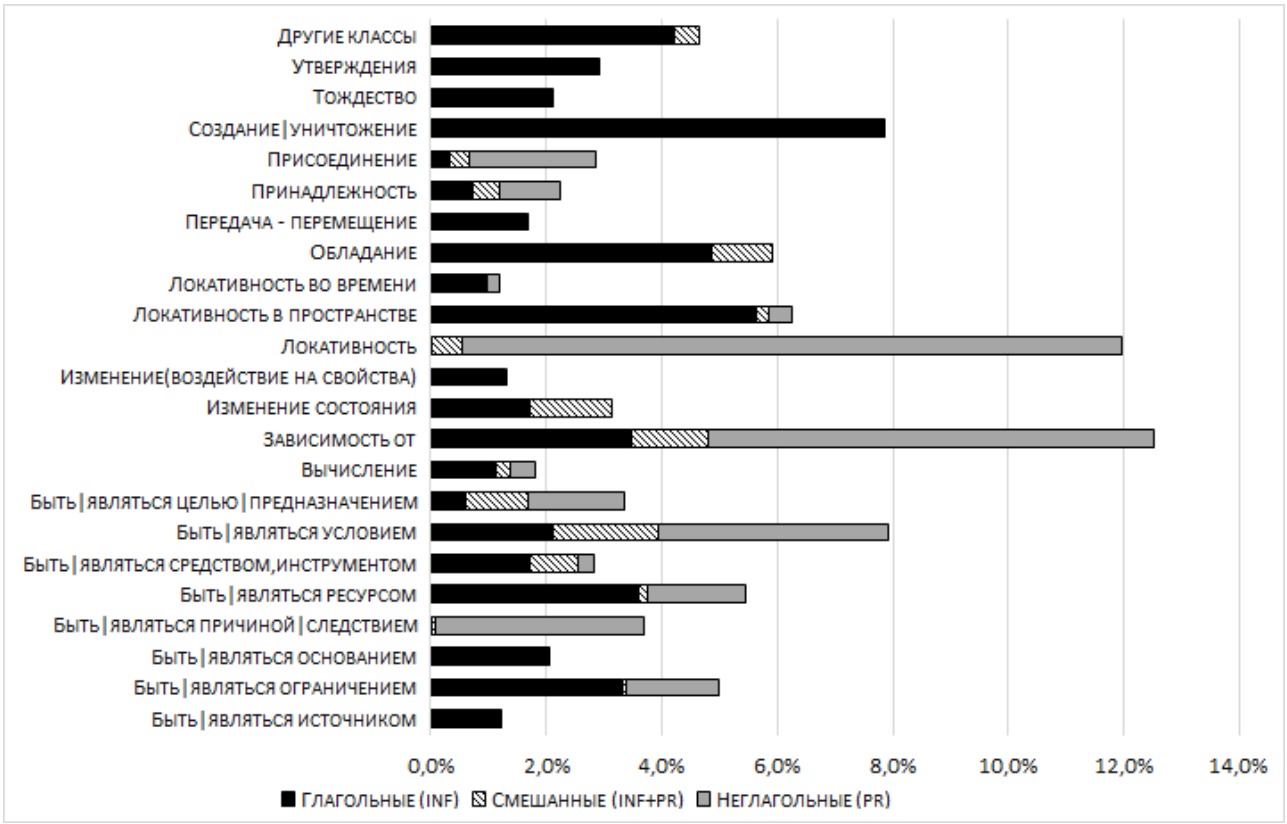

Рис. 1. Распределение троек с типизированными отношениями, извлечённых из текстов стандартов, по шаблонам связей

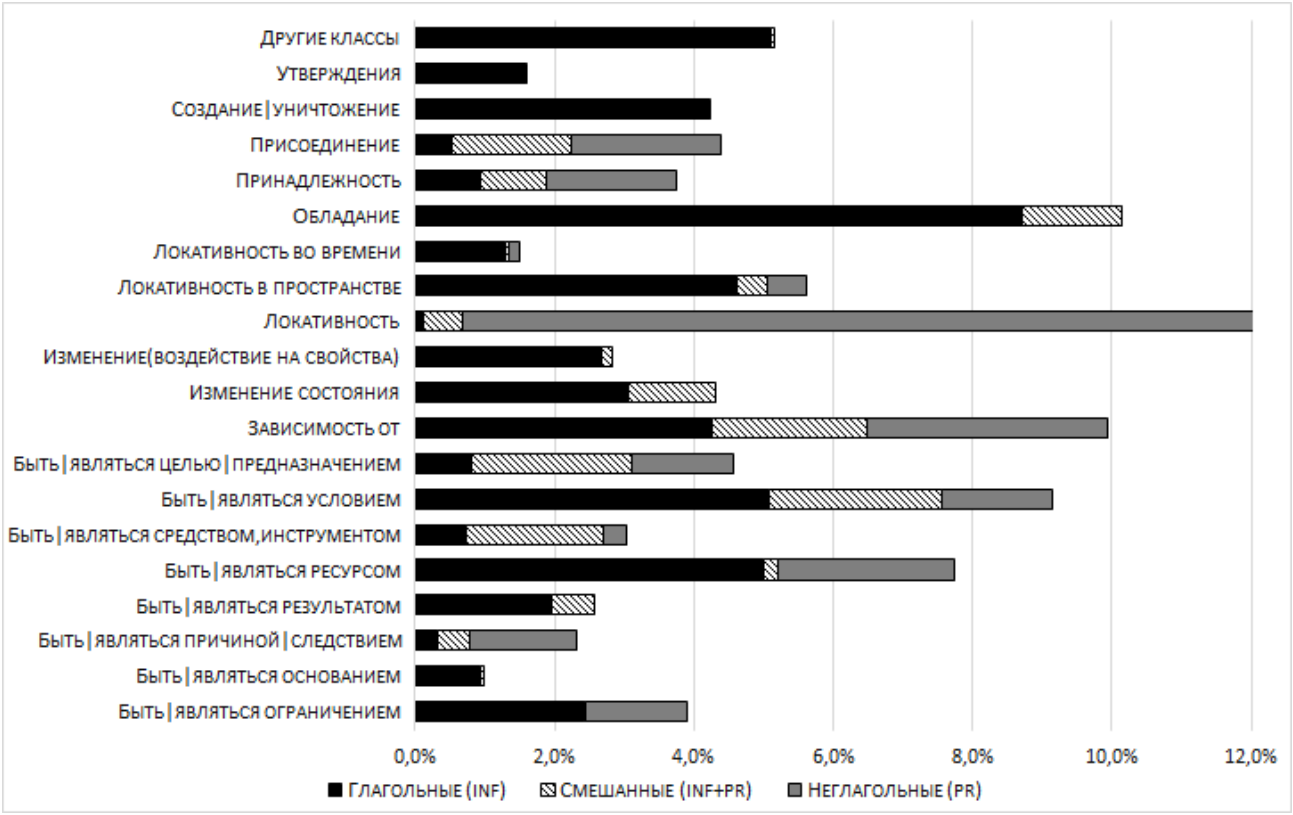

Рис. 2. Распределение троек с типизированными отношениями,
извлечённых из текстов проектно-конструкторской документации, по шаблонам связей

ЗАКЛЮЧЕНИЕ

При помощи решетки характеристических признаков построены классификация и таксономия семантических отношений верхнего уровня (не включающая семантические связи, специфические для отдельных предметных областей).

На основе массива автоматически выделенных в корпусе текстов научно-технических документов лингвистических конструкций, представляющих семантические связи между понятиями, а также словарей глаголов, причастий и предлогов, для каждого класса отношений выявлена предикатная лексика и построены учитывающие морфологические признаки слов и морфемы шаблоны, используемые для идентификации отношений, представленных в тексте разными конструкциями. Предложенная типизация позволяет соотносить сходные по смыслу отношения (и, следовательно, количественно оценивать их семантическую близость), задаваемые в тексте различными вербальными конструкциями. Это, в итоге, позволит производить основанные на вычислениях операции над онтологиями, учитывающие не только близость понятий, но и близость связей, которые есть между ними. На массиве текстов стандартов и проектно-конструкторской документации проведены экспериментальные исследования и получены статистические данные о распределении классов отношений в зависимости от вида документа.

Представляется перспективным углубление таксономии отношений, накопление лингвистических конструкций, уточнение шаблонов.

СПИСОК ЛИТЕРАТУРЫ

1. Van Renssen A. Gellish: A Generic Extensible Ontological Language. – Delft: Delft University Press, 2005. – 238 p.
2. СИНТОЛ // Сборник переводов по вопросам информационной теории и практики. – М.: ВИНИТИ, 1968. – С. 36–47; 50–52; 66–72; 76–80.
3. Скороходько Э.Ф. Лингвистические проблемы обработки теистов в автоматизированных информационно-поисковых системах // Вопросы информационной теории и практики. – 1974. – № 25. – С. 5-120.
4. Виноград Т. Программа, понимающая естественный язык. – М.: Мир, 1976. – 51с.
5. Максимов Н.В. Методологические основы онтологического моделирования документальной информации // Научно-техническая информация. Сер. 2. – 2018. – №3. – С. 6–22; Maksimov N.V. The Methodological Basis of Ontological Documentary Information Modeling // Automatic Documentation and Mathematical Linguistics. – 2018. – Vol. 52, №3. – P. 57–72.
6. Голицына О.Л., Максимов Н.В., Окропишина О.В., Строгонов В.И. Онтологический подход к идентификации информации в задачах документального поиска // Научно-техническая информация. Сер. 2. – 2012. – № 5. – С. 1–9; Golitsyna O.L., Maksimov N.V., Okropishina O.V., Strogonov V.I. The ontological approach to the identification of information in

tasks of document retrieval // Automatic Documentation and Mathematical Linguistics. – 2012. – Vol. 46, № 3. – P. 125–132.
7. Гордей А.Н. Теория автоматического порождения архитектуры знаний (ТАПАЗ-2) и дальнейшая минимизация семантических исчислений // Открытые семантические технологии проектирования интеллектуальных систем = Open Semantic Technologies for Intelligent Systems (OSTIS-2014). Материалы IV Междунар. науч.-техн. конф. (Минск, 20-22 февр. 2014 г.). – Минск: БГУИР, 2014. – С.49-64.
8. FrameNet project – URL: https://www.framenet.icsi.berkeley.edu/fndrupal/
9. Hirtz J. et al. A functional basis for engineering design: reconciling and evolving previous efforts // Research in engineering Design. – 2002. – Vol. 13, №. 2. – P. 65– 82.
10. Johansson I. et al. Functional anatomy: A taxonomic proposal //Acta biotheoretica. – 2005. – Vol. 53, №. 3. – P. 153–166.
11. Smith B., Grenon P. The Cornucopia of Formal-Ontological Relations // Dialectica. – 2004. – Vol. 58, № 3. – P. 279–296.
12. Kitamura Y., Sano T., Namba K., Mizoguchi R. A functional concept ontology and its application to automatic identification of functional structures // Advanced Engineering Informatics. – 2002. – Vol. 16(2). – P. 145–163.
13. Kitamura Y., Mizoguchi R. Ontology-based systematization of functional knowledge // Journal of Engineering Design. – 2004 – Vol. 15(4). – P. 327–351.
14. Kitamura Y., Koji Y., Mizoguchi R. An Ontological Model of Device Function: Industrial Deployment and Lessons Learned // Journal of Applied Ontology (Special issue on “Formal Ontology Meets Industry”). – 2006. – Vol. 1, № 3–4. – P. 237–262.
15. Апресян Ю.Д. Фундаментальная классификация предикатов /Языковая картина мира и системная лексикография. – М.: Языки славянских культур, 2006. – С. 76–110.
16. Васильев Л.М. Теоретические проблемы общей лингвистики, славистики, русистики: Сборник избранных статей. – Уфа: РИО БашГУ, 2006. – С.171–175.
17. Падучева Е.В. Динамические модели в семантике лексики. – М.: Языки славянской культуры, 2004. – 608 с.
18. Найханова Л.В. Основные типы семантических отношений между терминами предметной области // Известия высших учебных заведений. Поволжский регион. Серия Технические науки. Информатика, вычислительная техника и управление. – 2008. – № 1. – С. 62–71.
19. Власов Д.Ю., Пальчунов Д.Е., Степанов П.А. Автоматизация извлечения отношений между понятиями из текстов естественного языка // Вестник НГУ. Серия: Информационные технологии. – 2010. – Том 8, № 3. – С. 23–33.
20. Бексаева Е.А. Исследование и разработка алгоритма извлечения семантической информации из текста для формирования онтологии предметной области // Нечеткие системы и мягкие вычисления. Промышленные применения // Сб. науч. трудов IV Всероссийской научно-практической мультиконференции с международным участием "Прикладные информационные системы (ПИС-2017)". – Ульяновск: УлГТУ, 2017. – С. 72–79.
21. Котельников Д.С., Лукашевич Н.В. Итерационное извлечение шаблонов описания событий по новостным кластерам // Труды 14-й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» — RCDL-2012, Переславль-Залесский, Россия, 15-18 октября 2012 г. – Переславль-Залесский: Университет города Переславля, 2012. – С. 353–359.
22. Большакова Е.И. Язык лексико-синтаксических шаблонов LSPL: опыт использования и пути развития // Программные системы и инструменты: Тематический сборник. – 2014. – №. 15. – С. 14–22.

Материал поступил в редакцию 18.07.18.

Сведения об авторах

МАКСИМОВ Николай Вениаминович – доктор технических наук, профессор, ведущий специалист ВИНИТИ РАН, Москва
e–mail: NV-MAKS@YANDEX.RU

ГАВРИЛКИНА Анастасия Сергеевна – аспирант Национального исследовательского ядерного университета «МИФИ» (НИЯУ МИФИ), Москва
e–mail: asgavrilkina@yandex.ru

АНДРОНОВА Вероника Владимировна – студент кафедры Системного анализа НИЯУ МИФИ
e–mail: vero-nika_09@mail.ru

ТАЗИЕВА Ирина Албековна – студент кафедры Системного анализа НИЯУ МИФИ
e–mail: i.tazieva@mail.ru

ВНИМАНИЮ ПОДПИСЧИКОВ!

С 2018 года возобновляется издание информационного бюллетеня «Иностранная печать об экономическом, научно-техническом и военном потенциале государств-участников СНГ и технических средствах его выявления» серии «Экономический и научно-технический потенциал» (56741) взамен информационного бюллетеня «Экономика и управление»

Периодичность выхода – 12 номеров в год. Объем 48 уч.-изд. л. в год.

В бюллетене освещаются материалы иностранной печати по широкому спектру вопросов, касающихся сфер экономического и научно-технического развития России и стран СНГ: общие вопросы, финансы, промышленность, рынки, сельское хозяйство, космос, транспорт и связь, природные ресурсы, трудовые ресурсы, внешние торгово-экономические и научные связи

Оформить подписку на информационный бюллетень, начиная с любого номера, можно в ВИНИТИ РАН по адресу: 125190, Россия, Москва, ул. Усиевича, 20,

Телефоны: (499) 151-78-61; (499) 155-42-85

Факс: (499) 943-00-60;

E-mail: contact@viniti.ru; sales@viniti.ru

ВСЕРОССИЙСКИЙ ИНСТИТУТ НАУЧНОЙ И ТЕХНИЧЕСКОЙ ИНФОРМАЦИИ РОССИЙСКОЙ АКАДЕМИИ НАУК

предлагает научным работникам, аспирантам и другим специалистам в области естественных, точных и технических наук, желающим быстро и эффективно опубликовать результаты своей научной и научно-производственной деятельности, использовать способ публикации своих работ через *систему депонирования.*

Депонирование (передача на хранение) – особый метод публикации научных работ (отдельных статей, обзоров, монографий, сборников научных трудов, материалов научных конференций, симпозиумов, съездов, семинаров), разрешенных в установленном порядке к открытому опубликованию.

Подготовка и передача на депонирование научных работ происходит в соответствии с «Инструкцией о порядке депонирования научных работ по естественным, техническим, социальным и гуманитарным наукам» (М., 2014).

Депонированные научные работы находятся на хранении в депозитарии ВИНИТИ РАН, копии работ предоставляются заинтересованным организациям и специалистам на бумажном и электронном носителях и являются официальной публикацией.

Информация о депонированных научных работах включается в информационные издания ВИНИТИ РАН: Реферативный журнал, Базу данных и Аннотированный библиографический указатель «Депонированные научные работы».

Направить научную работу на депонирование можно, обратившись в Группу депонирования ЦНИО ВИНИТИ РАН по адресу:

125190, Москва, ул. Усиевича, 20.
ВИНИТИ РАН, Группа депонирования ЦНИО
Тел.: 499-155-43-28, 499-155-43-76, 499-155-42-43, Факс: 499-943-00-60,
E-mail: cnio@viniti.ru, dep@viniti.ru

С инструкцией о порядке депонирования можно ознакомиться на сайте ВИНИТИ РАН: **http://www.viniti.ru**